Tennessee TECH

OAK RIDGE National Laboratory

Proceedings of the 16th International Conference on Cyber Warfare and Security

Tennessee Tech University and Oak Ridge National Laboratory
Cooksville, Tenessee, USA
25-26 February 2021

Dr. Juan Lopez Jr., Dr. Ambareen Siraj and Dr. Kalyan Perumalla

aci

A conference managed by ACI, UK

Smart Semi-Supervised Accumulation of Large Repositories for Industrial Control Systems Device Information

Kimia Ameri[1], Michael Hempel[1], Hamid Sharif[1], Juan Lopez Jr.[2] and Kalyan Perumalla[2]
[1]Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, USA
[2]Oak Ridge National Laboratory, Tennessee, USA
kameri2@huskers.unl.edu
mhempel@unl.edu
hsharif@unl.edu
lopezj@ornl.gov
perumallaks@ornl.gov
DOI: 10.34190/IWS.21.064

Abstract: Industrial Control Systems device manufacturers frequently add new features to improve their product performance. Oftentimes, these changes are mainly vendor-driven initiatives, and customers may not be aware of the full impact of these new capabilities on their cybersecurity posture. In the energy sector, this can lead to considerable dissonance between vendor-provided cybersecurity claims and a customer's responsibility for Operation Technology cybersecurity compliance. Thus, the resulting dynamic verification burden is shifted towards the customer and may pose a significant cybersecurity risk to the energy sector landscape. We found that there is very limited research into cybersecurity auditing for Operational Technology. However, a solution is needed for vetting the vendor-supplied feature claims and their adherence to cybersecurity requirements and standards. We are presently engaged in an effort to develop such a system. This paper demonstrates one vital aspect of this effort in proposing an end-to-end framework to accumulate a large repository of ICS device information for this vetting system, curate the dataset, and conduct extensive processing. This framework is designed to use web scraping, data analytics and Natural Language Processing (NLP) techniques to identify vendor websites, automate the collection of website-accessible documents and automatically derive metadata from them for identification of product documents relevant to the repository. We have found that this automated approach to vendor identification, document extraction into a product repository, and NLP pre-processing is unique and has not been previously presented in the literature. The preliminary work shows that this is feasible and can produce reliable results with minimum supervision. Future work will be built upon this foundation in order to achieve semi-supervised vetting of device technical information – a vital capability for ensuring that vendor-claimed device cybersecurity capabilities match industry requirements.

Keywords: cybersecurity, industrial control systems, vendor discovery, NLP, web scraping, repository automation

1. Introduction

In the field of Operational Technology (OT), cybersecurity auditing is playing an increasingly crucial role. OT and Industrial Control Systems (ICS) underpin many critical infrastructure segments, including the energy sector. Post-deployment, OT solutions are expected to work for years, and need to provide robust and reliable cybersecurity. However, ICS vendors always equip their products with new features to provide reasons for system upgrades, and consumers are often unaware of the impact of these features on their cybersecurity posture and regulatory compliance.

International standards bodies and industry societies define and codify cybersecurity requirements (CR) in human-readable formats. Any vendor-supplied features (VSF) can (1) satisfy and match with a corresponding CR, (2) enhance and go beyond the related CR or (3) violate or contradict some of the CRs. This task is a challenging process. OT operators have the difficult task to interpret vendor claims about supported features, conduct extensive manual evaluations of features of interest to obtain a match rating between claim and capability, and reconcile that against industry requirements and their own customer requirements, in order to determine if an OT device poses risks by weakening their cybersecurity posture. This is further complicated through the extensive number of industry requirements and their complex nature, the variety of available devices and device documentation, and the associated assessment of Installation Qualification (IQ), Operational Qualification (OQ), and Performance Qualification (PQ). Traditionally, this work requires trained experts, is costly, and highly error prone. The result is subjective to human error and must be repeated periodically.

To address this challenge, our team is in the process of researching an automated vetting approach for cyber-physical security assurance (CYVET) to match, reconcile, and tally the VSFs against the relevant CR, as described

by Perumalla et al. (2020). The main goal of this research project is to enhance the current industry capabilities and enable verification and validation of OT infrastructure cybersecurity claims both pre- and post-deployment. The vetting system will provide insights into ICS systems, match capabilities against requirements, and greatly simplify compliance analysis and reporting. An overall flow for CYVET is shown in Figure 1.

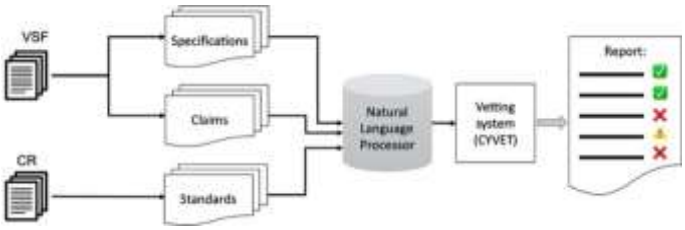

Figure 1: CYVET overall workflow

A cornerstone of the CYVET architecture, shown in Figure 1, is the availability of documents relevant to the vetting process. Since vetting is expected to encompass the gamut of deployed devices from numerous vendors, this would indicate the need for an extensive repository of vendor product documents, and industry documents describing industry standards and requirements, processing of these documents to classify them by type, language, device association, etc., and thus prepare them for the vetting process. This document repository curation and preparation is the foundation upon which CYVET's cutting edge NLP and cybersecurity vetting processes are being built, and a vital first step towards achieving CYVET's objectives. CYVET's capabilities address a critical need for OT operators, particularly in critical infrastructure sectors such as the energy sector.

However, this automated domain-specific information retrieval and document classification from online resources is a challenging task in itself. Identifying ICS vendors, analyzing websites to determine ICS product offerings and identifying documents based on their content for relevant product claims all require a generalized and robust approach to satisfy a wide range of scenarios. NLP techniques may present a solution to these challenges once a repository of documents is available. However, to the best of our knowledge no publicly available dataset for training such cybersecurity domain-specific models currently exists.

In this paper the focus is on presenting our research work on establishing a systematic data gathering process. This enables the generation of an extensive collection of relevant documents by applying automated decision-making paradigms to a large pool of candidate documents searched and extracted from vendor websites. Our framework presented in this paper addresses four key aspects of this effort: 1) automating the identification of potential ICS vendors and associated websites, 2) vendor-of-interest classification using a web content-based scoring metric along with Latent Dirichlet Allocation (LDA), 3) customized procedures for automated web scraping and downloading of documents from ICS vendor websites, and 4) content-based selection, organization and storage of product-related documents in preparation for the larger vetting process.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 describes the system architecture followed by a description of our approach to ICS device information extraction. Section 4 presents and discusses our current results, and in Section 5 our conclusions and future work are presented.

2. Related work

The amount of information available online and on websites is growing at an exponential rate. Along with that comes the need to be able to find information of interest. This drives research efforts into finding approaches for more intelligent information indexing and retrieval, to produce results that match the intent behind user search queries, for example. This is relevant to a number of application domains, including structured query processing and text mining in domain specific applications. For example, Shashirekha and Murali (2007) introduced an automated approach to extract information from unstructured, incomplete sentences, and an unordered sequence of words. This paper describes a keywords/constants pattern matching to extract the patterns and generate a structured text representation regarding a domain-specific ontology. Chandrika et al. (2020) proposed a method to extract unstructured information over the web using Python and parse the data to retrieve the information. The extracted data then is transferred into a relational database hosted on the internet to keep data up to date. All these algorithms are focused on only scraping and crawling websites, but they do not address the need for identifying domain-specific content.

There are limited scientific publications focusing on scraping web content for domain-specific areas. Two specific efforts are Modi and Jagtap (2020) and Anglin (2019) in the educational field, focused on identifying and classifying content with Machine Learning (ML) and NLP techniques. Anglin (2019) proposed a semi-automated framework to study local policy variation in school dress codes using web crawlers and NLP techniques. The web-scraping starts with a pre-determined list of schools, then scrapes their websites to collect all links leading to documents with several different formats. A Convolutional Neural Network (CNN) was then used to select policy-relevant documents and NLP techniques are applied to the relevant documents to identify policy nuances. In Modi and Jagtap (2020) the authors proposed a multi-pattern matching algorithm using NLP and Word Sense Disambiguation techniques to recognize educational websites. This algorithm relies on a set of keywords extracted from HTML content and multimedia tags of pre-defined allowed URLs and blocked URLs. These research works focused on developing a semi-automated web-scraper and NLP-based content classifier, which cannot be used in other domains due to their reliance on pre-defined sets of URLs and domain-specific training.

Recently, Koeva, Obreshkov, and Yalamov (2020) recommended an NLP framework to annotate Bulgarian Legislative Data. The proposed framework developed a web-scraping algorithm to scan the online Bulgarian State Gazette. This scraper will detect new entities and process the HTML to extract metadata information. This metadata will then be transferred into a non-relational database server. This framework is helpful for finding and pre-processing web content but does not address classifying webpages based on their content.

From our review of published scientific efforts, we can observe that none addresses the specific challenges mentioned in section 1. Therefore, we found that the need persists to research automated vendor product identification and data extraction for our CYVET architecture, driven by unsupervised classification models for websites and documents and develop an architecture to automate domain-specific data repository generation and curation from online resources.

3. System architecture

3.1 Overall design

In this paper we introduce an end-to-end framework for collecting ICS vendor documents and extracting structured device specification data. The framework creates a pipeline that produces an aggregated, structured dataset of ICS device information. The main benefit of this approach is the establishment of an automated and systematic process for data collection. This is the first step towards our primary goal of building the proposed cybersecurity vetting system. This framework automates identifying vendors website and extracting documents to build up a library of product manuals, catalogs, and brochures. This provides the basis for a subsequent step in which a document pre-processor prepares the data for CYVET's NLP processing workflow to detect and extract claims and specifications that are then vetted against each other. Figure 2 summarizes the overall system architecture for building our data repository. In the following sections, we will provide detailed descriptions of the core aspects of this architecture.

Figure 2: Our overall system architecture

3.2 Vendor and homepage identification

The first part of the framework's processing pipeline consists of two approaches that allow us to generalize the ICS vendor and homepage identification process:

Approach 1: In our first approach, for the purpose of vendor identification our algorithm conducts a scrape of CISA's ICS-Cert website at https://us-cert.cisa.gov/ics, in order to find potential names for ICS vendors. We established our ICS vendors homepage identification under the assumption that each entity has at least one homepage and according to Definition 1:

Definition 1. *A webpage's link is related to the ICS vendor if the vendor name appears in the NetLoc or Domain portions of the URL.*

We employ a Google search to locate the top ten most relevant links for each ICS vendor name query. Based on Definition 1, if the search results satisfy the name match condition the vendor name and its corresponding homepage will be added to the database.

Approach 2: Our second approach is broader and conducts a generalized search utilizing a list of pre-defined keywords. It leverages multiple search engines and collates the results. We currently utilize Google, Bing, AOL, and the Chinese search engine Baidu, but other search engines can be integrated as well. The search queries are conducted to find links related to industrial IoT systems. In order to make sure that this list of pre-defined keywords covers most IoT product-related links NLP pre-processing and text mining are used to refine the initial keyword set. To achieve this refinement, unique links retrieved from the keyword-driven approach undergo a text pre-processing step that consist of tokenization and lemmatization, and the most frequently occurring words from those links are then used to update the keyword list. With the updated keywords, the process is repeated and new homepages from the retrieved links will be added into our database.

Our algorithm internally manages a blacklist that avoids processing webpages typically associated with references to articles or content mentioning our keywords of interest. Examples of blacklisted pages include Wikipedia, amazon, and YouTube, which are not directly useful for building our repository even though resources such as books and blogs in those sites match our keywords.

3.3 Homepage classification

To ensure that a candidate homepage is related to an ICS vendor, each homepage identified in the previous section needs to be classified. This process employs a scoring metric and Latent Dirichlet Allocation (LDA) topic modeling in order to arrive at a label for each homepage. A homepage will only be labelled as ICS vendor homepage if both the scoring metrics and LDA label it as such. The initial step for homepage labeling is the extraction of several key pages of the specific homepage, to obtain the text content of the homepage, info page, and sitemap (if they exist). Along with this task, all PDFs and links are stored in the database for future steps.

3.3.1 Topic modeling

Topic modeling in general is a compelling approach in text mining to demonstrate the semantic relationships among words in a corpus. LDA is an unsupervised Bayesian probabilistic model that can be used for topic modeling, and was first introduced by Blei, Ng, and Jordan (2003). LDA proposes that the corpus can be regarded as a combination of multiple topics. Similar to probability distributions, each topic produces certain words more frequent than others. For example, in a webpage's text content, certain words such as "Technology", "Security", and "IoT" are more likely to be observed in ICS webpages than in Non-ICS webpages. Thus, LDA is capable of grouping websites that contain words with similar frequency as belonging to the same topic. For CYVET's preprocessing we use LDA to determine if a website likely belongs to an ICS/OT device vendor or not.

In LDA, each topic can represent the polynomial distribution of all words in the corpus. The probability density in an n-dimensional Dirichlet random variable Θ for the given parameter a is calculated as:

$$p(\Theta|a) = \frac{\Gamma(\sum_{i=1}^{n} a_i)}{\prod_{i=1}^{n} \Gamma(a_i)} \Theta_1^{a_1-1} \ldots\ldots \Theta_n^{a_n-1} \quad (1)$$

The joint distribution of a topic mixture Θ, a set of N topics Z and a set of N words in w document is:

$$p(\Theta, \mathcal{Z}, w) = p(\Theta|a) \prod_{n=1}^{N} p(\mathcal{Z}_n|\Theta)\, p(w_n|\mathcal{Z}_n, b) \quad (2)$$

where the marginal distribution is calculated as:

$$p(\mathrm{w}|a, b) = \int p(\Theta|a) \left(\prod\nolimits_{n=1}^{N} \sum_{\mathcal{Z}_n} p(\mathcal{Z}_n|\Theta)\, p(w_n|\mathcal{Z}_n, b) \right) d\Theta \quad (3)$$

LDA will calculate the probability of a corpus *D* as:

$$p(\mathrm{D}|a, b) = \prod_{d=1}^{M} \int p(\Theta_d|a) \left(\prod_{n=1}^{N_d} \sum_{\mathcal{Z}_{dn}} p(\mathcal{Z}_{dn}|\Theta_d) p(w_{dn}|\mathcal{Z}_{dn}, b) \right) d\Theta_d \quad (4)$$

The parameters a and b are corpus level, variable Θ_d is document level, and variables ω_{dn} and z_{dn} are word-level, and all are sampled once in every document.

In this paper, we use the online variable Bayes algorithm for LDA (online LDA), proposed by Hoffman, Bach, and Blei (2010). This algorithm uses online stochastic optimization with a natural gradient step that converges to a local optimum of the variant Bayes function. The online LDA algorithm's input is a sparse matrix of weighting scores for each token in each sentence in a corpus.

After removing all punctuations, digits and English stop words in the text content, each unique word (token) needs to be transformed into its Term frequency-inverse document frequency (Tf-idf) weighting vector. Tf-idf weights indicate the normalized frequency of a word in a corpus, which is calculated as:

$$w_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \quad (5)$$

Here, $tf_{t,d}$ is the frequency of term t in document d, df_t is the number of documents containing term t, and N indicates the total number of documents. Online LDA uses the Tf-idf score matrix to cluster the most related documents together.

3.3.2 Context-Dependent scoring metric

In this section, we present our Context-Dependent scoring metric algorithm based on the appearance of defined keywords and phrases in the text content of the webpages. This algorithm starts by building a pre-defined set of keywords and phrases to search in the webpage text content, utilizing the *FlashText* algorithm. *FlashText* algorithm was introduced by Singh (2017) to match keywords in the text. *FlashText* builds an internal dictionary of keywords based on the "Trie" (prefix tree) data structure, which makes it much faster than *Regex* and it will only find the exact match, not any subset of it in the word corpus.

The Context-Dependent scoring metric algorithm starts with finding a unique set of words or phrases within the text content of each webpage and determines if any of these match predefined keywords or phrases. Each predefined such key phrase is associated with a particular score, and the aggregate score across all matching key phrases for a webpage provides that webpage's score. For example, "copyright" and "automation" are both keywords, but "automation" has a higher score than the other, as it is more likely to occur on webpages of ICS vendors whereas copyright is more general. Based on an experimentally chosen scoring metric threshold, in our tests determined to be 20, we can then label this webpage. Algorithm 1 summarizes this scoring process.

Algorithm 1: Homepage Scoring
Input: Text content from homepage, info page and sitemap
Output: Normalized score
Define k as set of keywords;
Define v as a decimal fraction less than one based on the importance of k
Define keyword matrix with assigned values as M={k:v};
Define a set of phrases as P;
K_score = sum(v|if k in input)/Len(k);
P_score= Count the number of phrases in the input/Len(P);
Normalized_score = (P_score + K_score) *100.

3.3.3 Neural network classifier

Recognizing the vendors homepages with LDA and scoring metric require human supervision for labelling the topics of homepages. To minimize such interactions, we need a classifier to automatically label the homepages.

The Word embedding technique, first introduced by Mikolov et al (2013) is a pre-trained neural network that assigns a spatially close vector to the words that appear in similar contexts. This technique is useful to maintain the words' semantic meaning in the training corpus. Therefore, a fully connected NN with a word embedding and two dense layers is trained on the text content of the labelled webpages. This trained model can then be used to classify the future search links for identifying new ICS vendors. The flow diagram of the proposed NN model is shown in Figure 3. The input data are the padded Tf-idf vectors of words with maximum length of 300, which are extracted as part of the LDA topic modeling. A Global Average Pooling layer calculates the average output of each feature vector in the previous layer. As the final processing step, through the dense layers, the newly calculated features are mapped to the class labels.

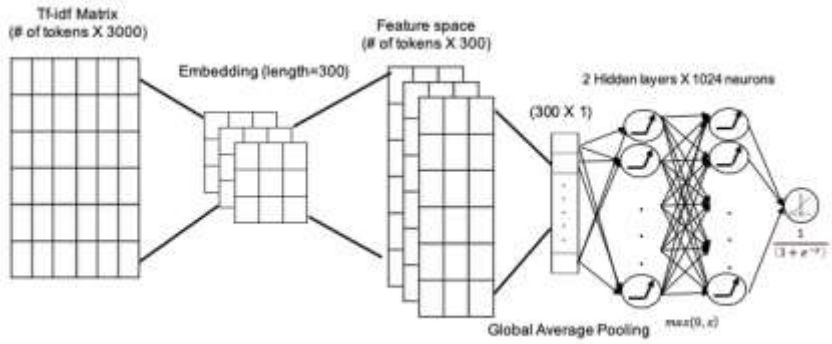

Figure 3: Architecture of the neural network

We then used the information collected in the previous steps and pooled them as our sample set. We then split that into training and test sets and trained the NN using the available information, the details and results of which are presented in our results section further below.

3.4 Document classification

Figure 4 represents the flow diagram for the document classifier. The PDFs that we extract from vendor homepages are stored in our database during the homepage Classification procedure. Subsequently, the text contained within every PDF is extracted and pre-processed for determining key terms, key segments, and key phrases. Finally, we employ a matching algorithm to match these key elements against a pre-defined set of important phrases in order to separate documents containing ICS device information, such as claims about its features, from other types of PDFs such as annual reports and other unrelated files.

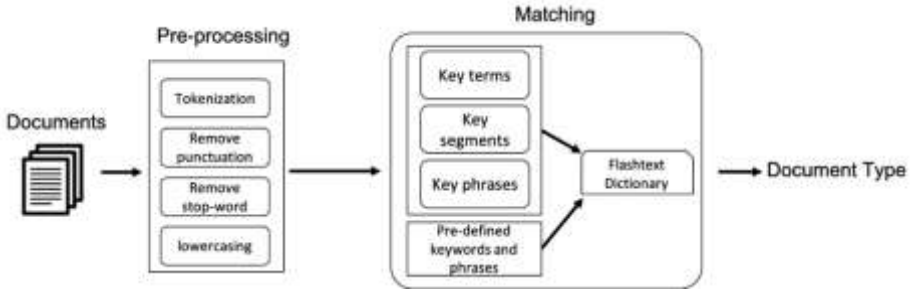

Figure 4: Flow diagram for document classifier

4. Results and discussion

To support evaluating our research work in this domain, all steps of our framework were implemented in Python to provide a fully automated framework for vendor identification and document extraction into a product repository. In this section, the results of each step, shown in the Figure 2, are presented.

Step 1: Vendor Identification. As of the time of this publication, we could extract 340 vendor candidate names by scraping the CISA website. These are predominantly ICS vendors, but the CISA website also includes non-ICS vendor listings. Separately, the keyword-driven search process initiated with a list of 7 keywords and phrases such as “industrial IoT”, “Industrial control system”, and “IoT advisory” utilizing the four search engines mentioned earlier in the paper resulted in 1092 links. For relevance we limit the number of retrieved search results for each keyword search from any search engine to the first 20 returned pages. After removing duplicates, we could obtain 1066 unique links. The *NLTK* python package was used to pre-process and tokenize these links and to determine the most frequently occurring words in order to update the list of pre-defined keywords. This text mining process showed that security, manufacturing, product and technology are highly repeated in these links, even though they were not a part of our initial keyword set. This process was then repeated iteratively, and in the end resulted in 340 vendor names and 1590 unique links. Figure 5 shows the number of matches grouped by search engine for the 11 final keywords. This figure indicates that Google and Bing could provide 93% percent of the overall extracted links.

Table 1: Vendor identification statistics in repository construction

Category		Number	Description
Vendor names		340	Scraped from CISA website (ICS and non-ICS vendors)
Keyword driven		1590	Unique and deduplicated search result links using 11 keywords and phrases
	Root keywords	7	Keywords and Phrases: Industrial IoT, Industrial Control System, IoT Advisory, Embedded IoT Devices, IoT Certificate, IoT Supplier, IoT Security
	Unique Links	1066	Links obtained using root keyword search
	Updated keywords	4	Keywords and Phrases: IoT Technology, IoT Manufacture, IoT Cybersecurity, IoT Product
	Unique Links	582	Links obtained using updated keyword search

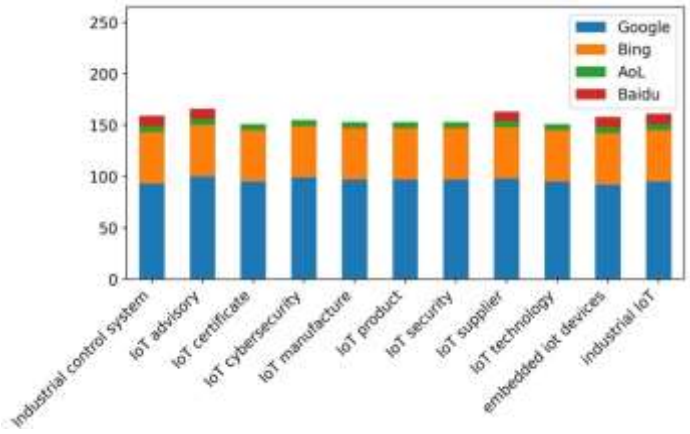

Figure 5: Number of matches for 11 keywords grouped by search engine

Step 2: Homepage identification. All 1590 unique links from the previous step were then parsed, and 1084 unique homepages could be extracted. For any vendors for which we could obtain a name in step 1 but for which we did not also obtain a corresponding homepage, we then conducted a separate homepage search using the vendor name, via a Google search. Using this process, we found an additional 420 vendor homepages based on the initially determined names. In the end, we thus had a collection of 1457 unique vendor homepages that we could identify using both approaches. Table 2 represents the number of homepages stored in the repository that were obtained using steps 1 and 2.

Table 2: Homepage identification statistics in repository construction

Category	Number	Description
Homepages from step 1 after cleanup	1084	Found via keyword-driven search queries
Homepages from step 2 after cleanup	420	Found via vendor name search queries
Total unique homepages	1457	After deduplicating combined set from steps 1 & 2

Step 3: Vendor classification. The set of vendors includes a mixture of ICS and non-ICS vendors at this point. Thus, we conduct vendor classification to filter out non-ICS vendors from the set. The 1457 websites obtained from the previous steps were scraped in order to collect the text content from their homepage, info-page, and sitemap. Based on the proposed scoring metric discussed in Section 3.3.2, coupled with the LDA topic modelling, all websites were then categorized. We reviewed the results of the website topic modelling and removed any inconsistencies between the LDA method and scoring metric algorithm, the cases where their respective labels did not agree. Thus, 286 ICS vendor websites and 290 non-ICS vendor websites were selected in order to create a balanced dataset for ML. The vendor classification output for each model is summarized in Table 3.

Table 3: Vendor classification statistics in repository construction

Category	#ICS Vendors	#Non-ICS Vendors	Description
LDA	578	879	Topic Modelling with LDA on website text content
Scoring Metric	850	607	Label websites with context-dependent scoring metric
Combined Set Size	286	637	Combined LDA and Scoring model results
NN Set Sizes	286	290	website selected for training NN

We then implemented a NN model with the structure shown in the Figure 3 using *Keras,* in order to learn the features of the website content. We split the NN set with the size shown in table 3 into a training and test set, using 70% and 30% of the samples, respectively. The NN will use 403 websites as training set to fit the classifier model and the remaining 173 samples in the test set to provide an unbiased evaluation of the classifier after training the model, and was measured using accuracy, precision, recall, and F1-score. The obtained results are shown in Table 4 and the training progression is shown in Figure 6.

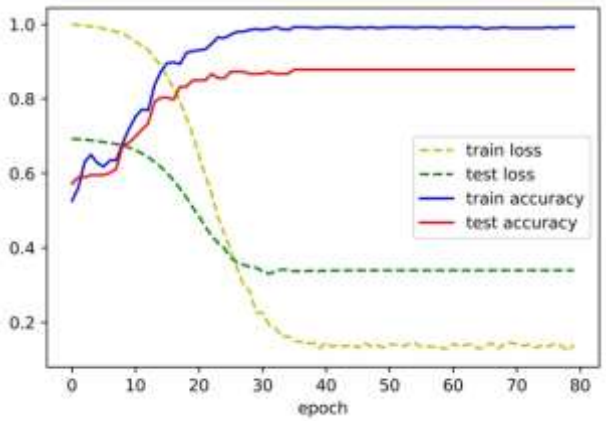

Figure 6: Vendor classifier training progression

Table 4: Vendor classifier accuracy, precision, recall, and F1-score for training and testing

		Precision	Recall	F1-score	Accuracy
Training Phase	ICS vendors websites	0.98	1.00	0.99	0.99
	Non-ICS vendors websites	1.00	0.98	0.99	
Test Phase	ICS vendors websites	0.90	0.90	0.89	0.88
	Non-ICS vendors websites	0.93	0.89	0.90	

Step 4: Document classification. Our framework then scraped the identified websites for documents. This process yielded 19,793 documents. After reviewing the downloaded documents, we determined that 5.6 percent are scanned and the remaining documents are text-based. The *PyMuPDF* python package was used to extract text from readable documents and the *pytesseract* python package was employed for optical character recognition (OCR) from the scanned PDFs. Using our keyword-driven approach, all downloaded files were then labelled. Out of the 19,793 documents we scraped across the identified vendor websites, 63 percent were found to be ICS product-related documents. From those 12581 product-related documents that we found, 71% were classified as “brochure”, 17% were classified as “manual” and 11% were “catalogs”. On average, each of the product-related documents contained 31.2 pages, whereas across all other document classes we found an average page count of 5.1 pages. The detailed output for each type of documents from document classification is summarized in Table 5.

Table 5: Document classification statistics in repository construction

Total number of documents		19,793
ICS product-related documents		12,581
	Manual documents	2248
	Brochure documents	9326
	Catalog documents	1007
Non product-related documents		7,212

Repository Composition and Operational Challenges. Based on the currently identified vendors, their associated websites, the framework, and our current web scraper implementation we could obtain the statistics shown in Table 6 for our data repository:

Table 6: Statistics in repository construction

Database Entities	Number of websites	1457
	Number of ICS vendors	286
	Number of links across all websites	~3.8 million
	Disk space utilized for all websites’ home page, info page and sitemap text content	~30 MB

	Number of documents in ICS vendor websites	~20,000
	Disk space utilized for documents	~40 GB

In our web scraper implementation, we utilized a commonly encountered user-agent value derived from Google's Chrome browser, in order to maximize the probability that the targeted server returns useful content for the requested URL, including during Google search queries.

We also incorporated a 10-second delay for Google search queries in order to reduce the probability of being temporarily blocked by Google. However, this additional delay results in an approximately 80% increase in Google query completion time when searching and retrieving 20 pages of search results, on average.

To avoid re-visiting the same pages in our website scraper when extracting all documents and links we utilize page counts and sitemaps, if available, for each targeted website.

On a computer with a 2.7 GHz Dual-Core Intel Core i5, the vendor and homepage identification steps require approximately two hours total for scraping the CISA website as well as searching for 11 keywords across all targeted search engines in order to identify 340 vendor names and 1590 unique links. The homepage classification step requires approximately 15 hours to download and process the text content of each website's homepage, info page, and sitemap across all 1457 targeted websites, and an additional two hours to classify the content using LDA and our scoring metric. The most time-consuming step is the scraping of websites to find all links as well as download all documents. It takes nearly 50 hours to download 20,000 documents from 286 ICS vendor websites. Classifying these documents, once downloaded, as manuals, catalogs, brochures, and unrelated documents requires approximately 6 hours. The overall process duration for each step as measured on the aforementioned test computer is summarized in Table 7.

Table 7: Overall processing load for ICS website and document identification

	Number of entities	Time	Description
Vendor Identification	~1500 websites	~2 h	Scraping of CISA website and searching for 11 keywords
Homepage classification	~1500 websites	~15 h	Downloading and processing content for all home pages, info pages and sitemaps
		~2 h	ICS vendor website classification based on web page content using LDA and Scoring metric
Document classification	~20,000	~50 h	Downloading all PDFs from 286 vendor websites
		~6 h	Identifying manuals, catalogs, and brochures among all downloaded documents with scoring metric

5. Conclusion and future work

In this paper, we presented our automated framework for ICS vendor identification. We found that our approach is not only feasible, but it can obtain reliable and useful results for ICS document repository construction with low supervision. Our work resulted in the following contributions:

- Websites are labelled based on the combination of the proposed scoring metric and LDA topic modelling. According to our results, the text content in the homepage, info-page, sitemap provides sufficient data in order to determine whether a website belongs to an ICS vendor or not.
- A Sequential Neural Network is trained based on the labelled website content. This trained model can later be used for more easily detecting new ICS vendors.
- By evaluating and comparing keywords obtained from 19,128 PDFs from downloaded 273 websites – a subset of the total vendor set we identified - we realized that using the text data contained in titles, headings, and bold-styled words, we can sufficiently distinguish product-related documents from other unrelated documents.

Our work provides the foundation to achieve semi-supervised vetting of device information system to determine how well vendor-claimed device cybersecurity capabilities adhere to expected capabilities for industry standards compliance. Although the repository is being built based on motivation from our current cybersecurity projects, we envision the repository to be useful for a variety of other purposes, such as tracking the trends of evolution in ICS devices, and detecting the emergence of various market trends. For our future work, we will utilize this data repository-based approach to train domain-specific NLP models and utilize those models for claims

extraction and specification identification. These form the core building blocks of the CYVET architecture for vetting cybersecurity claims in ICS devices and the associated implications for an OT installation's cybersecurity posture.

References

Anglin, K. L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. *Journal of Research on Educational Effectiveness*, *12*(4), 685–706. https://doi.org/10.1080/19345747.2019.1654576

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Chandrika, G. N., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Web scraping for unstructured data over web. In *Embedded Systems and Artificial Intelligence* (pp. 853–859). Springer.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 856–864). Curran Associates, Inc. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf

Koeva, S., Obreshkov, N., & Yalamov, M. (2020). Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. *Proceedings of The 12th Language Resources and Evaluation Conference*, 6988–6994. https://www.aclweb.org/anthology/2020.lrec-1.863

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. https://www.aclweb.org/anthology/W04-3252

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Modi, S. S., & Jagtap, S. B. (2020). Multimodal Web Content Mining to Filter Non-learning Sites Using NLP. In A. P. Pandian, T. Senjyu, S. M. S. Islam, & H. Wang (Eds.), *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018)* (pp. 23–30). Springer International Publishing. https://doi.org/10.1007/978-3-030-24643-3_3

Perumalla, K., Lopez, J., Alam, M., Kotevska, O., Hempel, M., & Sharif, H. (2020). A Novel Vetting Approach to Cybersecurity Verification in Energy Grid Systems. *2020 IEEE Kansas Power and Energy Conference (KPEC)*, 1–6.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, *1*, 1–20.

Shashirekha, H. L., & Murali, S. (2007). Ontology Based Structured Representation for Domain Specific Unstructured Documents. *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, *1*, 50–54. https://doi.org/10.1109/ICCIMA.2007.255

Silva, L. C., Almeida, H. O., Perkusich, A., & Perkusich, M. (2015). A Model-Based Approach to Support Validation of Medical Cyber-Physical Systems. *Sensors*, *15*(11), 27625–27670. https://doi.org/10.3390/s151127625

Singh, V. (2017). Replace or Retrieve Keywords In Documents at Scale. *ArXiv:1711.00046 [Cs]*. http://arxiv.org/abs/1711.00046