Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020

Matching Possible Mitigations to Cyber Threats: A Document-Driven Decision Support Systems Approach

Martha Wagner McNeil
Dakota State University
martha.mcneil@trojans.dsu.edu

Cherie Bakker Noteboom
Dakota State University
cherie.noteboom@dsu.edu

Jun Liu
Dakota State University
jun.liu@dsu.edu

Omar El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Thomas Llansó
Johns Hopkins University Applied Physics Laboratory
thomas.llanso@jhuapl.edu

Abstract

Despite more than a decade of heightened focus on cybersecurity, the threat continues. To address possible impacts, cyber threats must be addressed. Mitigation catalogs exist in practice today, but these do not map mitigations to the specific threats they counter. Currently, mitigations are manually selected by cybersecurity experts (CSE) who are in short supply. To reduce labor and improve repeatability, an automated approach is needed for matching mitigations to cyber threats. This research explores the application of supervised machine learning and text retrieval techniques to automate matching of relevant mitigations to cyber threats where both are expressed as text, resulting in a novel method that combines two techniques: support vector machine classification and latent semantic analysis. In five test cases, the approach demonstrates high recall for known relevant mitigation documents, bolstering confidence that potentially relevant mitigations will not be overlooked. It automatically excludes 97% of non-relevant mitigations, greatly reducing the CSE's workload over purely manual matching.

1. Introduction

Cyber systems are ubiquitous in all aspects of society. Meanwhile, breaches to cyber systems continue to be front-page news and, despite more than a decade of heightened focus on cybersecurity, the threat continues to evolve and grow. Symantec reported that "Cyber attackers revealed new levels of ambition in 2016, a year marked by extraordinary attacks, including multi-million-dollar virtual bank heists, overt attempts to disrupt the US electoral process by state-sponsored groups, and some of the biggest distributed denial of service (DDoS) attacks on record powered by a botnet of Internet of Things (IoT) devices" [1]. Regrettably, subsequent years have not been less exciting on the cybersecurity front. The years 2017 and 2018 saw a dramatic rise in ransomware along with rapid adoption of cloud and Internet of Things technologies for which mitigations[1] are still immature [2].

To address possible impacts due to cyber threats, information system (IS) stakeholders must first assess the threats they face, then prioritize the risks. After completing the risk assessment, stakeholders must determine mitigations to counter the threats that pose unacceptably high risk. A number of threat-informed cyber risk assessment methodologies are described in the literature and in use today (e.g. [3]–[5]). At the other end of the cyber risk assessment spectrum, numerous authors have tackled the problem of mitigation optimization analysis; that is, taking a longer list of possible mitigations then prioritizing or down-selecting to a shorter list based on a set of defined objectives (e.g. [6]–[9]).

Most cyber risk assessment methods stop short of recommending mitigations. Meanwhile, optimization approaches universally assume that a starting set of possible mitigations exists on which to apply the optimization techniques. Several mitigation catalogs exist in practice today, including the Payment Card Industry Data Security Standard (PCI DSS) [10], Health Insurance Portability and Accountability Act (HIPAA) technical controls [11], and NIST 800-53 Security and Privacy Controls for Federal Systems [12]), but these do not map controls to the specific threats they counter. Likewise, a variety of threat frameworks exist in practice [13]. Of these, the

[1] In this paper we use the term "mitigation" synonymously with "countermeasure" and "security control" to mean a tool or technique that may counter a cyber threat.

URI: https://hdl.handle.net/10125/64522
978-0-9981331-3-3

Common Attack Pattern Enumeration and Classification (CAPEC) [14], and Carnegie-Mellon taxonomy of operational cyber security risks [15] contain representative mappings of mitigations to threats; however, there is currently no comprehensive source of threat-mitigation mappings.

A method to produce this starting set, the initial mapping of potential mitigations to cyber threats, is the gap this research fills. Accordingly, this research aims to devise an automated or semi-automated method for matching mitigations to cyber threats expressed as English language text documents using machine learning and text retrieval techniques in support of cyber risk assessment.

The primary contribution of this research to theory is the artifact, a novel machine learning method for matching mitigations to threats. From a practical perspective, an automated approach to matching mitigations to threats benefits all threat-informed cyber risk assessment approaches by aiding decision-making and reducing workload for CSEs whose job it is to mitigate the identified cyber threats. Moreover, an automated approach can support development and maintenance of a knowledge base to make mitigation selection more repeatable, facilitate knowledge reuse, and extend the reach of cybersecurity experts. The approach will be extensible to accommodate the continued evolution of both cyber threats and mitigations. The selection of mitigations applicable to each threat can serve as inputs into analyses of alternatives, both automated and manual, thereby bridging the gap between cyber risk assessment and final mitigation selection.

The remainder of this paper is organized as follows. In the next section, we discuss related literature. Then we discuss our research methodology, which is grounded in the principles of the Design Science Research Methodology (DSRM) [16]. Per the DSRM, we identify objectives of a solution to our stated research problem, then we discuss the design of the solution artifact drawing from the knowledge base of applicable research. We demonstrate use of the artifact to solve a real-life problem, discuss the results, and evaluate success. Finally, we discuss contributions and limitations of the present research and propose future work.

2. Literature review

During the literature review, we noted a lack of existing published research dealing specifically with automated matching of mitigation documents to cyber threats. Absent this, our literature review instead considers supportive analogous research. Casting our problem as an information retrieval (IR) problem gives rise to three veins of DSS research for investigation: (1) using classification to judge whether each item in the mitigation corpus should be included in or excluded from a particular threat's mitigation set, (2) using a retrieval model such as commonly used in search engines to enumerate mitigations ranked according to their degree of relevance to the threat, and (3) combinations of the two. We survey research for these alternatives in the next three sections.

2.1. Classification

Classification is a supervised machine learning technique in which a new item is assigned to its appropriate category by a classifier, an algorithm or model which has been trained to make such decisions after learning from training data consisting of items whose categories are already known. Classification-based document selection has been researched extensively in the context of medical systematic reviews (SRs) underpinning evidence-based medicine [17]–[24]. The process for SRs demands high recall to ensure all relevant research is considered, but is less stringent about precision, tolerating a few false positives. The document selection process for updating SRs bears stark similarities to the present research problem in which we have a large corpus of continually-evolving, highly technical cybersecurity literature and we want to present the relevant mitigation documents for a given threat while omitting those that are extraneous. Moreover, like SRs, threat-mitigation matching operates on an imbalanced corpus of candidate mitigations where only a small percentage are relevant to any particular threat. A key similarity between selecting literature for an SR and selecting mitigations for a threat may be the value judgment that high recall is more important than high precision. We elect to favor recall in the precision-recall tradeoff for the same reason this choice was made in the case of medical SRs and we assume that a few false positives can be manually screened out if necessary.

2.2. Ranked retrieval

Commonly used in search engines, ranked retrieval considers relevance between a query and a document as a matter of degree. A retrieval model assigns a relevance score to each query-document pair via a ranking function. When ordered in descending sequence by the relevance scores, those documents at the top of the list are the documents deemed to be most relevant to the query. Unfortunately, in order to make binary relevant/non-relevant decisions using ranked results, one must determine a cut-off point in the ordered list. This is a challenging problem because, in

general, the number of relevant results expected is not known in advance [25].

The Vector Space model and Latent Semantic Analysis (LSA) are similarity-based text retrieval models. LSA-based text retrieval demonstrates substantial improvement over keyword-based retrieval, in part because it accommodates language intricacies such as synonymy. Moreover, an LSA-transformed matrix is considerably smaller than the corresponding keyword-based term-document matrix, requiring only 50-150 factors compared with the hundreds or thousands of words typical of a large document corpus [26].

A few authors presented ranked retrieval research analogous to our research problem. Swanson et al. developed an automated method based on keyword searching for linking complementary sets of articles in the MEDLINE database [27]. Goldrich et al. applied search engine technology, to match cybersecurity requirements to descriptions of research projects in order to highlight research aligned with the requirements [28]. Foltz used LSA to filter a corpus looking for new relevant documents based on an existing profile of documents that had been previously deemed relevant [29].

2.3. Hybrid approaches

A few authors have explored combinations of classification and ranked retrieval techniques in text mining. For example, Manning et al. [25] and Nakamoto [30] discuss classification based on text retrieval features, such as cosine similarity and PageRank. Wiener et al. [31] utilized LSA for feature reduction to identify topics using a neural network classifier in a corpus consisting of more than 11,000 unique terms. Gee [32] described a method for classifying email as spam or not-spam using an LSA-inspired ensemble classifier implemented in three stages.

3. Research methodology and objectives

Our research is framed within the Design Science Research Methodology [16]. The DSRM is appropriate for this research because we want to create an IT artifact to solve a challenging problem for which a solution will contribute to theory and practice. The DSRM emphasizes design and evaluation rigor through building upon existing research from the literature.

In the DSRM, defining the objectives of a solution to the research problem at hand is an important predecessor to artifact design because it previews the desired end state. Objectives also provide the foundation on which to build an evaluation strategy. Our artifact will: (1) process existing English language text documents where each separately describes either a threat or a mitigation, (2) provide an automated method for recommending relevant mitigations when presented with a threat, (3) accommodate new and evolving threats and mitigations, and (4) match a high percentage of relevant mitigations for a given threat, while avoiding selection of irrelevant mitigations.

3.1. Theoretical background of the artifact

Because knowledge about threats and mitigations is largely expressed in unstructured or semi-structured text documents, our idea is to cast the threat-mitigation matching problem as an information retrieval problem, using the threat as a query and the mitigation documents as the corpus to be searched, and then build upon applicable DSS research. Applying techniques described in the literature we considered artifact designs from three categories for the threat-mitigation matcher: (1) classification drawing from medical SRs research [17]–[24], (2) ranked retrieval drawing from [27]–[29], and (3) hybrid approaches that combine techniques from ranked retrieval in conjunction with classification: drawing from [25], [30]–[32]. We ultimately arrived at a hybrid approach described in section 4.

3.2. Evaluation Approach

Evaluating the effectiveness of the artifact is a hallmark of the DSRM. Moreover, the ability to evaluate the effectiveness of a machine learning approach is crucial to ensuring that the results are not just a manifestation of chance. As we have cast our research as an IR problem, we apply IR evaluation methods to judge success. A full treatment of such methods is beyond the scope of this paper. Instead, we focus here on the methods we elected to use for evaluation of our artifact. Precision and recall, defined in Equations 1 and 2, are among the most common measures of IR effectiveness. In contexts where the objective is to correctly identify all positive instances, recall is a primary evaluation metric [33]. Specificity is the extent to which actual negative instances are classified as such [34]. It is a measure used in fields such as medicine and behavioral science to judge the effectiveness of diagnostic tests. In contexts where the objective is to rule out large swaths of negative instances, such as in medical SRs and mitigation selection, specificity can be an effective evaluation measure. The false positive rate is the probability that a non-relevant document will be retrieved. According to Raghavan et al. the “usefulness of a retrieval system is

determined to a great extent by how closely it can characterize the dichotomy" of relevant vs non-relevant documents for its intended purpose [35].

Eq. 1 $$Precision = \frac{TP}{TP + FP}$$ Eq. 2 $$Recall = \frac{TP}{TP + FN}$$

TP=retrieved/relevant; *FP*=retrieved/not relevant; *FN*=relevant/not retrieved

4. The method

In this section, we discuss our method (artifact). We introduce the data source, summarize the design iterations which lead to the artifact, discuss the machine learning techniques used, and explain the rationale for the method ultimately selected. To arrive at this method, we explored a number of designs, including various classifiers, feature sets, and feature reduction techniques. We used precision, recall, and the rate of false positives to judge the merits of each design. We emphasized recall (i.e. to present all relevant mitigations) and tolerated a few false positives.

4.1. Data source

We used version 2.11 of the Common Attack Pattern Enumeration and Classification [14] as the data source for this research. CAPEC is an existing corpus of descriptions of attack patterns (threats) expressed in English language documents. Although mitigation mapping is not the focus of CAPEC, some CAPEC attack patterns include illustrative mitigations, providing a convenient source of labeled data for our research. CAPEC is a hierarchical representation of attack patterns (i.e. threats), consisting of meta, standard, and detailed patterns. For our purpose, we focused on the standard patterns because their level of specificity is similar to that of threats in cyber risk assessments. There are approximately 125 standard threats in CAPEC. There are approximately 600 mitigation texts in the corpus. The number of mitigations mapped to each standard threat varies from 0 to about 10.

CAPEC has existed in the cybersecurity community since at least 2007. We consider the CAPEC threat-mitigation mappings to be ground truth. We recognize that data quality is key to our results and we accept CAPEC's heritage as an indicator of sufficient quality for this proof of concept research. By personal inspection, we searched CAPEC for threats which had at least a paragraph of descriptive text and about ten relevant mitigations for use as labeled data. We selected five threats and associated mitigations which were suitable test cases for our experimentation, starting with threat 49, password brute force guessing. We also added a few additional mitigation documents drawn from the Internet to bring the number of relevant mitigation instances up to about 20 for each test threat.

4.2. Summary of designs considered

We had an intuition that the best approach for one threat would also work for other threats. Figure 1 shows a summary of the precision, recall, and false positive rates (cross-validation statistics) for five designs for threat 49. The designations [C], [TR], and [H] in the design names indicate the design concept: classification, text retrieval, or hybrid. For the classification and hybrid approaches, we show the cross-validation statistics for both the relevant (R) and non-relevant (NR) classes. For the text retrieval designs, it is customary to evaluate based solely on relevant results retrieved.

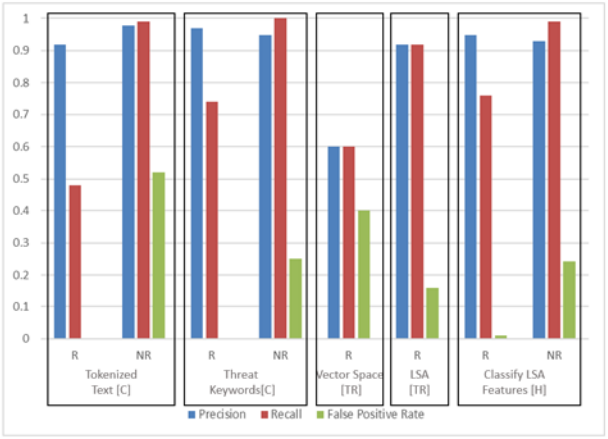

Figure 1. Summary of Designs

For our **classification** designs, we initially tested several classifiers before finally deciding on support vector machines (SVM) for reasons discussed in section 4.4. We experimented with two classification strategies, one using the full text of the mitigations (tokenized and represented via the Vector Space Model) and the other using threat keywords/phrases extracted with TextRank [36]. As shown in Figure 1, the full text model had high precision (0.92), no false positives, but unacceptably low recall (0.48) on the R class. The keyword/phrase model had high precision (0.97), no false positives, and improved recall (0.74) on the R class. On the NR class, precision and recall were very high (>0.99) for both but with high false positive rates.

We investigated two **ranked retrieval** approaches to matching, one based on the Vector Space Model and the other based on LSA. LSA outperformed the Vector

Space Model, retrieving 23 of 25 relevant items versus 15 of 25. To calculate precision and recall, we cut the ranked list off at 25 entries and applied the formulas in section 3.2. The main issue with this approach was lack of a general strategy for implementing the R vs NR cut-off point as discussed in section 2.2.

Finally, we experimented with a **hybrid** approach that combined ranked retrieval and classification. This design was ultimately the one we selected for our artifact. We discuss it in greater detail in section 4.3.

4.3. Artifact design

Our artifact is designed to leverage SVM combined with LSA. Model building is a three-step process as illustrated in Figure 2. Note that a model is built for each threat; thus, the mitigation documents input into the **indexing** stage are labeled as R/NR to the specific threat. For each mitigation text, we removed stop words, then tokenized, lower-cased, and stemmed the text. We computed a TF-IDF (term frequency-inverse document frequency) representation of the corpus, then transformed it to an LSA semantic space. The semantic space and the R/NR labels were saved for reuse.

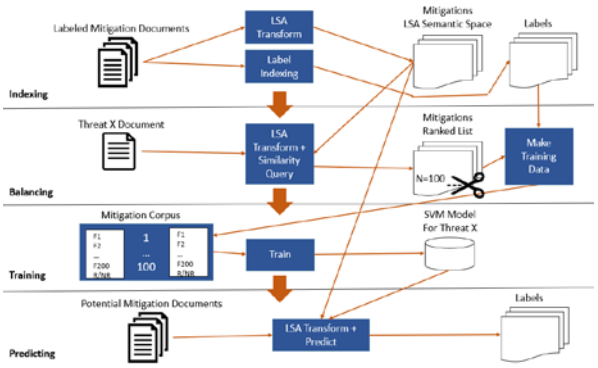

Figure 2. Artifact Design and Flow

In the **balancing** stage, we utilized LSA similarity scores as a means to balance the training data. We queried the mitigation LSA space using the full text of the threat document (tokenized, stemmed, lower-cased, and transformed to the semantic space) as the query. Then, the top 100 mitigation entries based on similarity to the threat text were retained. This balanced the data for input into training by reducing the number of NR instances.

During iterative design, we observed that the corpus was highly imbalanced in favor of NR instances. We initially experimented with methods to improve the balance, such as random undersampling of the NR class and oversampling of the R class with the Synthetic Minority Oversampling Technique (SMOTE) [37], [38]. SMOTE creates new instances of the minority class by drawing and combining features from K nearest minority neighbors. We intuited that our LSA-based balancing approach was better than simply undersampling at random because undersampling at random could drop relevant instances of which we already have too few. We also posited that ingesting the most similar entries during training, i.e. keeping the relevant entries plus the non-relevant entries that were most difficult to discriminate, aided the classifier in finding a good decision boundary.

In the **training** stage, we built an SVM classifier using the top 100 instances from the balancing stage and saved the model for later use to predict the classes of new unlabeled potential mitigations on a per-threat basis. Using LSA afforded a feature reduction from 1,500 features in the plain text to 200 LSA topics. This fits with optimal LSA dimensionality findings in Bradford [39]. We selected the low end of Bradford's range because our corpus is much smaller than his. Thus, the training data per document consisted of the 200 LSA topics augmented with the relevant or not relevant labels. We saved the models.

Finally, we utilized the saved model in the **predicting** stage to classify new potential mitigations as relevant or not relevant to the threat associated with the model. To do so, the new text was first transformed to LSA features relative to the saved LSA space. Then the saved threat-specific classifier was applied to label the LSA-transformed mitigations. A demonstration and evaluation of the method is discussed in section 5.

4.4. Design rationale

In the following paragraphs we provide insights into why we selected particular elements in the design of our artifact.

Why LSA? LSA has been shown to outperform retrieval of relevant documents from a corpus [26] when compared to keyword search because LSA accounts for inherent complexities of natural language, including the issue of synonymy, by evaluating the entire corpus for word patterns. In our experiments, we observed that LSA improved the matching of mitigations to threats over keyword-based matching.

Why Classification? Two-class classification of text documents has been successfully demonstrated in the medical SRs literature as well as in our experiments for threat-mitigation matching. Moreover, classification does not have the ambiguous cut-off problem encountered in text retrieval.

Why SVM? SVM has been shown to perform favorably for text classification, especially when the number of positive instances per category is small [40] and the feature set is large [41]. None of our

experiments with other classifiers gave us reason to go against these findings.

Why combine LSA and SVM? We used LSA in combination with SVM in our artifact for three reasons: (1) to balance the training data so as to reduce the tendency of the NR class to dominate the model (from >99.99% NR before balancing to about 75% NR after), (2) as a feature reduction technique (from >1500 features before LSA to 200 features after), and (3) because the LSA features are semantically richer, accounting for synonymy.

We crafted this design for the above reasons and selected it because of its high precision and recall and low false positive rate on the R class based on cross-validation statistics, excellent discrimination of the NR class, and the ability to fully automate construction of the "per threat" classifiers. The latter is a practical consideration related to scalability; if we operationalize this approach, we will have to build more than a few classifiers and we prefer not to do so manually. Building a classifier, i.e. the indexing, balancing, and training stages illustrated in Figure 2, can be scripted to run automatically per threat. The process for a single classifier takes less than a minute and the process for building many per-threat classifiers will scale linearly.

5. Demonstration and evaluation of results

5.1. Analysis of the text

Success in classifying textual data is heavily influenced by the characteristics of the text itself. During design we used diagnostic tools to identify the mitigations commonly misclassified for threat 49. We investigated these false positives (FP) and false negatives (FN) to better understand how they differed from the correctly classified instances. One thing the correctly classified instances had in common was that they contained text explaining how the mitigation addresses the threat. The false negatives lacked this explanatory text. The false positives fell into two categories: (1) some described password vulnerabilities but not specifically password brute force guessing and (2) others dealt with brute force guessing but not of passwords.

We hypothesized that improving the mitigation texts to include an explanation of how each one addresses the threat would improve the match results by reducing the FNs. Doing so also has practical benefits, allowing the CSE to better understand the reason a mitigation is relevant to the threat, to determine its applicability in context, and to better convey the rationale to the decision-makers who fund mitigations.

A comparison of the cross-validation statistics for models trained on the unimproved and improved text is shown in Figure 3 and discussed in section 5.3. In general, models trained with the improved text demonstrated better precision and recall in cross-validation statistics than models trained on the unimproved text.

5.2. Results for threat 49

We built an SVM classifier for threat 49 (and later for 4 other threats), inputting the top 100 instances (200 LSA-transformed features) from the balancing stage and their relevant or not relevant labels into the learning process. We saved the models for later use to predict the classes of new unlabeled potential mitigations. For threat 49, the model offered high precision (0.96) and recall (0.92) with minimal positives (1%) on the R class based on cross-validation statistics. On the NR class, precision and recall were very high (0.97 and 0.99 respectively) with an 8% false positive rate.

5.3. Extensibility to Other Threats

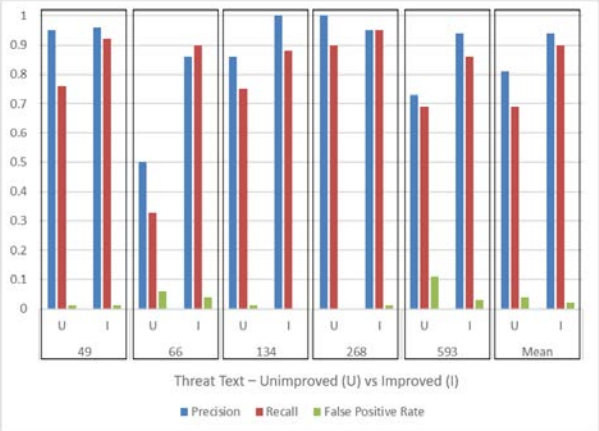

Figure 3. Comparison for 5 Threats

Having seen promising cross-validation statistics for threat 49 for the selected design, we wanted to know if this outcome would extend to other threats. We chose threats 66 (SQL injection), 134 (email injection), 268 (audit log manipulation), and 593 (session hijacking) as additional test cases. We compared cross-validation statistics for models trained for the five test cases before and after the text improvement discussed in section 5.1. Cross-validation statistics for models built for the five test threats are shown in Figure 3.

The left-most five sets of bars in Figure 3 show the precision, recall, and false positive rates for models

trained for the five test threats before and after the text improvement. In the figure, "U" and "I" stand for unimproved text and improved text, respectively. The rightmost set of bars shows the mean precision, recall, and false positive rate averaged across the five test threats. At a glance, this figure shows that the cross-validation measures are better after the text improvement, except for threat 268. Because threat 268 had 1.0 precision before the text improvement, precision declined slightly when recall went up after the text improvement.

As illustrated in Figure 3, precision is between 0.86 and 1.0 and recall is between 0.86 and 0.95 for all five test threats for improved text with false positive rate of 4% or less. Overall, although not a guarantee of generality, these classifier cross-validation statistics are favorable for the test threats and improved text.

6. Discussion

Demonstration and evaluation work together to show that the artifact effectively solves the problem. In the present research, we demonstrated and evaluated the artifact by applying instantiations of it for five test threats to predict the labels for new potential mitigation documents that were held aside and not used during training. The test data set consisted of 276 documents, 261 of which were extracted from the CAPEC mitigations for threats other than 49, 66, 134, 266, and 593. The remaining 15 were drawn from the Internet - 3 new relevant mitigations for each of the five test threats. We discuss the evaluation of the artifact in the next few paragraphs by revisiting the solution objectives (Section 3). Quantitative machine learning and IR performance metrics are shown in Figure 4.

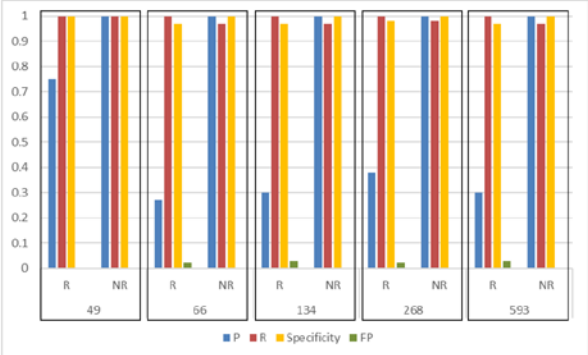

Figure 4. Test Results – Improved Text

Objective: Process existing English language text documents where each separately describes either a threat or a mitigation. By testing, we demonstrated that the method accepts English language text documents, such as from CAPEC.

Objective: Provide an automated method for recommending relevant mitigations when presented with a threat. By testing, we demonstrated that the method proposes matching mitigations for a threat.

Objective: Accommodate new and evolving threats and mitigations. The method can accept new mitigations which it will match to existing threats using a stored model trained from labeled data. The method can also accommodate new threats with the caveat that adequate labeled data consisting of known relevant mitigations for the threat would be needed so that a threat-specific model can be trained.

Objective: Match most of the relevant mitigations for a given threat while avoiding selection of irrelevant mitigations. We experimented with several artifact designs to see which attained the best performance. Thus, we needed some objective measures for comparison. During iterative design, we used 10-fold cross-validation, comparing recall, precision, and false positive rate to decide which designs to advance or leave behind. Although suitable for comparing models, cross-validation measures are not definitive for new document instances.

During the evaluation stage, we re-evaluated the classifiers on test data held aside and not used during training as is customary in machine learning evaluation. We computed the recall, precision, false positive rate, and specificity by comparing the predicted and actual labels for the test instances. As Powers points out, focusing solely on precision and recall tends to obscure a method's prowess in correctly identifying non-relevant instances [33]. This is measured using specificity, and we think it is important for threat-mitigation mapping because ruling out non-relevant instances can lead to substantial workload reduction for the CSE over purely manual matching.

Figure 4 shows the test results on the improved text for five threats. Because the training measures yielded precision and recall greater than 0.93 for the NR class, this foreshadowed excellent discernment of the NR class. Although we are most interested in the R class, the model's ability to discriminate NR instances is also a benefit. Test results for precision and recall on the NR class lived up to the promises made by the training statistics. In addition, all five models had high specificity (97-100%) on the NR class meaning at least 97% labor savings for the CSE in ruling out non-relevant mitigations when compared to totally manually matching efforts. With recall of the R class registering 1.00 on test data for all five models, we can be confident that the model will not overlook relevant mitigations. This is desirable because we do not want to obscure any relevant mitigations from the CSE.

Precision is lower than we anticipated (between 0.27 and 0.75, mean 0.40) and with this comes a few (up to 3%) false positives. In our application, a false positive means we will recommend a mitigation that does not actually counter the threat resulting in a false sense of security. In practice, this shortfall can be addressed by providing CSE screening of the recommended matches before they are made available for use in a knowledge base.

In summary, excellent recall on the R class combined with high precision (1.00), recall (>0.97), and specificity (1.00) of the NR class means the models will not overlook relevant mitigations and will accurately eliminate 97% of the non-relevant mitigations without manual intervention, greatly reducing the CSE workload when compared to purely manual matching while leaving just a few false positives for the CSE to remediate.

7. Conclusions

Mitigation selection to remediate cyber threats has heretofore been primarily a manual process done by human experts using disparate textual sources. Reliance solely on human experts brings issues of scalability, consistency, and repeatability. The ongoing shortage of cybersecurity experts combined with a burgeoning cyber threat landscape compelled us to look for a way to improve this situation.

In this research, we set out to devise a method for matching mitigations to cyber threats expressed as English language text documents using machine learning and text retrieval techniques in support of cyber risk assessment. We ultimately arrived at a matching method that achieves the stated objectives and we instantiated five examples as SVM "per threat" classifiers based on LSA. We rigorously evaluated the instantiations for our five test cases and were encouraged by the results.

This research contributes to theory by taking steps towards a novel machine learning method for automatically mapping mitigations to threats, both expressed as English language text, and by demonstrating instantiations of the method. This method fills a research gap in the cyber risk assessment literature by providing a semi-automated method to produce a starting list of possible mitigations to cyber threats which can flow into mitigation optimization techniques. It is extensible to accommodate the continued evolution of both cyber threats and mitigations, an important consideration in light of the dynamic cyber landscape. We have also demonstrated a way to improve the textual descriptions of threats and mitigations to better support automated matching.

In practice, an automated approach to matching mitigations to threats benefits all threat-informed cyber risk assessment approaches by providing a means to recommend relevant mitigations to remediate specific threats thereby aiding decision-making for IS stakeholders and CSEs. This is important because under-mitigating the actual threats provides a false sense of security, while over-mitigating is costly and wasteful.

When operationalized into a knowledge base that can save and reuse matches, this method may make mitigation selection more repeatable, facilitate knowledge reuse, reduce manual labor, and extend the reach of CSEs. The list of mitigations applicable to each threat can serve as input into analyses of alternatives, both automated and manual, enabling practitioners to leverage a large body of mitigation optimization research. Finally, because this method can respond to the evolutionary nature of cyber threats and mitigations, while also reducing the time and effort required for manual matching, it may improve overall security of cyber systems when used as part of a risk assessment and mitigation cycle by making more frequent reassessments of cyber systems more feasible.

8. Limitations and future work

For this initial work, we bounded the scope, providing ample opportunities for incremental improvements. First, the artifact we developed is based on English language documents. It would be interesting to extend it to other languages. Also, we made no effort to address redundant threats and mitigations from our corpus. In order to ingest documents from additional sources, the method should be preceded by an automated approach for dealing with duplication.

The method could be improved by exposure to more threat and mitigation sources. In addition, further analyses of the structure and semantics of threat and mitigation documents from various sources could lead to discovery of more ways to improve the document content and by extension the matching method.

We only dealt with existing threats and corresponding labeled mitigation data. This work could be extended by investigating semi-supervised learning classification techniques to build classifiers for new threats where labeled data does not yet exist. Moreover, it is possible that semi-supervised learning could also be used to improve the classifiers initially trained for existing threats by taking into account new matches that come about as new mitigation documents are added.

We focused our research on the defensive aspect of cybersecurity, starting from the threat and identifying

relevant mitigations. It is possible that our method may be applicable or extensible to the "white hat" offensive cybersecurity characterized by starting with a mitigation and then considering threats against it, such as to better understand attacker behavior or residual exposure. While we established a degree of utility by demonstrating that the artifact solves the problem for five examples, survey research to investigate the perceived utility by actual CSEs would be beneficial.

Finally, we identify several long-range goals. Improving the ways that threat and mitigation texts are written could make them more amenable to automated matching, such as by devising standards for how to rigorously express mitigations. For example, in section 5.1, we noted that improving the mitigation texts to include an explanation of how each one addresses the threat improved the match results by reducing the false negatives. Furthermore, devising a robust structure (e.g. [13], [42], [43]) to capture the intricacies of threat/mitigation relationships would offer great potential to improve the matches, helping to tease out complexities such as overlapping threats and one to many mitigation-threat mappings. This structure could be used as metadata to improve the matching models. In the long term, we envision the artifact as a component of an overarching architecture with a reusable, continually evolving, peer-reviewed knowledge base of threat-mitigation mappings with contributions coming from many sources, including threat frameworks, mitigation catalogs and vendor literature.

9. References

[1] K. Chandrasekar *et al.*, "Symantec Internet Security Threat Report," 2017.

[2] Cisco Systems, "Cisco 2018 Annual Cybersecurity Report." 2018.

[3] T. Llanso, G. Tally, M. Silberglitt, and T. Anderson, "Mission-Based Analysis for Assessing Cyber Risk in Critical Infrastructure Systems," in *International Conference on Critical Infrastructure Protection*, 2013, pp. 201–214.

[4] S. Fenz and A. Ekelhart, "Formalizing information security knowledge," *Proc. 4th Int. Symp. Information, Comput. Commun. Secur. - ASIACCS '09*, p. 183, 2009.

[5] National Institute of Standards and Technology, "National Institute of Standards and Technology Special Publication 800-30 R1: Guide for Conducting Risk Assessments," 2012.

[6] O. F. El-Gayar and B. D. Fritz, "A web-based multi-perspective decision support system for information security planning," *Decis. Support Syst.*, vol. 50, no. 1, pp. 43–54, 2010.

[7] T. Llansó, M. W. McNeil, and C. Noteboom, "Multi-Criteria Selection of Capability-Based Cybersecurity Solutions," *Hawaii International Conference on System Sciences*. 2019.

[8] T. Llanso, "CIAM: A Data-driven Approach for Selecting and Prioritizing Security Controls," *IEEE Int. Syst. Conf. SysCon 2012*, vol. March, pp. 1–8, 2012.

[9] J.-J. Lv, Y.-S. Zhou, and Y.-Z. Wang, "A Multi-criteria Evaluation Method of Information Security Controls," *2011 Fourth Int. Jt. Conf. Comput. Sci. Optim.*, pp. 190–194, 2011.

[10] PCI Security Standards Council, "PCI DSS Quick Reference Guide - Understanding the Payment Card Industry Data Security Standard Version 3.2.1," 2018.

[11] Centers for Medicare and Medicaid Services, "HIPPA Security Standards: Technical Safeguards," 2007.

[12] National Institute of Standards and Technology, "SP-800-53Ar4 Security and Privacy Controls for Federal Information Systems and Organizations Security and Privacy Controls for Federal Information Systems and Organizations," 2014.

[13] S. Launius, "Evaluation of Comprehensive Taxonomies for Information Technology Threats," *SANS Information Security Reading Room*. SANS, 2018.

[14] MITRE, "Common Attack Pattern Enumeration and Classification," 2017. [Online]. Available: https://capec.mitre.org/index.html. [Accessed: 04-Feb-2018].

[15] J. J. Cebula, M. E. Popeck, and L. R. Young, "A Taxonomy of Operational Cyber Security Risks Version 2," *Carnegie-Mellon Univ Software Engineering Inst*. pp. 1–47, 2014.

[16] K. Peffers, T. Tuunanen, M. a. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, Dec. 2007.

[17] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews," *BMC Bioinformatics*, vol. 11, 2010.

[18] T. Bekhuis, E. Tseytlin, K. J. Mitchell, and D. Demner-Fushman, "Feature engineering and a proposed decision-support system for

systematic reviewers of medical evidence," *PLoS One*, vol. 9, no. 1, pp. 1–10, 2014.
[19] I. Shemilt *et al.*, "Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews," *Res. Synth. Methods*, vol. 5, no. 1, pp. 31–49, 2014.
[20] O. Frunza, D. Inkpen, and S. Matwin, "Building systematic reviews using automatic text classification techniques," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 303–311.
[21] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O'Blenis, "A new algorithm for reducing the workload of experts in performing systematic reviews," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 4, pp. 446–453, 2010.
[22] G. Kontonatsios *et al.*, "A semi-supervised approach using label propagation to support citation screening," *J. Biomed. Inform.*, vol. 72, pp. 67–76, 2017.
[23] A. M. Cohen, W. R. Hersh, K. Peterson, and P. Y. Yen, "Reducing workload in systematic review preparation using automated citation classification," *J. Am. Med. Informatics Assoc.*, vol. 13, no. 2, pp. 206–219, 2006.
[24] J. Liu, P. Timsina, and O. El-Gayar, "A comparative analysis of semi-supervised learning: The case of article selection for medical systematic reviews," *Inf. Syst. Front.*, pp. 1–13, 2016.
[25] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*, Online edi. Cambridge University Press, 2009.
[26] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[27] D. R. Swanson and N. R. Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," *Artif. Intell.*, vol. 91, no. 2, pp. 183–203, 1997.
[28] L. Goldrich, S. Hamer, M. McNeil, T. Longstaff, R. Gatlin, and E. Bello-Ogunu, "REQcollect: Requirements collection, project matching and technology transition," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 4887–4894, 2014.
[29] P. W. Foltz, "Using latent semantic indexing for information filtering," *Proc. ACM SIGOIS IEEE CS TC-OA Conf. Off. Inf. Syst.*, vol. 11, no. 2–3, pp. 40–47, 1990.
[30] Y. Nakamoto, "A Short Introduction to Learning to Rank," *IEICE Trans. Inf. Syst.*, vol. E94-D, no. 1, pp. 1–2, 2011.
[31] E. Wiener, J. O. Pedersen, and A. Weigend, "A neural network approach to topic spotting," *Proc. 4th Annu. Symp. Doc. Anal. Inf. Retr.*, pp. 317–332, 1995.
[32] K. R. Gee, "Using latent semantic indexing to filter spam," *Proc. 2003 ACM Symp. Appl. Comput. - SAC '03*, p. 460, 2003.
[33] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation." School of Informatics and Engineering - Flinders University, 2007.
[34] D. G. Altman and J. M. Bland, "Diagnostic tests 1: sensitivity and specificity," *BMJ*, vol. 308, no. June, p. 1552, 1994.
[35] V. V Raghavan, G. S. Jung, and P. Bollman, "A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 205–229, 1989.
[36] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," *Proc. EMNLP*, vol. 85, pp. 404–411, 2004.
[37] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
[38] P. Timsina, J. Liu, and O. El-Gayar, "Advanced analytics for the automation of medical systematic reviews," *Inf. Syst. Front.*, vol. 18, no. 2, pp. 237–252, 2016.
[39] R. B. Bradford, "An empirical study of required dimensionality for large-scale latent semantic indexing applications," *Proceeding 17th ACM Conf. Inf. Knowl. Min. - CIKM '08*, p. 153, 2008.
[40] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. 1998.
[41] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings ECML*, 1998, pp. 137–142.
[42] J. Ferdinand and R. Benham, "The Cyber Security Ecosystem : Defining a Taxonomy of Existing, Emerging and Future Cyber Threats." pp. 1–39, 2017.
[43] Enclave Security, "Open Threat Taxonomy." 2015.