Towards Improved Network Security Requirements and Policy: Domain-Specific Completeness Analysis via Topic Modeling

Jane Huffman Hayes, Jared Payne, Emily Essex, Kelsey Cole, Joseph Alverson
Computer Science, University of Kentucky, Lexington, Kentucky, USA
{hayes, jared.payne, Emily.essex, Kelsey.cole, joseph.alverson}@uky.edu

Alex Dekhtyar, Dongfeng Fang, Grant Bernosky
Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo
{dekhtyar, dofang, gbernosk}@calpoly.edu

***Abstract* - Network security policies contain requirements – including system and software features as well as expected and desired actions of human actors. In this paper, we present a framework for evaluation of textual network security policies as requirements documents to identify areas for improvement. Specifically, our framework concentrates on completeness. We use topic modeling coupled with expert evaluation to learn the complete list of important topics that should be addressed in a network security policy. Using these topics as a checklist, we evaluate (students) a collection of network security policies for completeness, i.e., the level of presence of these topics in the text. We developed three methods for topic recognition to identify missing or poorly addressed topics. We examine network security policies and report the results of our analysis: preliminary success of our approach.**

***Index Terms*—Requirements quality, completeness, machine learning, network security, empirical evaluation**

I. Introduction

With our world relying more than ever on virtual communication due to the current COVID-19 crisis, and with universities switching to distance education, campus networks, as the conduit for connecting faculty and students in virtual classrooms, have become even more vital. Campus network security is thus of paramount importance. This study views network security policies, as requirements documents, and applies the requirements engineering methodology to their analysis. Such an approach allows us to understand a variety of important properties of the network security policy documents, that are typically well understood in the context of requirements specifications: completeness, consistency, understandability, to mention a few [1].

We choose network security policy completeness for our initial study. This work limits the analysis to Acceptable Use Policies (AUPs) for university campus network services[1]. Specifically, we assess *de-facto completeness* of an AUP document. Informally, we define a network policy document to be *complete* if the document describes *all issues necessary* for a wireless network acceptable use policy, and if each issue is discussed to a sufficient level of detail.

But what issues/topics *are necessary* for an AUP? This question can be answered in two ways. Prior work [2,3,4] concentrated on specifying what topics an AUP ***shall*** contain in a proscriptive manner. At the same time, a specific AUP can ignore one or more topics from such lists.

We adopt a different, complementary approach. We assume that while individual documents in a large AUP collection may contain missing topics, any *truly necessary topic* would be covered in an abundance of documents in the collection. We thus ask two questions:

- Given a collection of AUPs what is the *de-facto* set of topics that constitutes a *complete campus* AUP document? That is - what topics are *covered in abundance* in a collection of AUPs?
- Can we determine automatically how well a specific topic is covered in a given AUP document?

Research approach. We undertake a feasibility study to determine if an automated unsupervised topic modeling technique called Latent Dirichlet Analysis (LDA) [5] (a) can properly identify true aspects (semantic topics) of network security policies, and (b) can predict presence or absence of topics in AUP documents. LDA is a method of analysis of textual documents. It assumes that each unit of text (sentence, paragraph, chapter, etc.) in each document in a given collection has been generated to address a specific latent "topic": one out of a given number of topics. Each topic yields its own probability distribution over the words, and LDA uses maximum likelihood estimation to learn the best probability distributions for each latent topic. Given a document, LDA provides information about the percentage of words that it attributes to each topic (so called "topic loadings"). It also explicitly surfaces the probability distribution of words for each learned topic. Because LDA looks to assign topics to cohesive

[1] This is due to (a) expected/desired uniformity of these documents, and (b) their availability for collection. In the future we plan to extend our study to network security policies of other organizations.

units of text, the latent topics it learns often correspond to semantically meaningful themes/topics (that experts can name). In our analysis, we combine the automated construction of topics for an AUP collection using LDA, with manual examination of word clouds that LDA associates with each topic. This method allows us to convert some of latent topics surfaced by LDA into actual network security policy topics/issues discussed in prior work [2,3,4], such as, e.g. "network access" , "violations of policy", or "electronic communications". We then look at topic loadings returned by LDA for each AUP. We interpret higher topic loading values as evidence of substantial coverage of the topic in a document, and lower topic loading values as evidence of insufficient coverage or absence of coverage of a topic.

To test the predictions to topic coverage made by the LDA-based analysis described above, we picked a sample of our AUP documents and asked a group of experts (students in a wireless security course) to read them and determine how well each topic from the list we collected is represented.

We compare how well LDA estimates absence/presence of specific topics in AUPs to two baseline methods, the *frequency count method* and the *section headings* method.

The *frequency count* uses the output of the LDA process. Specifically, it looks at the top 25 words associated by LDA with each topic. For each document and each topic, we count the total combined number of occurrences of top 25 words. This count is then normalized by the overall size of the document and compared to the mean normalized count for the collections. Documents where the normalized frequency of words representing a topic is 25% below the mean are considered to have insufficient coverage of the topic.

The *section headings method* is straightforward: most documents "advertise" topics discussed in them via section headings. The method determines for each AUP what constitutes a heading, and extracts all words found in the headings. If a phrase/sentence containing words from a specific topic is found in a heading, we declare the topic as sufficiently represented.

The rest of the paper is organized as follows. Section II provides related work. Section III details data collection and the topic extraction process. Section IV discusses evaluation and Section V shows the results of our validation study. Section VI provides future work.

II. Related work

Doherty et al. define a list of topics that AUPs *should cover*: access management, acceptable behavior, unacceptable behavior, license compliance, roles and responsibilities, user monitoring, sanctions for policy violations, and policy management [2]. Of the AUPs studied, most topics were only covered in about half of them [2]. Our analysis of university AUP documents indicates that this is a partial list of topics, and that not all AUPs follow this list. Gallagher et al. also defined seven "features" that acceptable use policies should have: aims and objectives, eligibility, scope, illegal use, unacceptable use, service commitments, and user commitments" [5].

Our work automated the process of defining the topics an AUP should cover to be considered "complete." Yono et al. applied a Joint Sentiment Topic (sJST) model, joining LDA and JST, by combining text and numerical data to determine a sentiment of market (risk on or risk off). They used their results to predict foreign exchange market price movement [6]. Our work applies similar topic modeling methods to network security policies. Malhotra et al. used a semi-automated approach to determine the completeness and consistency of security features in general software [2]. In contrast to this work, our study focuses on network security policies, and, even more domain specific, AUPs.

III. Topic Extraction

For this study, we collected over 200 AUP documents from the web sites of US Universities. The documents were collected in a PDF format, then converted into plain text using an off-the-shelf PDF conversion Python library. This process converts the document's text as a whole; no unnecessary text such as dates, headers, or page numbers are removed. Several documents failed to convert to a meaningful quantity of recognizable text, typically because of a text encoding error although encrypted files and files containing malformatted text were also causes of errors. These documents were excluded from further analysis. The collected documents present a fairly broad, albeit homogenous (limited to universities), set of policies. The final data set that was analyzed contained 231 documents. Individual policy documents in our dataset ranged in size from one-two pages to 15-20 pages[2].

We used LDA (see Section I) to analyze our collection of AUPs. We obtained: (1) a list of most relevant words for each latent topic – we kept the size of the list to 25 words, and (2) a vector of topic loadings for each individual document. We ran LDA several times to calibrate the number of topics and the level of a "semantic unit" (sentence/paragraph) considered to belong to a single topic. For several LDA runs we examined the lists of most relevant words for each topic.

Multiple co-authors manually examined the list of relevant terms independently, and proposed names (and semantics) for each latent topic. After that, we selected on LDA run which identified the best and the most diverse collection of topics to be used in the rest of the study[3]. The "winning" LDA run used sentences as semantic units and produced 20 latent topics. Some of the 20 latent topics (nine total) did not represent any meaningful themes associated with network use policies. The remaining 11 topics, are listed in Table I with some of the "indicator" terms and represented clearly identifiable themes.

For the purposes of the rest of our preliminary study, we consider the topics indicated in Table I to be a complete list of topics/themes/issues that a wireless network AUP must have.

[2] We are working on making the dataset we assembled (including some secondary artefacts we generated) available to the research community.

[3] We fully recognize the human factors that went into this decision as a threat to validity and discuss it in Section IV.C.

IV. EVALUATION

This section presents the study design, measures, and threats to validity.

A. Study Design

The research question is evaluated using student review treated as expert and ground truth-yielding. A total of 23 policy documents randomly selected from our collection of 231 documents were read and evaluated by 25 students from an undergraduate wireless security course at CalPoly[4]. The students have sufficient understanding of security requirements of wireless networks since they were studying wireless security, which covers the topics of wireless vulnerabilities and security requirements. The students also studied security analysis based on different scenarios, such as sensor networks and cellular networks. The students were asked to indicate the level of coverage (absent, insufficient, sufficient, over-represented) for each of the 11 identified topics. In our study, we take the first two levels of coverage to mean that a topic is *insufficiently covered* and the last two levels of coverage to mean that a topic is *sufficiently covered* in a document. We consider participants in this exercise to be experts and treat their responses as ground truth.

A. Measures

In our evaluation, we look at the accuracies of predicting whether a topic is sufficiently or insufficiently present in the document. We utilize both the full accuracy measure (percentage of correct predictions for both classes) as well as *positive recall, precision and f1 measures* - that is, recall, precision and f1 measure for the "sufficiently present" prediction, and *negative recall, precision and f1 measures:* the recall, precision and f1 measure for the "insufficiently present" prediction. In general, we want our methods to accurately indicate topics that are *missed* in the policy, to study completeness. Thus, we examine and, where appropriate, optimize for the highest possible values of the negative measures.

B. Threats to Validity

Our work may have suffered from selection bias. To mitigate this threat, we randomly selected policies for the students to evaluate. A separate selection bias occurred when the co-authors examined multiple LDA models and their latent topics and selected one model believed to have the best coverage of topics. The LDA runs examined but rejected by the co-authors resulted in several latent topics mapping to a single "real" issue. It is possible that the co-authors examined insufficient number of LDA runs -- this will be mitigated in the future work. Threats to external validity included the limited number of policies that were evaluated in order to validate our work. It is possible that our topic list does not fully express all topics that should be in a policy, so there are still possible construct validity threats.

V. RESULTS AND DISCUSSION

We confirmed all 11 topics identified by LDA and human experts' method via an independently conducted literature survey, which did not uncover any topics of significance that were missed from our list. Table I compares the topics learned to those in literature [2,3,4]. The topics in the literature were more general than the topics identified by LDA but could be mapped. For example, the learned topics "email accounts," 'wireless network access," "electronic communications," "network access," and "authentication" fall under the broader literature topic "Service Commitments." All general topics were mapped to the learned topic list in this manner.[5]

Results. We examined the accuracy of the LDA as well as the two baseline method predictions for each of the 11 topics individually. We report the accuracy numbers in Table 1. For the LDA method, we considered a topic expressed in a document if its loading exceeded a threshold level that optimized *negative F1 measure.* Specifically, we selected the smallest threshold from a grid of thresholds with values ranging from 0.001 (0.1% of the text) to 0.249 (24.9% of the text) with a step of 0.004, that yielded the best negative F1 for each topic.

Our results can be briefly summarized as follows. Seven of 11 topics had best possible negative F1 of 0.75 or above; two additional topics (POLICY ADMINISTRATION AND RETENTION OF RECORDS and APPROPRIATE USE OF RESOURCES) had best possible negative F1 values of 0.687 and 0.615, while the remaining two topics (VIOLATIONS OF POLICY and COPYRIGHTED MATERIALS….) had negative F1 scores of 0.47 and 0.4, respectively. The threshold values at which best negative F1 scores were observed varied from 0.005 (0.5% of all text) for the VIOLATIONS OF POLICY topic to 0.245 (24.5% of all text) for INCIDENTS AND INCIDENT RESPONSE, with a mean of 0.13, a median of 0.15, and a standard deviation of 0.092.

The frequency method outperformed the LDA method for two topics (EMAIL ACCOUNTS and NETWORK ACCESS). The headings method did significantly better than the LDA method on two topics: INCIDENTS and INFORMATION RETENTION, and tied it on three more topics (See Table I). Overall, though, the LDA method has done better than the baseline techniques.

B. Discussion

We can make several positive observations about our proposed approach of using LDA for determining the completeness of a network policy document. Foremost, even in our very simplified and imperfect topic modeling exercise, LDA properly identified a set of topics that can be successfully used to judge the completeness of network policy documents.

[4] We selected 30 documents initially and assigned 20 students to examine a single document, and five students to examine two short documents. Several students opted out of participation in the exercise. As a result, the total number of documents for which we obtained ground truth information is 23.

[5] The topics are labeled by number in the list and their corresponding source (so (1)[4] is first in the list in the related works section from paper [4]).

The LDA completeness method shows definite promise: higher accuracy potential of predictions on sufficient/insufficient topic coverage than the baseline methods. We are able to tune the performance of the method to optimize any of the collected measures, and it seems that optimizing negative F1 does not hurt the overall accuracy too much (does not come at the cost of losing correct positive predictions). Due to space restrictions, a set of figures and tables can be found here: https://networkrequirements.wixsite.com/appendix.

TABLE 1. LEARNED TOPIC LIST, KEY TERMS, ACCURACY OF THE THREE METHODS ON TOPIC BASIS, COMPARISON TO LITERATURE SURVEY TOPICS [2,3,4]

Topic	Name	Keywords	Citations	LDA	Heading	Freq
Topic 1	**Violations**	Users, violation, violations, access, law . . .	(3)[3], (5) [3], (2)[4], (7)[2]	0.783	0.652	0.409
Topic 2	**Appropriate use of resources**	Responsible, appropriate, resources, use, electronic. . .	(7)[3], (2) [4], (1) [4] (2)[2], (5)[2]	0.652	0.652	0.636
Topic 3	**Incidents**	Incident, response, security, procedures, integrity . . .	(2)[4], (1)[4], (7)[2]	0.522	0.870	0.455
Topic 4	**Network access**	Computer, network, account, access, permission . . .	(5)[3], (2)[3], (1)[2], (3)[2[	0.739	0.478	0.773
Topic 5	**Policy administration and records retention**	Records, policy, retention, president, vice . . .	(1)[3], (6)[3], (6)[2], (8)[2]	0.826	0.696	0.636
Topic 6	**Authentication**	Password, administrator, authentication, user, secure . . .	(6)[3]	0.696	0.435	0.500
Topic 7	**Wireless network access**	vpn, wireless, campus, network, use . . .	(2)[3], (6)[3], (1)[2]	0.783	0.565	0.727
Topic 8	**Information retention/policy compliance**	security, data, policy, protect, compliance . . .	(6)[2], (1)[5]	0.348	0.826	0.455
Topic 9	**Email accounts**	Account, password, address, email, messages . . .	(6)[3], (2)[3], (5)[2], (3)[2]	0.609	0.609	0.682
Topic 10	**Copyright/Copyright violations**	Copyright, copyrighted, downloading, material, content . . .	(5)[3], (4)[3], (2)[4], (3)[2], (4)[2]	0.696	0.348	0.545
Topic 11	**Electronic communication**	Mail, email, communications, faculty, staff . . .	(6)[3], (2)[3]	0.696	0.696	0.500

VI. FUTURE WORK

This is a very preliminary study with a number of major limitations which we hope to correct in subsequent in-depth study of our proposed approach. First, our ultimate goal is to analyze an unseen network policy and predict correctly what topics are insufficiently covered in it. Our results concentrated on topic-by-topic, rather than document-by-document, accuracy analysis of the LDA method due to the need to properly calibrate thresholds for determining whether a topic is sufficiently represented. Document-to-document comparison remains for future work. Our preliminary study did not include automated threshold selection methods needed for building an automated network policy analysis tool. This will be addressed in the subsequent study.

ACKNOWLEDGMENT

We thank NSF for partially funding this work under grant CICI 1642134. We thank the class of CalPoly students for undertaking the study. We thank Dr. Tingting Yu for helpful discussions on experimental design.

REFERENCES

[1] Alan M. Davis. 1990. Software requirements: analysis and specification. Prentice Hall Press, USA.

[2] R. Malhotra, A. Chug, A. Hayrapetian and R. Raje, "Analyzing and evaluating security features in software requirements," 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), Noida, 2016, pp. 26-30.

[3] Gallagher, C., McMenemy, D. and Poulter, A., "Management of acceptable use of computing facilities in the public library: avoiding a panoptic gaze?", Journal of Documentation, Vol. 71 No. 3, pp. 572-590. 2015.

[4] Robinson, Elaine, and David McMenemy, "'To Be Understood as to Understand': A Readability Analysis of Public Library Acceptable Use Policies." Journal of Librarianship and Information Science, 2019.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation," 3(Jan):993-1022, 2003.

[6] K. Yono, K. Izumi, H. Sakaji, H. Matsushima and T. Shimada, "Extraction of Focused Topic and Sentiment of Financial Market by using Supervised Topic Model for Price Movement Prediction," 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), Shenzhen, China, 2019, pp. 1-7.