Topic Modeling of Significant Concepts and Terminologies in Cybersecurity and Data Science and Their Potential Guidance to Seed Future Research Direction

Dr. Tamir BECHOR
Center for Information Systems & Technology, Claremont Graduate University
Claremont, California 91711, USA

Bill JUNG
Center for Information Systems & Technology, Claremont Graduate University
Claremont, California 91711, USA

ABSTRACT

Arguably, the two domains closely related to information technology recently gaining the most attention are 'cybersecurity' and 'data science'. Yet, the intersection of both domains often faces the conundrum of discussions intermingled with ill-understood concepts and terminologies. A topic model is desired to illuminate significant concepts and terminologies, straddling in cybersecurity and data science. Also, the hope exists to knowledge-discover under-researched topics and concepts, yet deserving more attention for the intersection crossing both domains. Motivated by these, this study attempts to take on a challenge to model cybersecurity and data science topics clustered with significant concepts and terminologies, grounded on a text-mining approach based on the recent scholarly articles published between 2012 and 2018. As the means to the end of modeling topic clusters, the research is approached with a text-mining technique, comprised of key-phrases extraction, topic modeling, and visualization. The trained LDA Model in the research analyzed and generated significant terms from the text-corpus from 48 articles and found that six latent topic clusters comprised the key terms. Afterwards, the researchers labeled the six topic clusters for future cybersecurity and data science researchers as follows: Advanced/Unseen Attack Detection, Contextual Cybersecurity, Cybersecurity Applied Domain, Data-Driven Adversary, Power System in Cybersecurity, and Vulnerability Management. The subsequent qualitative evaluation of the articles found the LDA Model supplied the six topic clusters in unveiling latent concepts and terminologies in cybersecurity and data science to enlighten both domains.

Keywords: Cybersecurity, Data Science, Topic Modeling, Text Mining, Research

1. INTRODUCTION

1.1 General Background:

Like many practical domains, cybersecurity is seeing ever-increasing use of data science, such as machine learning (ML), data mining (DM), and artificial intelligence (AI). As an exemplar, Chen et al. [1] summarized the applications, data, analytics, and impacts of "BI&A (Business Intelligence and Application)" in the security and public safety domains.

Both cybersecurity and data science are monumental in terms of significance and popularity, respectively. Armerding [2] said "over the past decade," cybersecurity has become as important as "military or law enforcement security". Related to such claim, former U.S. President Barack Obama stressed that "cybersecurity is one of the most important challenges we [the U.S.] face as a Nation, and for more than seven years he has acted comprehensively to confront that challenge" [3]. Then, he put effort into action by "directing his Administration to implement a Cybersecurity National Action Plan (CNAP)" [3]. Moreover, the EU (European Union) consider "the [cyber-] security and stability of the net, as well as the integrity of data flows" tremendously significant, as "the digital age" provides enormous benefits in "wealth, knowledge and freedom" [4].

On the other hand, "defined as an interdisciplinary field in which processes and systems are used to extract knowledge or insights from data" [5], data science is growing huge popularity as firms are recognizing its potential and impacts to their operations [5]. If job demand equates popularity, the popularity of data science can be gauged by the immense demand for data scientists, as [6] calls "Data Scientist: The Sexiest Job of the 21st Century".

However, there are three potential issues to consider. First, the cybersecurity community needs to understand concepts and terminologies of data science applied in the domain. Secondly, both domains would want to avoid the inadvertence of overlooking significant concepts. Lastly, because popular terms tend to attract more attention, both need to circumvent lost opportunities to the less popular constructs worth another looks. Due to these, the community needs to shed light on topic models, projecting significant, related concepts. This will render to the community a summary view of most researched topics or phenomena associated with both domains from recent scholarly literature. It will also become a potential seed to guide information systems and technology (IST) researchers for future research and to ultimately enlighten them to contribute to the existing body of knowledge across both domains.

Actuated by the above, the purpose of this research is to model topics of cybersecurity and data science clustered with significant concepts and terminologies discovered using a text-mining method based on recent scholarly articles published between 2012 and 2018.

1.2 Theoretical Background: Latent Dirichlet Allocation (LDA) Method

Blei et al. argued the necessity of considering mixture models representing words and documents' exchangeability while extending the de Finetti theorem in that "any collection of exchangeable random variables has a representation as a mixture distribution - in general an infinite mixture" [7]. Then, they demonstrated to "capture significant intra-document statistical structure via the mixing distribution" [7]. Furthermore, they

argued that, while their paper concentrated on "bag-of-words", the LDA methods were usable to larger bodies of text, such as paragraphs [7].

The current paper's research selects LDA as the approach, instead of other methods, for reducing dimensionality of text collections in topic modeling form because of its simplicity and "useful inferential machinery in domains involving multiple levels of structure" [7], such as the text-corpus of the current research.

1.3 Topic Modeling

As the dimensionality of applied concepts and terminologies from data science increases and becomes more complex as applied in the cybersecurity, topic modeling produces profound benefits. [8] reasoned, with more available information, finding and discovering of needed information become harder and also argued for new devices in organizing, searching, and comprehending enormous information volumes. Also, [8] summarized topic modeling as approaches to organize, understand, search, and summarize a large corpus of digital texts automatically; additionally, this approach can discover the *hidden* themes pervading the collection. In explaining Topic Models Vs. Unstructured Data, [9] posited topic models provide potent approaches in exploring and understanding otherwise disorderly information and in discovering latent structures in documents and laying down relations among them.

With the urgency of topic models of cybersecurity and data science and the aforesaid topic modeling benefits, by conducting this scientific research approach using text-mining, we aim to strengthen the aforementioned justifications for research and to contribute to the body of knowledge.

1.4 Research Problems:

This research ultimately aims to address the following primary question:

- In recent scholarly articles on the topic of 'cybersecurity' and 'data science' published between 2012 and 2018, what have been the significant terminologies and other related nomenclature most frequently mentioned around these terminologies?

With the above primary research question raised, the secondary research questions of the current study are as follows:

- How distinguishable are clusters from the topic modeling result? Are they clearly separable, or do they considerably overlap?
- How reliable is the result of document-clustering into the topic models of cybersecurity and data science?

The subsequent organization of this paper is as follows: Section 2 reviews the relevant literature using topic modeling approaches in the cybersecurity and data science domains. Then, section 3 describes the research methods, followed by section 4 describing the research results. After these results are described, the six topic models resulted from the analysis are evaluated in section 5. Finally, section 6 discusses the research implications, and the paper ends with a conclusion in section 7.

2. LITERATURE REVIEW

Through the search engine Google Scholar [10], we searched relevant literature using the following advanced search terms:

- Entered cybersecurity or cyber security for the field, all of the words
- Entered "topic modeling" for the field, with the exact phrase
- Selected the choice, "anywhere in the article", for the field "where my words occur"
- Entered 2012 and 2018 for the field: "Return articles date between"

After the search hits, we skimmed the title, abstracts, and sections of the articles and selected twelve suitable studies for the literature review.

In chronological order of publication year, Table 1 below lists and summarizes the reviewed articles and breaks down their topic modeling approach, key topics researched, and gap analysis comparing the articles in question to the existing research.

Table 1. Summary of Literature Review and Comparison with the Research in the Current Article

Article	Topic Modeling Approach	Key Topic(s) Researched	Difference Compared to Current Research
[11]	Quantitative security risk assessment model using vulnerability scanners and the impact score and frequency values based on the empirical data derived from NVD	Exploration of a software product evaluation method	Methodological difference in topic modeling
[12]	WL-LDA for better obtainment of results via vector space generation on themes and HT-SVM for better leveraging of the prior knowledge of vulnerability distribution	Automated classification of vulnerability through ML	Key topical difference; Methodological difference in topic modeling – using WL-LDA and HT-SVM to extend LDA
[13]	LDA to knowledge-discover from big data	Intrusion detection	Key topical difference
[14]	The CS Gibbs sampling algorithm to apply the probabilistic generative model based on LDA	Cybercriminal networks from online social media	Methodological difference in topic modeling
[15]	LDA to cluster topics related with IP address via SSH authentication logs	Classifying SSH logs to identify and differentiate brute-force attackers from normal users	Key topical difference

[16]	LDA as the main method to extract topic clusters and to understand the hacker assets	Hacker assets, such as source code postings, tutorial postings, and postings with attachments	Key topical difference
[17]	LDA feature selection and DM algorithms	Detection of malware	Key topical difference; Methodological difference in topic modeling – the use of additional DM algorithms
[18]	LDA to cluster toipcs, DTM to discover trending topics, and ATM to identify the key hackers in each topic cluster	Exploration of key hackers and cyber threats in Chinese hacker communities	Key topical difference; Methodological difference in topic modeling – using DTM and ATM to extend LDA
[19]	LDA to analyze topic model of information security issues of Korea, the US, and China	Analysis of information security awareness	Key topical difference
[20]	Nonparametric supervised topic model (NSTM)	Identification of high quality carding services in the supply chain of the underground economy and adapting the heterogeneity and precariousness of cybercriminals' customer reviews	Methodological difference in topic modeling
[21]	LDA to identify topic clusters of hacker code from online hacker forums	Cyberthreat intelligence and malware analysis	Key topical difference
[22]	Clustering, topic modeling, and LDA algorithm to find comparison and contrast among the NCSs and latent topics discovered in the NCSs	The 60 national cybersecurity strategies NCSs to compare and contrast among the NCSs and implicit topics found	Key topical difference

Note. The key topical difference in the above table means the main topics of research were different in comparison to the current research.

According to the literature reviewed, the preponderant number of research used LDA as the topic modeling approach, with varying key topic discussions. However, the literature review indicates that there has been no attempted study to data-mine recent scholarly articles with main discussion topics of 'cybersecurity' and 'data science'. Therefore, we believe this is the first attempt through systematic research to elucidate latent themes of cybersecurity and data science from the recent scholarly literature in the form of topic modeling using LDA.

3. METHODS

We searched scholarly articles published between 2012 and 2018 with the two main topics: 'cybersecurity' and 'data science'. Initially, few relevant articles were found with the search terms "cybersecurity" and "data science" via widely-known databases, such as ACM Digital Library, Web of Science, and ABI/Inform. Then, we searched Google Scholar [10] using the two terminologies and after finding relevant articles we searched more using a snowball approach. At the end of the snowball search process, we found a total of 50 scholarly articles. However, after validating the relevancy, we excluded two articles, as we mistakenly included one article with its publication year out of range and questioned the fitness of the other article for limited contribution to our topic modeling effort. Therefore, a total of 48 articles became the subsequent text-mining's sources.

To provide the readers the conceptual roadmap of the research, Figure 1 below presents the overall process flow of research methods.

Figure 1. Flowchart of Overall Research Methods

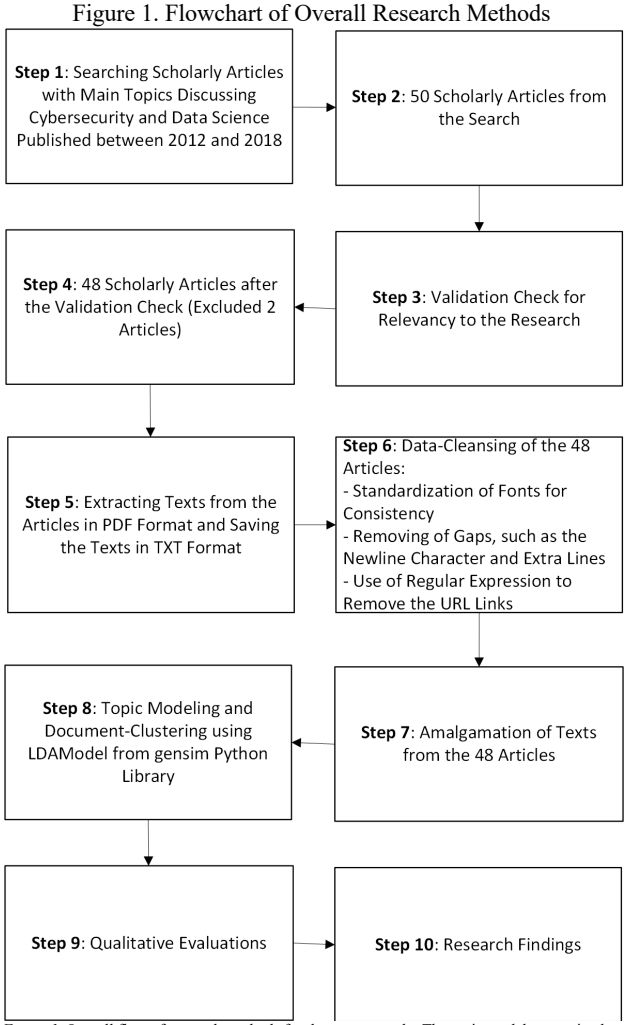

Figure 1. Overall flow of research methods for the current study. The topic model was trained to result most separable topic clusters.

3.1 Text-Mining with Key-phrases Extraction, Topic Modeling, and Document-clustering

After gathering the 48 scholarly articles in PDF format, we used a software converter, PDFMiner [23] to extract and convert texts from the PDFs into plain text (.TXT) to process and text-mine the text in Python 2.71 environment. Then, we pre-processed the text data from the articles as described in Figure 1.

We referenced the Python notebook's code example featured at Kaggle [24] to run the text-mining with topic modeling, document-clustering, and visualization. We prepared and stored the text data from the plain text (.TXT) files, consisting of the 48 articles' titles, author(s), and main text-corpus, into a DataFrame of pandas Python library [25]. Then we tokenized the corpus, removed numbers, lemmatized the words in the corpus, computed bigrams and trigrams, removed rare and common tokens, and lastly vectorized the text data. To use and train the LDAModel of Gensim [26], the following parameters were set:

- Number of topics: 6
- Chunk size (size of the documents looked at every pass): 10
- Passes: 50 (number of passes through documents)
- Iterations: 400

These parameters were chosen to train Gensim's LDAModel after experimenting with varying numbers of topics, ranging from 3 to 10 topics; 6 seemed to separate the topics well, as a too small number, such as 3, resulted in clusters of too small numbers while a too big number, such as 10, resulted in clusters of too many numbers with the topic clusters overlapping with one another (note: the descriptions of the parameters are from the Python notebook [24]).

As the Python notebook demonstrated [24], we used pyLDAvis [27] to visualize the results from the topic modeling. The results from the topic modeling method are discussed in detail in the Results section.

4. RESULTS

4.1 Text-Mining with Key-phrases Extraction, Topic Modeling, and Document-clustering

After the training of Gensim's LDAModel with the aforementioned training parameters, the model analyzed and generated significant terms from the text-corpus. As the model was unsupervised and generated only numerical labels for each topic, we reviewed each of the six topics and provided labels to each as follows:

- Advanced/Unseen Attack Detection
- Contextual Cybersecurity
- Cybersecurity Applied Domain
- Data-Driven Adversary
- Power System in Cybersecurity
- Vulnerability Management

These labels were determined based on the content analysis of the most frequent terms in each cluster. After the following analysis — the sub-topics and the corresponding frequencies resulted from Gensim's LDAModel in Table 4 in the Appendix — and then via internal discussions between the researchers, we finalized naming the labels. After labeling the six topics, we quantified the labels to see which topics were most prevalent and the percent of each topic's tokens. Table 2 below lists the six topics in the percent of tokens' order.

Table 2. - Six Topic Clusters and Percent of Tokens

Topic	Percent of Tokens
Advanced/Unseen Attack Detection	22.9%
Contextual Cybersecurity	19.9%
Cybersecurity Applied Domain	18.5%
Data-Driven Adversary	11.7%
Power System in Cybersecurity	7.9%
Vulnerability Management	19%

Note. The topic Advanced/Unseen Attack Detection had the largest size with 22.9% of total tokens, followed by Contextual Cybersecurity (19.9%), Vulnerability Management (19%), and Cybersecurity Applied Domain (18.5%). As the percentages revealed, the proportions of the four aforementioned topics were similar around 20%. The rest of the topics, Data-Driven Adversary (11.7%) and Power System in Cybersecurity (7.9%), combined made about another 20%. Thus, it could be stated that the main topics of the corpus of the 48 scholarly articles with cybersecurity and data science published between 2012 and 2018 in this research were evenly spread and clustered around 5 topics, with Data-Driven Adversary and Power System in Cybersecurity conceptually combined into one topic.

Brief analysis of the result of the topic modeling is provided in the alphabetical order of topic names as follows:

Topic 1: Advanced/Unseen Attack Detection: The cluster with 22.9% of total tokens was not obvious to label initially. However, after examining the terms in the topic and also inspecting the abstracts of the articles in the cluster, we determined this cluster's label.

Topic 2: Contextual Cybersecurity: The cluster with 19.9% of total tokens did not initially suggest an obvious label. However, concerning cybersecurity, the terms, such as situational_awareness and contextual_information, seemed to suggest 'context' and 'situation' uniquely applied to the cybersecurity settings.

Topic 3: Cybersecurity Applied Domain: This topic comprised 18.5% of total tokens. After inspecting the top ten terms in the cluster, we concluded the range of terms was a diverse mix of applied domains and labeled it as such.

Topic 4: Data-Driven Adversary: With 11.7% of total tokens, this cluster was dominated by similar 'adversary'-associated terms and shown the cluster was about data-driven adversary in the cybersecurity.

Topic 5: Power System in Cybersecurity: This cluster included 7.9% of total tokens. The most salient term, power_system, with the dominant frequency of .041 within the topic, stood out from the rest of the terms and hinted that the topic was about power system in the cybersecurity context.

Topic 6: Vulnerability Management: The cluster contained 19% of total tokens and was predominated by the vulnerability-related terms. Thus, we decided to label this cluster as Vulnerability Management to remediate cyber threats, such as botnets and malware.

Also, Gensim's LDAModel document-classified the 48 articles into the six topic clusters. This classification result was used as the data source of the qualitative evaluation in section 5. Table 3 below is the result of the document-clustering into the six topic models.

Table 3. Document-clustering of the 48 articles into the six topic models

Topic	Article
Advanced/Unseen Attack Detection	[28] [29] [30] [31] [32] [33] [34] [35]
Contextual Cybersecurity	[36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47]
Cybersecurity Applied Domain	[48] [49] [1] [50] [51] [52] [53] [20] [54] [55]
Data-Driven Adversary	[56] [57] [58] [59]
Power System in Cybersecurity	[60] [61] [62] [63] [64]
Vulnerability Management	[65] [66] [67] [68] [69] [70] [71] [72] [73]

Note. The 48 articles have been ordered by the topic names in alphabetical order. Within each topic, the articles have been ordered by the authors' names in alphabetical order. The researchers labeled each topic's names after examining the salient terms within each topic.

5. EVALUATIONS OF THE RESULTS

To evaluate the results, we revisited all 48 articles and validated to see whether the text-mining and the categorization results match with their analysis. We assessed each article's fitness to the topic cluster to which it had been categorized and performed qualitative evaluation. Each sub-section below provides analytic evaluations bifurcated into positive and negative facets of the text-mining results within cybersecurity and data science realms. Table 5 provides the 48 articles' evaluation summary.

5.1 Advanced/Unseen Attack Detection

Evaluated towards Positivity: [32] claimed a classification model's good performance in detecting uncommon and sophisticated attack types, while [35] built prediction models to cope with high false-positive rates from the system sensors in detecting intrusion. Similarly, [29] discussed anomaly detection and unseen attacks in a literature survey, whereas [33] claimed an algorithm's performance improvement in detecting unseen network intrusion. Likewise, [30] sought to forecast an organization's breach chances contingent upon its network attributes, whilst [28] discussed previously unobserved malicious events and activities in using synthetic data.

Evaluated towards Negativity: [31] applied ML techniques to identify IoT devices on a network, but the study seemed tangential to the current topic. Likewise, [34] focused the SCADA system protection of the Power System topic.

5.2 Contextual Cybersecurity

Evaluated towards Positivity: [40] discussed Big Data analytics built on diverse data types, while [37] discussed heterogeneous datasets and data correlation used in Big Data applications. Also, [38] integrated external data sources in the cybersecurity data warehouse and explored diverse dataset aspects, when [39] focused on extracting cybersecurity, contextual concepts. Comparably, [47] discussed intrusion detection system (IDS) based on heterogeneous types of big data, whilst [42] analyzed real-time tweets to gain threat intelligence in temporal, contextual events. Likewise, [41] incorporated contextual data elements in a dashboard development. Then, [43] mapped vulnerabilities into threats in the post-attack forensics, providing a different context array. Similarly, [36] provided a review with the techniques implementing contextual information for intrusion detection.

Evaluated towards Negativity: [44] seemed tangential to the current topic, by centering on IDS, while [45]'s main theme was to review IDS and explain related ML terminologies. Also, [46] seemed irrelevant to this cluster because of its theme categorizing ML techniques into either "AI-based" and "computational intelligence-based (CI-based)" methods.

5.3 Cybersecurity Applied Domain

Evaluated towards Positivity: [52] covered the security aspects of mobile banking, while [51] illustrated cybersecurity and data application in the "digital forensics" domain. Correspondingly, [48] featured the "hacker communities" domain. Moreover, [49] illustrated cybersecurity applied domains, while [50] used the "black hat hackers community" domain. Furthermore, [20] profiled the "key sellers in the underground economy" domain, whilst [1] presented a research framework and its applications in various domains.

Evaluated towards Negativity: [53]'s main theme was about the "adversarial nature" in data, model, and ML, while [55] discussed three aspects of the "Science of Cybersecurity" topic. Then, [54] concentrated on the issues in big data security and privacy.

5.4 Data-Driven Adversary

Evaluated towards Positivity: [57] clearly supported the current cluster, by focusing its discussion on an ML library "in adversarial settings", while [59] concentrated on "attacks and defenses" by researching recent findings in "ML security and privacy". Finally, [58] focused "black-box attacks" leveraging "adversarial examples".

Evaluated towards Negativity: [56] provided a literature review and emphasized ML approaches in addressing "wireless sensor network" issues more relevant to the Power System in Cybersecurity.

5.5 Power System in Cybersecurity

Evaluated towards Positivity: [61] developed a "cyber-physical test bed" and presented its architecture, while [62] evaluated ML approaches to "detect malicious SCADA communications". Also, [63] discussed the ML approaches for "power system disturbance and cyber-attack discrimination", when [64] studied "stealthy false data injection using machine learning in smart grid".

Evaluated towards Negativity: [60] focused on reviewing "current advances" in using cybersecurity "benchmark datasets" related to the cybersecurity domain.

5.6 Vulnerability Management

Evaluated towards Positivity: [73] discussed the vulnerability issues in detecting "botnet traffic". Correspondingly, [67] presented a metric applicable to deep learning models for the vulnerability of the "unintended memorization" and "extraction of secrets", while [72] added values to vulnerability management towards "trustworthiness of documents and actors". Moreover, [66] contributed to vulnerability management by studying "the trends of the ML and SC [soft computing] methodologies for ICT [Information and Communication Technology] security".

Evaluated towards Negativity: [65] seemed limited to "a learning system for discriminating variants of malicious network traffic", while [70] aimed to detect unknown Android

malware. Similarly, [68] aimed to "detect new malware samples", rather closely linked to the Advanced/Unseen Attack Detection, while [69] mainly argued the families of malware. Finally, [71] used an unsupervised ML technique to data-analyze "unknown traffic to detect botnets".

Overall, the above qualitative review finds a total 69% positivity, meaning the review has agreed 69% that Gensim's LDAModel clustered the articles into the proper topics. The Advanced/Unseen Attack Detection has 75% positivity (six out of eight articles contributing towards positivity), 75% for the Contextual Cybersecurity (nine out of twelve), 70% for the Cybersecurity Applied Domain (seven out of ten), 75% for the Data-Driven Adversary (three out of four), 80% for the Power System in Cybersecurity (four out of five), and 44% for the Vulnerability Management (four out of nine). Thus, compared to the other categories, Gensim's LDAModel seemed inaccurately clustering the articles into the Vulnerability Management, as the model seemed confused with the five studies focused on the Advanced/Unseen Attack Detection. However, with the overall 69% positivity we find Gensim's LDAModel helps categorize the articles into the six topic clusters of cybersecurity and data science.

Table 5. Evaluation Results of the 48 Articles

Topic	Total Number of Articles	Evaluated towards Positivity	Evaluated towards Negativity	Percent of Positivity
Advanced/Unseen Attack Detection	8	6	2	75%
Contextual Cybersecurity	12	9	3	75%
Cybersecurity Applied Domain	10	7	3	70%
Data-Driven Adversary	4	3	1	75%
Power System in Cybersecurity	5	4	1	80%
Vulnerability Management	9	4	5	44%*
Total	**48**	**33**	**15**	**69%***

Note. Outcomes of qualitative evaluation compared to Gensim's LDAModel's evaluation. "Evaluated towards positivity" means the qualitative review has agreed that Gensim's LDAModel clustered the article into the appropriate topic. In contrast, "evaluated towards negativity" means the qualitative review has *not* agreed that Gensim's LDAModel clustered the article into the appropriate topic. Percent of positivity denotes the number of articles in that topic evaluated towards positivity divided by the total number of articles.

* Rounded up to have no decimal.

6. DISCUSSION

The goal of this research was to model topics of cybersecurity and data science clustered with significant terms and concepts, and the researchers accomplished the goal by the text-mining approach consisted of key-phrases extraction, topic modeling, and visualization.

To answer the primary research question, the researchers searched and collected the 48 scholarly articles published between 2012 and 2018 and then text-mined and analyzed the articles by topic modeling and document-clustering using the LDAModel from the Gensim library [26]. The findings have been supplied in Table 2 and Table 4 in the Appendix. Gensim's LDAModel consequently resulted in the six latent topics, and the appropriate labels for the six clusters were provided, bottomed-up from the sub-topics within each cluster found in Table 4. Furthermore, the researchers analyzed the topic modeling's result and significant terminologies and provided a qualitative review of the findings.

The result and accompanying analysis of this study also address the two secondary research questions. Regarding the question of the separability, degree of separation, and degree of overlap of clusters from the result, the six clusters in Table 2 were overall well-separated from one another, while, as Table 4 has noted, there are overlaps of some terms appearing in multiple clusters. While Table 4's notation helps understand this research question, the current research does not provide measurements of the clusters' separations and overlaps. To answer the next question of the reliability of the result, the analysis of the Evaluation reveals that Gensim's LDAModel did not always cluster the source articles into clearly distinguishable topics, particularly for the Vulnerability Management cluster. Some seem better candidates for labeling with multiple topic clusters. Also, as the review evaluated, some appear misclassified.

One further technical limitation of the research was the topic model document-clustered each article into one topic only. Conversely, the topic model did not support multi-labeling of the articles to the clusters. While multi-labeling may increase the document-clustering's accuracy, it may increase complexity of the labeling and complicate the result's evaluation. Nevertheless, multi-labeling the articles to observe potentially different results is worthwhile.

By providing answers to the aforementioned research questions, this study can now clearly advance the fields of cybersecurity and data science. Regarding this study's contributions, they are twofold. First, the topic modeling approached using text-mining makes the cybersecurity domain unearth the terminologies that make IST researchers investigate further, as Gensim's LDAModel's finding results in the six clusters with the sub-topics of the most frequent terminologies in the selected literature. Thus, the current research's findings become a research seed. Secondly, using the result of the current project's analysis, IST researchers can decide terms of interest and further investigate the articles that supplied the terms. Therefore, the research seed becomes and makes an impact as a guidance for future research direction.

Inspecting each cluster and the sub-topics modeled within the clusters could provide insight worthy of further investigation. For instance, there are six topics, and the ten sub-topics within each topic as shown in Table 4. Choosing one particular topic, inspecting the sub-topics within the topic, and observing the sub-topics labeled 'Appears under multiple topics' may help the readers link multiple topics and build a model with relations based on the shared sub-topics. Also, we conjecture adding more articles that are not in the 48 articles to the data sources may diversify the concepts discovered and increase opportunities of unearthing concepts deserving more attention, as the current study is limited to the 48 articles.

7. CONCLUSION

The main contribution of this research project is the identification of key concepts in the topic clusters and text-mining key-phrases from the recent scholarly articles focusing on cybersecurity and data science. The approach is unique because of the application of probabilistic topic modeling (e.g. LDA Model) of most frequent terms from the articles. Also, the identification of the key concepts empowers IST researchers to further survey the areas unearthed.

Regarding contributions to the broader audience, the research contributes to multiple communities:

- Research: Towards achieving the goal of building a theory in the cybersecurity domain, the research has supplied a classification model in theory building, and this becomes a precursor to building a model with defining relationships in theory building process [74].
- Business: The research presents the logical, scientific topic model, and the outcomes. Professionals can apply

these findings to understand the most frequent terms from the research and correlate with the counterparts in the real-world to discover deeper insight.

- Technology: The research has provided a topic modeling approach using text-mining and analytics using a well-received Python library specializing topic modeling [26]. This method benefits the technology sector by illustrating a sounding approach to discover relevant, frequent terms in two related disciplines.

In this research, we used the popular LDA model [7] to perform the topic modeling. We encourage fellow IST researchers to adapt other models to perform topic modeling to see whether outcomes would be different. Also, both cybersecurity and data science are wide-ranging disciplines with numerous sub-topics within each discipline. Relating sub-topics from each of these disciplines makes studies more challenging. Perhaps focusing one topic cluster from the current research, such as Vulnerability Management, would provide IST researchers opportunities to conduct more focused research. This research has a couple implications for future research. First, the most frequent terms show future researchers the key-phrases in each cluster and enable them to deep-dive into more focused research arenas. Secondly, the documents clustered into the six clusters can guide fellow researchers to conduct focused literature reviews in their pursuing topics and become the seed for their future research.

8. REFERENCES

[1] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.

[2] T. Armerding, "Obama's cybersecurity legacy: Good intentions, good efforts, limited results," *CSO Online*, 31-Jan-2017. [Online]. Available: https://www.csoonline.com/article/3162844/security/obamas-cybersecurity-legacy-good-intentions-good-efforts-limited-results.html. [Accessed: 22-Aug-2018].

[3] The White House Office of the Press Secretary, "FACT SHEET: Cybersecurity National Action Plan," *whitehouse.gov*, 09-Feb-2016. [Online]. Available: https://obamawhitehouse.archives.gov/the-press-office/2016/02/09/fact-sheet-cybersecurity-national-action-plan. [Accessed: 22-Aug-2018].

[4] European Union, "Cyber-Security - EU Global Strategy - European Commission," *EU Global Strategy*, 22-Aug-2018. [Online]. Available: /globalstrategy/en/cyber-security. [Accessed: 22-Aug-2018].

[5] J. Goulart, "Data Scientist: The Number One Job In America," *edX Blog*, 20-Jan-2016. .

[6] T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, 01-Oct-2012. [Online]. Available: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. [Accessed: 22-Aug-2018].

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[8] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[9] G. Anthes, "Topic models vs. unstructured data," *Communications of the ACM*, vol. 53, no. 12, pp. 16–18, 2010.

[10] "Google Scholar." [Online]. Available: https://scholar.google.com/. [Accessed: 24-Sep-2017].

[11] R. Das, S. Sarkani, and T. A. Mazzuchi, "Fast Abstract: Software Selection Based on Quantitative Security Risk Assessment," in *High-Assurance Systems Engineering (HASE), 2012 IEEE 14th International Symposium on*, 2012, pp. 171–172.

[12] B. Shuai, H. Li, M. Li, Q. Zhang, and C. Tang, "Automatic classification for vulnerability based on machine learning," in *Information and Automation (ICIA), 2013 IEEE International Conference on*, 2013, pp. 312–318.

[13] J. Huang, Z. Kalbarczyk, and D. M. Nicol, "Knowledge discovery from big data for intrusion detection using LDA," in *Big data (BigData Congress), 2014 IEEE international congress on*, 2014, pp. 760–761.

[14] R. Y. Lau, Y. Xia, and Y. Ye, "A probabilistic generative model for mining cybercriminal networks from online social media," *IEEE Computational intelligence magazine*, vol. 9, no. 1, pp. 31–43, 2014.

[15] K. Aswani, A. Cronin, X. Liu, and H. Zhao, "Topic modeling of SSH logs using latent dirichlet allocation for the application in cyber security," in *Systems and Information Engineering Design Symposium (SIEDS), 2015*, 2015, pp. 75–79.

[16] S. Samtani, R. Chinn, and H. Chen, "Exploring hacker assets in underground forums," in *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, 2015, pp. 31–36.

[17] G. G. Sundarkumar, V. Ravi, I. Nwogu, and V. Govindaraju, "Malware detection via API calls, topic models and machine learning," in *Automation Science and Engineering (CASE), 2015 IEEE International Conference on*, 2015, pp. 1212–1217.

[18] Z. Fang *et al.*, "Exploring key hackers and cybersecurity threats in Chinese hacker communities," in *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, 2016, pp. 13–18.

[19] T.-H. Lee, W.-K. Sung, and H.-W. Kim, "A Text Mining Approach to the Analysis of Information Security Awareness: Korea, United States, and China.," in *PACIS*, 2016, p. 69.

[20] W. Li, J. Yin, and H. Chen, "Identifying high quality carding services in underground economy using nonparametric supervised topic model," 2016.

[21] S. Samtani, K. Chinn, C. Larson, and H. Chen, "AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis," in *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, 2016, pp. 19–24.

[22] F. Kolini and L. Janczewski, "Clustering and Topic Modelling: A New Approach for Analysis of National Cyber security Strategies," *PACIS 2017 Proceedings, Malaysia*, 2017.

[23] Y. Shinyama, "PDFMiner," 2014. [Online]. Available: https://euske.github.io/pdfminer/index.html. [Accessed: 08-Apr-2018].

[24] Yasmin, "LDA and T-SNE Interactive Visualization," *Kaggle*, 14-Sep-2017. [Online]. Available: https://www.kaggle.com/ykhorramz/lda-and-t-sne-interactive-visualization. [Accessed: 03-Apr-2018].

[25] The pandas project, "Python Data Analysis Library — pandas: Python Data Analysis Library," 2017. [Online]. Available: https://pandas.pydata.org/. [Accessed: 08-Apr-2018].

[26] "gensim: topic modelling for humans." [Online]. Available: https://radimrehurek.com/gensim/tut2.html. [Accessed: 03-Mar-2018].

[27] B. Mabey, *pyLDAvis: Python library for interactive topic model visualization. Port of the R LDAvis package*. 2018.

[28] S. Abt and H. Baier, "A Plea for Utilising Synthetic Data when Performing Machine Learning Based Cyber-Security Experiments," 2014, pp. 37–45.

[29] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[30] Y. Liu *et al.*, "Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents.," in *USENIX Security Symposium*, 2015, pp. 1009–1024.

[31] Y. Meidan *et al.*, "ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis," 2017, pp. 506–509.

[32] H. H. Pajouh, G. Dastghaibyfard, and S. Hashemi, "Two-tier network anomaly detection model: a machine learning approach," *Journal of Intelligent Information Systems*, vol. 48, no. 1, pp. 61–74, Feb. 2017.

[33] C. T. Symons and J. M. Beaver, "Nonparametric semi-supervised learning for network intrusion detection: combining performance improvements with realistic in-situ training," in *Proceedings of the 5th ACM workshop on Security and artificial intelligence*, 2012, pp. 49–58.

[34] S. Yasakethu and J. Jiang, "Intrusion detection via machine learning for SCADA system protection," in *Proceedings of the 1st International Symposium for ICS & SCADA Cyber Security Research*, 2013, pp. 101–5.

[35] L. Zomlot, S. Chandran, D. Caragea, and X. Ou, "Aiding intrusion analysis using machine learning," in *Machine Learning*

and Applications (ICMLA), 2013 12th International Conference on, 2013, vol. 2, pp. 40–47.

[36] A. Aleroud and G. Karabatis, "Contextual information fusion for intrusion detection: a survey and taxonomy," *Knowledge and Information Systems*, vol. 52, no. 3, pp. 563–619, Sep. 2017.

[37] R. Alguliyev and Y. Imamverdiyev, "Big data: big promises for information security," in *Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on*, 2014, pp. 1–4.

[38] B. D. Czejdo, M. D. Iannacone, R. A. Bridges, E. M. Ferragut, and J. R. Goodall, "Integration of external data sources with cyber security data warehouse," 2014, pp. 49–52.

[39] C. L. Jones, R. A. Bridges, K. M. T. Huffer, and J. R. Goodall, "Towards a Relation Extraction Framework for Cyber-Security Concepts," 2015, pp. 1–4.

[40] T. Mahmood and U. Afzal, "Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools," in *Information assurance (ncia), 2013 2nd national conference on*, 2013, pp. 129–134.

[41] S. McKenna, D. Staheli, C. Fulcher, and M. Meyer, "BubbleNet: A Cyber Security Dashboard for Visualizing Patterns," *Computer Graphics Forum*, vol. 35, no. 3, pp. 281–290, Jun. 2016.

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

[43] S. Noel, E. Harley, K. H. Tam, M. Limiero, and M. Share, "CyGraph: Graph-Based Analytics and Visualization for Cybersecurity," in *Handbook of Statistics*, vol. 35, Elsevier, 2016, pp. 117–167.

[44] J. Singh and M. J. Nene, "A Survey on Machine Learning Techniques for Intrusion Detection Systems," vol. 2, no. 11, p. 7, 2013.

[45] D. P. Vinchurkar and A. Reshamwala, *A Review of Intrusion Detection System Using Neural Network and Machine Learning*. IJESIT, 2012.

[46] M. Zamani and M. Movahedi, "Machine learning techniques for intrusion detection," *arXiv preprint arXiv:1312.2177*, 2013.

[47] R. Zuech, T. M. Khoshgoftaar, and R. Wald, "Intrusion detection and Big Heterogeneous Data: a Survey," *Journal of Big Data*, vol. 2, no. 1, Dec. 2015.

[48] V. A. Benjamin and H. Chen, "Machine learning for attack vector identification in malicious source code," in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, 2013, pp. 21–23.

[49] M. Brundage *et al.*, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[50] T. M. Georgescu and I. Smeureanu, "Using Ontologies in Cybersecurity Field," *Informatica Economica*, vol. 21, no. 3/2017, pp. 5–15, Sep. 2017.

[51] A. Guarino, "Digital forensics as a big data challenge," in *ISSE 2013 securing electronic business processes*, Springer, 2013, pp. 197–203.

[52] W. He, X. Tian, J. Shen, and Y. Li, "Understanding Mobile Banking Applications' Security risks through Blog Mining and the Workflow Technology," 2015.

[53] A. D. Joseph, P. Laskov, F. Roli, J. D. Tygar, and B. Nelson, "Machine Learning Methods for Computer Security (Dagstuhl Perspectives Workshop 12371)," Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany, 2013.

[54] B. Thuraisingham, "Big Data Security and Privacy," 2015, pp. 279–280.

[55] B. Thuraisingham *et al.*, "A data driven approach for the science of cyber security: Challenges and directions," in *Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on*, 2016, pp. 1–10.

[56] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014.

[57] N. Papernot *et al.*, "cleverhans v2. 0.0: an adversarial machine learning library," *arXiv preprint arXiv:1610.00768*, 2016.

[58] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, p. 13, 2016.

[59] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

[60] A. I. Abubakar, H. Chiroma, S. A. Muaz, and L. B. Ila, "A Review of the Advances in Cyber Security Benchmark Datasets for Evaluating Data-Driven Based Intrusion Detection Systems," *Procedia Computer Science*, vol. 62, pp. 221–227, 2015.

[61] U. Adhikari, T. Morris, and S. Pan, "WAMS Cyber-Physical Test Bed for Power System, Cybersecurity Study, and Data Mining," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2744–2753, Nov. 2017.

[62] J. M. Beaver, R. C. Borges-Hink, and M. A. Buckner, "An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications," 2013, pp. 54–59.

[63] R. C. Borges Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari, and S. Pan, "Machine learning for power system disturbance and cyber-attack discrimination," 2014, pp. 1–8.

[64] M. Esmalifalak, Nam Tuan Nguyen, Rong Zheng, and Zhu Han, "Detecting stealthy false data injection using machine learning in smart grid," 2013, pp. 808–813.

[65] J. M. Beaver, C. T. Symons, and R. E. Gillen, "A learning system for discriminating variants of malicious network traffic," 2013, p. 1.

[66] F. Camastra, A. Ciaramella, and A. Staiano, "Machine learning and soft computing for ICT security: an overview of current trends," *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 2, pp. 235–247, Apr. 2013.

[67] N. Carlini, C. Liu, J. Kos, Ú. Erlingsson, and D. Song, "The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets," *arXiv preprint arXiv:1802.08232*, 2018.

[68] Y. Fan, Y. Ye, and L. Chen, "Malicious sequential pattern mining for automatic malware detection," *Expert Systems with Applications*, vol. 52, pp. 16–25, Jun. 2016.

[69] E. Gandotra, D. Bansal, and S. Sofat, "Malware Analysis and Classification: A Survey," *Journal of Information Security*, vol. 05, no. 02, pp. 56–64, 2014.

[70] S. Hou, A. Saas, L. Chen, Y. Ye, and T. Bourlai, "Deep Neural Networks for Automatic Android Malware Detection," 2017, pp. 803–810.

[71] D. C. Le, A. N. Zincir-Heywood, and M. I. Heywood, "Data analytics on network traffic flows for botnet behaviour detection," in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, 2016, pp. 1–7.

[72] M. Mayhew, M. Atighetchi, A. Adler, and R. Greenstadt, "Use of machine learning in big data analytics for insider threat detection," in *Military Communications Conference, MILCOM 2015-2015 IEEE*, 2015, pp. 915–922.

[73] M. Stevanovic and J. M. Pedersen, "Machine learning for identifying botnet network traffic," *Networking and Security Section, Department of Electronic Systems, Aalborg University, Tech. Rep.*, 2013.

[74] P. R. Carlile and C. M. Christensen, "The cycles of theory building in management research," 2004.

APPENDIX

Table 4. Topics Modeled with 10 Most Frequent Terms within 6 Topics

Topic: Advanced/Unseen Attack Detection (22.9%; 23% in the pie-chart)	Summary: This topic cluster reveals latent terms related attack types that are not seen before, such as semi_supervised (as attacks are unseen before, there is need for review of manual human analysis), false_alarm (the researchers conjecture that there will be much

	false alarms associated with these types of attacks), synthetic data (data is not of natural origin), unknown attack, and time_series (because attacks are unseen before, collecting time series-based data will be fundamental in detecting these types of attack). Therefore, the researchers label this topic cluster as Advanced/Unseen Attack Detection.	
Sub-Topics	**Frequency**	**Notes**
semi_supervised	0.010	
data_set	0.009	Appears under multiple topics.
malicious_activity	0.008	
incident	0.007	
false_alarm	0.007	
detection_rate	0.006	
synthetic_data	0.006	
unknown_attack	0.006	Appears under multiple topics.
training_testing	0.005	
time_series	0.005	
Topic: Contextual Cybersecurity (19.9%; 20% in the pie-chart)	Summary: The terms closely associated with contextual data analysis, such as heterogeneous, situational_awareness, knowledge_base, contextual_information, correlation, and alert_correlation appear under this topic cluster. Therefore, the researchers label this topic cluster as Contextual Cybersecurity.	
Sub-Topics	**Frequency**	**Notes**
feature_selection	0.012	
heterogeneous	0.010	
situational_awareness	0.009	
analyst	0.009	
data_mining	0.008	Appears under multiple topics.
knowledge_base	0.008	
contextual_information	0.007	
correlation	0.007	
data_set	0.006	
alert_correlation	0.006	
		Appears under multiple topics.
Topic: Cybersecurity Applied Domain (18.5%; 19% in the pie-chart)	Summary: This topic cluster is named as Cybersecurity Applied Domain because the terms, such as mobile, social_medium, computer_security, and banking, are prevalent.	
Sub-Topics	**Frequency**	**Notes**
analytics	0.011	
mobile	0.010	
social_medium	0.009	
computer_security	0.006	
data_driven	0.006	
hacker	0.006	
text	0.006	
banking	0.006	
social_network	0.006	
hacker_community	0.006	
Topic: Data-Driven Adversary (11.7%; 12% in the pie-chart)	Summary: Except the data science-related terms, the terms related to adversary prevail in this topic cluster. Thus, the researchers label this topic cluster as Data-Driven Adversary.	
Sub-Topics	**Frequency**	**Notes**
adversarial_sample	0.024	
adversarial	0.023	
adversarial_example	0.017	
adversary	0.015	
substitute	0.013	
learning_algorithm	0.010	
oracle	0.008	
substitute_model	0.008	
model_trained	0.008	
logistic_regression	0.007	
Topic: Power System in Cybersecurity (7.9%; 8% in the pie-chart)	Summary: This topic cluster is dominated by industrial terms related to national infrastructure for utility. Therefore, the researchers label it as Power System in Cybersecurity.	
Sub-Topics	**Frequency**	**Notes**
power_system	0.041	
power	0.024	
smart_grid	0.012	
total_number	0.011	
scada_system	0.011	
command	0.009	
data_mining	0.009	Appears under multiple topics.
injection	0.009	
measurement	0.007	
cyber_crime	0.007	
Topic: Vulnerability Management (19%; 19% in the pie-chart)	Summary: This topic cluster is predominated by the terms associated vulnerabilities or threats, such as botnet, malware, and detection. Thus, the researchers label this topic cluster as Vulnerability Management.	
Sub-Topics	**Frequency**	**Notes**
botnet	0.019	
botnet_detection	0.015	
botnets	0.008	
botnet_traffic	0.007	
naive_bayes	0.007	
secret	0.007	
detection_rate	0.006	Appears under multiple topics.
evasion	0.006	
numeric	0.006	
malware_detection	0.006	

Note. The six topics are listed in alphabetical order. The summaries of each topic cluster are provided and also the terms appearing in multiple topics, such as data_mining, data_set and detection_rate, are noted in the table. The column "Topic" means each of the six topic clusters originally resulted in numeric value from Gensim's LDAModel and subsequently labeled by the researchers; "Summary" means a summary of each topic cluster denoting what each one represents, approached using a bottom-up analysis of the constituent sub-topics; "Sub-topics" mean ten most frequent terms within each topic cluster discovered by Gensim's LDAModel; "Frequency" means a percent of the sub-topic in the distinct terms of the entire text-corpus; and "Notes" mean the sub-topic appears in multiple topics.