Computers & Security 120 (2022) 102821

Contents lists available at ScienceDirect

Computers & Security

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

What changed in the cyber-security after COVID-19?

Rajesh Kumar*, Siddharth Sharma, Chirag Vachhani, Nitish Yadav

Department of Computer Science and Information systems, Birla Institute of Technology and Science, Pilani, India

a r t i c l e i n f o

Article history:
Received 20 April 2022
Revised 24 June 2022
Accepted 28 June 2022
Available online 5 July 2022

Keywords:
Cyber-security trends
Topic modeling
Trend analysis
COVID-19 pandemic
Latent Dirichlet Allocation
Unsupervised machine learning

a b s t r a c t

This paper examines the transition in the cyber-security discipline induced by the ongoing COVID-19 pandemic. Using the classical information retrieval techniques, a more than twenty thousand documents are analyzed for the cyber content. In particular, we build the topic models using the Latent Dirichlet Allocation (LDA) unsupervised machine learning algorithm. The literature corpus is build through a uniform keyword search process made on the scholarly and the non-scholarly platforms filtered through the years 2010-2021. To qualitatively know the impact of COVID-19 pandemic on cyber-security, and perform a trend analysis of key themes, we organize the entire corpus into various (combination of) categories based on time period and whether the literature has undergone peer review process. Based on the weighted distribution of keywords in the aggregated corpus, we identify the key themes. While in the pre-COVID-19 period, the topics of cyber-threats to technology, privacy policy, blockchain remain popular, in the post-COVID-19 period, focus has shifted to challenges directly or indirectly brought by the pandemic. In particular, we observe post-COVID-19 cyber-security themes of privacy in healthcare, cyber insurance, cyber risks in supply chain gaining recognition. Few cyber-topics such as of malware, control system security remain important in perpetuity.

We believe our work represents the evolving nature of the cyber-security discipline and reaffirms the need to tailor appropriate interventions by noting the key trends.

1. Introduction

Worldwide, the COVID-19 pandemic continues to cripple human life Atzrodt et al. (2020). Along with the overburdening of health systems of many countries, the crisis has compelled a transition to a new normal. E-learning, work-from-home, e-healthcare form this new normal McKinsey (2020). Symmetrically, this growing digital transition has resulted in a greater attack surface Manadhata and Wing (2010). According to F5blog, there has been a 220% increase in phishing-related computer crimes during the period (2019-2020) Warburton (2020). Authors in Burgess (2020) and Muthuppalaniappan and Stevenson (2021) note that the attackers, in past have not only targeted the organizations but also the national critical infrastructures such as of the healthcare services and the banking platforms. As a result of the increased cyber-threats, a number of government bodies around the world have emphasized on cyber-awareness campaigns, recommended useful cyber-hygiene techniques, and resorted to cyber-nudges EUA (2021)DHS (2021)Interpol (2020).

The aim of this paper is to critically highlight and compare the cyber-security trends before and since the start of the COVID-19 pandemic, thereby enabling the organizations to proactively respond to the emerging cyber-threats. To highlight the cyber-security trends, we search for the cyber-content in the literature published between the years 2010-2021 using the classical information retrieval techniques. The literature corpus is build using a uniform keyword search process made on the scholarly and the non-scholarly platforms. The scholarly platform yielded a corpus that consists of a total of 10,680 prominent peer-reviewed literature abstracts. The non-scholarly platform yielded 11,032 documents collected from four prominent security blogs. Notably, in this study, we use blogs as primary source of non-peer literature. While the peer-reviewed literature remains a reliable source of the information, we claim blogs as a reasonable choice to outline the differentiated cyber-themes. Related in the context of this research, where we are examining emerging cyber-trends, we believe blogs are a source of quick, updated, and insightful source Dinu (2022). Researchers in Gernhardt and Groš (2021); Kumar et al. (2022), note that many cyber-incidents are reported first in the blog posts. Similarly, authors in Lallie et al. (2021) use blog articles and social media posts to build a time-line of cyber-attacks related to the COVID-19 pandemic. In the same line, authors in Lemay et al. (2018) note that a comprehensive and updated ac-

* Corresponding author.
E-mail address: rajesh.k@pilani.bits-pilani.ac.in (R. Kumar).

https://doi.org/10.1016/j.cose.2022.102821

count of attack techniques, tools and procedures is typically available at non-scholarly platforms.

Technically, this paper adopts probabilistic topic modeling algorithms to identify emergent cyber-security themes. In particular, we use topic modeling which is an unsupervised machine learning technique, that uses topic models (algorithms) to uncover the hidden or the latent thematic structures (a.k.a topics) from a large collection of documents Syed and Spruit (2017). A number of topic models exist, for example, the Latent Semantic Analysis (LSA) Dumais (2004), Probabilistic Latent Semantic Analysis (PLSA) Hofmann (2013), Non-negative Matrix Factorization (NMF) Lee and Seung (1999), Dirichlet Multinomial Regression (DMR) Mimno and McCallum (2008), etc. Topic models have been popularly used in the cyber-security context in many domains. For example, in Alagheband et al. (2020), authors use topic modeling to perform time-based gap analysis of cyber-security trends in academic and digital media. In Bechor and Jung (2019), authors use topic modeling to identify key concepts from the scholarly articles focusing on cyber-security and data-science. We use Latent Dirichlet Allocation (LDA) topic model Blei et al. (2003) in this paper. LDA is a probability-based model and can reveal hidden connections which cannot be found by looking only at frequency-based methods, for example, LSA Kang et al. (2019). Specifically, with our analysis, we answer the following:

- Emergent cyber-security themes. We divide the corpus into two periods: the Pre-COVID-19 period [2010-2019] and the Post-COVID-19 period (2019-2021]. For each corpora, we distinguish the cyber-security themes.
- Evolution of cyber-security themes. We perform a year-wise trend analysis of the identified cyber-security themes.

Interesting results emerge from our study. While in the peer-reviewed literature prior to the beginning of the COVID-19, the technical security services, the security attributes, and the analytical methods such as of machine learning-based intrusion detection systems, game-theory based attack models, and software vulnerabilities are notable; since the pandemic, new specific cyber-themes are prominently discussed (for example, the ransomware and the privacy risks on health care, supply chain disruptions, cyber-awareness, and digital banking). Compared to the peer-reviewed literature, in the non-peer reviewed literature, topics related to the financial security such as of the fraud, the device skimmers, and the network privacy are the focal themes before to the pandemic. Since, the onset of pandemic, newer threats to the digital infrastructure such as of privacy and security in healthcare, as well as software vulnerability analysis have been more actively researched These cyber-security themes, allow researchers to poisition their research on active topics, thereby improving the overall posture of the cyber-security. Furthermore, these trends aid industry practitioners in rationally adopting business strategies, budgeting, and inventory management.

Related work. Numerous papers in the field of cyber-security have used topic modeling techniques to analyze emerging and evolving trends. In Dhillon et al. (2021), authors use text mining and Delphi-based analysis of questionnaires to highlight the existing gaps between the academic research and industrial practices. Authors in Wu et al. (2021) compare the cyber-security content in the non-peer-reviewed literature (news, blogs, and websites) from 2009 to 2019 to identify trending topics using LDA-based text mining models. In Alagheband et al. (2020), authors conduct a time-based gap analysis of cyber-security trends identified by using the LDA topic model on academic and digital media disseminated between the years 2008 to 2018. Adam et al. use topic modeling to link attack patterns to system models. This was accomplished by using the LDA topic model to get the topic distributions of the text in the system's model and then locating attack patterns with a similar topic distribution Adams et al. (2018). Lu and Li examined the peer-reviewed literature published between the years 2013 to 2017 to identify the research trends in Internet of Things (IoT) security Lu and Da Xu (2018). In Sleeman et al. (2021), authors used dynamic topic modeling to show the evolution of key themes in a time-stamped collection of documents.

Besides cyber-security, topic modeling have been traditionally used in many other domains such as healthcare, digital media, biochemistry, etc. Coventry and Branley examined research trends in healthcare cyber-security and highlighted the threats, faced by the healthcare industry Coventry and Branley (2018). Kawata and Fujiwara used LDA topic models to extract the topics from the non-peer-reviewed literature with an aim to determine the seasonality of the trending topics Kawata and Fujiwara (2016). In Kang et al. (2019), authors used topic modeling to identify research topics in the field of biochemistry over the last twenty years. Chen et al. proposed a topic-based technological forecasting approach for determining the trends of specific topics in massive patent claims Chen et al. (2017).

Our paper has a similar analytical engine as the previously mentioned papers with the idea of creating LDA topic models, albeit our focus is investigating COVID-19-induced cyber-security transition. Similar research focus as present paper is of interest in Bari (2021),Nabe (2021), where authors identify Post-COVID cyber-security themes. However, the process is largely based on intuition and/or surveys. Unlike previous approaches, we rigorously use the peer- and the non-peer-reviewed literature exploiting the state-of-the-art text mining techniques to have a comprehensive understanding of the COVID-19- upheaval on cyber-security.

The rest of the paper is as follows: In Section 2, we give a background about topic modeling, data collection, and the process of labeling the topic distributions. Section 3 explains our methodology which is a three-phase process. In Section 4, we explain the results. Finally, in Section 5, we give concluding statements.

2. Background

2.1. Topic Modelling Preliminaries

Topic modeling is an unsupervised machine learning technique popularly used to identify latent or hidden topics in a corpus. Topics are the central themes that uniquely identifies the document. Typically, they are not always detectable using the traditional keyword searches Srivastava and Sahami (2009). In this paper, we adopt the popular topic model of Latent Dirichlet Allocation (LDA), introduced by Blei et al. Blei et al. (2003). LDA belongs to the class of hierarchical Bayesian model Syed and Spruit (2017). The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Technically, LDA begins by randomly designating each word in a document as one out of K topics. Then, it calculates the conditional probabilities for every topic in each document through an iterative process Blei et al. (2003). LDA assumes the following process for each document of N words in a corpus D:

1. For every topic $k = \{1,...,K\}$, randomly sample the topic-word distribution, $\beta_{d,k}$ from a Dirichlet distribution with the parameter η .
2. For every document d, randomly sample the document-topic distribution θ_d from a Dirichlet distribution with the parameter α .
3. For each document d and each word w contained within the document d,
   (a) Randomly sample a word-topic assignment $z_{d,n}$ from a multinomial distribution θ_d , where $z_{d,n} \in \{1,...,K\}$.
   (b) Randomly sample a d^{th} word in n^{th} document $w_{d,n}$ from a multinomial distribution of β_k, $z_{d,n}$.

Given the Dirichlet parameters α and η, the joint distribution of the latent variables β_K (topics), θ_D (topic mixtures), and z_D (word-topic assignments) along with the observed variable w_D (words in the document) is expressed in (1). Here, an observed variable refers to the visible words in the document. Furthermore, a latent variable is estimated from the observed variable and Dirichlet parameters using following posterior.

$$P(\beta_K, \theta_D, z_D, w_D \mid \alpha, \eta) = \prod_{k=1}^{K} P(\beta_k \mid \eta) \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{n=1}^{N} P(z_{d,n} \mid \theta)\, P(w_{d,n} \mid z_{d,n}, \beta_{d,k}) \quad (1)$$

To infer the hidden structure with statistical inference, we must condition the latent variables on the only observed variable i.e, words in the document Syed and Spruit (2017). This posterior is given by (2):

$$P(\beta_K, \theta_D, z_D \mid w_D) = \frac{P(\beta_K, \theta_D, z_D, w_D)}{P(w_D)} \quad (2)$$

The posterior distribution mentioned above is intractable Blei et al. (2003). However, an approximation to the true posterior can be obtained using inference techniques such as of sampling-based method Porteous et al. (2008) or the variational-based method Hoffman et al. (2010). In this paper, we adopt the variational-based method. After approximating the LDAs posterior distribution, the K topics are represented as a multinomial distribution over vocabulary V (distinct words in a document) to get the topic distributions.

A topic distribution is a collection of all the words in the document. However, each word has a different probability. Words with a high probability in the topic distribution, tends to co-occur more frequently. The quality of topic distributions can be estimated using the metrics of perplexity Wallach et al. (2009) and/or coherence Röder et al. (2015). Perplexity measures the ability of a topic to characterize a document. Coherence measures the degree of similarity displayed by some words in the topic distribution. We choose coherence as a metric to evaluate our topic model because models that achieve higher predictive perplexity are often less comprehensible from a human standpoint Chang et al. (2009).

2.2. Data collection

The literature corpus is build using a uniform keyword search process made on the scholarly (Scopus academic database) and the non-scholarly platforms (four cyber-security blogs: Microsoft Security Microsoft (2022), Schneier on Security Schneier (2004), The Last Watchdogs Acohido (2000), and Krebs on Security Krebs (2000)). We divide the corpus into two sub-corpora based on its scholarly peer review process status – the peer-reviewed literature (PR – 9343 documents) and the non-peer reviewed literature (NPR – 10,679 documents). These corpora are termed as the all-time peer reviewed All-PR and the all-time non-peer reviewed All-NPR as they are aggregated irrespective of the COVID-19 pandemic, the year 2019 of its onset.

2.3. Manual inference of a topic from a weighted distribution of keywords

This paper adopts probabilistic topic modeling algorithms to identify the emergent cyber-security themes. As mentioned in Section 2.1, the LDA topic model generates topic distributions as the output. The high probability words, typically the top ten, are used to manually infer the topics. In Table 1, we demostrate on how these topic distributions look like utilizing an example document corpus of the all-time non-peer-reviewed literature. Here, the different cyber-themes outputted by running LDA algorithm are christened as T0,T1,T2,T3 and T5 Each topic is a weighted distribution of the keywords. For example, the topic T4, indicated with red text, is a combination of the top ten keywords (in decreasing order). Based on a brainstorming process, next we manually annotate the topic T4 as "Customer credentials". Following the same process, we annotate each topic, by running LDA algorithms on different corpora. In Appendix 6, we document these different keyword distributions and the annotated cyber-themes. In Table 4, we present for the peer-reviewed corpus. In Table 5, we tabulate the inferred topics in the non-peer-reviewed literature. Notably, such manual inference of topics in absence of an automated aid is common practice that is witnessed in several prominent works on topic models, for example in Alagheband et al. (2020) and Adams et al. (2018).

3. Methodology

Figure 2 provides a bird-eye view of our pipeline framework. It consists of three overarching phases: the data collection and pre-processing, the phrase modeling, and building the tuned LDA topic model.

Phase-1: Data collection and pre-processing. The first phase is of data collection. In Section 2.2, we have discussed the organization of our corpus. The peer-reviewed literature is collected using a unform keyword search made on the Scopus academic database. Following keywords of {"Cybersecurity, Cybersecurity"} and its combinations are iteratively used for this purpose. Filter of "year of publication" is used to discard the literature falling outside the year span of 2010-2021. For each chosen paper, we collect its corresponding meta-data, that includes the year of publication, the title of the publication, and the complete textual abstract. We choose the abstracts vis–vis the full-text because an abstract is high word density textual matter appropriate for the LDA model. This is also in accordance to the best practices followed in document analytics Gatti et al. (2015). In addition, compared to the full-text, abstracts are summary capturing much information while leveraging the benefits of less pre-processing and computational time.

Once, the literature corpus is aggregated and bifurcated into differented sub-corpus: Pre-PR, Post-PR, All-PR, Pre-NPR, Post-NPR, and All-NPR, we perform a generic data pre-processing on them. During this process, we remove all the punctuation and the URLs. Next, we lowercase the texts. Subsequently, we perform a tokenization. Tokenization is the process of decomposing a sentence int words (a.k.a the tokens) Patel and Arasanipalai (2021). In the next step, we perform stemming Bird et al. (2009) and lemmatization Thanaki (2017). In the process of stemming, we remove the suffixes from a word. For example, all the words – eats, eating, eaten after stemming is retained as eat. During lemmatization, we convert a word to its canonical form. For example, the word "caring" after lemmatization is retained as "care".

Phase-2: Phrase modeling. During the tokenization process, a sentence is decomposed into tokens. Due to this, the word order, that exist in a sentence is lost. For example, the word "white house" post-tokenization is decomposed into two disassociated words of {"house", "white"}. There is a likelihood, that both these tokens are assigned as separate topics during topic modeling. If these two tokens appear in a collocation, it stands to reason that they should be assigned to the same topic. Collocation are phrases with more than one word that occur more frequently in a given context than their individual word parts McKeown and Radev (2000). To preserve collocation, we retain the word order with bi-grams and tri-grams. Bi-grams are two-word phrases, such as "white house" Wang and Manning (2012). Tri-grams are three-word phrases, such as "out of business" Khanbhai et al. (2022). Collocation improves the interpretability of topics in the LDA topic model Wang et al. (2007). Therefore, to increase the interpretability, we use the "Point-wise Mutual In-

Table 1
Inferred topics for the all-time non-peer-reviewed corpus from LDA generated topic distribution.

Topic	Topic Distributions	Label
T0	network, time, business, privacy, company, year, attack, technology, secure, healthcare	Network privacy and security in healthcare
T1	story, bank, group, fraud, site, money, attack, week, case, botnet	Financial fraud
T2	software, malware, system, vulnerability, code, today, version, computer, flaw, program	Software system vulnerabilities
T3	card, credit, breach, company, payment, customer, encryption, fraud, source, investigation	Credit card breaches
T4	service, number, account, email, access, phone, address, site, name, password, customer	Customer credentials
T5	device, surveillance, location, iPhone, skimmer, machine, cash, camera, wireless, reader	Device skimmers

Table 2
Cyber-security themes in the peer-reviewed literature corpus. The text marked in red represents the themes common to both the pre-COVID-19 and the post-COVID-19 period.

Category	Trends
Pre-COVID-19 cyber-security themes	"Malware detection", "Control system security", "Intrusion detection system", "Software vulnerability analysis", "Attack models based on game-theory", "Blockchain and cryptocurrency","Cyber risk management"
Post-COVID-19 cyber-security themes	"Cyber insurance", "Cyber-security in healthcare", "Control system security", "Intrusion detection system", "Cyber-resilience in supply chain", "Malware detection", "Cyber risk management", Security awareness

Table 3
Cyber-themes in the non-peer-reviewed corpus. The red text represents themes common to both the pre-COVID-19 and the post-COVID-19 eras.

Category	Trends
Pre-COVID-19 cyber-security themes	"Privacy Protection", "Fake news", "Malware in business process", "Cyber-crime in banking", "Credit card frauds"
Post-COVID-19 cyber-security themes	"Social-engineering", "Side channel attacks", "Software vulnerability analysis", "Malware in cloud services"

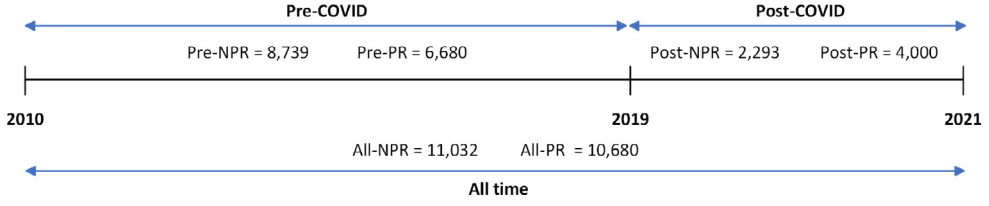


Fig. 1. Number of documents organized by a timeline. Pre-NPR refers to the pre-COVID-19 non-peer-reviewed corpus, Post-NPR refers to the post-COVID-19 non-peer-reviewed corpus, All-NPR refers to the all-time non-peer-reviewed corpus, Pre-PR refers to the pre-COVID-19 peer-reviewed corpus, Post-PR refers to the post-COVID-19 peer-reviewed corpus, and All-PR refer to the all-time peer-reviewed corpus

formation (PMI)" score to identify the co-occurrences of words that form meaningful bi-grams and tri-grams Bouma (2009). This metric assesses the likelihood of the words co-occurring in contrast if they were considered independently.

Phase-3: Tuned LDA topic model. The pre-processed data along with the phrase models serves as input to the LDA algorithm. Alongside, we pass the hyper-parameters to tune the model. Hyper-parameters are machine learning model settings that needs to be finely-tuned aiming at improving the model interpretability George et al. (2017). In our work, we identify four hyperparameters: the num_topics, the chunksize, the passes, and the iterations. num_topics refers to the number of topics extracted from a corpus. chunksize determines how many documents should be loaded into the computer memory to train the algorithm. passes are the number of training passes required to train the algorithm. Finally, the iterations are the minimum number of iterations required through the corpus to infer the topic distribution.

The coherence score is used as a performance metric to determine the optimal values of these parameters. Values are selected that result in highest coherence score obtained by iteratively configuring one parameter at a time. Figure 3 shows the optimal values of the hyperparameters obtained for the non-peer-reviewed all-time corpus. The optimal values for num_topics (refer Figure 3a), chunksize (refer Figure 3b), passes (refer Figure 3c), and iterations (refer Figure 3d) are 6, 900, 90, and 600 respectively. These hyperparameter values, along with the dataset and Dirichlet parameters (α and η), are fed into the LDA topic model as input. The final LDA topic model is the tuned LDA topic model.

4. Results

All the experiments are performed on an intel i5, 8th generation processor. The peer- and the non-peer-reviewed corpora were pre-processed and cleaned using the nltk[1] library and using customized python scripts. Next, the Gensim[2], is used to convert the pre-processed data to semantic vectors, implement LDA, and evaluate coherence scores. Subsequently, the topics distributions that are generated using the LDA, are visualized as weighted keywords using the LDAvis[3] tool. We analyze our findings putting it under the two

[1] Natural Language Toolkit, available at https://www.nltk.org
[2] Gensim python library, available at https://pypi.org/project/gensim/
[3] Python LDAvis tool, available at https://anaconda.org/conda-forge/pyldavis

Table 4
Topic distribution for the peer-reviewed corpus. Here PR refers to the peer-reviewed literature taken from Scopus academic database.

Corpus Name	Topic	Topic Distributions	Label
Pre- COVID (2010-	T2	detection, classify, malware, accuracy, feature, detect, anomaly_detection, module, memory, computation	Malware detection
	T7	system, analysis, approach, model, framework, attack, process, risk, network, problem	Cyber-risk management
	T9	software, vulnerability, defense, test,code, damage, cyberattack, functionality, malware, execution	Software vulnerability analysis
	T3	system, control, power, communication, access, secure, energy, architecture, operation, authentication	Control system security
	T4	network, traffic, technique, dataset, intrusion_detection, graph, ransomware, intrusion, honeypot, password	Network intrusion detection
	T5	game, influence, prediction, deterrence, theory, exercise, deception, uncertainty, file, aviation	Attack models based on game-theory
	T6	blockchain, insurance, learning, workshop, cryptocurrency, employment, shortage, literacy, index, spectrum	Blockchain and cryptocurrency
	T8	design, development, knowledge, field, engineering, computer, education, training, project, program	Cybersecurity engineering
Post- COVID	T1	risk, management, safety, business, solution, value, organization, adoption, technology, supply_chain	Cyber-risk management
(2020-2021)	T1	privacy, health, healthcare, science, crisis, care, innovation, cryptography, period, staff	Privacy in healthcare
(PR)	T2	investment, wireless, channel, disinformation, transaction, banking, aviation, author, radiation, bank	Digital banking
	T4	awareness, program, course, web, computer_ science, exposure, online, university, coverage, book	Security awareness
	T10	network, model, detection, attack, performance, method, approach, problem, traffic, intrusion	Network intrusion detection
All- time	T1	technology, development, analysis, threat, framework, approach, work, industry, level, environment	Cyber-governance
(2010- 2021)	T1	software, safety, engineering, education, design, program, project, vehicle, hardware, control	Control systems security
	T6	network, detection, model, attack, performance, method, system, intrusion, machine, time	Intrusion detection models
	T7	malware, cloud, service, evidence, event, time, group, framework, series, contribution	Malware detection
	T10	risk, management, business, assessment, authentication, resilience, trust, compliance, organization, supply_chain	Cyber risk management
	T10	malware, health, healthcare, ransomware, fraud, smart_city, consumer, payment, breach, device	Cyber-security in healthcare
	T10	risk, management, business, assessment, authentication, resilience, trust, compliance, organization, supply_chain	Supply chain cyber-risk

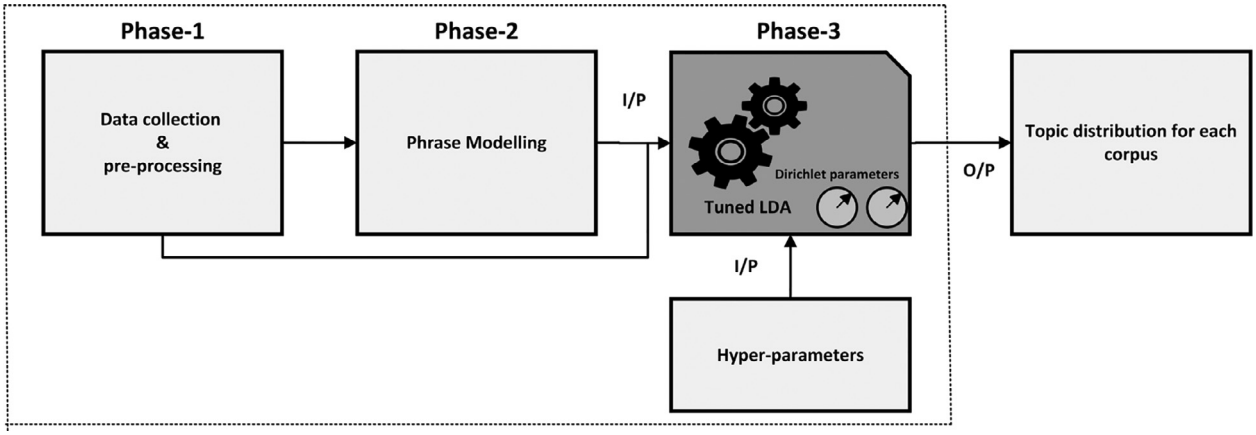


Fig. 2. Framework to generate topic distributions for each corpus.

categories of the emergent cyber-security themes and the evolution of cyber-security themes.

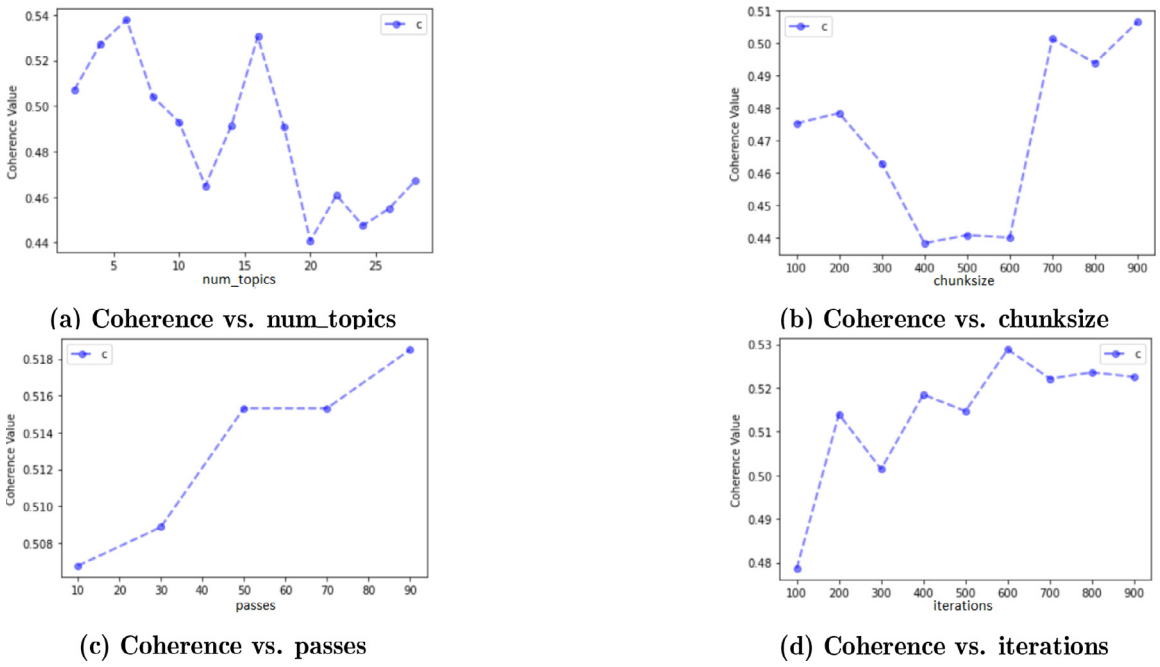
4.1. Cyber-security themes

We use the LDA topic model to generate the topic distributions for both the peer reviewed literature (see Appendix Table 4) and the non-peer reviewed literature (see Appendix Table 5). Next, we manually annotate each of these weighted distribution of words as a cyber-security theme.

4.1.1. Cyber-security themes in the peer-reviewed literature corpus

Table 2 presents the inferred cyber-themes for the peer-reviewed corpus.

" Pre-COVID-19 cyber-themes. Prior to the pandemic, we observe that the research community much focused on the threats on evolving technology. Focus was on to develop analytical attack



(a) Coherence vs. num_topics (b) Coherence vs. chunksize

(c) Coherence vs. passes (d) Coherence vs. iterations

Fig. 3. Coherence score metric vs. Hyper-parameters to determine the optimal values for the All-time non-peer-reviewed corpus.

models, for example using machine learning and game-theory, ensuring network security and customer privacy. Categorically, from the tuned LDA model, we observe specific cyber-security themes, for example, the "Malware detection, the "Blockchain and cryptocurrency, "Attack models based on game-theory" appear as significant research themes. A perpetual research theme is of building novel network intrusion detection mechanisms. We believe this is due to a continuous evolution of the cyber-threats Zhou and Jiang (2012). In the same line, industrial control systems have undergone a tremendous transformation with an enhanced connectivity to the public networks and the adoption of cloud infrastructure Bhamare et al. (2020). As a result, their security is a subject of intense research. Similarly, complex digital infrastructure necessitates anticipating malicious activity and instigating response in time Khraisat et al. (2019). Machine learning-based intrusion detection systems have been proved to be useful.

" Post-COVID-19 cyber-themes. Since the pandemic, the research community has focussed on novel cyber-security themes. In particular, certain application domains have gained prominence such as of healthcare, digital banking, supply-chain vulnerabilities. Categorically, from our LDA model, we observe that cyber-themes such as of 'Cyber-security in healthcare", "Cyber-resilience in supply chain", and "Security awareness" have piqued the researchers' interest. We believe that such themes have gained prominence as a spill-over of the pandemic necessities, for example, the increasingly tele-healthcare services Hill et al. (2021), IoT-based healthcare infrastructure Vedaei et al. (2020), and use of robotics Khan et al. (2020). Consequently, cyber risk management has continued to remain a prominent topic. Alongside, other cyber-themes, such as of "Malware detection", "Control system security", and "Intrusion detection system" have remained well-researched in perpetuity.

4.1.2. Cyber-themes in non-peer-reviewed corpus

Table 3 presents the inferred cyber-themes for the non-peer-reviewed corpus.

" Pre-COVID-19 trends. Prior to the pandemic, we observe that the practitioners emphasize on certain cyber-themes such as of the common security vulnerabilities, malware incidents, and data breaches. Interestingly, financial cyber-crimes such as fraud and its implication on organizational health are well-published. We believe a reason for such trend is that attacks on sectors such as banking, government institutes and critical infrastructures brings media attention. Bloggers with an aim of inciting huge followers many times tend to paint stories inciting human curiosity with information on cyber-attack victim, shadowy espionage groups, and tactical attacks.

" Post-COVID-19 trends. Since the pandemic, we observe that in the non-peer-reviewed literature, bloggers have authored on a few distinct cyber-themes. One such example is of healthcare SonicWall (2020), which is well researched topic in the post-pandemic peer reviewed literature too. Categorically from the tuned LDA models, we observe that the newer attack tactics such as of "Social engineering and Side channel attack have picked the interest. With a few recent incidents of ransomware and its crippling impact on organization, bloggers have also authored extensively on the topic. Malware has been extensively emphasized in both the pre-and the post-COVID-19 literature. Notable, whereas in the pre-COVID-19 literature, the focus was its implication on business processes, in the post-COVID-19 literature, the focus has shifted towards malware threats on evolving digital infrastructures such as of cloud services.

4.2. Evolution of cyber-security themes

In addition to the identification of distinct cyber-security themes, in this paragraph, we present its year-wise evolution. We do this by tabulating the cumulative count of the documents underlying a cyber-security theme. The aim behind this activity is to understand the temporal popularity of the cyber-security themes. To accomplish the task, we use the all-time literature corpus, see Table 4 for cyber-security themes in peer-reviewed literature and Table 5 for cyber-security themes in non-peer-reviewed literature.

Evolution of cyber-themes in the peer-reviewed literature. Figure 4 shows the different plots taking into account the peer-reviewed literature and the individual cyber-security themes. In these plots, the x-axis represents the year of publication and the y-axis represents the cumulative count of published documents in that

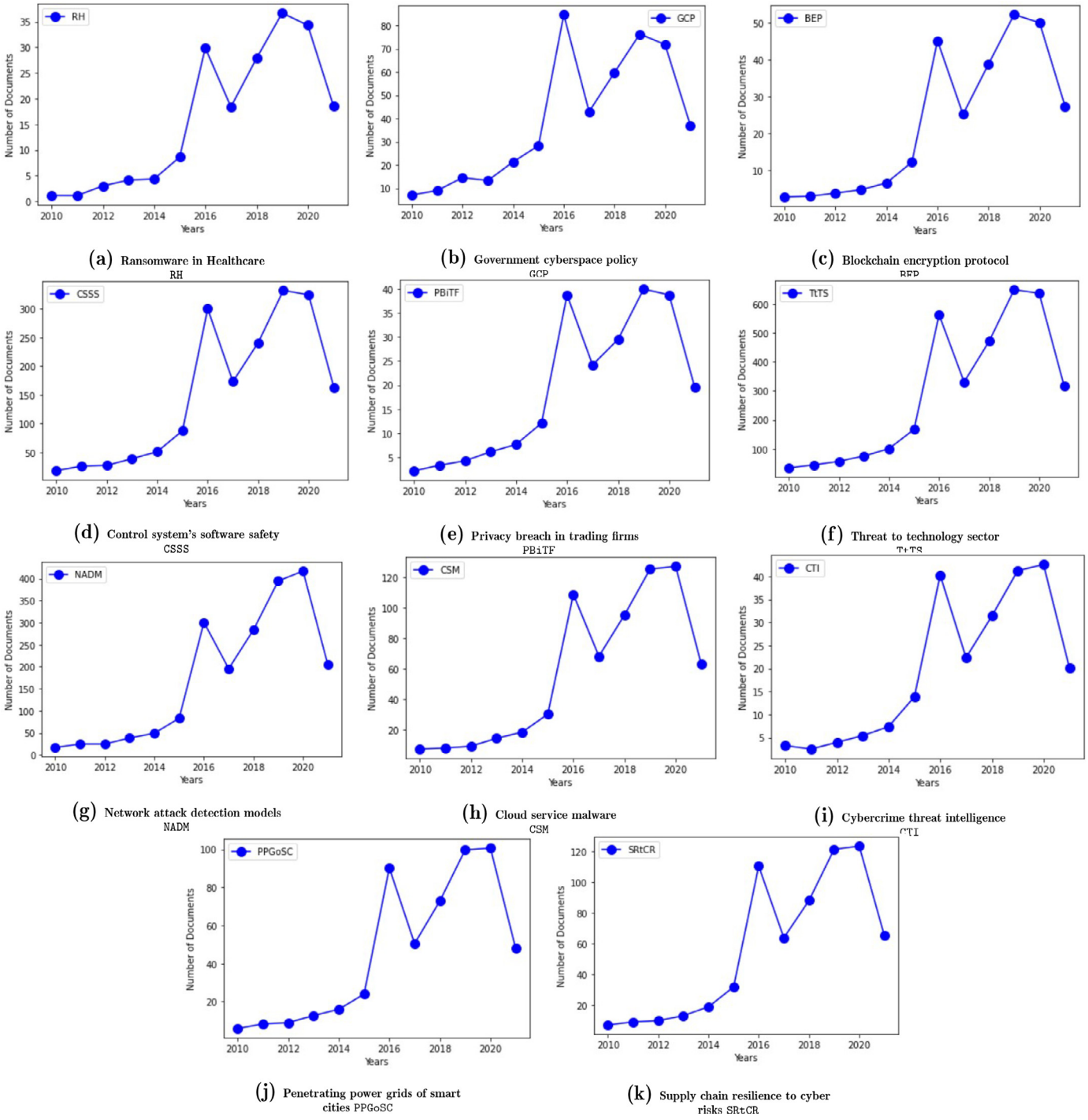


Fig. 4. Number of documents vs. time plot for each cyber-theme in peer-reviewed corpus. The year 2020 marks the beginning of the COVID-19 pandemic (shown by a vertical dotted red line).

year. Glancing over these plots reveal interesting observations. We find that all the research themes, namely, "Cybersecurity in healthcare", "Cyber-security risk management"," Control system security", "Cyber-governance", "Intrusion detection models" and "Malware detection" – show a steady uptick in the number of publications till the year 2016. Compared to the year 2018, post the year 2019, the year of COVID-19 pandemic, a few topics such as "Cyber-security risk management", "Cybersecurity in healthcare" and "Control-system security" have shown a declined trend. Comparatively, other cyber-themes such as of "Cyber-governance", "Intrusion detection" and "Malware" have shown an uptick beyond the year 2019 with an intermittent decline. The publication trends put in these plots specific to the year 2021 may not reflect the true picture as all the document may not been archived in the databases at the time of information retrieval of the articles.

In addition to the temporal evolution of the cyber-themes, we investigate the spatial popularity of a theme in the corpus. We do this by calculating a ratio of the number of publications investigating a theme vs the cumulative count of all the publications in the corpus. Interesting observations follow. We observe that the topics such as Control system securityCyber-governance, and Intrusion detection models are the dominating topics in both the pre-COVID

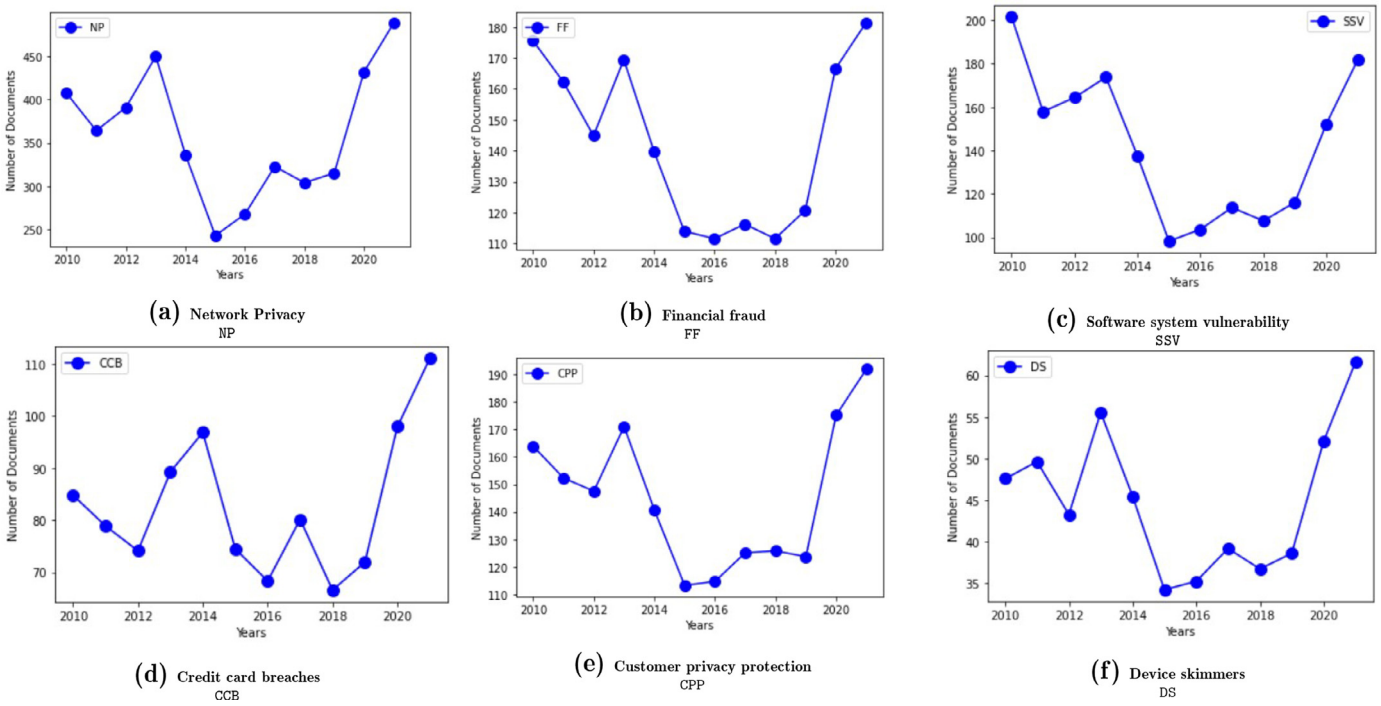


Fig. 5. Year vs Number of documents for each cyber-theme in non-peer-reviewed corpus. The year 2020 marks the beginning of the COVID-19 pandemic (shown by a vertical dotted red line).

19 and the post-COVID-19 period. In the case of "Cyber-security in healthcare, we observe that prior to the pandemic, merely a aggregate of 129 documents in the corpus collected over a ten year period. However, in the two years since the pandemic, the total number of documents have risen to 53. This supports our claim that research in the field of healthcare cyber-security is gaining prominence. In addition to this, we observe that cyber-themes such as of "Control system security", "Cyber security risk management", and "Intrusion detection system" are heavily researched when compared with other cyber-themes. These themes account for 77.8% of documents in the peer-reviewed literature. Also notable is that a few popular cybersecurity research sub-domains such as "Ransomware in healthcare", "Blockchain and cryptocurrency, Privacy breach" are loosing their grounds. One possible explanation may be a lack of technology pull or legislation for furthering expanding the research in these sub-domains.

Evolution of cyber-themes in the non-peer-reviewed literature. Figure 5 shows the different plots taking into account the non peer-reviewed literature and the individual cyber-security themes. We observe that the number of documents for all the topics in the pre-COVID-19 time-line (2010-2019) follow an inconsistent pattern. However, for all the topics, the number of documents have increased in the post-COVID-19 timeline (2020-2021]. Similar to peer-reviewed literature, we obtain a spatial distribution of cyber- security themes. We observe that, Network privacy is the dominating topic in both the pre- and post-COVID-19 epochs. It accounts for 36.8% of all documents in the corpus. Research on network privacy is trending, which may be a reflection of the recent focus on work-from-home models and transition to cloud infrastructure by the industry, consequently a leading avenue of research. Also notable is that cybersecurity in sub-domains such as Device Skimmers and Credit Card Breaches are showing the signs of decline. One possible explanation of this trend is an increased awareness on cyber-threats to these systems.

Evolution of cyber-themes in non-peer-reviewed literature. Figure 5 captures the number of documents vs. time plot for each topic in the non-peer-reviewed corpus. Prior to the pandemic, we see an inconsistent trend in the number of documents for all cyber-themes such as of "Financial fraud", "Software system vulnerabilities", "Customer credentials", "Credit card breaches", "Device skimmers", and "Network privacy and security in healthcare". However, since the pandemic, the number of documents for all cyber-themes has increased, indicating that these cyber-themes have received more attention. Furthermore, when compared to other cyber-themes, "Network privacy and security in healthcare" has received the most attention in the non-peer-reviewed literature. It accounts for 36.8% of all documents in the corpus.

Albeit, in the non-peer-reviewed literature, prior to the pandemic, overarching topics such as device skimmers, financial fraud and data breaches were the focal themes. Since the pandemic, privacy and security in healthcare, as well as software vulnerability analysis are more researched. Both peer-reviewed and non-peer reviewed literature point to a differentiated new normal in the cyber-security. We believe our work is a vital fulfillment on scientifically capturing the impact of pandemic upheaval on cyber-security. It also reinforces the need to quickly adapt with the changing societal needs.

5. Conclusion

The COVID-19 pandemic upheaval has induced the society to an unforeseen new normal. With an increased digital adoption, we witness newer forms of cyber-security challenges. In this paper, we scientifically examine this pandemic-induced affect on cyber-security by identifying the key themes, examining their temporal and spatial evolution. Technically, to do this, we utilize the state-of-the-art unsupervised machine learning algorithm of topic models. We build a massive corpus of over twenty thousand literature taken from the years 2010-2021 including both the peer-reviewed and the non-peer reviewed literature sources. Interesting observations we find is that post COVID-19 pandemic, health-care, cyber-resilience are the newer added cyber-security themes researched in the peer-reviewed litera-

Table 5
Labelled topics for the non-peer-reviewed corpus from LDA generated topic distribution. Here NPR= Non-peer-reviewed literature taken from four cybersecurity blogs (refer Section 3).

Corpus Name	Topic	Topic Distributions	Label
Pre- COVID	T0	firm, service, account, phone, address, program, password, vulnerability, protection, order	Common security vulnerabilities
(2010- 2019)	T2	government, part, customer, group, news, case, name, page, fact, encryption	Fake news
	T3	time, access, privacy, business, malware, email, today, industry, consumer, code	Malware in business processes
	T4	network, technology, bank, banking, cloud, process, approach, tech, community, cybercrime	Cybercrime in banking
	T6	breach, data, report, problem, issue, web, law, database, company, gang	Data breach reports
	T7	company, software, card, site, credit, system, control, fraud, management, payment	Credit card frauds
	T8	use, story, money, tool, organization, hacker, patch, domain, ability, victim	Hacker's victims' story
Post- COVID	T0	time, attack, site, email, use, activity, discussion, bank, organization, encryption	Side channel attacks
(2020- 2021) (NPR)	T1	network, today, group, customer, someone, credit, authentication, something, hack, digital_transformation	Network security
	T2	time, attack, site, email, insider, activity, discussion, bank, organization, encryption	Social engineering attack
	T2	company, software, service, malware, system, computer, code, password, day, file	Software vulnerability analysis
	T3	malware, fraud, work, cloud, identity, compromise, investment, campaign, services, member	Malware in cloud services
	T4	technology, ransomware, reading, week, traffic, process, government, accompanying_podcast, analysis, infrastructure	Ransomware
All- time	T0	network, time, business, privacy, company, year, attack, technology, report, government	Network Privacy
(2010- 2021)	T1	story, bank, group, fraud, site, money, attack, week, case, botnet	Financial fraud
	T3	card, credit, breach, company, payment, customer, encryption, fraud, source, investigation	Credit card breaches
	T4	service, number, account, email, access, phone, address, site, name, password, customer	Customer credentials
	T5	device, surveillance, location, iPhone, skimmer, machine, cash, camera, wireless, reader	Device skimmers
	T5	software, malware, system, vulnerability, code, today, version, computer, flaw, program	Software system vulnerability

ture. Typically, in pre-COVID-19 period, the focus was on building novel attack models, intrusion detection systems, among others. In the non-peer literature, post COVID-19, we observe newer forms of attacks such as of social engineering, side channel have spurred the interest. Typically, this community in the pre-COVID-19 period focused much on financial cyber-crimes. Alongside, a few cyber-themes, such as network security, malware detection, intrusion control systems and industrial control system security remain important in perpetuity. Interestingly, while doing temporal analysis of the cyber-themes in all-time-peer-reviewed literature, we observe that blockchain and privacy breaches have become been less popular as compared to the previous years. Similarly, research on network privacy is trending in all-time non-peer reviewed literature, which may be a reflection of the recent focus on work-from-home models and transition to cloud infrastructure by the industry, consequently a leading avenue of research. Also notable is that cybersecurity in sub-domains such as Device Skimmers and Credit Card Breaches are showing signs of decline.

We believe our work provide crucial insights on cyber-security trends and challenges useful for both academicians and practitioners. It will aid business in identifying their own strategic gaps, thus rationally recognizing their cyber-security needs, hire manpower and inventory management. Academicians will benefit from our work by appropriating their research on upcoming cybersecurity challenges and trending niche domains. Future work is to make the corpus more inclusive to include media reports. Another interesting extension is to automatically infer the label from a topic-distribution.

6. A non-comprehensive list of Labeled topics for peer and non-peer-reviewed literature

Declaration of Competing Interest

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

CRediT authorship contribution statement

Rajesh Kumar: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing, Investigation, Methodology. Siddharth Sharma: Data curation, Formal analysis, Investigation, Writing – original draft. Chirag Vachhani: Formal analysis, Validation, Investigation, Writing – original draft. Nitish Yadav: Data curation, Formal analysis, Validation, Investigation, Writing – original draft.

References

Acohido, B. V., 2000. The last watchdog. Last accessed: January 17, 2022, https://www.lastwatchdog.com/.
Adams, S., Carter, B., Fleming, C., Beling, P.A., 2018. Selecting system specific cybersecurity attack patterns using topic modeling. In: 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, pp. 490–497.
Alagheband, M., Mashatan, A., Zihayat, M., 2020. Time-based gap analysis of cybersecurity trends in academic and digital media. ACM Trans. Manag. Inf. Syst. (TMIS) 11, 1–20.
Atzrodt, C., Maknojia, I., McCarthy, R., Oldfield, T., Po, J., Ta, K., Stepp, H., Clements, T., 2020. A guide to covid-19: A global pandemic caused by the novel coronavirus sars-cov-2. The FEBS J. (17) 3633–3650.
Bari, Y. D., 2021. 2021 cybersecurity trends report. Last accessed: January 19, 2022, https://www.infosys.com/iki/insights/.
Bechor, T., Jung, B., 2019. Current state and modeling of research topics in cybersecurity and data science. Syst. Cybernet Inform. 17, 27.

Bhamare, D., Zolanvari, M., Erbad, A., Jain, R., Khan, K., Meskin, N., 2020. Cybersecurity for industrial control systems: A survey. Comput. Secur. 89, 101677.
Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit.
Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Bouma, G., 2009. Normalized (pointwise) mutual information in collocation extraction. Proc. GSCL 30, 31–40.
Burgess, M., 2020. Hackers are targeting hospitals crippled by coronavirus. Last accessed: January 26, 2022, https://www.wired.co.uk/article/.
Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D., 2009. Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009., pp. 288–296.
Chen, H., Zhang, G., Zhu, D., Lu, J., 2017. Topic-based technological forecasting based on patent data: A case study of australian patents from 2000 to 2014. Technol. Forecast. Soc. Change 119, 39–52.
Coventry, L., Branley, D., 2018. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas 113, 48–52.
Dhillon, G., Smith, K., Dissanayaka, I., 2021. Information systems security research agenda: Exploring the gap between research and practice. J. Strategic Inf. Syst. 30, 101693.
DHS, 2021. Advisory memorandum on ensuring essential critical infrastructure workers' ability to work during the covid-19 response. Last accessed: January 26, 2022, https://www.cisa.gov/publication/.
Dinu, C., 2022. Top cyber-security blogs to follow. Last accessed: March 22, 2022, https://heimdalsecurity.com/blog/.
Dumais, S.T., 2004. Latent semantic analysis. Annu. Rev. Inf. Sci. Technol. 38 (1), 188–230.
EUA, 2021. Enisa threat landscape 2021. Last accessed: January 26, 2022, https://www.enisa.europa.eu/publications/.
Gatti, C., Brooks, J., Nurre, S., 2015. A historical analysis of the field of or/ms using topic models. ArXiv preprint.
George, C.P., Doss, H., et al., 2017. Principled selection of hyperparameters in the latent dirichlet allocation model. J. Mach. Learn. Res. 18 (1), 5937–5974.
Gernhardt, D., Groš, S., 2021. Use of a non-peer reviewed sources in cyber-security scientific research. arXiv preprint arXiv:2106.06000.
Hill, W.E., Murphy, M.S., Hills, K.T., 2021. Assessment of virtual healthcare: Predictors of access and utilization before, during, and after the covid-19 pandemic. Med. Sci. Pulse 15.
Hoffman, M., Blei, D., Bach, F., 2010. Online learning for latent dirichlet allocation. In: Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010., pp. 856–864.
Hofmann, T., 2013. Probabilistic latent semantic analysis. CoRR abs/1301.6705.
Interpol, 2020. Global landscape on covid-19 cyberthreat. Last accessed: January 26, 2022, https://www.interpol.int/en/Crimes/Cybercrime/COVID-19-cyberthreats.
Kang, H., Kim, C., Kang, K., 2019. Analysis of the trends in biochemical research using latent dirichlet allocation (lda). Processes 7 (6), 379.
Kawata, S., Fujiwara, Y., 2016. Constructing of network from topics and their temporal change in the nikkei newspaper articles. Evol. Inst. Econ. Rev. 13, 423–436.
Khan, Z.H., Siddique, A., Lee, C.W., 2020. Robotics utilization for healthcare digitization in global covid-19 management. Int. J. Environ. Res. Public Health 17 (11), 3819.
Khanbhai, M., Warren, L., Symons, J., Flott, K., Harrison-White, S., Manton, D., Darzi, A., Mayer, E., 2022. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int. J. Med. Inf. 157, 104642.
Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J., 2019. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity 2, 1–22.
Krebs, B., 2000. Krebs on security. Last accessed: January 17, 2022, https://krebsonsecurity.com/.
Kumar, R., Kela, R., Singh, S., Trujillo-Rasua, R., 2022. Apt attacks on industrial control systems: A tale of three incidents. Int. J. Crit. Infrastruct. Protect. 37, 100521. doi:10.1016/j.ijcip.2022.100521.
Lallie, H.S., Shepherd, L.A., Nurse, J.R.C., Erola, A., Epiphaniou, G., Maple, C., Bellekens, X.J.A., 2021. Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. Comput. Secur. 105, 102248.
Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.
Lemay, A., Calvet, J., Menet, F., Fernandez, J.M., 2018. Survey of publicly available reports on advanced persistent threat actors. Comput. Secur. 72, 26–59.
Lu, Y., Da Xu, L., 2018. Internet of things (IoT) cybersecurity research: A review of current research topics. IEEE Internet Things J. 6, 2103–2115.
Manadhata, P.K., Wing, J.M., 2010. An attack surface metric. IEEE Trans. Softw. Eng. 37, 371–386.
McKeown, K.R., Radev, D.R., 2000. Collocations. Handbook of Natural Language Processing. Marcel Dekker 1–23.
McKinsey, 2020. The path to the next normal. Last accessed: January 25, 2022, https://www.mckinsey.com/.
Microsoft, 2022. Microsoft security. Last accessed: January 15, 2022, https://www.microsoft.com/security/blog/.
Mimno, D.M., McCallum, A., 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In: UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, pp. 411–418.
Muthuppalaniappan, M., Stevenson, K., 2021. Healthcare cyber-attacks and the covid-19 pandemic: an urgent threat to global health. Int. J. Qual. Health Care 33 (1), mzaa117.
Nabe, C., 2021. Impact of covid-19 on cybersecurity. Last accessed: January 20, 2022, https://www2.deloitte.com/ch/en/pages/risk/articles/.
Patel, A.A., Arasanipalai, A.U., 2021. Applied Natural Language Processing in the Enterprise.
Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M., 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 569–577.
Röder, M., Both, A., Hinneburg, A., 2015. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408.
Schneier, B., 2004. Schneier on security. Last accessed: January 16, 2022, https://www.schneier.com/.
Sleeman, J., Finin, T., Halem, M., 2021. Understanding cybersecurity threat trends through dynamic topic modeling. Front. Big Data 4.
SonicWall, 2020. Healthcare cybersecurity in the pandemic. Last accessed: January 15, 2022, https://www.sonicwall.com/medialibrary.
Srivastava, A., Sahami, M., 2009. Text mining: Classification, clustering, and applications. CRC press.
Syed, S., Spruit, M., 2017. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics, pp. 165–174.
Thanaki, J., 2017. Python natural language processing. Packt Publishing Ltd.
Vedaei, S.S., Fotovvat, A., Mohebbian, M.R., Rahman, G.M., Wahid, K.A., Babyn, P., Marateb, H.R., Mansourian, M., Sami, R., 2020. Covid-safe: an iot-based system for automated health monitoring and surveillance in post-pandemic life. IEEE access 8, 188538.
Wallach, H., Murray, I., Salakhutdinov, R., Mimno, D., 2009. Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning, Vol. 382, pp. 1105–1112.
Wang, S.I., Manning, C.D., 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 90–94.
Wang, X., McCallum, A., Wei, X., 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), pp. 697–702.
Warburton, D., 2020. Phishing attacks soar 220% during covid-19 peak as cybercriminal opportunism intensifies. Last accessed: January 27, 2022, https://www.f5.com/company/news/features/.
Wu, T., Ma, W., Wen, S., Xia, X., Paris, C., Nepal, S., Xiang, Y., 2021. Analysis of trending topics and text-based channels of information delivery in cybersecurity. ACM Trans. Internet Technol. (TOIT) 22, 1–27.
Zhou, Y., Jiang, X., 2012. Dissecting android malware: Characterization and evolution. In: 2012 IEEE symposium on security and privacy. IEEE, pp. 95–109.