Supplementary Materials for: Text Mining in Cybersecurity: A Systematic Literature Review

LUCIANO IGNACZAK, GUILHERME GOLDSCHMIDT, CRISTIANO ANDRÉ DA COSTA, and RODRIGO DA ROSA RIGHI, Laboratory of Software Innovation, Unisinos University, Brazil

1 RESEARCH METHODOLOGY

This study has applied the methodology of **Systematic Literature Review (SLR)** [61], which consists of a method to extract, synthesize, and summarize information based on a collection of studies. An SLR is a means to identify, evaluate, and interpret all available research relevant to a particular research question, topic area, or phenomenon of interest [39]. We performed this study according to the framework provided by Kitchenham et al. [40], which is arranged in three phases: (i) Plan Review: related to the design of the research protocol, comprises tasks such as the research questions, the search strategy, and the criteria to select the studies; (ii) Conduct Review: involves the tasks to identify the SLR corpus and which type of information needs to be extracted; (iii) Document Review: comprehends the writing report that needs to be formatted as a paper and must show the contribution to the research area.

1.1 Planning the Review

The first step is to determine the research questions, because they guide the selection of studies and define what type of data needs to be extracted. This step is based on the SLR objective and can be divided into several questions. This work aims to identify different applications of text mining to address issues related to cybersecurity, and it defines the following research questions:

- RQ1: How would be the taxonomy representing the application of text mining in the cybersecurity domain?
- RQ2: In which contexts has cybersecurity applied text mining?
- RQ3: What strategies of text mining have been utilized?
- RQ4: How has text classification performed in the cybersecurity domain?
- RQ5: How has the cybersecurity industry been applying text mining in real-world solutions?

The RQ1 aims to enumerate several cybersecurity activities that use text mining tasks to deal with security challenges. To organize the taxonomy, we introduce five cybersecurity domains and associated cybersecurity activities with them. The RQ2 proposes the analysis of four aspects concerning the application of text mining in the cybersecurity domain. One aspect is the types of content analyzed in the studies. Other aspects comprise the industries and technologies targeted by the text mining application. The last aspect addresses the languages of the datasets analyzed in the studies.

The RQ3 presents the strategies in the use of text mining. The strategy comprises three aspects related to the application of text mining tasks. The first aspect analyzes which text mining tasks are applied alone. The RQ3 also investigates which tasks were combined to evaluate text mining

0360-0300/2021/06-ART140 $15.00
https://doi.org/10.1145/3462477

Table 1. Keywords Used in the Slr's Search Query

Search Topic	Keywords
Cybersecurity	cybersecurity, cybersecurity, cyber-security, information security
Text Mining	text mining, information retrieval, natural language processing, information extraction, text analysis, text classification, text clustering, text summarization

performance. Last, the question evaluates the use of neural networks to support text mining in the cybersecurity domain. The RQ4 evaluates the text classification performance in different activities presented in the proposed taxonomy. The last research question (RQ5) investigates the use of text mining in real-world applications by the cybersecurity industry.

The protocol also involves the search strategy used to discover the studies related to the research questions. This study defines a broad strategy aiming to cover, as much as possible, the relevant research associated with it. The keywords are split into two search topics connected with the main areas of the study. The keywords related to the cybersecurity area comprehend the possible variations of the term. We also select studies related to information security, because there is much in common between the concepts. Other keywords try to cover the different expressions seen in text mining studies. Table 1 presents the groups of keywords. The search query was built using the logical operator "OR" between keywords of the same group and the logical operator "AND" between the groups.

The search strategy limits the use of search queries in some widely used scientific databases in computer science. The authors chose the sources based on a reference study [16] and other SLR conducted in the research field [53, 78]. The databases used are ACM,[1] IEEE Xplore,[2] Science Direct,[3] Scopus,[4] Springer,[5] and Web of Science.[6] The protocol compares the search query with the title, keywords, and abstract fields in each database. The search should span 10 years.

The planning also establishes the criteria to exclude or include a study in this SLR. According to Kitchenham et al. [40], criteria are essential to get evidence that the studies contribute to answering the research questions. This study defines six **exclusion criteria (EC)** and four **inclusion criteria (IC)** that are shown in Table 2. After executing the search queries in the different databases, the first task is to put together all results in a single file and delete duplicated entries. These studies are considered the candidate papers, and the next step is the application of the exclusion criteria.

The first exclusion criterion aims to remove all grey literature found in the result, because the authors decided to use just peer-reviewed studies. The second criterion removes short papers, and the third deletes secondary papers. The fourth criterion keeps just papers written in English, and the fifth holds just the complete version of each candidate paper. The last exclusion criterion deletes all candidate papers whose authors cannot access.

The study protocol defined four inclusion criteria to analyze candidate papers related to the application of text mining in the cybersecurity domain and assure that text mining is used to extract knowledge from unstructured content. The first criterion limits the candidate papers related to studies applying text mining to solve a cybersecurity problem. This criterion allows us to consider

[1] https://dlnext.acm.org/.
[2] https://ieeexplore.ieee.org/.
[3] https://www.sciencedirect.com/.
[4] https://www.scopus.com/.
[5] https://link.springer.com/.
[6] https://webofknowledge.com.

Table 2. Criteria Applied to Include or Exclude Studies in the SLR

ID	Inclusion/Exclusion Criteria
IC1	The study is related to the application of text mining to solve a cybersecurity challenge
IC2	Text mining must be the main method used to solve a cybersecurity challenge
IC3	Text mining techniques must be applied in unstructured content as a document
IC4	The study must have an experiment or case study, and present consistent results
EC1	The study not published in a journal or conference
EC2	The study is shorter than eight pages
EC3	The study is a literature review
EC4	The study is not written in English
EC5	The study is a shorter version of another study
EC6	Authors cannot access the full paper

exclusively papers focused on cybersecurity and remove general studies that only perform an evaluation using security data. The next criterion restricts the selection to candidate papers applying text mining as the primary method. This criterion is needed, because experiments can use several methods, and text mining can have a secondary role.

The third criterion limits candidate papers applying text mining to extract knowledge from unstructured content. For example, studies applying text mining in logs, URLs, and passwords are not included in the SLR. The last inclusion criterion keeps studies with a comprehensive and coherent presentation of the results based on some experiment or case study.

The protocol determined the application of the inclusion criteria in three rounds. The first round comprises the reading of the title and abstract. In the first round, authors just removed a paper if they are pretty sure that it is not related to the research questions. The second round comprises an analysis comprehending the introduction and the sections associated with methodology and experiment/case study. This round offers the confidence level that only papers strongly associated with the SLR research questions will be selected for the third round. The last round comprises full-reading and extraction of the information to answer the research questions.

Another decision took during the planning phase is about the quality of the studies. Kitchenham et al. [40] defined the quality assessment as an essential part of the process, because this step can improve the value of SLR, giving the reviewers the option to exclude papers that do not reach the expected quality level. This study uses the h5-index[7] score as a quality criterion, excluding all proceedings or journal papers with a score minor than 10.

1.2 Conducting the Review

The second phase determines the implementation of the protocol, following the decisions taken previously. The authors performed the search queries in the databases starting April 12, 2019, to April 17, 2019, and found 2.472 candidate papers. As defined in the protocol, the search sought the string in the title, keywords, and abstract. The only exception was Springer, which does not have this option. In this case, the search sought the expression in the whole paper. The next task located

[7] https://scholar.google.com/intl/en/scholar/metrics.html#metrics.

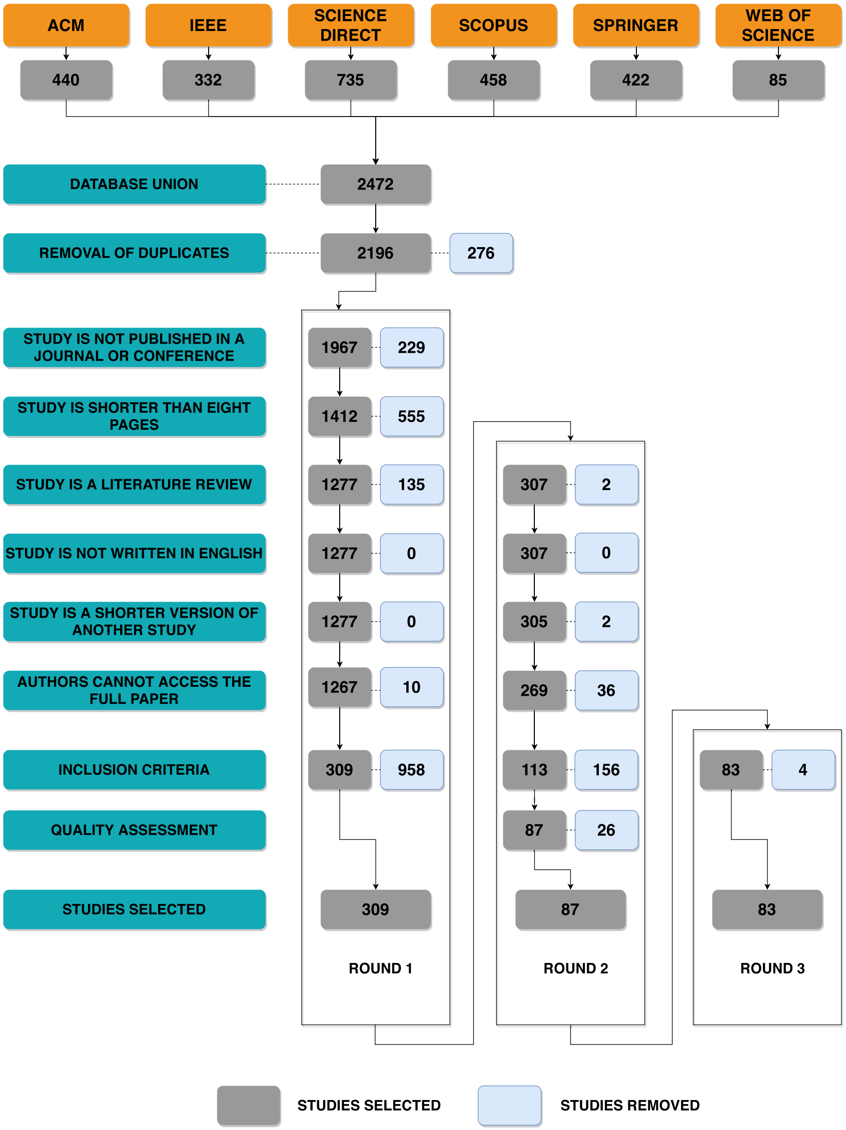

Fig. 1. The process applied to select the studies included in the SLR.

276 candidate papers with two or more occurrences and removed the duplicates. Some duplicate studies were found in different databases citing distinct years of publication. In these situations, the authors decided to remove the entry with the older publication year.

After the exclusion of duplicate studies, the remaining 2.196 candidate papers were applied to exclusion criteria. In the first round, the exclusion criteria removed 929 candidate papers, and the authors analyzed the abstract and title of the 1.267 remainings. After the analysis, 309 studies were selected for a more in-depth investigation on the second round. The 784 candidate studies were not selected because it was evident that they were not related to cybersecurity or text mining. Figure 1 presents the steps performed to select the studies. In the figure, the grey box, on the left, informs the number of studies selected to advance in the selection, and the blue box, on the right, presents the number of studies removed based on the criterion.

To accomplish the second round, the authors needed to download the remaining candidate papers, check if they do not match some exclusion criteria, and read the sections defined in the protocol. The first and second criteria were not applied because they would not exclude additional studies. Considering the initial set of 309 studies, we analyzed only 269. Forty candidate papers were removed, since they matched an exclusion criterion. After analyzing each candidate paper, the authors selected 113 studies to continue in the process.

The authors decided to apply the quality assessment in round 2. This decision was taken due to the number of candidate papers. The protocol established the h5-index as a threshold of quality, and there is no reason for a detailed reading of a set of papers that quality criterion will remove. The application of all journals and conferences to check the h5-index metric was performed from June 28 through June 30 of 2019. It resulted in the exclusion of 26 studies, so the number of candidate papers reaching the next round was 87. After the complete reading, the authors decided to exclude four studies. The remaining 83 papers were used in the systematic review.

2 THREATS TO VALIDITY

This SLR answers some research questions related to the application of text mining tasks to support cybersecurity activities. Nonetheless, potential limitations can impact the results exposed in this study, and they are discussed in this section.

To identify the primary studies to this SLR, we used search strings based on keywords extracted from a group of studies related to text mining and cybersecurity. However, different keywords can add studies associated with this SLR research topic not included in our corpus. To avoid excluding relevant primary studies, we analyzed 269 studies to select the corpus properly. The final selection was based on our view about the type of unstructured content that should be included in the SLR, so we did not consider studies restricted with short strings or studies that used just preprocessing text mining methods.

We analyzed several secondary studies, and we did not identify a consensus about the criteria applied in the quality assessment. Some secondary studies did not even mention the use of a quality assessment in the methodology section. In this SLR, we decided to use objective criteria, defining a metric and the cutoff score. We chose the h5-index because we believe that a metric based on 5-year citations can measure conferences and journals' quality. However, we are aware that any metric can present fails and include questionable studies.

We organized the proposed taxonomy based on cybersecurity domains, and all selected studies were connected to only one domain. We established the relationship between the study and the domain according to our understanding after reading the study. However, we identified that some studies could connect with more than one domain. For example, a study can use text mining to detect an attack on a computer system, so we understand the study also comprehend a cybersecurity incident and could also be related to the domain. In these cases, we decided to link the study with the domain more emphasized in the paper, even if we disagree about the domain bias included in the study. We think other researchers can establish different relations between studies and domain, so we consider this a threat to validity.

In the SLR, we presented the strategies adopted by the studies and explained technical aspects of the experiments' methods. The descriptions reflect our best efforts to comprehend the methods introduced by the studies' authors. In our view, some studies lacked details about the reasons related to the methods selection and implementation. Additionally, studies informed customized implementations based on related works but did not introduce the differences. Finally, we limited the data extraction to information present in the studies to answer the research questions. The only data inferred by the authors were the language of some datasets based on indirect information. Therefore, the SLR results can present inaccurate results case the studies' authors omitted some information used in their work.

3 SUMMARY OF THE SELECTED STUDIES

The studies selected in the final of the research methodology process are shown in Table 3. They are sorted alphabetically by author and related to their domain.

Table 3. Studies selected for the SLR

Reference in the Taxonomy	Authors and Year	Title	Domain
[4]	[Abuhamad et al. 2019]	Code authorship identification using convolutional neural networks	Incident
[5]	[Adams et al. 2018]	Selecting System Specific Cybersecurity Attack Patterns Using Topic Modeling	Attack
[6]	[Adewole et al. 2017]	SMSAD: a framework for spam message and spam account detection	Control
[11]	[Al-Rowaily et al. 2015]	BiSAL - A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security	Threat
[12]	[Aldwairi and Alwahedi 2018]	Detecting Fake News in Social Media Networks	Attack
[13]	[Alneyadi et al. 2013]	Adaptable N-gram classification model for data leakage prevention	Control
[15]	[Alohaly et al. 2018]	A Deep Learning Approach for Extracting Attributes of ABAC Policies	Control
[18]	[Amato et al. 2019]	Analyse digital forensic evidences through a semantic-based methodology and NLP techniques	Incident
[19]	[Amato et al. 2015]	An integrated framework for securing semi-structured health records	Control
[20]	[An and Kim 2018]	A Data Analytics Approach to the Cybercrime Underground Economy	Threat
[24]	[Ban et al. 2018]	A performance evaluation of deep-learnt features for software vulnerability detection	Vulnerability
[25]	[Banik and Bandyopadhyay 2018]	Novel Text Steganography Using Natural Language Processing and Part-of-Speech Tagging	Control
[26]	[Beebe et al. 2011]	Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies	Incident
[27]	[Beebe and Liu 2014]	Clustering digital forensic string search output	Incident
[28]	[Benjamin et al. 2016]	Examining Hacker Participation Length in Cybercriminal Internet-Relay-Chat Communities	Threat
[36]	[Chang and Clark 2014]	Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method	Control
[37]	[Chen et al. 2011a]	Assessing the severity of phishing attacks: A hybrid data mining approach	Incident
[38]	[Chen et al. 2011c]	Steganalysis against substitution-based linguistic steganography based on context clusters	Attack

(Continued)

Table 3. Continued

Reference in the Taxonomy	Authors and Year	Title	Domain
[39]	[Chen et al. 2011b]	Detection of substitution-based linguistic steganography by relative frequency analysis	Attack
[44]	[Cohen et al. 2016]	SFEM: Structural feature extraction methodology for the detection of malicious office documents using machine learning methods	Control
[45]	[Confente et al. 2019]	Effects of data breaches from user-generated content: A corporate reputation analysis	Incident
[47]	[Deliu et al. 2018]	Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks	Threat
[49]	[Deshpande et al. 2014]	The Mask of ZoRRo: preventing information leakage from documents	Control
[52]	[Edwards et al. 2017]	Panning for gold: Automatically analysing online social engineering attack surfaces	Attack
[53]	[El-Alfy and AlHasan 2016]	Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm	Control
[54]	[Fang et al. 2019]	Analyzing and Identifying Data Breaches in Underground Forums	Incident
[59]	[Gonzalez-Compean et al. 2019]	A policy-based containerized filter for secure information sharing in organizational environments	Control
[67]	[He et al. 2011]	An efficient phishing webpage detector	Control
[69]	[Hendler et al. 2018]	Detecting Malicious PowerShell Commands using Deep Neural Networks	Attack
[70]	[Holton 2009]	Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem	Attack
[72]	[Huang et al. 2016]	A study on Web security incidents in China by analyzing vulnerability disclosure platforms	Incident
[74]	[Huang et al. 2018]	A novel mechanism for fast detection of transformed data leakage	Control
[76]	[Hussain et al. 2018]	Towards ontology-based multilingual URL filtering: a big data problem	Control
[87]	[Joo et al. 2017]	S-Detector: an enhanced security model for detecting Smishing attack for mobile computing	Control
[88]	[Joshi et al. 2013]	Extracting Cybersecurity Related Linked Data from Text	Vulnerability
[89]	[Katz et al. 2014]	CoBAn: A context based model for data leakage prevention	Control
[91]	[Khandpur et al. 2017]	Crowdsourcing cybersecurity: Cyber attack detection using social media	Attack
[94]	[Kudugunta and Ferrara 2018]	Deep neural networks for bot detection	Threat
[99]	[Le Sceller et al. 2017]	SONAR: Automatic detection of cyber security events over the twitter stream	Incident
[101]	[Lee et al. 2016]	Sec-Buzzer: cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation	Threat

(Continued)

Table 3. Continued

Reference in the Taxonomy	Authors and Year	Title	Domain
[104]	[Li et al. 2017]	A comparison of classifiers and features for authorship authentication of social networking messages	Incident
[106]	[Li et al. 2016]	Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System	Threat
[107]	[Liu and Wang 2012]	Anonymizing bag-valued sparse data by semantic similarity-based clustering	Control
[108]	[Macdonald et al. 2015]	Identifying Digital Threats in a Hacker Web Forum	Threat
[110]	[Mashechkin et al. 2015]	Applying text mining methods for data loss prevention	Control
[112]	[Meral et al. 2009]	Natural language watermarking via morphosyntactic alterations	Control
[113]	[Milosevic et al. 2017]	Machine learning aided Android malware classification	Control
[118]	[Mittal et al. 2016]	CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities	Threat
[119]	[Moghimi and Varjani 2016]	New rule-based phishing detection method	Control
[121]	[Murnion et al. 2018]	Machine learning and semantic analysis of in-game chat for cyberbullying	Threat
[124]	[Nandhini and Sheeba 2015]	Online Social Network Bullying Detection Using Intelligence Techniques	Threat
[125]	[Narouei et al. 2017]	Towards a Top-down Policy Engineering Framework for Attribute-based Access Control	Control
[126]	[Narouei et al. 2018]	Automatic Extraction of Access Control Policies from Natural Language Documents	Control
[128]	[Nembhard et al. 2018]	A hybrid approach to improving program security	Vulnerability
[130]	[Noor et al. 2019a]	A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise	Attack
[131]	[Noor et al. 2019b]	A machine learning framework for investigating data breaches based on semantic analysis of adversarials attack patterns in threat intelligence repositories	Attack
[135]	[Park et al. 2018]	Detecting Potential Insider Threat: Analyzing Insiders' Sentiment Exposed in Social Media	Attack
[140]	[Pellet et al. 2019]	Localising social network users and profiling their movement	Threat
[141]	[Perera et al. 2019]	Cyberattack Prediction Through Public Text Analysis and Mini-Theories	Attack
[142]	[Phan and Zincir-Heywood 2018]	User identification via neural network based language models	Incident
[144]	[Posey et al. 2017]	Taking stock of organisations' protection of privacy: categorising and assessing threats to personally identifiable information in the USA	Incident
[147]	[Rout et al. 2016]	Deceptive review detection using labeled and unlabeled data	Control

(Continued)

Table 3. Continued

Reference in the Taxonomy	Authors and Year	Title	Domain
[149]	[Sapienza et al. 2017]	Early warnings of cyber threats in online discussions	Threat
[152]	[Shao et al. 2019]	Autonomic Author Identification in Internet Relay Chat (IRC)	Incident
[154]	[Slankas et al. 2014]	Relation extraction for inferring access control rules from natural language artifacts	Control
[160]	[Spanos and Angelis 2018]	A multi-target approach to estimate software vulnerability characteristics and severity scores	Vulnerability
[161]	[Suleiman and Al-Naymat 2017]	SMS Spam Detection using H2O Framework	Control
[164]	[Syed 2018]	Enterprise reputation threats on social media: A case of data breach framing	Incident
[165]	[Syed et al. 2018]	What it takes to get retweeted: An analysis of software vulnerability messages	Vulnerability
[168]	[Thakur et al. 2018]	Innovations of phishing defense: The mechanism, measurement and defense strategies	Control
[169]	[Thorleuchter and Van den Poel 2012]	Improved multilevel security with latent semantic indexing	Control
[171]	[Toor et al. 2018]	Visual Question Authentication Protocol (VQAP)	Control
[177]	[Vidros et al. 2017]	Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset	Threat
[183]	[Wang and Jin 2011]	Data leakage mitigation for discretionary access control in collaboration clouds	Control
[184]	[Wang et al. 2013]	The Association Between the Disclosure and the Realization of Information Security Risk Factors	Incident
[186]	[Wen et al. 2015a]	A novel automatic severity vulnerability assessment framework	Vulnerability
[187]	[Wen et al. 2015b]	ASVC: An automatic security vulnerability categorization framework based on novel features of vulnerability data	Vulnerability
[188]	[Williams et al. 2019]	Analyzing Evolving Trends of Vulnerabilities in National Vulnerability Database	Vulnerability
[190]	[Xiang et al. 2012]	Linguistic steganalysis using the features derived from synonym frequency	Attack
[194]	[Yu et al. 2019]	Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts	Threat
[195]	[Zardari and Jung 2016]	Data security rules/regulations based classification of file data using TsF-kNN algorithm	Control
[199]	[Zhu et al. 2016]	Beating the Artificial Chaos: Fighting OSN Spam Using Its Own Templates	Control
[201]	[Zitar and Hamdan 2011]	Genetic optimized artificial immune system in spam detection: a review and a model	Control

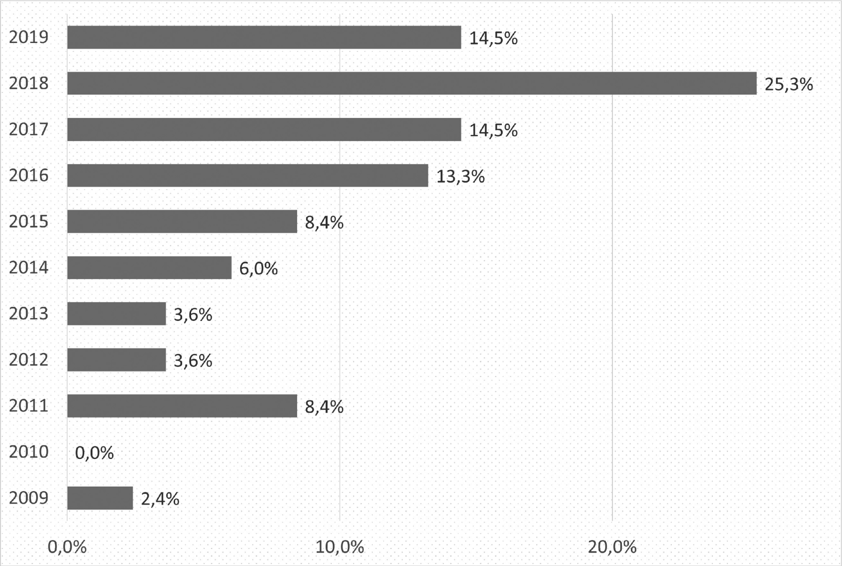

Fig. 2. Studies grouped by year.

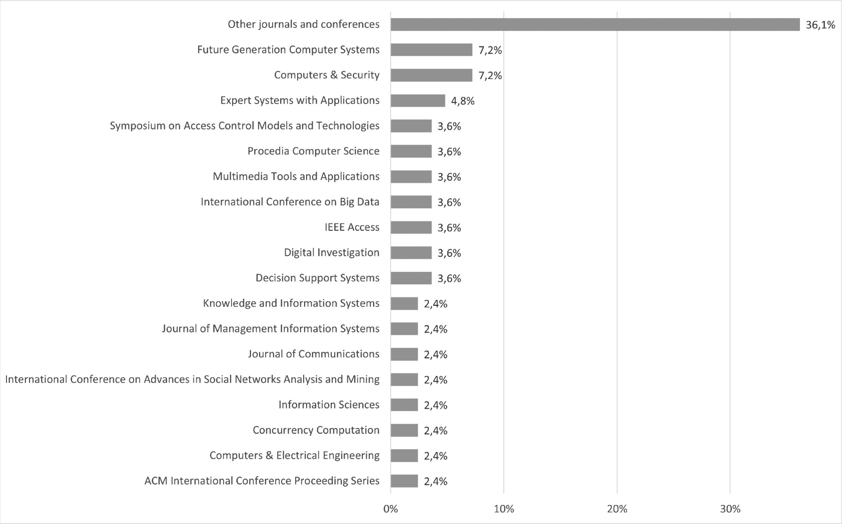

Fig. 3. Studies grouped by journals and conferences.

3.1 Publications Summary

This article aims to search for studies published since the early 2010s. Figure 2 shows the 83 selected studies grouped by the date of publication. Observing the research period, we can see the rise of publications since 2013, but we highlight that 67.5% of studies were published since 2016. Thus we understand the subject is relevant in the current technological and scientific scenario. Moreover, we collect the studies in the first half of 2019, and the total in this year reached the same number of publications of 2017, reinforcing the trend of new research linking text mining and cybersecurity.

The selected articles were distributed in 19 categories. These categories were defined by publishers and organized in descending order. Publishers who had only 1 article related were grouped as *others*. The *others* category includes journals like *Computers in Human Behavior*, *Knowledge-Based Systems*, *IEEE/ACM Transactions on Networking*, *European Management Journal*, *Journal of Systems and Software*, and *Information Systems Research*. The *others* category includes conferences such as *International Conference on Information and Knowledge Management*, *Conference on Computer and Communications Security*, *International Conference on Trust, Security and Privacy in Computing and Communications*, and *International Conference on Semantic Computing*.

Figure 3 shows that most studies come from journals or conferences with only one record in the database. The publications with the most significant number of studies are *Computers & Security* and *Future Generation Computer Systems*, both with six studies published. Additionally, the distribution of the publications points out that there is no preference by researchers to target journals and conferences strictly related to cybersecurity, since just 4 of 48 publishers included in this SLR have the focus in the area. Although we do not have evidence, this distribution could demonstrate the lack of interest from security researchers and publications, which are more focused on traditional security topics.

REFERENCES

[1] Mohammed Abuhamad, Ji su Rhim, Tamer AbuHmed, Sana Ullah, Sanggil Kang, and DaeHun Nyang. 2019. Code authorship identification using convolutional neural networks. *Fut. Gener. Comput. Syst.* 95 (2019), 104–115.

[2] Stephen Adams, Bryan Carter, Cody Fleming, and Peter A. Beling. 2018. Selecting system specific cybersecurity attack patterns using topic modeling. In *Proceedings of the 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering, (Trustcom/BigDataSE'18)*, 490–497. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00076

[3] Kayode Sakariyah Adewole, Nor Badrul Anuar, Amirrudin Kamsin, and Arun Kumar Sangaiah. 2017. SMSAD: A framework for spam message and spam account detection. *Multimedia Tools Appl.* 78, 4 (Jul. 2017), 3925–3960.

[4] Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian. 2015. BiSAL—A bilingual sentiment analysis lexicon to analyze dark web forums for cyber security. *Digit. Invest.* 14 (2015), 53–62.

[5] Monther Aldwairi and Ali Alwahedi. 2018. Detecting fake news in social media networks. *Proc. Comput. Sci.* 141 (2018), 215–222.

[6] Sultan Alneyadi, Elankayer Sithirasenan, and Vallipuram Muthukkumarasamy. 2013. Adaptable N-gram classification model for data leakage prevention. In *Proceedings of the 7th International Conference on Signal Processing and Communication Systems (ICSPCS'13).* https://doi.org/10.1109/ICSPCS.2013.6723919

[7] Manar Alohaly, Hassan Takabi, and Eduardo Blanco. 2018. A deep learning approach for extracting attributes of ABAC policies. In *Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies (SACMAT'18).* Association for Computing Machinery, New York, NY, 137-148. https://doi.org/10.1145/3205977.3205984

[8] Flora Amato, Giovanni Cozzolino, Vincenzo Moscato, and Francesco Moscato. 2019. Analyse digital forensic evidences through a semantic-based methodology and NLP techniques. *Fut. Gener. Comput. Syst.* 98 (2019), 297–307.

[9] Flora Amato, Giuseppe De Pietro, Massimo Esposito, and Nicola Mazzocca. 2015. An integrated framework for securing semi-structured health records. *Knowl.-Based Syst.* 79 (2015), 99–117. https://doi.org/10.1016/j.knosys.2015.02.004

[10] Jungkook An and Hee-Woong Kim. 2018. A data analytics approach to the cybercrime underground economy. *IEEE Access* 6 (2018), 26636–26652.

[11] Xinbo Ban, Shigang Liu, Chao Chen, and Caslon Chua. 2018. A performance evaluation of deep-learnt features for software vulnerability detection. *Concurr. Comput.* (2018).

[12] Barnali Gupta Banik and Samir Kumar Bandyopadhyay. 2018. Novel text steganography using natural language processing and part-of-speech tagging. *IETE J. Res.* 66, 3 (2018), 1–12. https://doi.org/10.1080/03772063.2018.1491807

[13] Nicole Lang Beebe, Jan Guynes Clark, Glenn B. Dietrich, Myung S. Ko, and Daijin Ko. 2011. Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies. *Decis. Support Syst.* 51, 4 (2011), 732–744.

[14] Nicole L. Beebe and Lishu Liu. 2014. Clustering digital forensic string search output. *Digit. Invest.* 11, 4 (2014), 314–322.

[15] Victor Benjamin, Bin Zhang, Jay F. Nunamaker Jr, and Hsinchun Chen. 2016. Examining hacker participation length in cybercriminal internet-relay-chat communities. *J. Manage. Inf. Syst.* 33, 2 (2016), 482–510. https://doi.org/10.1080/07421222.2016.1205918

[16] Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. 2007. Lessons from applying the systematic literature review process within the software engineering domain. *J. Syst. Softw.* 80, 4 (2007), 571–583. https://doi.org/10.1016/j.jss.2006.07.009

[17] Ching-Yun Chang and Stephen Clark. 2014. Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. *Comput. Linguist.* 40, 2 (Jun. 2014), 403–448.

[18] Xi Chen, Indranil Bose, Alvin Chung Man Leung, and Chenhui Guo. 2011. Assessing the severity of phishing attacks: A hybrid data mining approach. *Decis. Support Syst.* 50, 4 (2011), 662–672.

[19] Zhili Chen, Liusheng Huang, Haibo Miao, Wei Yang, and Peng Meng. 2011. Steganalysis against substitution-based linguistic steganography based on context clusters. *Comput. Electr. Eng.* 37, 6 (2011), 1071–1081. https://doi.org/10.1016/j.compeleceng.2011.07.004

[20] Zhili Chen, Liusheng Huang, and Wei Yang. 2011. Detection of substitution-based linguistic steganography by relative frequency analysis. *Digit. Invest.* 8, 1 (2011), 68–77.

[21] Aviad Cohen, Nir Nissim, Lior Rokach, and Yuval Elovici. 2016. SFEM: Structural feature extraction methodology for the detection of malicious office documents using machine learning methods. *Expert Syst. Appl.* 63 (2016), 324–343.

[22] Ilenia Confente, Giorgia Giusi Siciliano, Barbara Gaudenzi, and Matthias Eickhoff. 2019. Effects of data breaches from user-generated content: A corporate reputation analysis. *Eur. Manage. J.* (2019).

[23] Isuf Deliu, Carl Leichter, and Katrin Franke. 2018. Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data'17)*, 3648–3656. https://doi.org/10.1109/BigData.2017.8258359

[24] Prasad M. Deshpande, Salil Joshi, Prateek Dewan, Karin Murthy, Mukesh Mohania, and Sheshnarayan Agrawal. 2014. The Mask of ZoRRo: Preventing information leakage from documents. *Knowl. Inf. Syst.* 45, 3 (dec 2014), 705–730.

[25] Matthew Edwards, Robert Larson, Benjamin Green, Awais Rashid, and Alistair Baron. 2017. Panning for gold: Automatically analysing online social engineering attack surfaces. *Comput. Secur.* 69 (2017), 18–34.

[26] El-Sayed M. El-Alfy and Ali A. AlHasan. 2016. Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm. *Fut. Gener. Comput. Syst.* 64 (2016), 98–107.

[27] Yong Fang, Yusong Guo, Cheng Huang, and Liang Liu. 2019. Analyzing and identifying data breaches in underground forums. *IEEE Access* 7 (2019), 1–1.

[28] J. L. Gonzalez-Compean, Oscar Telles, Ivan Lopez-Arevalo, Miguel Morales-Sandoval, Victor J. Sosa-Sosa, and Jesus Carretero. 2019. A policy-based containerized filter for secure information sharing in organizational environments. *Fut. Gener. Comput. Syst.* 95 (2019), 430–444.

[29] Mingxing He, Shi-Jinn Horng, Pingzhi Fan, Muhammad Khurram Khan, Ray-Shine Run, Jui-Lin Lai, Rong-Jian Chen, and Adi Sutanto. 2011. An efficient phishing webpage detector. *Expert Syst. Appl.* 38, 10 (2011), 12018–12027.

[30] Danny Hendler, Shay Kels, and Amir Rubin. 2018. Detecting malicious powershell commands using deep neural networks. In *Proceedings of the ACM Symposium on Information, Computer and Communications Security (ASIACCS'18).* Association for Computing Machinery, New York, NY, 187–197. https://doi.org/10.1145/3196494.3196511

[31] Carolyn Holton. 2009. Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decis. Support Syst.* 46, 4 (2009), 853–864.

[32] Cheng Huang, JiaYong Liu, Yong Fang, and Zheng Zuo. 2016. A study on web security incidents in china by analyzing vulnerability disclosure platforms. *Comput. Secur.* 58 (2016), 47–62.

[33] Xiaohong Huang, Yunlong Lu, Dandan Li, and Maode Ma. 2018. A novel mechanism for fast detection of transformed data leakage. *IEEE Access* 6 (2018), 35926–35936.

[34] Mubashar Hussain, Mansoor Ahmed, Hasan Ali Khattak, Muhammad Imran, Abid Khan, Sadia Din, Awais Ahmad, Gwanggil Jeon, and Alavalapati Goutham Reddy. 2018. Towards ontology-based multilingual URL filtering: a big data problem. *J. Supercomput.* 74, 10 (Apr. 2018), 5003–5021.

[35] Jae Woong Joo, Seo Yeon Moon, Saurabh Singh, and Jong Hyuk Park. 2017. S-Detector: An enhanced security model for detecting Smishing attack for mobile computing. *Telecommun. Syst.* 66, 1 (Jan. 2017), 29–38.

[36] Arnav Joshi, Ravendar Lal, Tim Finin, and Anupam Joshi. 2013. Extracting cybersecurity related linked data from text, In *Proceedings of the IEEE 7th International Conference on Semantic Computing (ICSC'13)*, 252–259. https://doi.org/10.1109/ICSC.2013.50

[37] Gilad Katz, Yuval Elovici, and Bracha Shapira. 2014. CoBAn: A context based model for data leakage prevention. *Inf. Sci.* 262 (2014), 137–158.

[38] Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* In *Proceedings of the International Conference on Information and Knowledge Management, Proceedings*, Part F131841 (2017), 1049–1057. https://doi.org/10.1145/3132847.3132866

[39] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews.* Technical report. TR/SE-0401. Keele University.

[40] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. 2015. *Evidence-based Software Engineering and Systematic Reviews*. Vol. 4. CRC Press.

[41] Sneha Kudugunta and Emilio Ferrara. 2018. Deep neural networks for bot detection. *Inf. Sci.* 467 (2018), 312–322.

[42] Quentin Le Sceller, ElMouatez Billah Karbab, Mourad Debbabi, and Farkhund Iqbal. 2017. SONAR: automatic detection of cyber security events over the twitter stream. In *Proceedings of the International Conference on Availability, Reliability and Security (ARES'17)*. Association for Computing Machinery, New York, NY. https://doi.org/10.1145/3098954.3098992

[43] Kuo-Chan Lee, Chih-Hung Hsieh, Li-Jia Wei, Ching-Hao Mao, Jyun-Han Dai, and Yu-Ting Kuang. 2016. Sec-Buzzer: Cyber security emerging topic mining with open threat intelligence retrieval and timeline event annotation. *Soft Comput.* 21, 11 (Jul. 2016), 2883–2896. https://doi.org/10.1007/s00500-016-2265-0

[44] Jenny S. Li, Li-Chiou Chen, John V. Monaco, Pranjal Singh, and Charles C. Tappert. 2017. A comparison of classifiers and features for authorship authentication of social networking messages. *Concurr. Comput.* 29, 14 (2017).

[45] Weifeng Li, Hsinchun Chen, and Jay F. Nunamaker Jr. 2016. Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *J. Manage. Inf. Syst.* 33, 4 (2016), 1059–1086. https://doi.org/10.1080/07421222.2016.1267528

[46] Junqiang Liu and Ke Wang. 2012. Anonymizing bag-valued sparse data by semantic similarity-based clustering. *Knowl. Inf. Syst.* 35, 2 (Jun. 2012), 435–461.

[47] Mitch Macdonald, Richard Frank, Joseph Mei, and Bryan Monk. 2015. Identifying digital threats in a hacker web forum. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM'15)*. Association for Computing Machinery, New York, NY, 926-933. https://doi.org/10.1145/2808797.2808878

[48] I. V. Mashechkin, M. I. Petrovskiy, D. S. Popov, and Dmitry V. Tsarev. 2015. Applying text mining methods for data loss prevention. *Program. Comput. Softw.* 41, 1 (Jan. 2015), 23–30. https://doi.org/10.1134/S0361768815010041

[49] Hasan Mesut Meral, Bülent Sankur, A. Sumru Özsoy, Tunga Güngör, and Emre Sevinç. 2009. Natural language watermarking via morphosyntactic alterations. *Comput. Speech Lang.* 23, 1 (2009), 107–125.

[50] Nikola Milosevic, Ali Dehghantanha, and Kim-Kwang Raymond Choo. 2017. Machine learning aided Android malware classification. *Comput. Electr. Eng.* 61 (2017), 266–274. https://doi.org/10.1016/j.compeleceng.2017.02.013

[51] Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. 2016. CyberTwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'16)*, 860–867. https://doi.org/10.1109/ASONAM.2016.7752338

[52] Mahmood Moghimi and Ali Yazdian Varjani. 2016. New rule-based phishing detection method. *Expert Syst. Appl.* 53 (2016), 231–242.

[53] Vaia Moustaka, Athena Vakali, and Leonidas G. Anthopoulos. 2018. A systematic review for smart city data analytics. *ACM Comput. Surv.* 51, 5 (2018), 103.

[54] Shane Murnion, William J. Buchanan, Adrian Smales, and Gordon Russell. 2018. Machine learning and semantic analysis of in-game chat for cyberbullying. *Comput. Secur.* 76 (2018), 197–213.

[55] B. Sri Nandhini and J. I. Sheeba. 2015. Online social network bullying detection using intelligence techniques. *Proc. Comput. Sci.* 45 (2015), 485–492.

[56] Masoud Narouei, Hamed Khanpour, Hassan Takabi, Natalie Parde, and Rodney Nielsen. 2017. Towards a top-down policy engineering framework for attribute-based access control. In *Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies (SACMAT'17 Abstracts)*. Association for Computing Machinery, New York, NY, 103–114. https://doi.org/10.1145/3078861.3078874

[57] Masoud Narouei, Hassan Takabi, and Rodney Nielsen. 2018. Automatic extraction of access control policies from natural language documents. *IEEE Trans. Depend. Sec. Comput.* (2018), 1–1.

[58] Fitzroy Nembhard, Marco Carvalho, and Thomas Eskridge. 2018. A hybrid approach to improving program security.In *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI'17)*, 1–8. https://doi.org/10.1109/SSCI.2017.8285247

[59] Umara Noor, Zahid Anwar, Tehmina Amjad, and Kim-Kwang Raymond Choo. 2019. A machine learning-based FinTech cyber threat attribution framework using high-level indicators of compromise. *Fut. Gener. Comput. Syst.* 96 (2019), 227–242.

[60] Umara Noor, Zahid Anwar, Asad Waqar Malik, Sharifullah Khan, and Shahzad Saleem. 2019. A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories. *Fut. Gener. Comput. Syst.* 95 (2019), 467–487.

[61] Madhukar Pai, Michael Mcculloch, Jennifer D. Gorman, Nitika Pai, Wayne Enanoria, Gail Kennedy, Prathap Tharyan, and John M. Colford. 2004. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *Natl. Med. J. Ind.* 17, 2 (2004), 86–95.

[62] Won Park, Youngin You, and Kyungho Lee. 2018. Detecting potential insider threat: Analyzing insiders' sentiment exposed in social media. *Secur. Commun. Netw.* 2018 (2018). https://doi.org/10.1155/2018/7243296

[63] Hector Pellet, Stavros Shiaeles, and Stavros Stavrou. 2019. Localising social network users and profiling their movement. *Comput. Secur.* 81 (2019), 49–57.

[64] Ian Perera, Jena Hwang, Kevin Bayas, Bonnie Dorr, and Yorick Wilks. 2019. Cyberattack prediction through public text analysis and mini-theories. In *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data'18)*, 3001–3010. https://doi.org/10.1109/BigData.2018.8622106

[65] Tien D. Phan and Nur Zincir-Heywood. 2018. User identification via neural network based language models. *Int. J. Netw. Manage.* (2018). https://doi.org/10.1002/nem.2049

[66] Clay Posey, Uzma Raja, Robert E. Crossler, and A. J. Burns. 2017. Taking stock of organisations' protection of privacy: Categorising and assessing threats to personally identifiable information in the USA. *Eur. J. Inf. Syst.* 26, 6 (Nov. 2017), 585–604.

[67] Jitendra Kumar Rout, Smriti Singh, Sanjay Kumar Jena, and Sambit Bakshi. 2016. Deceptive review detection using labeled and unlabeled data. *Multimedia Tools Appl.* 76, 3 (Aug. 2016), 3187–3211.

[68] Anna Sapienza, Alessandro Bessi, Saranya Damodaran, Paulo Shakarian, Kristina Lerman, and Emilio Ferrara. 2017. Early warnings of cyber threats in online discussions. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW'17)*, 667–674. https://doi.org/10.1109/ICDMW.2017.94

[69] Sicong Shao, Cihan Tunc, Amany Al-Shawi, and Salim Hariri. 2019. Autonomic author identification in internet relay chat (IRC). In *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications (AICCSA'19).* https://doi.org/10.1109/AICCSA.2018.8612780

[70] John Slankas, Xusheng Xiao, Laurie Williams, and Tao Xie. 2014. Relation extraction for inferring access control rules from natural language artifacts. In *Proceedings of the 30th Annual Computer Security Applications Conference*, ACM International Conference Proceeding Series, 366–375. https://doi.org/10.1145/2664243.2664280

[71] Georgios Spanos and Lefteris Angelis. 2018. A multi-target approach to estimate software vulnerability characteristics and severity scores. *J. Syst. Softw.* 146 (2018), 152–166.

[72] Dima Suleiman and Ghazi Al-Naymat. 2017. SMS Spam Detection using H2O Framework. *Proc. Comput. Sci.* 113 (2017), 154–161.

[73] Romilla Syed. 2018. Enterprise reputation threats on social media: A case of data breach framing. *J. Strateg. Inf. Syst.* (2018).

[74] Romilla Syed, Maryam Rahafrooz, and Jeffrey M. Keisler. 2018. What it takes to get retweeted: An analysis of software vulnerability messages. *Comput. Hum. Behav.* 80 (2018), 207–215.

[75] Kutub Thakur, Juan Shan, and Al-Sakib Khan Pathan. 2018. Innovations of phishing defense: The mechanism, measurement and defense strategies. *Int. J. Commun. Netw. Inf. Secur.* 10, 1 (2018), 19–27.

[76] Dirk Thorleuchter and Dirk Van den Poel. 2012. Improved multilevel security with latent semantic indexing. *Expert Syst. Appl.* 39, 18 (Dec. 2012), 13462–13471.

[77] Andeep S. Toor, Harry Wechsler, Michele Nappi, and Kim-Kwang Raymond Choo. 2018. Visual question authentication protocol (VQAP). *Comput. Secur.* 76 (2018), 285–294 .

[78] Huy Tran, Uwe Zdun, et al. 2017. Systematic review of software behavioral model consistency checking. *ACM Comput. Surv.* 50, 2 (2017), 17.

[79] Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis, and Leman Akoglu. 2017. Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Fut. Internet* 9, 1 (Mar. 2017), 6.

[80] Qihua Wang and Hongxia Jin. 2011. Data leakage mitigation for discretionary access control in collaboration clouds. In *Proceedings of the 16th ACM symposium on Access control models and technologies.* Association for Computing Machinery, 103–112. https://doi.org/10.1145/1998441.1998457

[81] Tawei Wang, Karthik N. Kannan, and Jackie Rees Ulmer. 2013. The association between the disclosure and the realization of information security risk factors. *Inf. Syst. Res.* 24, 2 (Jun. 2013), 201–218. https://doi.org/10.1287/isre.1120.0437

[82] Tao Wen, Yuqing Zhang, Ying Dong, and Gang Yang. 2015. A novel automatic severity vulnerability assessment framework. *J. Commun.* 10, 5 (2015), 320–329. https://doi.org/10.12720/jcm.10.5.320-329

[83] Tao Wen, Yuqing Zhang, Qianru Wu, and Gang Yang. 2015. ASVC: An automatic security vulnerability categorization framework based on novel features of vulnerability data. *J. Commun.* 10, 2 (2015), 107–116. https://doi.org/10.12720/jcm.10.2.107-116

[84] Mark A. Williams, Sumi Dey, Roberto Camacho Barranco, Sheikh Motahar Naim, M. Shahriar Hossain, and Monika Akbar. 2019. Analyzing evolving trends of vulnerabilities in national vulnerability database. In *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data'18)*, 3011–3020. https://doi.org/10.1109/BigData.2018.8622299

[85] Lingyun Xiang, Xingming Sun, Gang Luo, and Bin Xia. 2012. Linguistic steganalysis using the features derived from synonym frequency. *Multimedia Tools Appl.* 71, 3 (Dec. 2012), 1893–1911.

[86] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2019. Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. *Comput. Secur.* 83 (2019), 106–121.

[87] Munwar Ali Zardari and Low Tang Jung. 2016. Data security rules/regulations based classification of file data using TsF-kNN algorithm. *Cluster Comput.* 19, 1 (Feb. 2016), 349–368.

[88] Tiantian Zhu, Hongyu Gao, Yi Yang, Kai Bu, Yan Chen, Doug Downey, Kathy Lee, and Alok N. Choudhary. 2016. Beating the artificial chaos: Fighting OSN spam using its Own templates. *IEEE/ACM Trans. Netw.* 24, 6 (Dec. 2016), 3856–3869.

[89] Raed Abu Zitar and Adel Hamdan. 2011. Genetic optimized artificial immune system in spam detection: a review and a model. *Artif. Intell. Rev.* 40, 3 (Nov. 2011), 305–377.