A Study of Classification of Texts into Categories of Cybersecurity Incident and Attack with Topic Models

Masahiro Ishii, Satoshi Matsuura, Kento Mori, Masahiko Tomoishi, Yong Jin and Yoshiaki Kitaguchi
Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan

Keywords: Text Classification, Seeded LDA, Topic Models, Data Mining, Cybersecurity Incidents, CERT.

Abstract: To improve and automate cybersecurity incident handling in security operations centers (SOCs) and computer emergency response teams (CERTs), security intelligences extracted from various internal and external sources, including incident response playbooks, incident reports in each SOCs and CERTs, the National Vulnerability Database, and social media, must be utilized. In this paper, we apply various topic models to classify text related to cybersecurity intelligence and incidents according to topics derived from incidents and cyber attacks. We analyze cybersecurity incident reports and related text in our CERT and security blog posts using naive latent Dirichlet allocation (LDA), seeded LDA, and labeled LDA topic models. Labeling text based on designated categories is difficult and time-consuming. Training the seeded model does not require text to be labeled; instead, seed words are given to allow the model to infer topic-word and document-topic distributions for the text. We show that a seeded topic model can be used to extract and classify intelligence in our CERT, and we infer text more precisely compared with a supervised topic model.

1 INTRODUCTION

Cybersecurity incidents have become more complicated as cyber attack methods evolve and latent vulnerabilities put organizations at risk. To mitigate the effects and damages caused by such incidents, computer emergency response teams (CERTs) and security operation centers (SOCs) have to analyze an enormous number of security alerts generated by security systems and devices and those stored in vulnerability databases. Security information and event management systems alone cannot protect against cyber threats because they are based on signatures or rigid rules and are thus insufficiently flexible to detect unknown attacks (Andrade and Torres, 2018). In addition, the management and customization of such systems are time-consuming and require expertise (Zhong et al., 2019).

In this paper, we use various topic models to classify incident reports in our CERT and blog posts by a security vendor and evaluate their performance using experiments. We start with a small trial to extract security intelligence useful incident handling. We focus on topic models because many data sources related to security incidents are in natural language form and the results of classifying text into categories are helpful for all CERT members, including non-experts.

The categories are based on incident and cyber attack types. Labeling all training data for predicting new text with models trained using a supervised algorithm is extremely time-consuming (Li et al., 2018). We perform experiments to compare a seeded latent Dirichlet allocation (LDA) topic model with a supervised topic model. For the seeded LDA (SLDA) model, text does not need to be labeled; instead, seed words are provided to classify the text into the designated topic categories. Our experimental results show that the SLDA topic model can classify our report and blog posts better than the supervised algorithm. We also discussed how we set the seed words to enhance the trained topic model by observing the clustering results and correlations between the clusters and the ground truth data

The rest of this paper is organized as follows. Section 2 quickly reviews studies on analyzing and utilizing cybersecurity intelligence extracted from data from various sources. Section 3 describes the LDA-based topic models used in the experiments. Section 4 presents the data and measures used for experiments on clustering security text and predicting new text, and the experimental results. We remark future tasks and the conclusions in Section 5.

Ishii, M., Matsuura, S., Mori, K., Tomoishi, M., Jin, Y. and Kitaguchi, Y.
A Study of Classification of Texts into Categories of Cybersecurity Incident and Attack with Topic Models.
DOI: 10.5220/0009099606390646
In *Proceedings of the 6th International Conference on Information Systems Security and Privacy (ICISSP 2020)*, pages 639-646
ISBN: 978-989-758-399-5; ISSN: 2184-4356

2 RELATED WORK

Data mining and machine learning have recently been applied to social media platforms, forums, and blogs for extracting security intelligence. Several studies have analyzed security intelligence and events using text classification based on machine learning or natural language processing models.

Deliu et al. (Deliu et al., 2017) performed experiments on text classification using convolutional neural network (CNN) and support vector machine (SVM) methods with text data from a hacker forum. They found that the SVM method outperformed the CNN method.

Nagai et al. (Nagai et al., 2018) classified security blogs from different vendors using a guided-topic model and a naive LDA model. They also analyzed cybersecurity events by setting seed words for various cyber threat categories.

Chambers et al. (Chambers et al., 2018) proposed a partially labeled LDA model for detecting early cyber attack discussions on Twitter. They used 2 million tweets consisting of data related to distributed denial-of-service (DDoS) attacks labeled binary value (attack or non-attack). Neither attack dictionaries nor seed words were used.

We remark that Nagai et al. (Nagai et al., 2018) also presented classification results using the SLDA model for security text. We evaluated not only clustering measurements but also classifying accuracy for test text. In addition, we tuned the seed words and evaluated the SLDA model in several steps. We furthermore performed the experiments to compare the SLDA model with the supervised topic model (labeled LDA).

3 TOPIC MODELS AND PREPROCESSING TEXTS

We adopted topic models based on LDA to analyze and classify Japanese text. Here, we briefly describe the features of the topic models and methods used to preprocess Japanese text for the models.

3.1 Topic Models based on LDA

LDA (Blei et al., 2003) is a basic unsupervised topic model algorithm that models a document with a mixture of topics. The model generates topic-word and document-topic distributions automatically, where the model explicitly assumes that each word is generated from one underlying topic. Figure 1a shows a graphical model of LDA. The topic-document distribution θ and the topic-word distribution ϕ are derived from the Dirichlet distributions with hyper parameters α and β, respectively. Then, a topic in documents and a word in topics are selected based on multinomial distributions θ and ϕ, respectively

Ramage et al. (Ramage et al., 2009) proposed the labeled LDA (LLDA), which is a supervised LDA-based topic model. In this model, document user tags are associated with latent topics. Figure 1b shows a graphical model of LLDA. The document-topic distribution is restrictedly drawn from the Dirichlet distribution with parameter α using the label set Λ. We note that LLDA is appropriate for documents with multiple labels.

SLDA (Jagarlamudi et al., 2012) is a guided-topic model that uses sets of seed words that users assume represent the underlying topics in a document. A graphical model of SLDA is shown in Figure 1b. This model uses seed words to improve the document-topic distribution by drawing the group variable g and the topic-word distribution by defining a topic as a mixture of seed topic distribution ϕ^s and regular topic distribution ϕ^r.

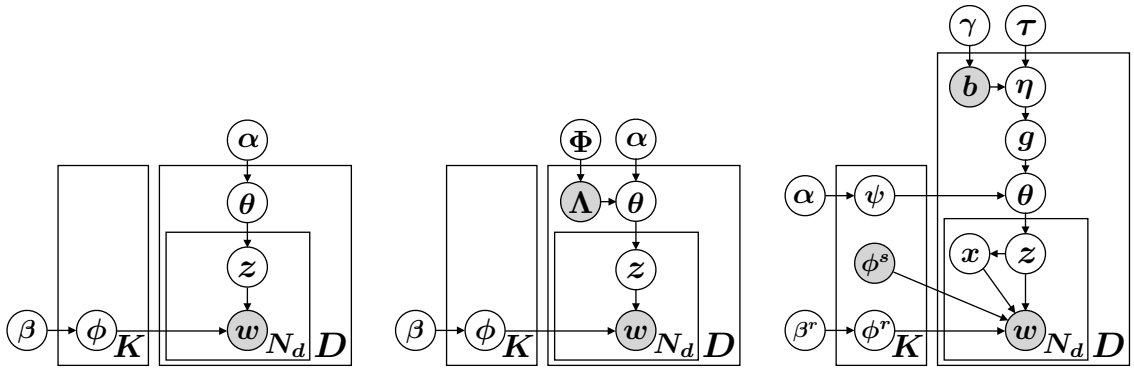
We used the library GuidedLDA (Singh and amrrs, 2017) for the above LDA-based models. GuidedLDA cannot be completely implemented by following the algorithm described in (Jagarlamudi et al., 2012, Section 2.3). Therefore, we implemented the SLDA algorithm based on GuidedLDA. Furthermore, we added the methods for computing LLDA model with labels of document on GuidedLDA.

3.2 Preprocessing of Japanese Text

MeCab (Kudo et al., 2004) is one of the most commonly used morphological analyzers for Japanese text. We used MeCab to parse Japanese text and extract morphemes, which are useful for training topic models. To extract proper nouns and recognized concepts, we used the dictionary *mecab-ipadic-NEologd*[1], which consists of words extracted from web resources. Stop words were set to enhance the topic models. We used the words provided in SlothLib[2] for Japanese words and those in the Natural Language Toolkit (Loper and Bird, 2002) for English words. We performed basic stemming of words by using the original form of words and converting the characters to lowercase and half-width characters for each Japanese and English word.

Many technical terms and entities appear in cybersecurity-related text, such as Internet Protocol (IP) addresses, Uniform Resource Locators (URLs),

[1] https://github.com/neologd/mecab-ipadic-neologd
[2] http://www.dl.kuis.kyoto-u.ac.jp/slothlib/



(a) LDA (Blei et al., 2003). (b) LLDA (Ramage et al., 2009). (c) SLDA (Jagarlamudi et al., 2012).

Figure 1: Graphical models of topic models.

hashes, and names of malware. In this work, we extracted IPv4 addresses and URLs from text using regular expressions and used them along with the abovementioned preprocessed words to train the topic models. In addition, we also extracted MD5, SHA1, SHA256, and SHA512 hashes of (malicious) files as IOC.

In addition to general stop words, we removed some words extracted using the LLDA model. When training models with LLDA for labeled text, the model assigns the corresponding label of text and the label *common_topic* to each token. Words that appear in the category *common_topic* also appear in other categories uniformly, and thus should not be set as seed words.

4 EXPERIMENTS

We evaluated LDA, LLDA, and SLDA models trained using labeled cybersecurity text. We conducted experiments on overlapping (or soft) clustering using multi-labeled text and predictions using test text (i.e., text not used to train the models).

4.1 Data

We used the following 405 security text from internal and external sources in the experiments.

- 45 security incident reports written from the beginning of 2016 to March 2019 managed in our CERT.
- 120 blog posts in Trend Micro Security Blog (in Japanese) [3].
- 120 blog posts in Cisco Japan Blog categorized as *security research* [4].
- 120 blog posts filtered by *LANGUAGE: Japanese, SYMANTEC BLOGS: Security Response* in Symantec Connect[5].

For the blogs by each security vendor, we collected the recent 120 blog text posted before the end of September 2019.

In our CERT, the members summarize the results of incident investigations and responses for incident handling interactions with other teams (SOC or the network operations center) or the person who contacted the CERT. The reports are stored as an Excel file. We analyzed the following types of text in the reports using the topic models: abstract of incident, status of response, cause of incident and possible attacks, information on the compromised system (e.g., IP address, network environment, OS version), attack (source) information (e.g., IP address/domain, name of malware), and measures implemented to prevent recurrence.

Next, we consider suitable categories for classifying security text with topic models. In this paper, we define categories of cybersecurity incident and attack that correspond to some standards that are commonly recognized.

Specifically, we set the following 6 categories of incident based on NIST SP800-61 Revision 1 [6] and FIRST [7] in Table 1, and 9 categories of attack based on MITRE CAPEC [8] in Table 2.

[3]https://blog.trendmicro.co.jp

[4]https://gblogs.cisco.com/jp/category/security/research/

[5]https://www.symantec.com/connect/search

[6]https://www.fismacenter.com/SP800-61rev1.pdf

[7]https://www.first.org/resources/guides/csirt_case_classification.html

[8]https://capec.mitre.org

Table 1: Incident categories.

Id	Category
0	Others
1	Unauthorized Access, Compromised Information/Asset, Unlawful Activity (Theft, Fraud, Human Safety, Child Porn), Espionage
2	DoS
3	Malware
4	Scan, Probe, Attempted Access, Reconnaissance
5	Improper Usage, Policy Violations

Table 2: Attack categories.

Id	Category
0	Others
1	Social Engineering (Phising, Targeted Attack, Information Elicitation, Pretexting, Identity Fraud)
2	DoS
3	Vulnerability Exploit,
4	Malware, File/Configuration/Environment Manipulation, Manipulation During Distribution
5	Inject Unexpected Items (XSS, Command Injection, Targeted Malware)
6	Employ Probabilistic Techniques (Brute Force, Fuzzing)
7	Collect and Analyze Information (Scan, Sniffing, Fingerprinting, Excavation
8	Subvert Access Control (Session Hijacking, Cross Site Request Forgery, MITM)

For the incident categories, we merged the categories by NIST and FIRST. We basically define each category of attacks corresponding to mechanisms of attack in MITRE CAPEC.

These categories are rather general and may be unsuitable for classifying incidents in detail, unlike those described in the guide [9] and the Cyber Threat Categories by SurfWatch Labs Inc. [10]. In the incident report, we must check the categories of incidents and attacks that can be used to label reports to classify them according to incident type. Because we focus on observing our security report data and classifying them, we use the required set.

Although each blog post is tagged, the tags are rather general, such as *cyber attack, attack methodology*, or too much detailed such as malware families for few special documents. Thus, it is thus appropriate for labeling according to distinct attack types and computing good classifier with topic models. We manually labeled the security text and set appropriate multi-labels at most 4. These multi-labels are regarded as ground truth data in our experiments.

[9]https://www.first.org/resources/guides/csirt_case_classification.html

[10]https://www.surfwatchlabs.com/threat-categories

4.2 Measurements

To measure overlapping clustering and predict test text with the topic models, we evaluated the models using purity, inverse purity, F1-score, and normalized mutual information (NMI). Specifically, we use macro (inverse) purity and macro F1-score which is the harmonic mean of purity and inverse purity. Let C be the cluster sets formed by clustering with the topic models, and C' be the ground truth class sets. Then the macro purity and inverse purity correspond to $F_{C,C'}$ and $F_{C',C}$, respectively where $F_{X,Y}$ is defined as Eq. 7 in (Lutov et al., 2019).

We also use normalized mutual information $NMI(C,C')$ to evaluate overlapping clusterings. Detailed formulae to compute NMI are described in (Lutov et al., 2019) (see Section IV.C.1 and Eq.11, 12). We note that we did not evaluate generalized NMI but naive NMI in the experiments.

4.3 Initial Seed Words

The quality of the SLDA model greatly depends on the defined seed words. We first observed words that frequently appeared for the given ground truth labels. In (Jagarlamudi et al., 2012, Section 2.4), the authors describe the technique to select initial seed words automatically by using information gain.

We computed the information gain of words in text and their ground truth labels, and extract top ranked 15 words for each category of incident and attack. The sampled seed words translated into English for the incident and attack topic categories are defined as shown in Tables 3 and 4. The categories are simplified and the words in bold and italic style were removed or added for evaluating effects of seed words in the experiments. The words which are not written by bold style are sampled extracted words with information gain. We describe how we set the seed words in more detail in Section 4.5.

4.4 Clustering Results

We conducted an experiment with overlapping clustering because the labeled text belongs to one or more categories. To classify the text into multi-categories,

Table 3: Incident categories and seed words.

Category	Seed words
Others	*ip, "blog author"*, supervision, *exe, windows, evangelist, system, microsoft, infection, malware*, **round up article, comprehension**
Unauthorized Access	indicator, *infection*, vulnerable, iocs, compromise, *"sha1 hash"*, dnspionage, suspicious, *micro, trend*, vulnerability, **disclosure of information**
DoS	*present, blog, denial*, research, prevent, *newsroom*, vmx, snortsnort
Malware	*infection, malware, indicator*, download, screenshot, ip, ioc, scan, exe, feature, **firepower**
Scan, Attempted Access	udp, dictionary attack, port, ssh, ip address, ddos attack, tcp, router, backdoor, jump server, flood, relay, **indicators, bounty**
Improper Usage	coinhive, rip, virus scan, *mac*, laboratory, isolation, free, subspecies, **confidential information, personal information, license, version, crack**, **"experimental product"**

we extracted the topic category for which the inferred topic ratio is higher than a given threshold in order of descending ratios, and we limited the number of inferred topics for each text. Table 5 shows the results of macro F1-score and NMI for overlapping clustering.

We denote the number of topic categories as K; $K = 6$ for incident categories and $K = 9$ for attack categories. We used the standard hyperparameters values $\alpha = 50/K, \beta = 0.01$ and $\pi_k = 0.7$ (Jagarlamudi et al., 2012, Section 2.1), and symmetric Dirichlet distributions. Although we evaluated the clustering results varying the maximum numbers of inferred topics from 1 to 4, we only showed the cases when then number is 1 or 4. The results for which the threshold is 0 and the number of inferred topics is 1 correspond to those of hard clustering.

We can see that SLDA outperformed LDA in all cases. When the maximum number of inferred topics is 4, the F1-scores for threshold $0.01/K$ and $0.1/K$ are higher than the F1-score for threshold $1/K$. Actually, the inverse purity become higher unreasonably since most of the samples are classified into 4 categories, and such a classifier is useless for identifying topics of security text. In these cases, the NMI values become lower that means the clusterings are not good comparing with the ground truth class sets. In our experiments, including the following evaluation for test text, we set the threshold to $1/K$ and the maximum number of inferred topics to 4.

4.5 Update Seed Words

Here, we describe experiments for updating the seed words by observing the clustering results. For the trained models with SLDA, we manually corresponded the clusters to the incident or attack categories by the document-topic distributions and the seed-topic distributions of the model since SLDA is an unsupervised algorithm. Then we can evaluate how the SLDA model classifies security text into multi-categories. We modified the seed words and evaluated clusterings in the following steps.

- Initial Automatic extraction by computing information gain using the ground truth labels of text.
- Step 1 We manually remove words that reduce overlapping clustering accuracy such as frequently used common words in several categories or peculiar words.
- Step 2 After training the SLDA model with seed words defined in Step 1, we select few words extracted by information gain with classification results.
- Step 3 We manually proper words based on heuristic knowledge such as an IP address, IOC, or cybersecurity proper words from security intelligences from internal and external sources.

In Tables 3 and 4, we underline the words that are actually used in the experiments. For the extracted words in the initial phase (the non-bold words), the italic words are removed in Step 1. We then added the bold words that are selected in Step 2 and 3 together. We note that the double-quoted words are obfuscated since these words indicate individual or sensitive informations.

Tables 6 shows the clustering results with F1-scores and NMI for the seed words defined in each step.

The evaluated values for the seed words in Step 1 are worse than those for initial seed words. However, it is necessary to remove some words in Step 1 since the initial seed words totally depend on ground truth labels of the security text. The F1-scores for the

Table 4: Attack categories and seed words.

Category	Seed words
Others	*file*, found, code, *vulnerability, exploit*, version, execution, *remote*, **round up article**
Social Engineering	suspicious, mail, vulnerable, cve, japan, attacker, *evangelist, "blog author"*, **personal information, disclosure**, **"IPv4 address (internal)"**
DoS	ddos attack, ddos, applet, not found, ip address, backdoor, *installation*, udp, tcp, dictionary attack, flood, bot, miori, mirai, **jump server**
Vulnerability Exploit	cve, vulnerability, *rule*, trick, arbitrarily, exploit, spotlight, *strategy, advisory*, bug, release
Malware	*infection, malware*, indicator, download, trojan, exe, feature, ioc, trojan, powershell, **firepower**
Inject Unexpected Items	electronic commerce, ticketmaster, magecart, dpi, security, *fashion*, constructor, **object, context**
Employ Probablistic Techniques	dictionary attack, pcastle, yowai, hakai, distributed, put, ssh, gafgyt, *damage*, *blog, page, ordinary*, port, **brute, polycom**
Collect and Analyze Information	layer, lack, exposure, port, udp, shodan, upnp, unit, home, *recent, install*, search engine, **ngips**, **disconnection, "public web server (internal)"**
Subvert Access Control	*dynamic*, "IPv4 address (external)", aptdnsdns, intersecdns, authority, Syria, cctld, rewrite, redirect, domain, electric company, **dnspionage, path, rewrite**

Table 5: F1-scores and NMI in parenthesis for various thresholds and maximum numbers of inferred topics.

Category	Model	(threshold, maximum number of inferred topics) $(0,1)$	$(0.01/K,4)$	$(0.1/K,4)$	$(1/K,4)$
Incident	LDA	0.514 (0.097)	0.694 (0.007)	0.66 (0.01)	0.551 (0.072)
	Seeded LDA	0.563 (0.103)	0.68 (0.014)	0.659 (0.021)	0.595 (0.074)
Attack	LDA	0.536 (0.137)	0.667 (0.018)	0.657 (0.021)	0.601 (0.072)
	Seeded LDA	0.597 (0.145)	0.761 (0.021)	0.761 (0.03)	0.654 (0.11)

words in Step 3 are slightly higher than those for the initial words.

4.6 Test Texts with Trained Models

SLDA is an unsupervised algorithm and thus an inferred topic cannot be identified with an actual category. In contrast, the inferred topics obtained with LLDA correspond to actual categories. By corresponding inferred topics to ground truth categories, we evaluated clustering in terms of precision, recall, and F1-score for SLDA and LLDA. We computed micro precision, recall, and F1-score for the evaluation. The text used for training and testing LLDA contained the label *common_topic*. We omitted this label when extracting inferred topics.

Here, we show the results of classification of the incident reports using the model trained using blog posts. For the incident reports and the blog posts, we used 270 text for training the model and 135 text as the test data. The results are shown in Table 7.

SLDA achieved higher precision, recall, and F1 values compared with those for LLDA for all cases. Nevertheless the classifying accuracy of SLDA is not at a practical level. We should therefore enhance the SLDA model by improving seed words and other hyperparameters.

4.7 Clustering Analysis

Here, we briefly describe the cluster analysis for the SLDA topic model and effects by the seed words. In this paper, we only show the correlations between the clusters and the ground truth labels for the attack categories because of limitations of space. Figure 2 shows the hierarchically-clustered heatmap with the matrix for ground truth multi-labels of text and clustered text with the SLDA model. We used Ward's method and the standard euclidean distance in hierarchical cluster analysis.

The text that are tagged ground truth labels "Malware", "Social Engineering", or "Vulnerability Exploit" by the SLDA with the initial seed are classified into similar cluster sets. Thus, it is hard to distinguish

Table 6: Comparing results of F1-score and NMI in parenthesis for each set of seed words.

Category	Model	Seed words			
		Initial	Step 1	Step 2	Step 3
Incident	LDA	0.551 (0.072)			
	Seeded LDA	0.595 (0.074)	0.579 (0.073)	0.588 (0.08)	0.644 (0.086)
Attack	LDA	0.601 (0.072)			
	Seeded LDA	0.654 (0.11)	0.653 (0.106)	0.629 (0.116)	0.673 (0.094)

Table 7: Evaluation of topic models using test text.

Seed words	Model	Incident categories			Attack categories		
		precision	recall	F1-score	precision	recall	F1-score
Initial	LLDA	0.175	0.089	0.118	0.115	0.065	0.083
	SLDA	0.298	0.354	0.324	0.314	0.43	0.363
Step 3	LLDA	0.241	0.133	0.171	0.137	0.085	0.105
	SLDA	0.312	0.411	0.355	0.361	0.46	0.404

clusters each other for such attack categories. For the results using the SLDA model with the Step 3 seed set, we can see that the clusters for the text in the category "Vulnerability Exploit" are slightly distinguishable from those for other categories "Malware" and "Social Engineering".

Though the effects by using different seed sets are not explicitly shown in the stochastic measurement results for the models and cluster analysis, we remark how we chose the seed words to be used. We added the some words for the attack category "Social Engineering" by observing frequently appeared words in our incident reports with computing information gain and heuristic knowledges. Furthermore, we could not classify the text well when we remove the ambiguous seed words which occurred in multiple categories such as "cve" in the categories "Social Engineering" and "Vulnerability Exploit". In our experiment, the clustering results are better when we set some seed words for related multi-categories.

5 CONCLUSION AND FUTURE WORK

We applied several LDA-based topic models to the classification of security text and extraction of intelligence from the text. This work was motivated by analyzing internal data (incident reports) with an SLDA model trained with external data (e.g., security blog posts). The results showed that SLDA overcomes naive LDA in terms of overlapping clustering, and SLDA classified the test text with higher accuracy than LLDA did. Although the results of SLDA are unsatisfactory to classify security text in a real-world situation. The followings are future works to enhance the SLDA model.

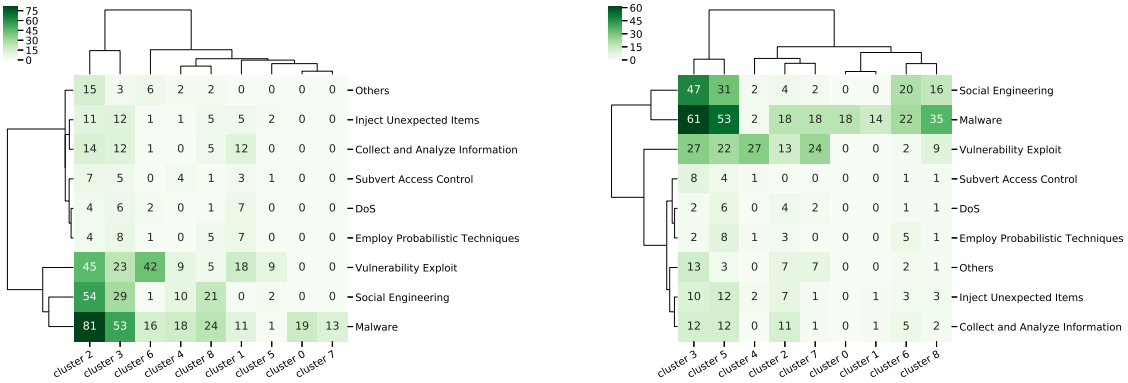
We should modify the stop words so that the SLDA model can classify security text more distinguishably. There are many words that have common meaning or concept all over the categories since the text used in our experiments are based on cybersecurity topics. We furthermore should consider how we can define categories of appropriate size to classify security text well. Although we used standard categories in this paper, we will need to consider subdivision of the categories and its hierarchical structure.

To enhance SLDA models, further experiments for various seed sets are needed. In addition, we should tune the hyperparameters of the Dirichlet distributions α and β and timing of updating these parameters in the sampling algorithms.

To make the models suitable for systems dedicated to handling incidents continuously, the seed words should be updated depending on the situation. Hu et al. (Hu et al., 2014) proposed a framework that users can use to iteratively refine topics by observing words inferred by LDA. It would be helpful to develop a framework that allows users to interactively refine the model with seed words. The well refined model and the extracted intelligences should lead to a system to handle cybersecurity incidents automatically.

ACKNOWLEDGEMENTS

This work was partially supported by JST CREST Grant Number JPMJCR1783 and Suematsu prize of



(a) Initial seed words (b) Step 3 seed words

Figure 2: Hierarchically-clustered heatmap for the attack categories for the initial and step3 seed words.

Tokyo Tech, Japan.

REFERENCES

Andrade, R. and Torres, J. (2018). Enhancing intelligence soc with big data tools. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1076–1080.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Chambers, N., Fry, B., and McMasters, J. (2018). Detecting denial-of-service attacks from social media text: Applying NLP to computer security. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1626–1635, New Orleans, Louisiana. Association for Computational Linguistics.

Deliu, I., Leichter, C., and Franke, K. (2017). Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3648–3656.

Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A. (2014). Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Jagarlamudi, J., Daumé, III, H., and Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 204–213, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Li, X., Li, C., Chi, J., Ouyang, J., and Li, C. (2018). Dataless text classification: A topic modeling approach with document manifold. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 973–982, New York, NY, USA. Association for Computing Machinery.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lutov, A., Khayati, M., and Cudré-Mauroux, P. (2019). Accuracy evaluation of overlapping and multi-resolution clustering algorithms on large datasets. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 1–8.

Nagai, T., Inui, T., Takita, M., Furumoto, K., Shiraishi, Y., Takano, Y., Mohri, M., and Morii, M. (2018). Clustering security blog posts using guided-topic model for threat analysis. In *Proceedings of Computer Security Symposium 2018*, volume 2018, pages 481–488.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Singh, V. and amrrs (2017). Guidedlda: Guided topic modeling with latent dirichlet allocation. https://github.com/vi3k6i5/guidedlda, Accessed: 2019-11-15.

Zhong, C., Yen, J., Liu, P., and Erbacher, R. F. (2019). Learning from experts' experience: Toward automated cyber security data triage. *IEEE Systems Journal*, 13(1):603–614.