Cyber-Attack Features for Detecting Cyber Threat Incidents from Online News

Mohamad Syahir Abdullah
School of Computing, Universiti Teknologi Malaysia,
Johor, Malaysia
msyahir38@utm.my

Anazida Zainal
School of Computing, Universiti Teknologi Malaysia,
Johor, Malaysia
anazida@utm.my

Mohd Aizaini Maarof
School of Computing, Universiti Teknologi Malaysia,
Johor, Malaysia
aizaini@utm.my

Mohamad Nizam Kassim
Cyber Security Responsive Division,
CyberSecurity Malaysia,
Seri Kembangan, Selangor, Malaysia
nizam@cybersecurity.my

Abstract—**There are large volume of data from the online news sources that are freely available which might contain valuable information. Data such as cyber-attacks news keep growing bigger and can be analyzed to gather informative insights of current situation. However, news is reported in many styles, added with the emerging of new cyber-attack and the ambiguous terms used have made the detection of the related news become more difficult. Thus, to handle these situations, the aim of this paper is to propose a scheme on detecting the related news about cyber-attacks. The scheme starts with identifying the cyber-attack features which will be used to classify the cyber-attack news. The scheme also includes a machine learning approach using Conditional Random Field (CRF) classifier and Latent Semantic Analysis (LSA) for further analysis. The results from this research should help people by showing the actual picture of cyber-attack occurrences in our surrounding and give valuable information to public thus raising social awareness about cyber-attack activities.**

Keywords— Online news, cyber-attacks, features identification, machine learning, CRF-classifier, LSA

I. INTRODUCTION

Internet is one of the important aspect in our life. It provides us with the access to real world information anywhere the network is available. The process of gaining information of real time or past incidents has evolved and become easier with online newspaper. Many news provider companies compete with each other to provide a better service to attract customers which in the meantime, benefits the customers to get more quality and resourceful information [1].

One of the issues that becoming more popular reported in the online newspaper nowadays is cyber-attack. Incidents that occurred recently, global emergence of ransomware such as "Wannacry" and "Petya" for instance are two of the ransomware incidents that had caused havoc and many people were affected and lots of resources were lost such as money and valuable data or personal information. Besides ransomware news, the increasing number of malicious cyber campaign on personnel or government sectors [2], online fraud and many more cyber security related stories also have surrounded our daily life [3]. However, even with lots of online newspaper articles report on these situations, many people still unable to see the whole picture that currently happening in the cyber world and as the huge volume of data is available [4], there is a need to process the news and produce an informative insight so that public awareness towards cyber-attacks can be increased.

However, to process the news, it can be difficult as online news has different reporting types or styles from one sources to another[3]. From the literature review and through the observation on the data collected, there are usually three types of news published by online media. The cyber threat facts – usually describing past events and explain the nature of cyber-attacks; technical bulletin report – in-depth and sometimes excessive details of the cyber-attacks; and cyber incidents report – report on the current cyber-attacks occurrence. From these three types of news format, only cyber incidents report and cyber threat fact can provide the necessary information such as the threat actors involved, method used and so on. As all three types of news use the same cyber-attacks keywords, it provides more distraction by having redundant details [10] that has affected the accuracy of detection of the news[5]. Therefore, cyber-attack features such as the threat types, threat actor name, malware names and even the event that have happened in the past need to be identified to characterize and describe each of the cyber-attack activities. Then, a features model which contain different set of cyber-attack features will be constructed to classify and distinguish the news types related to cyber-attack activities. This model will consist of features set from several level such as word-level and sentence-level for more accurate detection.

In addition, as technology keeps evolving in fast paces, the emergences of new cyber-attacks also contribute to another problem. The existing dictionary might not be able to pick up new terms used to describe the cyber-attacks. Thus, the names of the cyber-attacks need to be updated. The dictionary includes a popular or often used terms or keywords of a particular cyber-attack activities. These words are recorded based on the previous identified cyber-attack features. For example, for cyber-attack feature like threat name, new cyber-attack names such as "Wannacry", "Petya" and "NotPetya" will be recorded. Other than that, ambiguity problem also occurred when multiple news source reported the same news using different terms. Term ambiguity such as synonymy and polysemy are fundamental problems often occurred in natural

978-1-5386-7541-0/18/$31.00 ©2018 IEEE

language processing [2]. This situation leads to redundancy problem when retrieving the information. Therefore, to overcome the problem, we will use the suitable text analysis techniques such as Named Recognition Entity (NER), CRF-classifier and LSA to obtain valuable information from the online news.

In this paper, we focus into identifying the cyber-attack features and show the process to classify the online news into several cyber-attack news types. The cyber-attack features is one of the important steps in this research as it will help in characterizing and distinguish the news types which should increase the accuracy of detection of news that related to cyber-attack.

This paper consists of several sections. The next section discussed the related works that have been done beforehand by other researchers. The third section discussed about the approach taken to conduct this research that include all process and steps involve such as the data used, text-preprocessing, identifying the cyber-attack features, and classification using CRF and LSA. In the final section, section four, it will show early results discussion and also conclude this paper.

II. Related Works

A. NER

There are many researches that have been conducted using NER in multiple different domains. For example, in crime monitoring and prevention domain. Works done by [3] and [6] combine the use of NER and CRF to identify the crime locations. Their focus of work involves in thievery and they wanted the information of the crime location to be available to the public in which can be used to prevent the crime. Researchers such as [3] and [6] have explored different techniques in order to retrieve information from an unstructured data. For example, [3] use NER and CRF classifier to detect whether the online newspaper contains a crime location or not. The sentence level approach is deployed in the research where it compares the sentence with the features that describing the crime location. The CRF classifier at first able to produce high accuracy of detection when being tested with default local online news sources in their research. However, when the detection model is being used to other foreign source, it yields lower accuracy results. This situation can due to the reason that their work only focussing on sentence level analytics where the machine learning model is unable to detect the relation between different sentence. Compare to the work done by [6] where they deployed NER to detect crime. In their case, the relationship between sentence can be made as they used more complex features model involving word features and contextual features. In their work, they use term weighting approach using SVM and NB to detect the crime mentioned in the news.

B. Cyber Attacks

The work done by [2] to detect the cyber security threats in weblogs using probabilistic models which known as LSA that used term-document matrix (TDM) to provide the weight to the terms appeared in the text that can describe the patterns of cyber-attacks term appeared across the documents [2]. They want to distinguish between different categories (cyber-attacks, cyber war) of cyber-attacks terms appeared on online news as each news. Using LSA, they classify the news into the same categories based on the TDM, they able to detect the keywords of various topics related to cyber security threats. The plot they obtain clearly categorizes the groups of cyber threat issues published from the weblogs. From their findings, this research intends to implement LSA technique to solve the ambiguity problem faced in the news.

III. Approach

A scheme has been proposed to help the detection of the cyber-attacks news. There are three main phases involved in this scheme which are text pre-processing, identifying cyber-attack features and classification using CRF and LSA as we can see in the Figure 1.

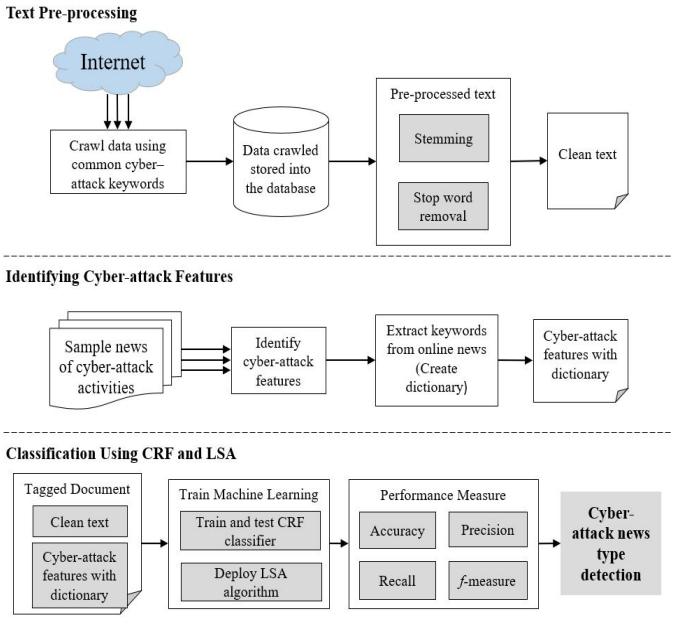

Figure 1. Cyber-attacks News Detection Scheme

There are two problems attempted to be solved by this study. First is to identify and detect the cyber-attacks related news from the online sources. As there are many news websites out there, the types of the news published also differs from each other and as the new cyber-attacks terms emerges, its affect the accuracy of the detection [3]. Furthermore, to avoid detecting the different news about the same topics reported and solve the ambiguity issue, LSA will be deployed to avoid getting duplicate data which can affect the analysis that will be made in future work. Thus, this research proposes to create features of cyber-attacks that will focus around the online news with the addition of new dictionary for each feature and scheme a machine learning consisted of CRF-classifier and LSA to help in the process.

In order to identify the features, data must be collected first. Data will be collected from the news article related to the cyber-attacks that include incidents or attacks that had happened, the threat actors, the method used and few others. From Figure 2, the news reported a new botnet, which is a treat type, called as JenX which is a threat name. Data can be obtained from the news websites such as Recorded Future.com, FireEye, Security Week, Micro Trend and more.

The websites mentioned are known for their reliability and they were recommended by security experts who used the reporting services provided by these websites. Data collected are in form of unstructured text. As the study has identified three news types that related to cyber-attack activities, all news needs to be collected in same amount. These data also will be used for updating the cyber-attack dictionary. Lastly, unnecessarily element such as html tag is left out before the data undergo text-pre-processing.

Figure 2. Example of Crawled Data

A. *Text Pre-processing*

All the news articles that have been collected will go through text pre-processing stage which includes additional process such as tokenization, stop word removal, and stemming [8]. This process will reduce the noise and clear the text so that it will be easier to be analyzed. Common word such as "a", "is", "the" etc. are eliminated which made the content of the text more meaningful. Meanwhile, stemming process will separates the root word from word that appear in the text in order to minimize the frequency. For an example, words like "attacking" and "attacks" will undergo a stemming process and later on, the result will be stored in database as "attack" in which preventing the redundancy of word in the database, saving spaces in memory. Porter Stemmer is often used in stemming process as excellent trade-off between speed, liability and accuracy. Next, the NER will be deployed. NER is a part of task in Information Extraction comprising of identifying and categorizing some types of information elements, called Named Entities (NE). The output from this step are the standardized structured form of text.

B. *Identifying Cyber-attacks Features*

Referring to Figure 3, the cyber-attack features will be identified from the collected data. Features such as the name of the cyber-attacks, its types, the respective threat actors, the organization involved and so on. These features are being compared to other researchers work and have it validated from the cyber security expert. In the same time, any related information obtained from the text are extracted and will be recorded. The information or we called it as terms or keywords are used for updating the dictionary of the cyber-attacks. All the terms currently present are collected through manual reading of the cyber-attack news from multiple sources. Then, to have a better news type classification, a features model is created consisting of several sets of cyber-attack features. This model will have two sets which are word level and sentence level. Each word from the cyber-attack features' dictionary belong to word level. For sentence level or better known as contextual level, the research will compare several words before and after the cyber-attack terms.

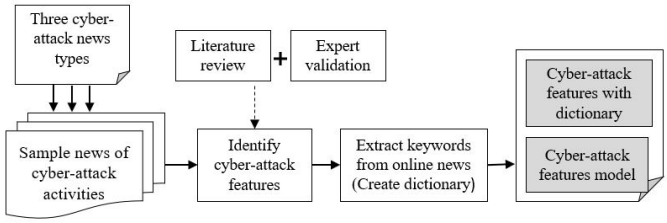

Figure 3. Identifying Cyber-attack Features

Text that have been pre-processed will be tagged using the proposed cyber-attack features model guided with the new dictionary. This process is done manually via Python programming. By creating a new dictionary on cyber-attacks, text from the news are manually labelled based on the existing features [7].

C. *Classification Using CRF and LSA*

The CRF-classifier will be trained and later used to classify the news. CRF is a statistical modelling technique for machine learning that is used to build a probabilistic model to label the data. It learns weights of the feature from the training datasets and produce a model in labelling the sign from the test dataset [3]. The identified features are used to assign the sentence with proper label stating whether the sentence contain any cyber-attacks incidents or not manually. Then, CRF classifier will be used to test the testing dataset. To access the performance of the experiment, this research will used f-measure approach which is a standard approach that is normally used in information processing measurement. The trained machine learning is eventually can detect the news related to the cyber-attacks incidents accordingly.

The text or news that have been detected will be analysed further in order to solve the ambiguity issues. Multiple news might be reporting the same incidents but using different terms to describe it. For example, a computer virus known as botnet or trojan might be called as malware, a more generalized term that is simple to understand by non-technical people. Thus, to classify malware news into a category that differ than social engineering attack (phishing, brute password attack) or network attack (DDoS, DNS attack), a latent semantic analysis approach has been taken. LSA uses a term document matrix in order to describe the pattern in a text [2]. Weight are given accordingly to the appearance of words across the text. News that involving malware might have different words to describe the incident than the words used in social engineered attack news. The result of the deployed LSA algorithm will show the plot chart of this particular group in which can show the distinctive group of incidents that occurred more compared to the rest.

IV. Results and Conclusion

As the research is still in progress, there are some phases that still need to be done. In the moment, the research has been able to identify the features for the cyber-attacks that also include all the keywords that has been collected as can be seen in Table 1. Some of the features such as threat type, threat name, threat actor has been updated with the terms related to the features. These terms or keywords will become the part of the cyber-attack dictionary for this study. The created features will also be used immensely in building the news detection scheme. The scheme will be implemented with two machine learning techniques which are CRF classifier and LSA. However, this part of scheme is some of the phases that are yet need to be completed. The research will continue with the building of the algorithm for the classifier and test it with actual data to make sure it works perfectly when it is time to be implemented. Same as the LSA algorithm. It needs to be constructed effectively in order to classify the news type comprehensibly.

TABLE I. Example of Cyber-Attack Features

Cyber-attack feature	**Terms**
Threat type	Malware, Botnet, Ransomware, Phishing, Trojan, Backdoor Spyware, Privilege Escalation Attack, Worm
Threat name	Reaper, Satori, Wannacry, Loapi, Notpetya, Andromeda
Threat Actor	The Shadow Brokers, Anonymous, Nexus Zeta, North Korea, Carbanak Gang, Lazarus Group
Organization Affected	Google, Microsoft, Apple
Platform Affected	Router, iOS, Windows, Play Store, Cryptocurency
Country Affected	South Korea, Brazil

The machine learning using CRF and LSA with addition of new cyber-attack features and dictionary should able to increase the accuracy of detection of the cyber-attack news [9]. The scheme is expected to help a better detection of the related cyber-attacks news and categories the threats accordingly. The output will eventually show the whole picture of cyber-attack activities that occur in our cyber space. Thus, it can help cyber security organization to generate report regarding cyber-attack activities or help in providing insight for any decision-making process. In the end, this research is trying to increase the awareness of public by showing the numbers and figures of the current cyber threats occurrences so that people will be more alert and able to avoid from being a victim.

Acknowledgment

This research is funded by CyberSecurity Malaysia under strategic collaboration with Cyber Threat Intelligence Lab, School of Computing, UTM.

References

[1] H. J. Carey and M. Manic, “HTML Web Content Extraction Using Paragraph Tags,” pp. 1099–1105, 2016.

[2] F. S. Tsai and K. L. Chan, “Detecting Cyber Security Threats in Weblogs Using Probabilistic Models,” pp. 46–57, 2007.

[3] B. Tony, R. Savarimuthu, and M. A. Purvis, “Extracting Crime Information from Online Newspaper Articles,” no. Awc, pp. 31–38, 2014.

[4] N. Kallus, “Predicting Crowd Behavior with Big Public Data,” 2013.

[5] X. Liu, “Extracting Addresses From News Reports Using Conditional Random Fields,” 2016.

[6] H. Shabat, N. Omar, and K. Rahem, “Named Entity Recognition in Crime Using Machine Learning Approach,” pp. 280–288, 2014.

[7] X. Qiu and X. Lin, “Feature Representation Models for Cyber Attack Event Extraction,” pp. 29–32, 2016.

[8] R. N. Zaeem, M. Manoharan, Y. Yang, and K. S. Barber, “Modeling and analysis of identity threat behaviors through text mining of identity theft stories,” *Comput. Secur.*, vol. 65, pp. 50–63, 2017.

[9] W. Zhang, S. X. Kong, and Y. C.Zhu, "Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach," *Cluster Computing*, pp.1-14, 2018.

[10] R. Wongso, F.A. Luwinda, B.C. Trisnajaya, and O. Rusli, News Article Text Classification in Indonesian Language. *Procedia Computer Science*, 116, pp.137-143, 2017.