Analyzing Coverages of Cyber Insurance Policies Using Ontology

Markos Charalambous
markos.charalambous@eecei.cut.ac.cy
Department of Electrical Engineering, Computer Engineering and Informatics Cyprus University of Technology
Limassol, Cyprus

Aristeidis Farao*
arisfarao@unipi.gr
Department of Digital Systems, University of Piraeus
Athens, Piraeus, Greece

George Kalantzantonakis
george@lstech.io
LSTech ESPANA
Madrid, Spain

Panagiotis Kanakakis
takiskanakakis@lstech.io
LSTech ESPANA
Madrid, Spain

Nikos Salamanos
nik.salaman@cut.ac.cy
Department of Electrical Engineering, Computer Engineering and Informatics Cyprus University of Technology
Limassol, Cyprus

Evangelos Froudakis
e.froudakis@ssl-unipi.gro
Department of Digital Systems, University of Piraeus
Athens, Piraeus, Greece

Evangelos Kotsifakos
ekotsifakos@lstech.io
LSTech ESPANA
Madrid, Spain

ABSTRACT

In an era where all the transactions, businesses and services are becoming digital and online, the data assets and the services protection are of utmost importance. Cyber-insurance companies are offering a wide range of coverages, but they also have exclusions. Customers of these companies need to be able to understand the terms and conditions of the related contracts and furthermore they need to be able to compare various offerings in order to determine the most appropriate solutions for their needs. The research in the area is very limited while at the same time the related market is growing, giving every potential solution a high value. In this paper, we propose a methodology and a prototype system that will help customers to compare contracts based on a pre-defined ontology that is describing cyber-insurance terms. After a first preliminary analysis and validation, our approach accuracy is averaging at almost 50%, giving a promising initial evaluation. Fine tuning, larger data set assessment and ontology refinement will be our next steps to improve the accuracy of our tool. Real user evaluation will follow, in order to evaluate the tool in real world cases.

KEYWORDS

Cyber-insurance, Ontology, Coverages, Exclusions, Premium, Weakest link

1 INTRODUCTION

As more and more businesses are going online – offering their products and services using online platforms, shared cloud and infrastructure [5] – the exposure to cyber-threats and the risk for breaches and business interruption is getting higher [12]. The cost [8] of such threats can be enormous, especially for small businesses that do not have the budget to build highly secure infrastructures or to recover from attacks– especially if this involves fines that they have to pay for not being able to protect their data [14]. At the same time, the cyber-insurance market [16] is growing and evolving at a fast pace trying to offer solutions that will safeguard the online businesses. Selecting the proper cyberinsurance policy is a difficult task; especially, trying to understand what they cover and what

they do not and comparing the offers as well as their prices. The evaluation of different policies and contracts is a manual and time consuming process, often requiring technical or legal knowledge. However, one of the biggest drawbacks is the Information Asymmetry that has a negative effect on the cyber insurance ecosystem and includes two components: (i) the inability of the insurer to distinguish between insureds of different (high and low risk) types, and (ii) insurers undertaking actions (i.e., reckless behavior) that affect loss probability after the insurance contract is signed, knowing that they would be insured. The reasons that lead to information asymmetry are the following: (i) insurers lacking vital information regarding applications, software products installed by insureds, and security maintenance habits, which correlate to the risk types of insureds, and (ii) insureds hiding information about their reckless behavioral intentions from their insurers, after they get insured, knowing that they would be compensated – irrespectively of their malicious behavior (e.g., being careless with security settings, etc.,) [15].

In this paper, we propose a prototype system for parsing cyber-insurance policies/contracts and extracting inclusions and exclusions, offering to the user a list of what is covered and what is not. In this way, the user will be able to easily compare several policies/contracts and to choose the one that fits he/she needs in a better way.

In the following sections we describe the background and the related research, we present our approach and we provide details on the architecture and our implementation. Finally, we present some preliminary results and we conclude with the future work.

2 BACKGROUND

As the cyber-attacks become more sophisticated targeting a broad range of companies and state or private institutions, the cyber-security is evolving too, together with the cyber-insurance. Cyber insurance is a rapidly developing area and an alternative way to deal with residual risks [4], [13]. Cyber-insurance is a powerful tool to incentivize the market towards protecting online businesses from information technology-related risks. The cyber insurance market is still immature facing several challenges on the way of becoming a common reality for online businesses and individuals [16], [15], [1]. Information asymmetry is one of the most prominent challenges and refers to the lack of information between the insurer and insured. First, as the cyber-insurance market is growing, it becomes more and more challenging for the insured to search and compare the various cyber-insurance policies (i.e., coverages and exclusions) that are offered by the market. In addition, the cyber insurance policies often list details about coverages and exclusions, using legal terms that can be difficult to be comprehended by the insured organization. Thus, moral hazard can occur where the insured organization could increase its exposure to risk, as well as the probability of loss during the contract period. Secondly, it is difficult for the insurer to distinguish between high and low risk businesses and individuals.

Although the cyber-insurance market is rapidly growing, few studies have been conducted in this area. The problem of identifying the coverages that an insurance company offers regarding cyber-security is relatively new and therefore not a lot of solutions are available. Analyzing the cyber-insurance contracts is mainly a problem about text analysis and keyword extraction, while being able to semantically distinguish what the insurance is covering and what is not.

Romanosky et al. [18] have presented qualitative research, of the current state of the cyber-insurance market. First, the authors collected insurance policies from state insurance commissioners in the United States. They collected over 235 policies from New York, Pennsylvania, and California, as well as policies posted publicly on various insurance companie's websites. Then they examined the composition and variation across three components: (i) the coverage and exclusions (ii) the security application questionnaires – by which an applicant's security risk level is estimated– and (iii) the rate schedules which define the method used to compute premiums. The finding depicts that there is a strong similarity regarding the covered losses, with more variation in exclusions. Bohme et al. [2] proposed a unifying framework to illustrate the parameters that should be included in the model of cyber insurance. The framework features a common terminology and deals with the specific properties of cyber-risk in a unified way. It unites phenomena such as interdependent security, correlated risk, and information asymmetries, in a common risk arrival process. Their framework offers a unified terminology to deal with specific properties of cyber risk and helps to alleviate discovered shortcomings.

The automatic ontology population from raw texts is a powerful procedure, since it extracts data from various documents which even if they contain irregular and ambiguous information, it is still able to enrich and assign the data with a precise structure and semantics. In this context, Ganino et al. [6] presented a methodology for the automatic population of predefined ontologies with data extracted from text and they proposed the design of a pipeline based on the General Architecture for Text Engineering system. Elnagdy et al. [3] presented the Semantic Cyber Incident Classification (SCIC) model, an ontology-based knowledge representation methodology for cyber-insurance. The method uses semantic techniques to provide a consistent knowledge representation for mapping the entities in the Cyber insurance system. Finally, other studies on populating ontology schema for legal text documents are: [7] for service level agreements and [10] for web service provider privacy policies.

Addressing the information asymmetry problem, one prominent approach is by parsing the various cyber-insurance policies and contracts that are offered by the insurance companies, to extract, and categorize the coverages and exclusions in a completely automatic way. One of the first studies that followed the aforementioned approach is the work of Joshi et al. [9]. The authors have presented a framework that automatically extracts keywords from cyber insurance policy documents and populates an ontology schema (or knowledge graph) to represent the extracted keywords as coverages and exclusions. The proposed cyber insurance ontology has been constructed by analyzing publicly available insurance policies from seven insurance providers. Moreover, the key ontology classes along with their relations are based on industry standards proposed by the United States Federal Trade Commission (FTC). Finally, they applied a grammar-based natural language parser using deontic expressions, to extract coverages and exclusions from the policy

documents. Deontic logic describes statements containing permissions, and obligations, whereas temporal logic describes time-based requirements. The use of domain-specific ontologies, is a popular approach to represent domain knowledge.

Our approach presented in this paper is different in several points from the one in [9]. First, the dataset used in [9]is not publicly available, hence, we were not able to use it in our model. Moreover, apart from the very limited research in cyber-insurance contract evaluation, there is neither commonly agreed list of coverages and exclusions that serves as an official terminology, nor official cyber-insurance ontology available. As Romanosky et al. [17] pointed out, there is lack of clarity in what is covered and excluded by a given policy, in the event of a security incident. Thus, the lack of comprehensibility of a policy rule often leads to courtroom discussion to determine the validity of coverage clauses. Many "ontology standards" exist, but none is explicitly defined as "information security ontology". For this reason, we have manually analyzed several available contracts and cyber - insurance policies from various companies to define our own list of terms and consequently to construct related information-security ontology. Furthermore, our approach is able to deal with large collection of documents due to the simpler text parsing and keywords extracting method. Therefore, our approach is scalable, time and memory efficiently.

3 SECONDO APPROACH

Our approach is based on the following main methodology. First, we parse the contract/ policy document and we extract the text that refers to the coverages and the text related to the exclusions. Then, using these two different texts as input, along with a generalized cyber-insurance terms ontology, we define which of the terms of the ontology are found in the coverages or in the exclusions. The use of the ontology allows us to be able to categorize coverages and to have a tree-like structure, where a category can include various coverages. This gives us the flexibility: (i) to include a set of coverages that are categorized, and they might not be mentioned by the exact wording in the policy; (ii) to allow the user to provide their own ontology (either defined manually or provided by an organization). One of our goals is to have an extensible tool so that the user will be able to use their one ontology-vocabulary. The final output of our approach is a table with the terms of the ontology and an indication whether this is covered or not by the specific policy. With this approach, we can also deal with the language problem, since the tool gets as input a manually created ontology file, that can be in any language and it matches the terms with the policy text in the same language. In other words, although we have evaluated our tool with policies in English, the tool is language-independent.

The first step of our process is to automatically extract the coverages from an original contract in .pdf format and depict them in such a way that it would be easier to analyze them in the next steps of the process. For designing this, we examined two approaches. The first approach we examined and the approach that we finally decided to implement was to make an automated process with python3 code that would take each original contract in .pdf format as input, it would map each line of text as a type of header or paragraph by the .html format to the file and output it in a .txt file. After that, another function would take as input the .txt file and remove unnecessary headers and footers, find keywords that show if some damage is or is not covered by the contract and list the covered and not covered damages in two final .txt files that are the final output from the program. This approach was easily executable and, the program could be easily evaluated, and micro adjustment could be made to work properly in all the possible formats of contracts (making the final outcome reliable).

Another approach we examined was that of automation by trained neural networks. The way that this approach would work is that we would make a fully or partially connected neural network that would take as input the contract in .pdf form and output a boolean value for all the possible coverages. In the training stages the output would be compared with the expected output and the distance (in the geometrical space that is defined by the vectors-coverages) between them would be the value of this performance. After each performance, a backward propagation function would make micro adjustments to the connections of the network in order to minimize that value for all the given contracts. This approach would not only be more universal, because all of the possible contract types would have been analyzed and trained on, but the further development and the adaptability of the process would be easier as we would not need to reprogram the whole program but to add some more specialized functions or continue the training in new sets of data. Nevertheless, the neural network approach was abandoned as there was not a fitting trained network in the bibliography. Another risk that this approach would pose is the credibility of the result as in those methods even the slightest unpredicted change could have an effect in the result.

4 SYSTEM ARCHITECTURE

Concern over cybersecurity is growing across all sectors of the global economy, as cyber risks have grown, and cyber criminals have become increasingly sophisticated. For insurers, cybersecurity incidents can harm the ability to conduct business, compromise the protection of commercial and personal data, and undermine confidence in the sector. The participants who take part in the cyber insurance market are the following: i) Insurer; ii) Insured; iii) Agent and iv) Broker.

Insurer: Insurers offer premiums that can cover a variety of cyber risks and incidents, such as phishing, data breaches, or malware that can affect companies and individuals. It can provide first-party coverages, such as damage on digital assets, business interruption, and incident response costs, as well as third-party coverage, such as privacy and confidentiality-related liabilities. Moreover, insurers provide policy holders with premiums and with the element of risk assessment, in case they fall victim to a cyber threat, providing technical, legal support in case of an incident. There is quite a lot of variation between the contracts, and this always depends on the needs of individuals or organizations. It also depends on the need for insurance coverage as well as the type and level of risks that will be exposed. Insurers offer cyber insurance policies as part of a contract or as a standalone product.

Insured is a person whose assets (tangible and intangible assets) are protected by an insurance policy; moreover, he is a person who contracts for an insurance policy that indemnifies him against loss.

In terms of cyber insurance individuals and organizations can benefit, as cyber incidents can evoke cyber risks. Aftermaths of a cyber threat may have a negative impact on individuals and businesses, including the loss of customers and revenue. Cyber insurance policies may change as an impact of the continuously changing market. Insurers nowadays are facing many challenges in the insurance industry such as, the need to find a trusted advisor, to find the proper insurance program, to find a broker or agent who addresses their specific and special insurance needs, a competitive insurance program in comparatively the current market environment and to find a tailor-made contract in their needs.

Agent: An insurance agent is a licensed person who has an important role to achieve an agreement and to conduct business on behalf of insurance companies. He is the professional who has the necessary knowledge needed to transmit the multifarious to the prospective clients. He is the intermediary who has undertaken the difficult role of approaching the client, informing him about the offered products of the insurance company, convincing him to buy them, and most important and the most difficult part is to acquire trust and become the person who will be interested in satisfying him, regarding the agreed claim that the insured has. However, the insurance agent is the one who must study the financial conjunctions, analyze them, predict the changes that affect the interested parties by all factors such as consumers, investors, those who are interested in savings plans, and all those who are interested to be insured.

Brokers organize and execute financial transactions on behalf of their respective clients for categories such as assets, stocks, forex, real estate, and insurance. For the orders he executes, the customers are charged with a commission according to the agreement of the contract. A broker can have an advising role on buying or selling products as some can provide their customers with market data analytics to help them make the right decision. The broker may be full-time or only for executions. To do the above he must be certified to provide the appropriate advice as well as the client's permission to perform any action.

As shown in Figure 1, the general structure of our tool is divided into two discrete main sub-modules: i) the Parser and ii) the Cyber Insurance Ontology. On the one hand, the Parser sub-module (as its name implies) is responsible to receive the contract that will be under process and in the end discretely present the coverages and exclusion that the aforementioned contract bears with. On the other hand, the Cyber Insurance Ontology contains lists regarding the common coverages and exclusion that the majority of cyber insurance contracts bear with, as well as it contains the cyber insurance ontology. The proposed ontology will be used between the Insured, the Broker and the Agent. We have to note the cyber insurance ontology is not standalone, but it is part of the SECONDO [4] architecture, which is responsible for providing a holistic security solution as a platform for organizations to fight cyber risks providing them with innovative security controls including risk transfer.

Our tool interacts with the following entities: i) the SECONDO handler and ii) the SECONDO end-user. At this point we have to note the term SECONDO handler contains the following entities [11]: i) Insurance company; ii) Insurance agent and iii) Insurance broker. This stakeholder is responsible to feed the tool with new

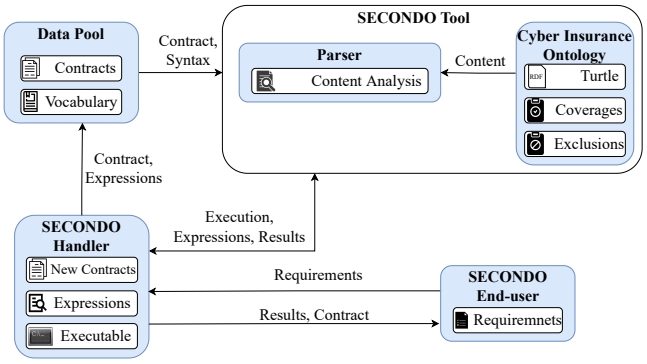

Figure 1: System architecture

cyber insurance contracts, expressions that are used in the contracts to express the existence of a coverage and exclusion, as well as to execute the tool. While the end-user could be a prominent insurer having specific requirements. Finally, there is a Data Pool that is responsible to securely store the vocabularies and the contracts that have been analyzed.

5 IMPLEMENTATION

Our tool implementation is a combination of bash scripting, python development and ardf ontology in turtle format. The bash environment helps us orchestrate the execution flow as it controls the input/output of the core environment, the python scripts. Our implementation is a pipeline of steps which contains contract reprocess, the core of our tool and result combination. Each step is given an input and extracts an output which is given to the next stage.

The first steps of our process are to clean our file from the different fonts and all the graphical parts that are useless to us. This happens with the two first functions *fonts()* and *font_tags()*. Specifically, the first function extracts and returns all the fonts and their usage. The second function takes as input and returns a dictionary with font sizes and tags as keys and values respectively. After that the function *headers_para()* takes all the headers and paragraphs from the .pdf file and with the help of the output of the *font_tag()* function and returns them as text with element tags.

The next step is to select the covered and not covered parts. First, we make all the characters lowercase for easier and better handling. Then we remove headers that came from headers and footers of the .pdf file and not from actual titles and subtitles with the function *remove_headers_footers()*. Those headers and footers do not contain any new information but are very confusing to the algorithm. The algorithm recognizes them because they are repeated on every page. After that we use the function *coverd_and_not_coverd()* to separate files that contain the covered and not covered damages by the contract. The algorithm finds the covered and not covered damages by searching for keywords as "cover", "covered", "coverage", "not covered" and "not cover" in the lines that came from subtitles and titles to recognize which paragraphs are talking about the coverages. The main core of our implementation is described by a python script file which is executed given the output of the previous step, the covered and not covered text, as long as the ontology file.

Our aim is the use of the well-defined ontology to find keywords in text files that will help us understand whether something is covered or not. To efficiently find the similar words, the input text files were tokenized to ngrams and stored in memory as python sets. As continuous sequences of words or tokens in a document, the n-grams in our case are defined in sets of two words. The choice of two words is based on the fact that our ontology contains mainly single terms and occasionally terms of two words. Thus, it is more efficient to compare the contracts text with the ontology terms. Subsequently, the ontology is turned to an in-memory RDF graph and using a sparql query the necessary information is obtained as a python set too. The final step of our algorithm is the creation of two new sets which will describe the covers and not covers. To obtain the covers, we need to intersect the covered set of ngrams with our ontology whereas to find the non covers we need to use the non-covered set of ngrams. Our results are written in an xlsx format file where every sheet is named by the main ontology class and contains all the subclasses along with a yes or no-depending on the insurance coverage.

Overall, the proposed implementation is able to receive a set of contracts at the same time that will be processed sequentially, and the output will be a set of files, one for each policy, with the coverages and exclusions of each policy.

6 PERFORMANCE-EVALUATION

In this section, we aim to evaluate the applicability and effectiveness of the proposed approach that has been introduced as a tool as well as its performance in terms of speed, resource consumption and scalability. For the proof of concept implementation, we have developed our own code (see Section 5), also, we have isolated cyber insurance policies from leading insurance companies to evaluate the proposed tool against their policies. The experiments were performed in an Ubuntu 18.04 desktop PC being equipped with an Intel Xeon(R) Silver 4114 CPU @ 2.20 GHz and 12GB RAM.

To evaluate our system, we performed an initial assessment. First, we defined an ontology with terms that we extracted from a set of insurance contracts from well-known companies, like AXA, Vero, RSA, Allianz, Tokio Marine, Travelers, Philadelphia, Delta, Hartford, Zurich and Hiscox. To achieve that we manually read and analyzed the contracts extracting a list of insurance terms. We consolidated the terms from the various contracts in order to obtain a generic list of terms that would suit all the contracts. Using this list, we created a table with the coverages and the exclusions of these contracts. This table is our "ground-truth", considering that we manually performed the semantic analysis of the contracts. The table contains terms that are under the categories of business interruption and cyber-crime. The terms that we included under the first category are the following: adware, brute force attack, cookies, Denial of service (DoS), Distributed denial of service (DDoS) attack, hacker attacks, key stroke loggers, logic bombs, Malicious code, Malware, past of present employee, phishing, spider ware, spyware, Trojan horses, Un-authorized access to a Computer System, Un-authorized access to data assets, Un-authorized used of a Computer System, Un-authorized used of data assets, virus, worms, zero-day. The terms under the second category are: fraudulent funds, theft loss, communications loss, fraudulent signature, vandalism loss, credit account, debit account, Telecom fraud, Social Engineering Fraud.

In the next step, we created an ontology using these terms and along with the analyzed contracts, we provided them as input to our tool. The ontology is manually created as an ascii file, with specific format. This is done ones and in the future amendments can be easily done. The output of the tool is a list of coverages and exclusions for each of the contracts. In the next steps we compared this list with the ground-truth table to see how many of the coverages and exclusions were correctly identified by our tool. Our tool utilizes the categorized terms as follows. If the name of a category is found in the coverages of a contract, it assumes that all the terms under this category are covered by the contract.

This initial evaluation showed that the accuracy of our tool varies, from 27% to 87% without any tuning. The average accuracy is 45%. In our approach, the accuracy of the results depends mainly on the definition of the ontology and how close the terms are defined in comparison with the actual policy wording. For this reason, it is expected that a more well-defined ontology, or a richer one, will give better results.

A second test has been performed using the same terms for the ontology but without classifying them under categories, having no hierarchies. This means that the algorithm will consider coverages only for the terms that are explicitly mentioned in the contracts, making it "stricter".

It is observed that in the case of the use of an ontology without hierarchy, the results are quite different in some of the contracts. The overall accuracy is also a bit better.The accuracy in this case variesalso from 27% to 87% but it differs for some of the contracts. The average accuracy here is 50%. What we can conclude by these two initial experiments is that the accuracy of the system depends on the ontology definition by the expert. In the case we have a very detailed ontology, the results should be better. On the other hand, having hierarchies in the ontology, although it is more appropriate semantically, it might not have the desired accuracy in our system.Therefore, more experiments should be performed using different ontology structures and definitions in order to conclude the most suitable one for most of the test contracts.

Regarding performance, we evaluated it in terms of speed, resource consumption and scaling, we performed several tests using a different number of contracts - pdf files, of various size. The experiments have been conducted 5 in the tested reported above. In particular, the experiments contained input with 10, 20, 100 and 200 discrete policies. The first experiment contained 10 contracts, and the time that the proposed ontology needed to complete the analysis was 62 seconds. During the second experiment, we fed the ontology with 20 unique policies, which the SECONDO achieved to successfully process them in 165 seconds. Later, the ontology assessed against 100 contracts and 200 individual contacts, the ontology spent 950 and 1451 second respectively to process them. We can observe that the time needed by the ontology to process the input is relatively linear in relation to the number of contracts it analyzed (see Figure 3). In addition, we have evaluated the resource depletion due to the ontology process. In terms of processing power, the program needed 15% CPU and 3% of the RAM regardless of the number of fields that feed the ontology (see Figure 2). This

occurs because our tool does not multi-process the policies, instead it processes one file per execution circle.

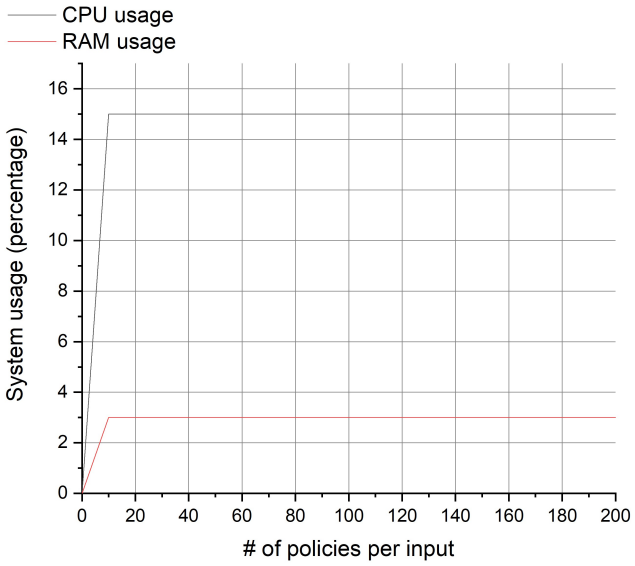

Figure 2: Resource usage

Moreover, we have assessed the ontology against a large pdf file (26.5 MB) and it terminated successfully after 594 seconds. We have to note that the size of a cyber insurance policy is not more than 1MB.Overall, we can observe that our proposed tool performs reasonably well, and the time needed to analyze the contracts is acceptable.

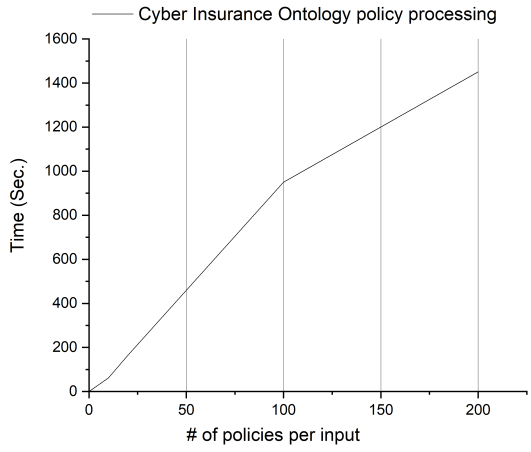

Figure 3: Cyber insurance ontology policy processing evaluation

Based on the above initial experiments we can identify the following advantages of our approach. Scalable (resource-depletion): The proposed solution is a tool that scales well without significant performance drawbacks in issues related to CPU and RAM consumption; it is a characteristic that leads to the fact that end-users can easily use it without specific hardware.

Scalable (words): The proposed solution is scalable regarding the wording. It is word-independent; by this, we mean that the proposed tool can be refined, re-edited and altered based on end-user requirements and desires. This allows the tolls to be updated any-time, a back-end feature.

Scalable (language): The proposed solution is scalable regarding the language. Currently, the tool works only for cyber insurance policies written in English. It is language-dependent; by this, we mean that the proposed tool can be refined, re-edited and altered based on end-users requirements and desires. For instance, correct words in different languages (Greek, Spanish, etc.) can be added to utilized vocabularies. This allows the tolls to be updated any-time, a back-end feature.

Time efficient: The proposed solution scales well regarding time management issues; we have already proven that the tool regardless of the size of the processed files performs well and is not a time-consuming tool.

Environment independent/ deployment: Currently the existing implementation is environment independent; by this, we mean that the proposed tool can work not only in a UNIX based environment (like the tested one, see Section 4 and 5), but also in a windows-based environment. The only requirement is the installation of Python in the working environment.

On the other hand, our approach also has some technical limitations that are listed below.

Contract parsing and formatting: We have tried to use a pdf parsing library that can analyze all the pdf contract files but since the contracts do not have a generic, globally accepted and defined structure or formatting, there is the possibility that a pdf cannot be analyzed correctly, giving wrong results. This issue cannot be addressed beforehand, but a mechanism to report any parsing errors can be developed.

Terms matching: The algorithm that does the matching between the ontology terms and the extracted terms from the contracts use exact word matching, meaning that if we have words that are not the same, the algorithm will not consider them a match. This limitation, though, can be overcome if we define an ontology using all the terms in all their possible forms. Since the ontology is to be defined once, this can be done initially and, in the future, it can be updated.

Different languages: Our system is flexible, and it can be used for different languages. Although, for each language we need to define the appropriate ontology, defining the terms that are used in each language. This of course, on the other hand has the advantage of not having to change the code or the algorithm in order to use it for any language.

Semantic contract analysis: Our system does not use an AI based approach to analyze the contracts or/and automatically define the ontology. It is probable that such a solution could have better results. Of course, in order to verify this, we have to compare our tool with another one that uses the AI-based approach. Ontology creation: Our solution requires manual ontology creation, by an expert. This is a step that has to be done initially and this also gives the possibility to easily extend and refine the ontology, having more control over

it. Since the ontology creation is done only once, this does not add a lot of complexity. An already defined ontology can be also used, as long as it can be extracted and then transformed in the format that our tool receives it as input.

7 CONCLUSION

An initial evaluation of our tool shows that our approach is valid and that the results are promising. Although it has been only assessed against a very limited number of documents and it has not been tuned to increase the accuracy and optimize the results. For this reason, the next step is to first optimize the ontology and the way our tool is using its terms to identify the coverages and exclusions of a contract. Another area of improvement is the parsing of the contracts and the extraction of the paragraphs that are mentioning the inclusions and exclusions. The text analysis is based on specific keywords and not in a semantic analysis of the document. While this seems to be accurate enough, more research is needed in order to validate it. Providing a broader list of terms or using a semantic analysis approach, may lead to better accuracy on extracting the parts of the document that are related to coverages and exclusions. Finally, a larger number of contracts has to be assessed and the list of terms and the ontology needs to be refined in order to be able to be more accurate in the coverages and exclusions extraction. Real user evaluation will follow, in order to evaluate the tool in real world cases.

In our future plans there is also the goal to define a generic ontology for the cyber-insurance domain which could be adopted by the major insurance companies. Finally, there is a provision to transform our tool to an online service providing an API that can be used to directly evaluate the various contracts, expand the contracts dataset and gain statistics insights to the cyber-insurance market.

ACKNOWLEDGMENTS

This research has been funded by the European Commission (Horizon 2020 Programme), and particularly by the project SECONDO (Grant Agreement no. 823997).

REFERENCES

[1] Tridib Bandyopadhyay, Vijay S Mookerjee, and Ram C Rao. 2009. Why IT managers don't go for cyber-insurance products. *Commun. ACM* 52, 11 (2009), 68–73.

[2] Rainer Böhme, Galina Schwartz, et al. 2010. Modeling cyber-insurance: towards a unifying framework.. In *WEIS*.

[3] Sam Adam Elnagdy, Meikang Qiu, and Keke Gai. 2016. Cyber incident classifications using ontology-based knowledge representation for cybersecurity insurance in financial industry. In *2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE, 301–306.

[4] Aristeidis Farao, Sakshyam Panda, Sofia Anna Menesidou, Entso Veliou, Nikolaos Episkopos, George Kalatzantonakis, Farnaz Mohammadi, Nikolaos Georgopoulos, Michael Sirivianos, Nikos Salamanos, et al. 2020. SECONDO: A platform for cybersecurity investments and cyber insurance decisions. In *International Conference on Trust and Privacy in Digital Business*. Springer, 65–74.

[5] Aristeidis Farao, Eleni Veroni, Christoforos Ntantogian, and Christos Xenakis. 2021. P4G2Go: A Privacy-Preserving Scheme for Roaming Energy Consumers of the Smart Grid-to-Go. *Sensors* 21, 8 (2021), 2686.

[6] Giulio Ganino, Domenico Lembo, Massimo Mecella, and Federico Scafoglieri. 2018. Ontology population for open-source intelligence: A GATE-based solution. *Software: Practice and Experience* 48, 12 (2018), 2302–2330.

[7] Aditi Gupta, Sudip Mittal, Karuna P Joshi, Claudia Pearce, and Anupam Joshi. 2016. Streamlining management of multiple cloud services. In *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*. IEEE, 481–488.

[8] IBM. 2022. Cost of a Data Breach Report 2021.

[9] Ketki Joshi, Karuna Pande Joshi, and Sudip Mittal. 2019. A semantic approach for automating knowledge in policies of cyber insurance services. In *2019 IEEE International Conference on Web Services (ICWS)*. IEEE, 33–40.

[10] Karuna P Joshi, Aditi Gupta, Sudip Mittal, Claudia Pearce, Anupam Joshi, and Tim Finin. 2016. Semantic approach to automating management of big data privacy policies. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 482–491.

[11] Security Magazine. 2021. SECONDO. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5dab106e5&appId=PPGMS7. [Online; accessed 19-June-2022].

[12] Security Magazine. 2022. 92% of data breaches in Q1 2022 due to cyberattacks. securitymagazine.com/articles/97431-92-of-data-breaches-in-q1-2022-due-to-cyberattacks. [Online; accessed 19-June-2022].

[13] Angelica Marotta, Fabio Martinelli, Stefano Nanni, Albina Orlando, and Artsiom Yautsiukhin. 2017. Cyber-insurance survey. *Computer Science Review* 24 (2017), 35–61.

[14] Antonio Muñoz, Aristeidis Farao, Jordy Ryan Casas Correia, and Christos Xenakis. 2021. P2ISE: Preserving Project Integrity in CI/CD Based on Secure Elements. *Information* 12, 9 (2021), 357.

[15] Ranjan Pal. 2012. Cyber-insurance in internet security: A dig into the information asymmetry problem. *arXiv preprint arXiv:1202.0884* (2012).

[16] Sakshyam Panda, Aristeidis Farao, Emmanouil Panaousis, and Christos Xenakis. 2019. *Cyber-Insurance: Past, Present and Future.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1–4. https://doi.org/10.1007/978-3-642-27739-9_1624-1

[17] Sasha Romanosky, Lilian Ablon, Andreas Kuehn, and Therese Jones. 2017. Content analysis of cyber insurance policies: How do carriers write policies and price cyber risk? *Available at SSRN 2929137* (2017).

[18] Sasha Romanosky, Lillian Ablon, Andreas Kuehn, and Therese Jones. 2019. Content analysis of cyber insurance policies: How do carriers price cyber risk? *Journal of Cybersecurity* 5, 1 (2019), tyz002.