

Machine learning based system for semantic indexing documents related to cybersecurity

Tiberiu-Marian GEORGESCU
The Bucharest University of Economic Studies
tiberiugeorgescu@ase.ro

This article presents a semantic indexing software system which uses natural language processing (NLP) techniques to understand documents related to cybersecurity. The purpose of this solution is to facilitate the cybersecurity documentation process as well as increasing cybersecurity awareness. The solution automatically collects documents related to cybersecurity available on the internet, keep relevant data, perform a cognitive analysis and enrich the documents, store the annotated documents and offer the possibility to access them according to users' choices. The paper describes the components of the system, the methods, technologies and tools proposed in order to implement the system. The solution includes a domain ontology and a machine learning (ML) model specialized in cybersecurity as well as a scraper to automatically download relevant data.

Keywords: Cybersecurity, Machine Learning, Natural Language Processing, Semantic Indexing

DOI: 10.12948/ei2019.01.01

1 Introduction

The speed of IT evolution, the constant changes of software technologies, tools or even paradigms bring many benefits to organizations and society in general, but they also bring challenges related to cybersecurity. These challenges can be considered from two perspectives: (1) technology is increasingly present in people's lives, hence the potential dangers intensify, (2) the volume of information and measures necessary to ensure an acceptable degree of cybersecurity is increasing, and specialists have difficulties in keeping up.

In order to maintain the managed systems properly configured and protected, the specialists are required to follow numerous sources of information on a regular basis. This process is proving to be a difficult task in practice, therefore facilitating the access to information could contribute to improving cybersecurity. Thus, this paper describes a semantic indexing solution which facilitate the information process.

In the literature review, various studies discussing solutions to improve the cybersecurity information process were identified. Article [1] recognizes the general increase of cyberattack surface and discusses about the

necessity of information sharing systems. A study of the main cybersecurity information sharing papers is performed and 82 relevant articles are identified and analyzed. Several papers discuss about using semantic tools to modeling cybersecurity domain, such as [2] and [3].

In paper [4] the authors selected 25 cybersecurity experts, collected over 70.000 of their Tweets and used analytics techniques to create a thesaurus on cybersecurity. By modeling the cybersecurity domain, such solutions can perform cognitive analysis and extract relevant information. Paper [5] presents a prototype system which collects and analyze cybersecurity related information posted on Twitter. Our work has a similar approach, but is not based only on information available on Twitter, it collects any relevant cybersecurity data available online.

A software system designed to facilitate cybersecurity information is described in detail. Its purpose is to automatically monitor the latest information relevant to cybersecurity, to filter them and to present them in an organized and structured manner according to the users' needs. The system automatically collects text data, analyzes it using NLP algorithms and stores the relevant documents on-

ly. Afterwards, the documents are annotated and semantically indexed. The annotated documents are available on a platform where the users can make semantic searches. The study is based on the author's PhD thesis, [6].

Section 2 illustrates the four-level architecture of the system and describes its components. Further sections detail each level, by discussing methods, technologies and tools necessary for implementation. Section 3 describes aspects about collecting data automatically from the internet. Custom-made web scrapers are used and a scraping solution implemented by us is presented. Section 4 discusses about cognitive analysis solutions for cybersecurity documents. An ontology and a ML based NLP model were developed. Section 5 presents databases designed to store documents annotated with semantic data and describe a solution we implemented. A cy-

bersecurity web platform containing annotated documents is presented in section 6. On this platform, users could consult relevant information and perform both syntactic and semantic searches.

2 The Architecture

This section display an overview of the proposed system. The architecture of the system is presented, along with a brief description of every level. Further sections detail each level. The architecture was built based on the following functional requirements: (1) automatic collection of data relevant to cybersecurity from the internet, (2) data analysis and semantic annotation through NLP tools, (3) storing the relevant documents along with semantic metadata, (4) presentation of the relevant data according to the users' options. Figure 1 illustrates the architecture of the semantic indexing solution.

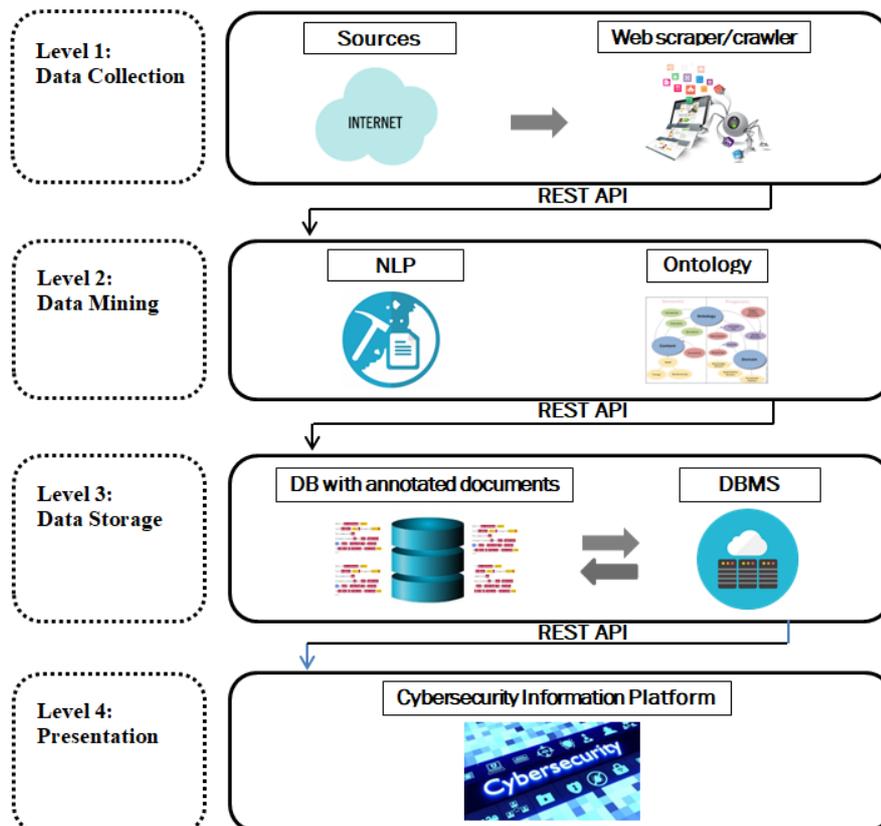


Fig. 1. The architecture of the semantic indexing system

Tier 1 architecture involves automatically collecting data related to cybersecurity. Data consists in text documents or HTML pages extracted from the internet. Custom web

scrapers are developed and used in order to continuously download data with high potential of being relevant. At this stage, information such as content, author, date, tags or

metadata are extracted and saved in JSON files.

The first two levels are connected through a REST API which facilitates the change of JSON files. Tier 2 performs data mining on the collected data. We propose the use of a ML based NLP service and a domain ontology. The main purposes of the NLP model are to assess a confidence score to the documents and enrich them with semantic annotation. Each data set has a relevance indicator attached, therefore we can set a threshold so only the documents with an indicator value over the threshold are kept and passed to the next level.

At level 3, annotated data is stored in NoSQL databases. The datastore needs to be scalable, so it is able to manage large volumes of data (up to millions of documents). A database with annotated documents is required, along with a database management system (DBMS).

Tier 4 deals with the presentation of the results. A web platform is proposed, where users can filter the documents and perform searches based on syntactic and semantic options. Within it, users can search for ideas, relations, entities and concepts. The web application can also be developed to provide e-learning facilities.

Below, each level is described in detail and technical solutions are proposed.

3 Data collection

Within level 1 of the architecture, relevant data is collected automatically. Most of the sources indicated do not have fetching possibilities through APIs, which is why scraping applications are developed and used.

Since our solution is required to handle and store large volumes of data, it is important to

start the filtering process from the tier 1 and eliminate noise as much as possible. Therefore, the web crawlers and scrapers used are designed to search only for cybersecurity information. Two approaches are taken into account. The first one consists in using a semantic crawler, which chooses which pages to index and which not to, by using ontologies or dictionaries. After the crawler selects relevant pages, the scraper downloads data from that pages. This approach is necessary when we want to extract particular data from multi-purpose websites. The second approach is to pre-select relevant websites and use custom-made scrapers to extract data. Its main advantage consists in the fact that the complexity of the crawler is reduced. For our system we consider the second approach to be good enough.

A pre-selection of relevant cybersecurity websites was conducted. In this regard, we are consulting dozens of cybersecurity professionals to find out which are their sources of information. Such a website, which was mentioned by more than 75% of the specialists consulted so far is <https://packetstormsecurity.com/>. Figure 2 illustrates the source code of a spider which automatically collects data from the website mentioned above.

This process was performed using the Scrapy framework. Scrapy is an open-source application used for automatically accessing websites and extracting unstructured data. Originally designed as a scraper, it can also be used to extract data using APIs or as a general-purpose web crawler [7]. We created a web spider for each of the websites from which data is downloaded. The data of interest is set through the selectors or xpaths.

```
import scrapy
class PacketStorm(scrapy.Spider):
    name = "packetone"
    start_urls = ["https://packetstormsecurity.com/files"]
    def parse(self, response):
        packet = response.xpath('//*[@class="file"]')
        for pack in packet:
            yield {
```

```

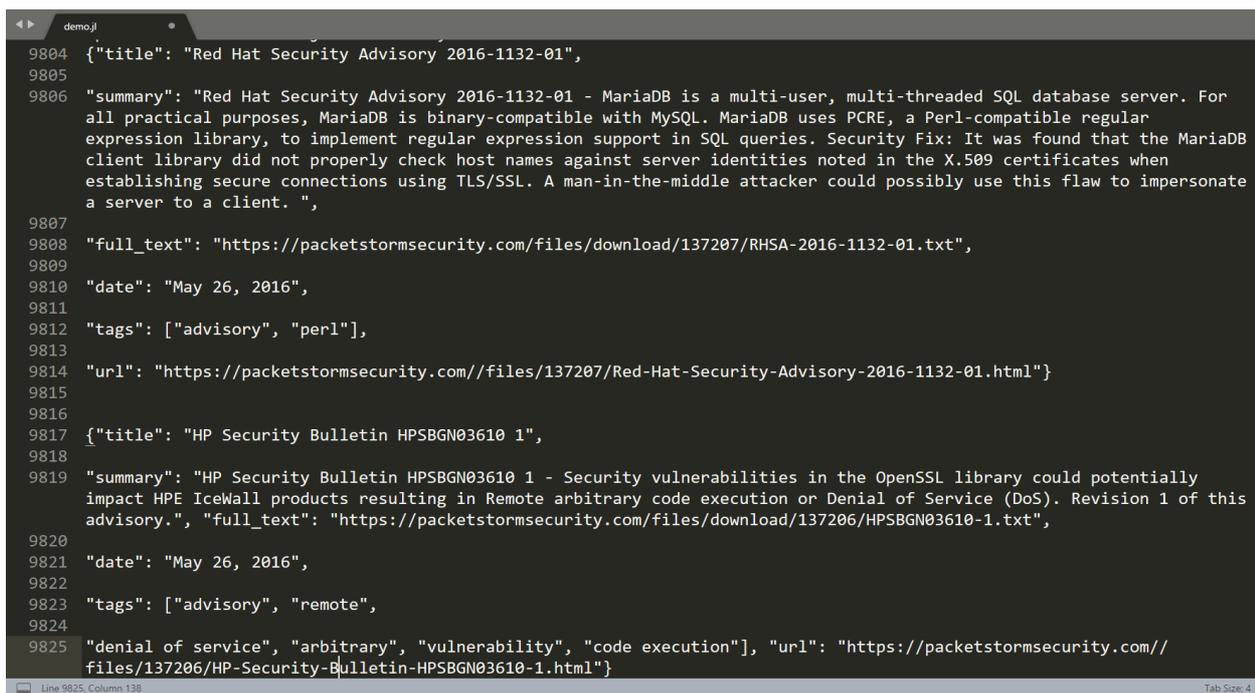
        'title': pack.xpath('//*[ @class="ico text-plain"]/text()').extract_first(),
        'summary': pack.xpath('//*[ @class="detail"]/p/text()').extract_first(),
        'full_text': 'https://packetstormsecurity.com' + str(pack.xpath('//*[ @class="act-
links"]/a/@href').extract_first()),
        'date': pack.xpath('//*[ @class="datetime"]/a/text()').extract_first(),
        'tags': pack.xpath('//*[ @class="tags"]/a/text()').extract(),
        'url': 'https://packetstormsecurity.com/' + str(pack.xpath('//*[ @class="ico text-
plain"]/a/@href').extract_first())
    }
    next_page_url = response.xpath('/html/body/div[2]/div/div[1]/div[3]/a[7]/@href').extract_first()
    absolute_next_page_url = "https://packetstormsecurity.com" + next_page_url
    yield scrapy.Request(absolute_next_page_url)

```

Fig. 2. The source code of the spider used to download data from <https://packetstormsecurity.com/>

The scraper is configured to automatically browse page by page until there are no more pages available. The xpath of the next page is stored in the variable `absolute_next_page_url`. The downloaded data is structured into the following categories: title,

summary, full text, date, tags and page URL. We downloaded approximately 120,000 posts and saved the data to JSON or JSON-LD files. Figure 3 illustrates a screenshot of a JSON document that contains data collected from <https://packetstormsecurity.com/>.



```

9804 {"title": "Red Hat Security Advisory 2016-1132-01",
9805
9806 "summary": "Red Hat Security Advisory 2016-1132-01 - MariaDB is a multi-user, multi-threaded SQL database server. For
all practical purposes, MariaDB is binary-compatible with MySQL. MariaDB uses PCRE, a Perl-compatible regular
expression library, to implement regular expression support in SQL queries. Security Fix: It was found that the MariaDB
client library did not properly check host names against server identities noted in the X.509 certificates when
establishing secure connections using TLS/SSL. A man-in-the-middle attacker could possibly use this flaw to impersonate
a server to a client. ",
9807
9808 "full_text": "https://packetstormsecurity.com/files/download/137207/RHSA-2016-1132-01.txt",
9809
9810 "date": "May 26, 2016",
9811
9812 "tags": ["advisory", "perl"],
9813
9814 "url": "https://packetstormsecurity.com//files/137207/Red-Hat-Security-Advisory-2016-1132-01.html"}
9815
9816
9817 {"title": "HP Security Bulletin HPSBGN03610 1",
9818
9819 "summary": "HP Security Bulletin HPSBGN03610 1 - Security vulnerabilities in the OpenSSL library could potentially
impact HPE IceWall products resulting in Remote arbitrary code execution or Denial of Service (DoS). Revision 1 of this
advisory.", "full_text": "https://packetstormsecurity.com/files/download/137206/HPSBGN03610-1.txt",
9820
9821 "date": "May 26, 2016",
9822
9823 "tags": ["advisory", "remote",
9824
9825 "denial of service", "arbitrary", "vulnerability", "code execution"], "url": "https://packetstormsecurity.com//
files/137206/HP-Security-Bulletin-HPSBGN03610-1.html"}

```

Fig. 3. JSON file with data automatically downloaded from the website <https://packetstormsecurity.com>

4 Data Mining

Once downloaded, two operations are necessary. The first consists in assessing a confidence score for each document in order to choose what to pass to the datastore and what to delete. The second consists in annotating the relevant document with semantic data. Various NLP solutions can be used. For our

approach we propose a ML based NLP solution and an ontology. In paper [8], we described an extensive ML based NLP model which is suitable for the architecture proposed in this article. Also, we develop a prototype called *Cybersecurity Analyzer*, available at [9], which can perform named entity recognition (NER) and relation

extraction to cybersecurity related documents. Also, it associates confidence scores to each document.

Both rule-based or ML approaches can be used. There are several articles that discuss the differences between rule-based and ML based NLP approaches, such as [10], [11] or [12]. For our system, we consider ML approaches to be preferred over the rule-based ones. The need to use ML tools comes from the limitations of rule-based models. Using ML algorithms, a properly trained model has two major advantages: (1) it can identify new instances, which have not been previously defined through dictionaries, (2) it can identify new surface forms of previously defined instances.

Choosing the best ML based NLP services

Globally, the leading cloud service providers are Amazon, Microsoft, Google and IBM. All four companies have also developed ML platforms-as-a-service. Amazon Machine Learning [13], Azure Machine Learning Studio [14], Google Cloud AutoML [15] and IBM Watson [16] are ML-as-a-service cloud solutions that provide full platforms for rapid model preparation and deployment. Although these companies offer ML solutions for various areas, from the perspective of our study only ML solutions for NLP are of interest. Article [17] illustrate a detailed comparison between the four NLP cloud services provided by the platform mentioned above.

The main services that we considered important for the development of the cognitive analysis solution described in this paper were: NER, relations extraction, sentiment analysis, intention analysis, personality analysis, syntactic analysis, POST, extracting key phrases, extracting topics, and extracting metadata. Based on these functional requirements, it was decided to use the cloud NLP solutions provided by IBM. The services provided by the IBM Watson suite of applications meet all functional requirements.

In our study [8], we described in full detail the development of a domain ontology and of a ML based NLP model for cybersecurity. First, an ontology which contained 18 types

of classes and 33 types of relations was created. The ontology structure was later used and implemented in a ML model, developed in IBM Watson Knowledge Studio [18]. Using this service, the following processes were realized:

- the implementation of a NLP based on ML model;
- the training of the model;
- the integration of Watson Knowledge Studio with other services;
- the calculation of the model's performance indicators in order to analyze and improve it.

The model was tested and improved until the performance indicators were considered satisfactory. Once a reliable model was developed in Watson Knowledge Studio, it was connected to Watson Discovery [19], a cloud service that allows the use of NLP services with ML components. The model was trained with documents totaling over 300.000 words and regular performance evaluations were performed. F1 score, precision and recall indicators were considered [20]. The F1 score obtained for NER, the main functionality, was 0.88, which showed the model's validity. Once the ML model is trained, it performs cognitive analysis on the documents. Uploaded data is automatically annotated by the model, returning additional information such as:

- metadata about entities and the classes to which the entities belong;
- the relations between the entities identified;
- confidence coefficients for each entity and relation identified, as well as for each document;

The decision to store or not a document is based on its confidence coefficient. The relevant documents are saved in JSON files along with their annotations and passed to the datastore.

5 Data storage

Tier 3 of the architecture deals with the management of the relevant documents. For this purpose, the use of document-based NoSQL databases is recommended. The functional requirements within this level are:

- storage of large volumes of data consisting

of JSON documents;

- development of functionalities that interpret the annotations associated with documents by the NLP model;
- development of functionalities that offer the possibility of conducting queries based on the enriched documents;
- development of APIs through which the datastore can be connected with a NLP solution and with a web interface.

In order to implement the level 3 components, two solutions are proposed, MongoDB, which is open-source and IBM Watson Discovery which is commercial.

In order to reduce the costs, we consider is MongoDB to be the most suitable open-source solution for data management specific to the designed software system. Unlike SQL solutions, MongoDB does function as a traditional databases, but stores the data through JSON files, which are organized according to dynamic schemas. MongoDB has a number of advantages such as:

- stores data as JSON documents, facilitating communication through APIs between level 3 of the described software system architecture and adjacent levels;
- allows full indexing of data;
- offers the possibility of performing flexi-

ble queries on documents;

- perform data partitioning;
- offers features for back-up, restoration, replication and availability;
- it is widely used, with high quality documentation;
- it is designed to be installed and configured on a server, allowing uninterrupted access to data [21].

Although MongoDB is suitable as a technical component of the software system proposed, the development of custom the features and functionalities mentioned above may involve the work of an entire programming team. Therefore, at this stage we opted to use Watson Discovery. Watson Discovery service meets all the functional requirements discussed above.

Watson Discovery store the metadata about each document. Each document is associated with a JSON file that contains all the annotations. Annotated documents are stored in a database that can be queried using a specific language called Discovery Query Language. Figure 3 illustrates part of the response received when loading a document called *Web Application Security*. The document is available on the homepage of the Cybersecurity Analyzer web application.

```
{
  "matching_results": 1,
  "session_token": "1_cxMn1KznDz77UIq3m33Lxcs5rH",
  "passages": [],
  "results": [
    {
      "id": "e8e2286c28dfb55c8e6030d45dbe6078",
      "result_metadata": {
        "confidence": 0.6414666332194946,
        "score": 0
      },
      "text": "no title\n\nWeb Application Security\n\n\n\nAs with any new class of technology, web applications have brought with them a new range of security vulnerabilities. The set of most commonly encountered defects has evolved somewhat over time. New attacks have been conceived that were not considered when existing applications were developed. Some problems have become less prevalent as awareness of them has increased. New technologies have been developed that have introduced new possibilities for exploitation. Some categories of flaws have largely gone away as the result of changes made to web browser software.\n\n\nThe most serious attacks against web applications are those that expose sensitive data or gain unrestricted access to the back-end systems on which the application is running. High-profile compromises of this kind continue to occur frequently. For many organizations, however, any attack that causes system downtime is a critical event. Application-level denial-of-service attacks can be used to achieve the same..."
    }
  ]
}
```

Fig. 3. The representation of the entities in JSON format

As can be seen in the figure above, the confidence coefficient for this document is 0.641, approximately 3 decimal places.

6 Data presentation

In order to achieve the highest level of utility, it is proposed to develop a web application that will serve as an interface to the semantic

indexing software system. Within it, users can search for current cybersecurity information. Simple or aggregate semantic queries can be performed by classes, entities, or relations, as well as syntactic searches by keywords. The interface returns the relevant paragraphs or documents.

The implementation of the software system

proposed is useful from two perspectives:

(1) it facilitates information for people concerned with cybersecurity by integrating on a single platform large volumes of relevant data extracted from multiple sources;

(2) it allows users to perform semantic searches, not just syntactic ones. Usually, for documentation, security specialists seek information by keywords. Searching by type of entity and relations between entities can lead to more accurate results. Following such searches, the solution returns either only the relevant paragraph/paragraphs or the entire document. Below, there is a semantic search written in Discovery Query Language. The query searches for documents which include the text *remote*, but only when it is part of a token which belongs to the class *Attacker*, the entity *FreeBSD*, when belongs to the class *Software*, as well as the text *buffer overflow* when it belongs to the class *Vulnerability*.

```
enriched_text.entities:(text:remote,type:
    Attack-
er),enriched_text.entities:(text:FreeBSD
,type:Software),en-
riched_text.entities:(buffer over-
flow,type:Vulnerability);
```

The implementation of e-learning solutions could further diversify the types of users. With the cognitive text analysis model behind it, a system can be developed that automatically builds a series of lessons based on the requirements of each user.

7 Conclusion and future work

The dangers to cybersecurity are continually diversifying, and security specialists need to make constant efforts to be informed about the latest vulnerabilities, threats, types of attacks, software system protection solutions and so on. This paper proposes an automated system for semantic indexing of cybersecurity documents that can facilitate the access to information. The system components, techniques and technologies through which it can be implemented are described in detail. Up to this point, we implemented separately the most important components. In the future we want to implement all the components together, as a self-contained software system.

Our study can also be used as an example for developing semantic indexing solutions for other domains. The same technologies, architecture and components as those proposed above can be used. The main differences consist in the development of ontologies and ML based NLP models which have to be customized according to the characteristics of the chosen fields.

Acknowledgement

This paper is based on the research the author made in his PhD paper "Modelarea bazată pe volume mari de date. Securitate Cibernetică în contextul Big Data" (eng. "Modelling based on large volumes of data. Cybersecurity in the context of Big Data") [6] and within the PN-III-P1-1.2-PCCDI-2017-0272 ATLAS project ("Hub inovativ pentru tehnologii avansate de securitate cibernetică / Innovative Hub for Advanced Cybersecurity Technologies"), financed by UEFISCDI through the PN III –"Dezvoltarea sistemului national de cercetare-dezvoltare", PN-III-P1-1.2-PCCDI-2017-1 program.

References

- [1] Ali, and Jun Zhuang Pala, "Information Sharing in Cybersecurity: A Review.," *Decision Analysis*, vol. 16, no. 3, pp. 172-196, 2019.
- [2] A. E., K. Węcel, and W. Abramowicz Aviad, "A semantic approach to modelling of cybersecurity domain," *Journal of Information Warfare*, vol. 15, no. 1, pp. 91-102, 2016.
- [3] Peter, Amila Silva, and Wei Lu. Phandi, "Semeval-2018 Task 8: Semantic Extraction from CybersecUrity REports using Natural Language Processing (SecureNLP)," in *Proceedings of The 12th International Workshop on Semantic Evaluation. 2018*, 2018.
- [4] Q., Bakare, S., Verma, A., Casasanta, C., White, C., Cotoranu, A., & Leider, A. Chen, "Analyzing Expert Cybersecurity Twitter Accounts by Using Thesaurus Methods for Text Analytics," in *Proceedings of Student-Faculty Research*

- Day, CSIS, Pace University, 2018.
- [5] Satyanarayan Raju, George Hsieh, and Kevin S. Nauer Vadapalli, "TwitterOSINT: Automated Cybersecurity Threat Intelligence Collection and Analysis using Twitter Data," in *Proceedings of the International Conference on Security and Management (SAM), The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, 2018.
- [6] Tiberiu-Marian Georgescu, *PhD thesis: Modelarea bazată pe volume mari de date. Securitate cibernetică în contextul Big Data (eng. "Modelling based on large volumes of data. Cybersecurity in the context of Big Data")*. Bucharest: The Bucharest University of Economic Studies, 2019.
- [7] Scrapinghub Ltd. (2019, January) Scrapy. [Online].
<https://docs.scrapy.org/en/latest/intro/overview.html>
- [8] Tiberiu-Marian Georgescu, "NLP model for automatically processing cybersecurity related documents, sent for peer review," *Informatica Economică*, vol. 23, no. 4, 2019.
- [9] Tiberiu-Marian Georgescu. (2019, January) Cybersecurity Analyzer. [Online].
<http://www.cybersecurityanalyzer.com/>
- [10] R. M., Fabbri, D., Denny, J. C., Rosenbloom, S. T., & Jackson, G. P. Cronin, "A comparison of rule-based and machine learning approaches for classifying patient portal messages," *International journal of medical informatics*, vol. 105, pp. 110-120, 2017.
- [11] W., Daud, A., Nasir, J. A., & Amjad, T. Khan, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait journal of Science*, vol. 43, no. 4, 2016.
- [12] Maryna Dorash. (2017) Machine Learning vs. Rule Based Systems in NLP. <https://medium.com/friendly-data/machine-learning-vs-rule-based-systems-in-nlp-5476de53c3b8>.
- [13] Amazon. (2019, October) Amazon Machine Learning. [Online].
<https://docs.aws.amazon.com/machine-learning/latest/dg/what-is-amazon-machine-learning.html>
- [14] Microsoft. (2019, October) Microsoft Azure. [Online].
<https://azure.microsoft.com/en-us/overview/ai-platform/>
- [15] Google. (2019, October) Google Cloud. [Online].
<https://cloud.google.com/automl/>
- [16] International Business Machines Corporation. (2019, October) IBM Watson Studio. [Online].
<https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/ml-overview.html>
- [17] Altex Soft. (2018) AltexSoft. [Online].
<https://www.altexsoft.com/blog/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/>
- [18] IBM. (2019, Februarie) IBM. [Online].
https://cloud.ibm.com/docs/services/watson-knowledge-studio?topic=watson-knowledge-studio-wks_tutintro
- [19] IBM (International Business Machines). (2019, January) IBM Cloud. [Online].
<https://console.bluemix.net/docs/services/discovery/index.html#about>
- [20] CoNLL 2018. (2018) Shared Task Evaluation. [Online].
<https://universaldependencies.org/conll18/evaluation.html>
- [21] MongoDB Inc. (2019, Februarie) MongoDB. [Online].
<https://www.mongodb.com/>



Tiberiu-Marian GEORGESCU graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2012. In 2015 he graduated the Informatics Systems for the Management of Economic Resources Master program. He completed his PhD program in Economic Informatics in September 2019 at the Bucharest University of Economic Studies. Currently, he is working as a Research Assistant in the Department of Economic Informatics and Cybernetics at the Bucharest University of Economic Studies. His main fields of interest are cybersecurity, machine learning and natural language processing.