Services Computing for Cyber-Threat Intelligence: The ANITA Approach

Daniel De Pascale[1], Giuseppe Cascavilla[2], Damian A. Tamburri[2], and Willem-Jan van den Heuvel[1]

[1] University of Tilburg & JADS, The Netherlands
[2] Eindhoven University of Technology & JADS, The Netherlands

Abstract. Major cybersecurity and threat intelligence analysts agree that online criminal activity is increasing exponentially. Technologies, newspapers, the internet, and social media made the dark web an accessible place to almost everyone. The ease of accessing the dark side of the web makes the problem more critical than ever. For this reason, the European Union financed the ANITA project, consisting of different tools for monitoring and fighting illegal criminal activities on the Dark Web. In the ANITA project, we propose different Big Data analytic tools for the analysis of all data extracted from illegal marketplaces. In this survey paper we present our developed tools for detecting trends and analyzing the incoming information with respect to illegal trafficking. The tool extracts information about specific trends, analytics and produces actionable insight on buying and transaction habits and user behaviors. The tool extracts statistics in order to support and guide investigators and law enforcement agencies for the detection of criminal activities.

1 Introduction

Over the recent years, online illegal trafficking activities have hugely elaborated and expanded so that to operate at global level with worldwide supply chains, production facilities and administrative offices, while their legal, economic and sustainability state is optimised [1]. For tackling these emerging challenges, a significant part of LEAs' (Law Enforcement Agencies) efforts has been invested on training activities to equip officers and practitioners with the necessary knowledge and skills related to this emerging and continuously/rapidly evolving scenery [2].

Accordingly, the European Union financed the EU H2020 ANITA project, consisting of different services for monitoring and fighting online criminal activities. ANITA's primary goal is twofold: a) to boost the LEA's investigation process and to significantly increase their operational capabilities, by introducing a set of innovative tools for efficiently addressing online illegal trafficking challenges (namely online data source analysis, blockchain analysis, Big Data analytics, knowledge modelling, incorporation of human cognitive function in the analysis pipelines, user-oriented intelligence applications), and b) to significantly facilitate the novice officers training process and to optimize the learning

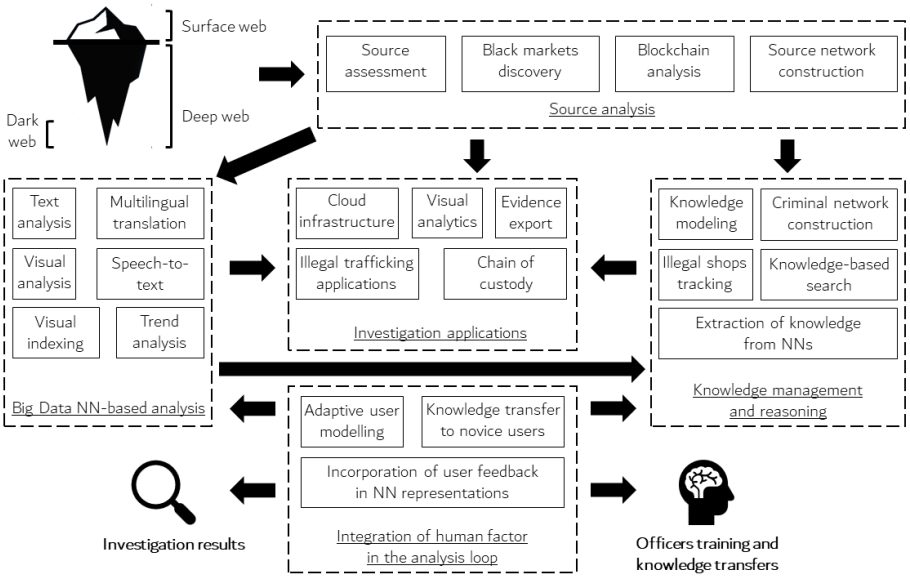

Fig. 1: Graphical representation of the ANITA action.

curve (by collecting, integrating and re-using knowledge from multiple expert officers and through the development of a recommendation functionality to transfer the acquired 'know-how' to the new officers). This paper discusses the different approaches part of the ANITA action.

2 Concept and Approach Overview

ANITA, as explained in the graphical representation in Fig. 1, is divided into five different steps:

1. *Source Analysis toolbox;*
2. *Big Data NN-based analysis;*
3. *Knowledge management and reasoning box;*
4. *Investigation application box;*
5. *Integration of human factor in the analysis loop.*

Source Analysis box: the first necessary steps in the analysis chain comprise the detection, assessment and analysis of potentially interesting sources that can be found on Surface/Deep/Dark Web. In particular, a new generation of data collection tools is developed, specific to LEAs needs. Specifically, dedicated services are responsible for: a) Anonymous identification of new relevant content with a balance between speed and precision; b) High performance download and storing in a secure repository; and c) Assessing the importance (i.e. level of relevance and dangerousness) of the examined Web sources, the discovery of

black markets, block chain analysis for revealing cues about illegal transactions tracking and the construction of a source network (that includes multi-level information for every source as well as the interconnections/interrelations among the identified sources).

Big Data NN-based analysis box: having identified and collected vast amounts of multimedia material related to illegal trafficking, ANITA applies a set of sophisticated Big Data analytic services for efficiently manipulating the acquired information and robustly detecting meaningful events. Below the list of the provided set of analysis.

- Text analysis services, delivered through the usage of a semantic based engine and capable of automatic categorization and entity extraction of the contents coming from Social, Surface/Deep/Dark Web, and other sources.
- Visual content analysis in order to identify potentially interesting pieces of information or evidence in the formed databases. For addressing the particular challenge deep hashing approaches are developed to recognise semantic entities at multiple scales.
- For supporting the processing of documents written in different languages a multilingual translation service is developed in order to automatic translate segments of speech to another language, audio streams, speech-to-text methodologies to perform the transformation of a speech segment to the form of a written document.
- Illegal trafficking trend analysis in order to extract information about specific trends, analytics and actionable insights on buying habits and user behaviours.

Knowledge management and reasoning box: apart from the inevitable large-scale data-driven analysis, the ANITA system is grounded on appropriate semantic knowledge structures that summarize explicit domain expert knowledge regarding the application field and which enable the realization of high-level semantic inference tasks (e.g. inconsistency checking, reasoning, outlier detection, etc.). The collected knowledge (i.e. application domain expertise) renders feasible the realization of complex and highly demanding tasks, like criminal network construction, illegal shops tracking and knowledge-based search and retrieval that are of vital for analysing different aspects of the illegal trafficking incidents.

Investigation application box: all the above-mentioned system functionalities drive the design and support the operation of a set of novel investigative applications to be delivered to the project stakeholders. In particular, the ANITA system is based on the design and implementation of a scalable and Big Data oriented infrastructure, able to analyse large volumes of data in near real time and to summarize analysis results to provide LEAs with relevant insights on illegal trafficking related phenomena.

Integration of human factor in the analysis loop box: overall, the fundamental consideration of the ANITA system to integrate the human user in the analysis pipeline serves the following two fundamental project goals (and simultaneously

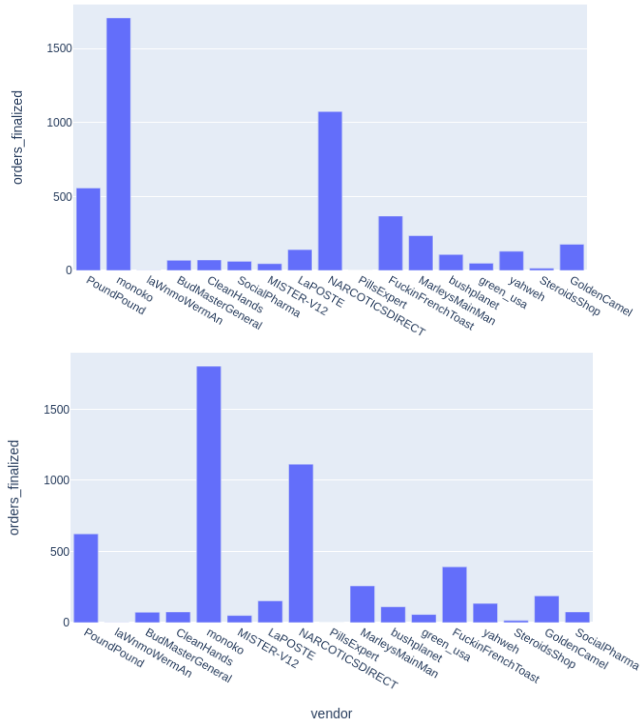

Fig. 2: Trend analysis on orders done in the Berlusconi market by different vendors in two different snapshots.

main outputs of the system): a) To significantly boost the efficiency of the investigation process, by continuously improving the robustness of the feature detectors through the incorporation of the explicit and implicit user feedback, while also updating and expanding the knowledge infrastructure for the selected application domain. b) To remarkably speed up the training process of new/novice investigators, practitioners and officers for the application domain at hand, by re-using and transferring knowledge that has been collected and combined from multiple expert users.

3 Big Data Services and Analytics

As described in Fig. 1 in the box of *Big Data NN-based analysis* the services computing part reflects mostly the large scale trend analysis. Our research approach in this direction focuses on developing services, methods, techniques, and approaches to extract useful information from Social, Surface, Deep, and Dark Web in order to develop a classification tool and a trend analysis tool.

On the one hand, the following text summarises the approach we followed to classify any given web page from the Surface, Deep, and Dark Web to give a clear

insight of its content and being independent from the language constraint. First, using a random forest machine learning approach we were able to predict with an accuracy 81.664%, whether a web page contains illegal activities, making use of the 3 activity classes (Suspicious, Unknown, Normal) as dependent variable. Second, our proposed approach showed an accuracy of 66.251% on 26 different classes of a specific activity type (i.e., Drugs, Hacking, Forum, Social-Network, Violence, Fraud, Counterfeit-Money, etc.). Finally, we rank recommendations to provide the best approach to predict the content of a web page is to use both website appearance and software quality parameters.

On the other hand, in order to classify different web sites we developed an architecture to collect, preprocess and model data. On top of that, we conducted a trend analysis on the classifications, in order to see if there are relations between observable factors (e.g., if more cocaine is being sold, there is less hash being sold). The process starts from the information gathering from different pages, in order to extract all possible information about vendors and products of different markets. This process is executed on the same pages in different interval of time. In this way, it is possible to analyze the trend of different variable, as price or number of visualization. The Fig. 2 shows the trend analysis applied on the order finalized by different vendors. The figure shows two snapshots: one related to the order finalized in 2019-09-11 and the other the order finalized in 2019-09-18; even from this single datapoint a trend in the time-series is evident.

3.1 Trend Analysis Module Architecture

The scraping tool consists of four consecutive steps. These steps will globally be discussed to give an overview of the design of the tool. In the next chapter will discuss each part in a little more depth. The overview is given in Fig. 3

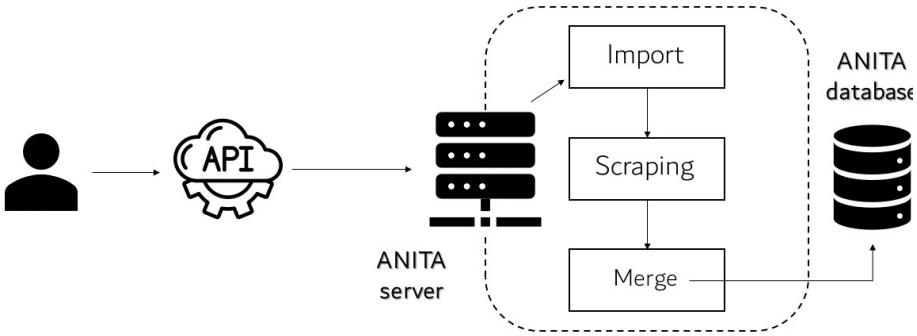

Fig. 3: General overview of the scraping part of the tool.

The goal of the complete tool is to extract features from darknet market and to structurally store this information to be able to find changes and trends in

the pages about vendors or products on these markets. This can be done by manually looking at the information for specific vendors or products or using the accompanying visualization tool.

The input of the tool consists of a dump of a market to be scraped in a ZIP format or as a simple folder. The tool can take the information of the market by the user himself or it can automatically extract the information inside the web pages included in the ZIP file.

1. **Import** The first phase imports and filters the useful files out of the provided dump of the market. The tool takes the HTML files and checks whether the HTML file is a page about a vendor or product. At the end of the entire process, all pages analyzed are removed from the server, except for the ZIP file, that is stored in the system for security purpose.
2. **Scrape** The second phase is focused on retrieving the features out of the HTML files. While every market is different, every market has a separate file that includes the information what to scrape from the page. In this phase the file simply scrapes this information to save in the database.
3. **Merge** The third phase focuses on merging the scraped information. While markets usually contain out of different pages that might contain the same information, there will be duplicate values in the data. For example, the profile page of a vendor might have information about the score of the vendor, while the page about feedback also contains this information. The merge step merges the information per page into information per vendor or product.
4. **Export into database** The last phase saves the information extracted in the scraping step and merged in the merge step into the database.

At the end of the process, the information extracted are saved into the db. In order to give access of those information to the end user, we provide some API services.

4 Conclusion

The ANITA project aims to define a tool for monitoring and fighting illegal trafficking activities on the Dark Web. To support this goal, ANITA provides a pipeline that, after having analyzed the Dark Web to discover different black markets and extract information from them, it applies different approaches to analyse them, from Big Data analysis to Knowledge management, allowing the interaction from all these components to improve results. In our case, we are focusing on the Big Data risks and trends analysis services experimentation and implementation. We implemented two different approaches, one based on the analysis of web pages using software quality, making the entire process language-independent. The other approach, starting from many product and vendor pages, provides a trend analysis among all variables extracted from these pages (e.g. price, order finalized, and trends).

References

1. Lewis, J., Baker, S.: The economic impact of cybercrime and cyber espionage. Tech. rep., Centre for Strategic and International Studies (2013)
2. Miller, C.H., Dunbar, N.E., Jensen, M.L., Massey, Z.B., Lee, Y.H., Nicholls, S.B., Anderson, C., Adams, A.S., Elizondo, J., Thompson, W., Wilson, S.N.: Training law enforcement officers to identify reliable deception cues with a serious digital game. IJGBL **9**(3), 1–22 (2019)