TIU-JADS - Status report

Period: 1st May 2018 – 1st October 2019 (M1-M18)

Resp.: Giuseppe Cascavilla (TIU-JADS)

# Activities completed during the 1st Reporting Period

|  |  |
| --- | --- |
| WP | Activities description and deviation |
| WP1 | activities have concentrated on leading technical work of tasks T5.1 and T6.5 as specified below. |
| WP4 |  |
| WP5 | In task 5.1 we conducted a systematic literature review (SLR) in the state of the art of deep/dark web-based classification of criminal activities. First we defined our Inclusion and Exclusion criteria in order to search our papers from the literature. We then defined the topic codes in order to codify the paper found. We then connected each paper to the right code based on the paper’s theme. We then applied topic modelling used to provide emerging themes in textual data, subsequently labelling the emerging themes which are visible and observable characteristics of potential online sources for criminal activity. Results were reported regularly as planned in D5.1. |
| WP6 | In task 6.5 we aimed at providing advanced tooling for fighting online illegal trafficking focusing on specific trend and risk analyses. We incepted appropriate tooling which, although in prototypical stage is already mature enough to provide valuable predictions of online criminal activity given specific target sources of analysis. Our prototype classifies and analyzes websites on the surface-, deep-, and dark-web, based upon website characteristics (e.g., the appearance as well as software quality parameters). We use indicators and parameters elicited from task 5.1 in order to incept learners based on those parameters and indicators as features of deep and darkweb sites. Subsequently, we conducted experimental research to select and test different Machine-Learning approaches in order to find the best classifier or combinations thereof. To build our dataset we used the Darknet Usage Text Address (DUTA). Moreover, we created our own web scraper, a tool able to retrieve the complete layout, descriptive text and Html source code of pages from the darkweb. This scraper is built by using Python Requests library over the TOR browser. |
| WP9 |  |
| WP10 |  |
| WP11 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| WP | Task | Activity | Results/Achievements |
| WP1 | **T1.2** | <Brief description of the completed task> | <Brief description of the achieved results and achievements in the task> |
| WP4 | **T4.2** |  |  |
| WP5 | **T5.1** | Task 5.1 aims at providing a synthesis of the state of the art in cybercrime threat intelligence, accounting for both grey and white literature on the matter with a systematic multi-vocal literature review. | We found a gap among grey literature and the white literature. The Grey one mainly discusses reported vulnerabilities as well as organisational/economical/financial consequences of being targeted by cybercriminal activity. The White one mainly focuses on offering scattered non-definitive attempts at predicting, avoiding, or protecting against specific criminal-activity types. The gap is related to the lack of scientific studies in Dark Web field related to Anonymous Crawling Policies, Software Quality and WebSite Appearance Parameter.  Second, we found that no single community encapsulates cybercrime-fighting software, tools, approaches and techniques, rather, these techniques or their relevant related work is scattered across as many as 30+ domain-specific communities (e.g., software security, data privacy, software engineering, distributed computing, artificial intelligence, and more). Third, finally, there is no one definitive solution towards assisting law-enforcement agencies in their cybercrime fighting activity. |
| WP6 | **T6.5** | Task 6.5 provides advanced tooling for fighting online illegal trafficking. To be more specific, the tool classifies and analyzes websites on the surface-, deep-, and dark-web, based upon website characteristics like the appearance as well as software quality parameters. In order to classify these websites, an architecture to collect, pre-process and model the data is developed and evaluated as well. We tested different type of machine learning models: K-nearest neighbor (KNN), Logistic regression (LR), Support Vector Machines (SVM) and Random Forest (RF). | From our testing we can assert that the Random Forest showed to be the best model to predict whether a website has suspicious activity (that is an activity that is likely to be illegal such as selling weapons or drugs) with an accuracy of 81.154%. This tool is thought to be used by Law Enforcement Agencies to get a clearer overview what kind of websites have illegal activities, without having to explore the website themselves. One of the best achievements from the proposed tool, and differently from the other available tools, is that there is no need to include any advanced text analysis, which means that any website in any language can be classified. |
| WP9 | **T9.5** |  |  |
| **T9.6** |  |  |
| WP10 | **T10.2** |  |  |
| WP11 | **T11.1** |  |  |
| **T11.2** |  |  |
| **T11.3** |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| WP | Task | Completed deliverables (including internal) | Achieved Milestones: |
| WP1 | **T1.2** | * Dx.y - <Title>   <any other produced result> | MSx - <Title> |
| WP4 | **T4.2** |  |  |
| WP5 | **T5.1** | * D5.1 - Data source risk assessment in the Surface Web, Deep Web and Dark Nets |  |
| WP6 | **T6.5** | * D6.5 - Illegal trafficking trend analysis |  |
| WP9 | **T9.5** |  |  |
| **T9.6** |  |
| WP10 | **T10.2** |  |  |
| WP11 | **T11.1** |  |  |
| **T11.2** |  |
| T11.3 |  |

# Activities in progress

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| WP | Task | Ongoing Activity | Expected Results/Achievements | Due Date |
| WP1 | **T1.2** | <Brief description of the activities in progress> | <Brief description of the expected results and achievements> |  |
| WP4 | **T4.2** |  |  |  |
| WP5 | **T5.1** | We built a questionnaire for practitioners in order to understand if what we found in task 5.1 shared also in the everyday environment. Based on this research we will write down a deliverable where we discuss the findings. | We expect to have a confirmation from the practitioners about our parameters and indicators found in task 5.1. |  |
| WP6 | **T6.5** | We built the Machine Learning models. We are now further testing them in order to have better performances. Moreover, we are going to test the trend analysis tool for the Dark Web markets. | The Random Forest classifiers on 28 classes classification reaches 65.3% of accuracy, on 3 classes classification reaches 81% of accuracy. We expect to improve these results adding the possibility to use a text classifier. |  |
| WP9 | **T9.5** |  |  |  |
| **T9.6** |  |  |  |
| WP10 | **T10.2** |  |  |  |
| WP11 | **T11.1** |  |  |  |
| **T11.2** |  |  |  |
| **T11.3** |  |  |  |

|  |  |  |
| --- | --- | --- |
| WP | Task | Work in progress deliverables (including internal) |
| WP1 | **T1.2** | * Dx.y - <Title>   <any other produced result> |
| WP4 | **T4.2** |  |
| WP5 | **T5.1** | * D5.6 - <Title> |
| WP6 | **T6.5** | * D6.5 – Final version of the previous deliverable including the trend analysis |
| WP9 | **T9.5** |  |
| **T9.6** |  |
| WP10 | **T10.2** |  |
| WP11 | **T11.1** |  |
| **T11.2** |  |
| **T11.3** |  |

# Risks

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Risk | Impact | Probability (L/M/H) | Mitigation action/plan | Involved partners |
| T1.2 | <Brief description of the risk> | <Brief description of the impact> |  |  |  |
| T4.2 |  |  |  |  |  |
| T5.1 |  |  |  |  |  |
| T6.5 | **Lack of dataset from the previous task** | Impact High, since that it will be too difficult to download a dataset by hands in order to test the tool for the trend analysis |  |  |  |
| T9.5 |  |  |  |  |  |
| T9.6 |  |  |  |  |  |
| T10.2 |  |  |  |  |  |
| T11.1 |  |  |  |  |  |
| T11.2 |  |  |  |  |  |
| T11.3 |  |  |  |  |  |

# WP\_ - Meetings/conference calls

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Task | Meeting/ teleconference | Date | Location | Purpose / Justification / Outcomes | Host | Partner Participants |
| WP1 |  |  |  |  |  |  |
| WP2 |  |  |  |  |  |  |
| WP3 |  |  |  |  |  |  |
| WP4 |  |  |  |  |  |  |
| WP5 |  |  |  |  |  |  |
| WP6 |  |  |  |  |  |  |
| WP7 |  |  |  |  |  |  |
| WP8 |  |  |  |  |  |  |
| WP9 |  |  |  |  |  |  |
| WP10 |  |  |  |  |  |  |
| WP11 |  |  |  |  |  |  |
| WP12 |  |  |  |  |  |  |

**Publication of articles**

* <Authors, *Title*, Journal/Conference/other>