Project Deliverable

D8.5
Results of the competitions on machine learning for security

Project Number	700692
Project Title	DiSIEM – Diversity-enhancements for SIEMs
Programme	H2020-DS-04-2015

Deliverable type	Report
Dissemination level	PU
Submission date	31st of August 2019 (M36)
Extended submission date	14th of September 2019

Responsible partner	FCiências.ID
Editor	Pedro Ferreira
Revision	1.0

The DiSIEM project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700692.

Editor
Pedro Ferreira, FCiências.ID

Contributors
Pedro Ferreira, FCiências.ID
Nuno Dionísio, FCiências.ID
Fernando Alves, FCiências.ID
Alysson Bessani, FCiências.ID
Olivier Thonnard, Amadeus
Pedro Dias Rodrigues, EDP
Mario Faiella, Atos

Version History

Version	Author	Description
0.9	Pedro Ferreira (FCiências.ID)	First version of the document for internal review
1.0	Alysson Bessani (FCiências.ID)	Minor improvements and submission to EC.

Executive Summary

This report discusses the work carried out in the context of Work Package 8 (*Dissemination, Communication and Exploitation*), Task 8.3 (*Security-related threat prediction competition*). It describes the two cybersecurity-related competitions that were organized in neural network and machine learning conferences. The challenges raised by these competitions were related to the work on the OSINT Threat Detector component, described in Work Package 4 deliverables. The report details the competitions problem statements, data sets, evaluation metrics and participation, which was much lower than we expected, despite many efforts to disseminate them. A reflection is provided on the possible reasons for the minimal participation achieved in the competitions.

Table of Contents

List of Figures

List of Tables

1 Introduction

To foster the dissemination of DiSIEM within the technical and scientific communities working on cybersecurity threat detection and prediction using Open Source INTelligence (OSINT) and machine learning, two open competitions were organized within two renowned international conferences: the *2018 International Joint Conference on Neural Networks*, IJCNN'2018 (co-located with the IEEE World Congress on Computational Intelligence); and the *Workshop on Machine Learning for CyberSecurity*, MLCS'2019 (co-located with the 2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECMLPKDD'2019).

1.1 Context

Cyber Threat Intelligence (CTI) is becoming a major topic in cyber defense. A study from February of 2018 estimated that the cost of cyber-crime may have cost up to 0.8% of the global GDP in 2016 [1]. Another study from July of 2019, found that data breach costs have increased by 12% over the past 5 years [2].

Given the current situation in the cyberspace, companies and organizations are striving for Cyber Threat Intelligence (CTI). Obtaining actionable information about the latest updates, patches, vulnerabilities and threats in a timely and precise manner results in informed decision making and a robust cyber threat awareness. Through CTI, useful information can be extracted and converted into knowledge, which can be used to prevent and mitigate attacks as well as support decision making. A report by the European Union Agency for Cybersecurity (ENISA), published in December of 2017, showed that although the adoption of SOCs and incident response plans are mainstream, CTI is still in its early adoption phase [3]. The main goal of CTI tools is to provide security analysts with timely information about security threats to the IT infrastructures under their responsibility. This translates into two important objectives:

- Maximize the amount of relevant information presented to the analyst;
- Minimize the amount of irrelevant information presented to the analyst.

Research has shown that OSINT provides a massive amount of information that could be leveraged to produce timely, relevant and critical intelligence feeds [4]. Currently, OSINT is the centerpiece of applications for CTI. These sources include not only security-specific media, such as security blogs, articles, vulnerability and exploit descriptions [5], but also social media platforms, e.g., Twitter, have gained the attention of the research community for their natural capability of aggregating multiple information sources [4, 6, 7, 8, 9].

The DiSIEM project explored the possibility of using OSINT information to fulfil the objectives stated. For that purpose, we continuously collect tweets from a tailored set of accounts, concerning the security of three case-study IT infrastructures specified by three industrial partners of the project: EDP, Atos, and Amadeus. Although there are many sources of OSINT, including security-related ones, Twitter was used for two main reasons. First, Twitter is a social media micro

blogging website well-recognized as an important information hub for short notices (almost in real-time) from multiple sources [10], about cutting edge information on events regarding many subjects. These include cybersecurity-related events as demonstrated by the highly active accounts of most security feeds and researchers, where they tweet security-related news. Second, since a tweet is limited to 280 characters (mostly 40–60 words), these messages are simple to process automatically.

In this context, and in alignment with the work developed in work package 4, two DiSIEM competitions have been organized:

- *WCCI 2018 Competition on Open Source Intelligence Discovery for Cybersecurity Threat Awareness*;
- *MLCS 2019 Competition on Multi-Task Learning in Natural Language Processing for Cybersecurity Threat Intelligence*.

1.2 Organization of the document

Chapter 2 describes the first competition organized within the WCCI'2018 IEEE Congress. Then, in Chapter 0, we describe the second competition, organized within the MLCS'2019 workshop co-located with ECMLPKDD'2019. Both chapters describe in detail the competitions, including how data was acquired and labelled, the task or tasks that participants were meant to tackle and how evaluation was conducted. Chapter 0 provides a reflection on the possible reasons behind low participation, and important lessons taken towards future improvements that may lead to more successful competitions. Finally, Chapter 0 presents a summary of the work and draws conclusions.

2 WCCI 2018 Competition

The first competition proposal, entitled *Competition on Open Source Intelligence Discovery for Cybersecurity Threat Awareness*, was submitted to the call for competitions [1] that was organized by IJCNN'2018 within WCCI'2018. [2] This conference is mainly directed to the Artificial Neural Networks (ANN) community. Besides availability, the conference was chosen because ANNs were considered the most promising technique to address the competition challenge.

Once the competition was accepted by the WCCI'2018 organization, a competition web page[3] (Figure 1) was created on the DiSIEM project web page. Additionally, the competition was also created and publicized[4] on the popular Kaggle[5] data science and machine learning competitions web site, to maximize dissemination and attract participants.

Figure 1 - Web page of the WCCI 2018 Competition.

The last date for the submission of the participation package by competition participants was June 24, 2018. Preliminary results were provided until July 1, 2018. The validation of the final results and of the submissions were concluded before July 5, 2018.

[1] http://www.ecomp.poli.br/~wcci2018/call-for-competition/
[2] http://www.ecomp.poli.br/~wcci2018/
[3] http://disiem-project.eu/index.php/wcci-2018-competition/
[4] https://www.kaggle.com/c/competition-on-open-source-intelligence-discovery
[5] https://www.kaggle.com/

2.1 Sponsorship

The competition was kindly sponsored by DiSIEM partner Amadeus, who sponsored the following prizes in Amazon gift cards:

- Winning team: € 500;
- 2nd place: € 250;
- 3rd place: € 100;
- 4th and 5th places: € 50.

2.2 Problem statement

A key part in a Twitter-based CTI tool is a binary classifier to distinguish cybersecurity-related (relevant) tweets from irrelevant ones, as extensively discussed and described in deliverable D4.4, in the context of the OSINT Threat Detector (OTD) component. Therefore, this competition consisted in using previously labelled tweet data sets concerning three case-studies, to design binary classification models, one for each of the case studies. The challenge presented to participants was the development of classification models that take tweets as input and produce the corresponding classification for each tweet: -1 (not relevant) or 1 (relevant).

2.3 Data sets

Both for the classifiers design stage (before the deadline for submission of results) and the classifiers evaluation stage (after the deadline for submission of results) three data sets were provided,[6] each corresponding to a different case study (case studies A, B, and C) and corresponding IT infrastructure. The IT infrastructures were specified by three partners of the project: Amadeus, Atos and EDP. Each infrastructure is defined by a set of assets, which in turn are specified by a variable number of keywords. The sets consist of tweets that have been manually labelled as relevant or not to the security of the IT infrastructures represented in the case studies.

For the design stage we call these sets Design Set A, B and C (DSA, DSB and DSC). The tweets in DSA, DSB and DSC have been collected from a set of Twitter accounts designated Account Set 1 (AS1). In the evaluation stage the classifiers proposed by the competition participants were tested using Evaluation Set A, B and C (ESA, ESB and ESC). The tweets in ESA, ESB and ESC have been collected from a set of Twitter accounts designated Account Set 2 (AS2).

ESA, ESB and ESC are such that all the tweets were posted in Twitter after the tweets in DSA, DSB and DSC. AS2 includes all accounts from AS1, but extends it with an additional set of accounts, having in total 223 tweeters. Hence, the evaluation procedure tests the classifiers generalization performance not only in future unseen data, but also considering data from additional Twitter accounts.

[6] http://disiem-project.eu/wp-content/uploads/2018/04/datasets.zip

The data sets are structured in line-based text files with lines containing a hyperlink, an integer number that uniquely corresponds to one tweeter, and one label. The hyperlink references the tweet, the integer number relates to the account that posted the tweet, and the label provides its class. A simple program written in Python was provided together with the datasets, that receives these files and outputs files with the actual tweets text, the corresponding labels and the integer referencing the tweeter account.

Participants were deemed to use only the data sets provided (DSA, DSB and DSC) to train and design their classifier(s). Two options were given:

- Train a single classifier for the three case studies, therefore using the data sets altogether;
- Train one classifier per case study, therefore using the data sets individually. In this case evaluation metrics were computed considering the aggregated results.

Participants were limited to freely available tools/frameworks to train and design their classification models. All the results had to be reproducible by using solely code provided by the participants and the data sets provided by the competition.

2.4 Evaluation metrics

The classifiers were evaluated by metrics reflecting the two main objectives:

- Maximize the amount of relevant information presented to the analyst. This means presenting the highest possible fraction of tweets that were correctly classified as relevant, which corresponds to maximizing the True Positive Rate (TPR) or sensitivity;
- Minimize the amount of irrelevant information presented to the analyst. This means presenting the smallest possible fraction of tweets that were wrongly classified as relevant, which corresponds to maximizing the True Negative Rate (TNR) or specificity.

Denoting the number of relevant tweets correctly classified as relevant (True Positives) by TP, the number of relevant tweets incorrectly classified as irrelevant (False Negatives) by FN, the number of irrelevant tweets correctly classified as irrelevant (True Negatives) by TN, and the number of irrelevant tweets incorrectly classified as relevant (False Positives) by FP. The metrics are given by:

$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{TNR} = \frac{TN}{TN + FP}$$

By using the evaluation data sets, participants were ranked from smallest to highest value in each case study, according to the Euclidean distance of their classifier (TPR, TNR) result to the ideal (1.0,1.0) result. Then a number of points from 1 to the number of participants, (n), was awarded according to the ranking. Finally, the participant team receiving the smallest total number of points considering all the case studies, was appointed winner of the competition.

2.5 Participation

Despite all efforts in publicizing the competition in national and international academic email lists, to the consortium partners relevant contacts, in the project web site, host conference web site, and Kaggle, the competition did not attract the expected participation. We had two manifestations of interest (one from Symantec Argentina), which resulted in one participation.

Team WXYZ from Huazhong University of Science and Technology, Wuhan, China, composed by Xiao Zhang, Zihan Liu, Yuqi Cui and Dongrui Wu, achieved the results presented in Table 1.

Table 1 – Evaluation results obtained by team WXYZ.

	Dataset		
	ESA	**ESB**	**ESC**
TPR:	0,812	0.858	0.839
TNR:	0,934	0.959	0.963

After the participation submission deadline, once the results were provided to participants, these were given access to the evaluation data sets ESA, ESB and ESC, allowing them to verify the results. At the same time, the organizing committee verified that the submissions complied with the competition rules. Then the results were considered final.

3 MLCS 2019 (ECMLPKDD'2019) Competition

By considering the level of participation in the WCCI'2018 competition, we sought the 2019 calendar of good cybersecurity conferences that could host a competition, but the search was not successful. Currently, hosting competitions is not a common practice in major cybersecurity conferences. In alternative, aiming to broaden the scope of the audience, we looked for the major machine learning conferences in the viable period for DiSIEM. We could only find KDD'2019[7] (25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining), which hosts the KDD Cup. By analyzing the venue location and the typical competitions hosted, we concluded that this would not be an option. KDD hosts a very limited number of competitions, sponsored with prizes in the range of tens of thousands of dollars by global companies.

Still aiming at a machine learning conference, we turned our attention to ECMLPKDD 2019 (2019 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases). Although this conference does not have a call for competitions, it has a call for workshops. Our strategy consisted in proposing a DiSIEM supported workshop hosting the intended competition. A workshop proposal was submitted, which was accepted by the ECMLPKDD'2019 organizers: a one-day workshop on Machine Learning for Cybersecurity (MLCS'2019).[8]

The second DiSIEM competition, entitled *Competition on Multi-Task Learning in Natural Language Processing for Cybersecurity Threat Intelligence*, hosted by MLCS'2019, will be organized on September 20, 2019 in Wurzburg, Germany, co-located with ECMLPKDD'2019.

Once MLCS'2019 and the competition were accepted by the ECMLPKDD'2019 organization, a competition web page[9] (Figure 2) was created on the DiSIEM project web page. Considering that the previous competition did not attract attention in Kaggle, we did not create the second competition on that platform. Kaggle most successful competitions are sponsored with significant money prizes by private institutions. Purely academic competitions do not seem to get traction.

[7] https://www.kdd.org/kdd2019/
[8] http://mlcs.lasige.di.fc.ul.pt/
[9] http://disiem-project.eu/index.php/mlcs-2019-competition/

Figure 2 – Web page of the MLCS'2019 competition.

The last date for the submission of the participation package by competition participants was August 24, 2019. Preliminary results were provided until August 31, 2019. The validation of the final results and of the submissions were concluded before September 7, 2019.

3.1 Problem statement

As discussed in deliverable D4.3, concerning the OTD component designed to provide end-to-end OSINT-based cybersecurity threat awareness, to increase efficacy a named entity recognizer is used to locate and identify valuable information items in the tweets that are classified as relevant by the classifier. This competition consisted in using previously annotated tweet data sets concerning the three mentioned case studies, to design multi-task models capable of performing both the classification and Named Entity Recognition (NER) tasks. Therefore, participants were challenged to develop models that take tweets as input and produce the corresponding classification for each tweet: 0 (not relevant) or 1 (relevant). Additionally, for the tweets considered relevant, the models must also output a set of entities (information items) found in the tweets.

The information items found must be classified as one of five possible entities:

- ORG: Organization/Company;
- PRO: Product/Asset;
- VER: Version numbers, likely corresponding to the asset;
- VUL: Vulnerability/Threat
- ID: Any useful identifier (e.g., CVE or NVD identifiers)

Similarly, to the previous competition, participants had different options on how to develop their models. These included:

- train a single classifier for each task, for all three case studies;
- train a multi-task classifier for each infrastructure, meaning a single model to perform both binary classification and NER;
- train one multi-task classifier for all infrastructures and tasks.

3.2 Data sets

As for the previous competition, we collected tweets related to IT infrastructures of the three mentioned partners of the project. Besides the four months of data originally labelled for the first competition, another three months of data were collected and labelled regarding their security relevance. For the additional NER task, all relevant tweets were manually annotated with the existing named entities. To standardize the tweets representation, we implemented a pre-processing stage which included turning every character into lower-case, removing hyperlinks and special characters, except for '.', '-', ' ', and ':', as they are often used in version numbers, component names, and to identify other entities of interest such as vulnerability identifiers.

The data was split into two sets, one for training and the other for evaluation. The data provided[10] displays six fields: pre-processed tweet, three Boolean fields identifying if the tweet contained a keyword from any of the infrastructure case studies (Amadeus, Atos, EDP), binary ground truth values for the classification task, and a sequence of named entities which are the ground truth for the NER task.

Participants were limited to freely available tools/frameworks to train and design their classification models. All the results had to be reproducible by using solely code provided by the participants and the data sets provided by the competition.

3.3 Evaluation metrics

Although the competition involves designing a multi-task model, the evaluation considered each task individually. For the final score, these metrics will be averaged across the three case studies. For the binary classification task, the models will be evaluated by the same metrics used in the first competition, i.e., the TPR and TNR, which will be summarized by the F1 score:

$$\mathrm{F1} = \frac{2 \times TPR \times TNR}{TPR + TNR}$$

For the NER task, the evaluation will also be carried out by the F1 score:

$$\mathrm{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

[10] http://disiem-project.eu/wp-content/uploads/2019/05/ecml_dataset.zip

where, for each entity, *Precision* is the percentage of correct classifications over all positive predictions, and *Recall* is the percentage of correct classifications over the total number of entries present for said entity.

The participants were ranked on two scoreboards, one for each task. Each participant received the number of points given by the sum of the rank in both scoreboards. The participant achieving less points wins the competition. In case of a draw, the winner is the one having the smallest Euclidean distance of their F1 scores to the ideal (1.0, 1.0) point.

3.4 Participation

Unfortunately, despite the different strategy and more complex problem when compared with the previous competition, the second edition had also a weak participation, with only one participant sending a submission.

The participant was Félix Iglesias from the Technical University of Wien, Austria. Table 2 presents the average results across the three case studies infrastructures, obtained in the two tasks.

Table 2 - Multi-task results averaged across the three infrastructures.

	Binary classification task	**NER task**
TPR / Precision	0,831	0,773
TNR / Recall	0,988	0,754
F1 score	0,902	0,763

As with the WCCI'2018 competition, once the results were provided to the participant, he was given access to the evaluation data sets, allowing verification of results.

4 Reflection on Low Participation

Although the DiSIEM competitions reached a very large audience via the hosting conferences and other publicizing means, both achieved minimal participation. While making a reflection on the reasons for this, we hypothesized the following possibilities:

- In the machine learning field, there are currently many competitions available both in Kaggle and in top conferences, all with significant money prizes. Purely academic competitions with no money prize have little chance to attract satisfactory participation.
- Academic cybersecurity conferences do not host competitions; therefore, this community becomes more difficult to target and attract to other fields conferences.
- Cybersecurity researchers and practitioners do not seem to have a culture of participating in competitions (other than classical capture the flag, which are very different from the ones we are proposing in DiSIEM) and many might not be proficient enough in machine learning techniques to feel inclined to participate in competitions like the ones we are organized. The low participation in a cybersecurity workshop indicates this.
- The DiSIEM-related tasks that motivated the competitions comprise text processing from social media and other OSINT sources, which are more related with Natural Language Processing techniques than cybersecurity or machine learning at large. However, the experts in this topic might not feel compelled to address a very specific cybersecurity problem.
- Arguably, deep learning and neural networks might be the best technique to address the proposed competition challenges, but these research communities seem to be more focused and attracted by machine vision problems in very active fields, like the automotive.

All in all, it is not easy to see how one could improve the participation on competitions like the ones we proposed. Having a big prize might help, but maybe the main issue is related with the lack of tradition of having competitions in the cybersecurity community and/or the different interests from machine learning researchers (e.g., machine vision).

5 Summary and Conclusions

This document presents the activities related to the organization of two cybersecurity-related machine learning competitions in the scope of the DiSIEM H2020 project. The first competition, on Twitter based cyber threat detection, was organized in a neural network focused conference, the IJCNN'2018 (within WCCI'2018). The second competition, on named-entity recognition in cybersecurity-related tweets, was organized in a cybersecurity-related workshop, MLCS'2019, organized in co-location with a machine learning conference, the ECMLPKDD'2019. Although these competitions were well structured, with clear challenges, large social media text data sets, and objective evaluation procedures, they attracted only minimal participation. Some possible reasons for this are discussed in the document. Still, the organization of the two competitions disseminated the DiSIEM project to very large audiences, through the conference web sites, the project web site, and national and international scientific email lists.

References

[1] Center for Strategic and International Studies (CSIS) and McAfee. "Economic Impact of Cybercrime — No Slowing Down Report".
https://www.csis.org/analysis/economic-impact-cybercrime
(accessed: Sep. 2019).

[2] IBM, "IBM Study Shows Data Breach Costs on the Rise; Financial Impact Felt for Years,"
https://newsroom.ibm.com/2019-07-23-IBM-Study-Shows-Data-Breach-Costs-on-the-Rise-FinancialImpact-Felt-for-Years
(accessed: Sep. 2019).

[3] ENISA, "Exploring the opportunities and limitations of current Threat Intelligence Platforms,"
https://www.enisa.europa.eu/publications/exploring-the-opportunities-and-limitations-of-current-threatintelligence-platforms
(accessed: Sep. 2019).

[4] C. Sabottke, O. Suciu, and T. Dumitras, "Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits," in Proc. of the 24th USENIX Security Symposium (USENIX Security 15). USENIX Association, 2015.

[5] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence," in Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS). 2016.

[6] N. Dionísio, F. Alves, P. M. Ferreira, and A. Bessani, "Cyberthreat Detection from Twitter using Deep Neural Networks," Proc. of the 2019 International Joint Conference on Neural Networks (IJCNN). July 2019.

[7] Q. Le Sceller, E. B. Karbab, M. Debbabi, and F. Iqbal, "SONAR: Automatic Detection of Cyber Security Events over the Twitter Stream," in Proc. of the 12th International Conference on Availability, Reliability and Security (ARES). Association for Computing Machinery, 2017.

[8] F. Alves, A. Bettini, P. M. Ferreira, and A. Bessani, "Processing Tweets for Cybersecurity Threat Awareness," arXiv e-prints, arXiv:1904.02072, April 2019.

[9] S. Zhou, Z. Long, L. Tan, and H. Guo, "Automatic identification of indicators of compromise using neural-based sequence labelling," 2018.

[10] A. Attarwala, S. Dimitrov, and A. Obeidi, "How efficient is Twitter: Predicting 2012 U.S. presidential elections using Support Vector Machine via Twitter and comparing against Iowa Electronic Markets," Intelligent Systems Conference, 2017.