MLCS 2019 – Workshop on Machine Learning for CyberSecurity

Competition on Multi-Task Learning in Natural Language Processing for Cybersecurity Threat Awareness

(MLCS 2019 is co-located with ECMLPKDD 2019)

Introduction

This competition is organized within the scope of the European Union H2020 project DiSIEM – Diversity Enhancements for Security Information and Event Management.

Please note:
Participation in the competition does not imply participation on the ECML PKDD 2019 conference. However, we strongly encourage participants to submit a paper to MLCS 2019 about their approach and early results.

About the DiSIEM project

The DiSIEM project objective is to improve Security Information and Event Management (SIEM) systems capabilities using diversity-related technolc gy. One e of

against the infrastructure monitored by the SIEM, hence **improving organizations cybersecurity threat awareness capabilities**.

Cybersecurity Threat Awareness

The main goal of cybersecurity threat awareness 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.

As an example, in DiSIEM we are continuously collecting tweets concerning the security of three case-study IT infrastructures specified by three industrial partners of the project. To fulfil the objectives above, two important tasks have to be carried out: classify tweets as relevant or not; detect security-related content in relevant tweets. A tweet is relevant if it mentions a threat to an element of an IT infrastructure (e.g., a vulnerability or an exploit) or a security measure to protect that element (e.g., an update or a software patch).
The detection of security-related content can be done using Named Entity Recognition (NER) techniques, where each word present in a tweet is tagged with a named entity. Due to the experimental nature of this work we defined five entities that, although general, are able to summarize the important information necessary to issue a security alert or to document an Indicator of Compromise (IoC).

Why Twitter?

Although there are many sources of OSINT, including security-related ones, Twitter was used for two main reasons. First, Twitter is well-recognized as an important information hub for short notices (almost in real-time) 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.

More importantly, our work also demonstrates that it is possible to obtain valuable security-related information from Twitter before it becomes available on established

were tweeted 9 (see publication) and 6 (see publication) days ahead of official confirmation on NVD.

Problem Statement

A key part in any software component designed to provide end-to-end OSINT-based cybersecurity threat awareness is a binary classifier that takes as input one piece of information, e.g., a tweet, and assigns it to one of the classes, relevant (1) or irrelevant (0).
In order to further maximize the usefulness of information, a named entity recognizer locates and identifies valuable chunks of information in the tweet. These chunks are classified as one of five possible entities:

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

This competition consists of using previously labelled tweet data sets concerning three case-studies, to design multi-task models capable of performing both the classification and named entity recognition tasks.
Therefore participants will develop models that take tweets as input and produce the corresponding classification for each tweet: 0 (not relevant) or 1 (relevant). Furthermore, for the tweets considered relevant, the models must also output a sequence of *n* entities, where *n* is the number of words in the tweet.

Datasets

Zip file with the training dataset.

The ZIP file contains a comma-separated value (CSV) with the training dataset for model design. Each line contains 6 columns:

1. ***clean_tweet***: contains the pre-processed tweet. We removed all special characters, except a few that are useful to identify version numbers and

(relevant) values.

3. ***entities***: holds a sequence of tags, with the same length as the *clean_tweet*, using the Inside-Outside-Beginning (IOB) scheme. If the *class* is 0, this column will display a *null* value.
4. ***A***: boolean identifying tweet as related to case study A infrastructure.
5. ***B***: boolean identifying tweet as related to case study B infrastructure.
6. ***C***: boolean identifying tweet as related to case study C infrastructure.

At least, one of the A, B, or C columns should display 1. There can be tweets related to more than one infrastructure.

In the evaluation stage, the models proposed by the competition participants will be tested by the organisers using a testing set. This set will be made public when releasing the competition results.

Evaluation Metrics

Although the competition involves designing a multi-task model, the evaluation will consider 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 metrics reflecting the two objectives stated above:

- *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.

These metrics will be summarized by the F1 score:

$F1 = 2 \times \frac{TPR \times TNR}{TPR + TNR}$, where TPR is the True Positive Rate and TNR is the True Negative Rate.

Precision+Recall

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

Ranking the results

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

Competition Rules

The participants must comply with the following rules:

1. Participants will use the dataset provided to train and design their multi-task models. Two options are available:
    - train a single classifier for the three case studies;
    - train one classifier per case study.
2. Pre-trained language models and other resources are allowed as long as they are open-sourced and publicly available online.
3. Participants must use only publicly available tools/frameworks to train and design their models.
4. All the results have to be reproducible by using solely the code provided by the participants.

Once the results are published on this web page, the participants will receive the evaluation dataset, allowing them to verify the results published. At the same time, the competition organizers will also verify that the participants' submissions comply with the competition rules. Only after these verifications will the results be considered final.

For planning purposes, we kindly ask all participants to send an email to Nuno Dionísio and Pedro M. Ferreira (please, see addresses below) mentioning their willingness to participate.

Before the submission deadline, participants will deliver an archive (e.g., a ZIP file) or share a code repository (e.g., Github repository) with the following components:

- A short report naming the team, the authors and their affiliation, and providing a contact person and corresponding email address, a description of the software tools/platforms employed to train and design the model(s), a description of the methodologies employed, and instructions on how to reproduce the design of the model(s) submitted;
- A computer program in an interpreted language (e.g., Java, Python, R, ...) for each model, that is able to take an input text file with one tweet per line and produce an output text file with the corresponding predictions for each tweet.
- A source code package that is able to take the training data set as input and reproduce the design and training of the models submitted. Instructions on how to execute the code will be given in the report mentioned above. Authorship of the code, procedures and methodology remains with the participants.

Important dates

The date for the submission of the participation archive is **August 24, 2019**.

The preliminary results will be published on **August 31, 2019**.

The validations of the results and of the submissions have to be concluded before **September 7, 2019**.

Organizers

Pedro M. Ferreira, Alysson Bessani, Nuno Dionísio
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Email addresses:
PMF: pmf (at) ciencias (dot) ulisboa (dot) pt
ND: ndionisio (at) lasige (dot) di (dot) fc (dot) ul (dot) pt

PROJECT INFORMATION

Project number: 700692
Start date: 2016-09-01
End date: 2019-08-31
Project cost: € 4.020.018
Project funding: € 3.445.875
Programme type: Horizon 2020 Innovation Action
Programme acronym: DS-04-2015

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