Digital tools for Law Enforcement – issue of training and testing data

On 28-29 March 2019, a workshop on digital tools and artificial intelligence (AI), aimed at supporting Law Enforcement Agencies (LEAs) in fighting crime, including cybercrime and terrorism, took place in Brussels, within the scope of the Community of Users' event. The core of the workshop participants were practitioners from LEAs as well as developers/researchers from 18 on-going projects of the security research strand of Horizon 2020 and from 4 on-going ISF-Police projects.

During the workshop, the issue of a lack of realistic, up-to-date and numerous enough training and testing data was tackled, amongst other. This issue has been regularly raised by the projects. Namely, the accuracy of tools, notably digital/AI ones, depends heavily on the quantity and on the quality of the training and testing data, including the quality of their structure and labelling, and how well the training and testing data represent the problem to be tackled.

This issue gets more emphasized in the security domain due to the sensitivity of the data, which complicates the access to real datasets or the creation of representative datasets at a national level.

In EU-funded projects, in the area of the fight against crime and terrorism, the problem of having a huge amount of up-to-date high-quality data needed to develop reliable digital/AI tools in support of LEAs becomes even more complex. Namely, training and testing data sets considered legal and used in one Member State have to be shared and accepted in other Member States, while simultaneously observing fundamental rights and substantial or procedural safeguards. The lack of legislation at the national and international level makes this particularly difficult.

After the workshop, we sent the following questionnaire to these 22 projects:

*1. **Is your project concerned by the issue** of training and/or testing data when developing digital tools to be used by LEAs?*
If yes:
*2. Which **types of data** (e.g., voice, images, videos, investigation reports,…) and/or digital evidence (e.g., emails, phone numbers,…) did/do you need for the project?*
*3. Which **main obstacles** did you encounter for training and/or testing your tools?*
*4. How did you **overcome these obstacles** (e.g., using fake data set, applying data anonymization techniques, working through LEAs, retrieving existing data sets,…)?*
*5. Which **recommendations** would you have for future (e.g., building specific common data sets, adjusting the legal framework,…)?*

We obtained the answers from all projects, which are summarized in the following.

1. Concerned by this issue?
All projects are concerned by this issue.

2. Data needed
Almost all types of data are needed and employed inside projects, most often cited were:

- **Videos** (including illegal)
- **Audio**
- **Images** – original or altered (including illegal)
- **Text** (multilingual) – articles, police reports
- **web** URL/contents, **social network** messages
- **Biometric** data
- **Financial transactions**, including virtual currency (e.g., bitcoin addresses involved in malicious activities)
- **Intelligence** exports from existing LEA systems (e.g., the Police National Database)
- + **Annotations** (ground-truth information for all these types of data)

3. Main obstacles
There have been several obstacles; the situation varies from Member State to Member State, and from LEA to LEA – most common ones are:

- **Difficulties to access the necessary data**
  - **From Law Enforcement**
    - **Lack of representative data**: Datasets are not made available because they are confidential.
    - Lack of representative **structured** datasets: even the structure (not just the content) is confidential.
    - **Lack of enough amount of data**.
    - The **limited nature of the data** that can be used to train or test the tools limits the machine learning processes limiting the full potential of the tools to develop.
    - Even when legally feasible, Data Protection Officers of LEAs are **afraid to authorize** the release of datasets, for privacy/data protection reasons, as research and innovation related testing and training of the tools are considered as falling outside lawful policing purposes under GDPR and Law Enforcement Directive.
    - Fictitious datasets available to LEAs (which are used for training) often do not belong to the LEAs but to training consulting companies (IPR issue).
  - **From open sources**
    - **Open source datasets** are not always representative,
    - Obtaining and retaining social media data whilst remaining compliant with GDPR and provider's policies.
- **Limits in using alternative datasets and enriching available ones**
  - **Anonymized** datasets (i.e. not just pseudonymized) are not helpful, as they break the links between different pieces of evidence.
  - **Pseudonymization** takes a lot of time and effort for LEAs.
  - **Generation of ground truth and labelled data** for comparison and benchmarking is complex and time consuming.

4. Mitigation measures
Case-by-case, we have used all of these mitigation measures:

- **Difficulties to access the necessary data**
  - **From Law Enforcement**

        - **Working through LEAs**: In general, a lot of negotiation and pressure is needed. Having some police academies helps, as they have datasets that they can share, or they can help to retrieve datasets from LEAs.
        - **On-premises solutions with LEAs**: however, not having cloud solutions duplicates efforts and costs.
        - Using approaches in which industry and research provide an algorithm developed on some data to **LEAs** to **evaluate on real data**.
        - Some LEAs found helpful datasets that are fictitious and complex enough (used for training). Some LEAs provided **typical structure and fictitious examples** of the datasets; technical partners can populate them further following these examples, and maintaining the same structure.
    - **From open sources**
        - There are open datasets that can be used. This is also typical for videos and images, where datasets exist for training and validation of some functions (e.g. object recognition). However, when building a data set from open sources there are three key issues:
            - Cost and time
            - Personal data and privacy risks
            - Infringement of terms of services

- **Avoiding duplication of efforts: Collaboration with other R&D projects** (e.g. exchange of best practices concerning the data management).

5. Recommendations for future

- **A common methodology for sharing and creating data sets in a lawful manner.** Currently there are many resources spent, and much effort in every project is devoted to this, whereas there should be, if not a common solution, at least clear guidelines.
- A **common approach in which industry and research provide an algorithm to LEAs** so that they can evaluate it on **real data** and give a feedback.
- A common approach for applying techniques that allow **on-premise training and classification**, e.g. federated learning.
- **Allowing access to some existing datasets** as well as **building specific representative** (in terms of quality and quantity) **common datasets** potentially involving:
    - a **controlled data lake in a sandbox**,
    - a **hosting platform** where any LEA could collaborate through developing and testing process.
    - pseudo-anonymized or anonymized data, depending on the necessity for the dataset.
    - publicly open dataset repositories can provide examples of practices.
- A **structured approach to semi-automated annotation of training data**. This will be vital as LEAs choose to adopt AI technologies, as they simply cannot add more people to train data manually.
- **An assessment and potentially adjustments of the legal framework** that allow for some means of sharing of data between LEAs and research. However, the challenge is not only legal as a **mindset** needs to change also.
- An appropriate lawful mechanism (or legal certainty) which would allow:
    - Access and usage of appropriate and relevant **illegal content within a controlled environment**.
    - Access and usage of relevant **illegal content data sources** (i.e. known dissemination points in Dark Web etc.).
    - LEAs have opportunity to utilise some of the **research exemptions** that academia and research institutes have to undertake such work.