Experiencing the Detection of Radicalized Criminals on Facebook Social Network and Data-related Issues

Andrea Tundis[1,*], Leon Böck[1], Victoria Stanilescu[2]
and Max Mühlhäuser[1]

[1]*Department of Computer Science, Technische Universität Darmstadt (TUDA), Darmstadt, Germany*
[2]*Siemens AG, Marburg, Germany*
E-mail: tundis@tk.tu-darmstadt.de; boeck@tk.tu-darmstadt.de; victoria.stanilescu@siemens.com; max@tk.tu-darmstadt.de
[*]*Corresponding Author*

Received 20 November 2019; Accepted 06 January 2020;
Publication 28 January 2020

Abstract

Online social networks (OSNs) represent powerful digital tools to communicate and quickly disseminate information in a unofficial way. As they are freely accessible and easy to use, criminals abuse of them for achieving their purposes, for example, by spreading propaganda and radicalising people. Unfortunately, due to their vast usage, it is not always trivial to identify criminals using them unlawfully. Machine learning techniques have shown benefits in problem solving belonging to different application domains, when, due to the huge dimension in terms of data and variables to consider, it is not feasible their manual assessment. However, since the OSNs domain is relatively young, a variety of issues related to data availability makes it difficult to apply and immediately benefit from such techniques, in supporting the detection of criminals on OSNs. In this perspective, this paper wants to share the experience conducted in using a public dataset containing information related to criminals in order to both (i) extract specific features and to build a model for the detection of terrorists on Facebook social network, and (ii) to highlight the current limits. The research methodology as well as the gathered

Journal of Cyber Security and Mobility, Vol. 9_2, 203–236.
doi: 10.13052/jcsm2245-1439.922

results are fully presented and then the data-related issues, emerged from this experience, are discussed $(^{*})$.

Keywords: Internet-based crimes, criminal detection, cyber-criminality, online social network analysis, radicalization, terrorist networks, machine learning.

1 Introduction

The process by which an individual or group comes to adopt increasingly extreme political, social, or religious ideals and aspirations that reject or undermine the contemporary ideas and expressions of the nation is called radicalization. It can be both violent and nonviolent, although most academic literature focuses on radicalization into violent extremism [35, 36]. Europe, and in particular the CEE region, is experiencing, for example, a rise in polarizing political and social movements characterized by Euroscepticism, chauvinism and xenophobia, radicalization and incidents of violent extremism. Far-right movements have gained momentum in Poland and Romania paralleled by increasing community tensions and radicalization of at-risk-groups. Germany's recent elections saw AfD achieve a vote-share not seen by a far-right party since the 1930s [51] reflecting a broader rise in populism and polarization also elsewhere in Europe [52]. There are multiple pathways that constitute the process of radicalization, which can be independent but are usually mutually reinforcing.

Thanks to the rapid growth of the information technology (IT) such phenomena are simplified and, especially, the use of OSNs in everyday life, makes the communication and interaction among people living in different geographical locations easier. Indeed, not only is it possible to get in touch with known people, but also to meet new users and establish connections, based on common interests and ideologies, that go from a simple knowledge and friendship to even business relationships [12].

Unfortunately, not only are the social networks used from inoffensive members of the society but also for supporting the organization and execution of illegal activities. There exist groups of users, who commit criminal acts against the law in order to achieve their personal goals using Internet-based technologies [56–58]. These groups learned to use the achievements of the technological progress for their own purposes by coordinating their actions as well as carrying out illegal operations online [30]. Organized Crime (OC)

$(^{*})$ This work is an extension of the paper [59].

and Terrorist Networks (TNs) represent the main criminal groups that benefit of IT advancements [1, 29]. Specifically, OC is mostly oriented to obtain economic gain on a large scale through smuggling goods, money laundering, credit card fraud, or online sex fraud. Whereas, TNs are usually motivated either politically or religiously as well as ideologically by spreading radical beliefs, false news and propaganda as well as by brainwashing people and recruiting new one.

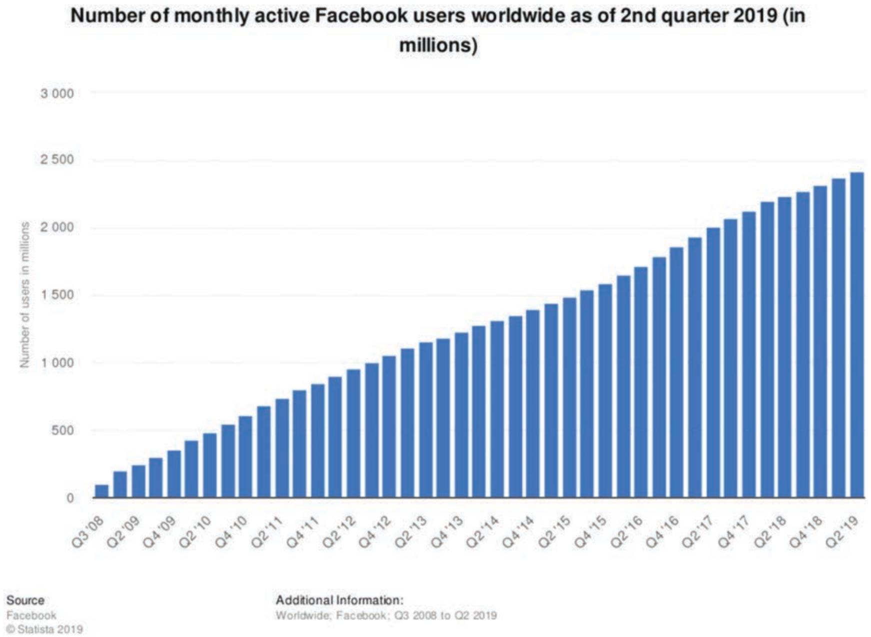
Facebook, as well as Twitter, for example, play the major role in this context because of their popularity [27, 28]. Millions of users publish daily billions of posts through them, as they are freely accessible and provide low publication barriers for both posting and viewing information. Because of the size and variety of the network, it is not trivial to identify online criminals who abuse of such digital tools for their purposes. Indeed, the identification of criminal's users is not a trivial task due to the huge amount of information and, in principle, because of the lack of proper models and tools [8].

In this context, the paper discusses the experience of using machine learning algorithms for supporting the detection of criminals on social networks, by highlighting the current limitations in defining detection models due to the lack of adequate data. Specifically, a feature-based model is first defined, on the basis of a public dataset related to terrorists called *John Jay & ARTIS Transnational Terrorism Database* [23], and then experimented on a Decision Tree, a Support Vector Machine, and Artificial Neural Networks [31–33]. The gathered results are presented as well as the main criticalities and current limits, that emerged, are highlighted.

The rest of the paper is structured as in the following: the growing popularity of Facebook over the year and its role in the society is discussed in Section 2; whereas in Section 3, the related works related to crime detection approaches on social networks as well as research projects are reported; the background on the exploited machine learning techniques is provided in Section 4. The proposed methodology and the defined detection model is described in Section 5; whereas results evaluation and data-related issues are discussed respectively in Section 6 and Section 7. Section 8 summarizes and concludes the paper.

2 The Role of Facebook in the Society

A media, in the classical sense, is a communication means that allows people to spread and disseminate news on a large scale, typically, according to a one-to-one or one-to-many model. With the introduction of the Web, there has been an evolution of this concept, giving rise to "Social media". It represents



Source
Facebook
© Statista 2019

Additional Information:
Worldwide; Facebook; Q3 2008 to Q2 2019

Figure 1 Popularity of Facebook in modern society [41].

a change in the way the people discover, read and share news, information and content, by allowing a more interactive communication, based on a many-to-many model [40]. A further evolution is represented by OSNs, which not only allow to generate and distribute information through electronic communication but they try to emulate the social life of people through the creation and maintenance of personal and business relationships online.

Facebook, for example, is one of the most popular social network and microblogging website. With 2.41 billion monthly active users as of the second quarter of 2019 (see Figure 1), it is the biggest social network worldwide, where 76% of users spend in average 35 minutes to check it [41].

The main advantages of Facebook rely on being free and providing low publication barriers for both posting and viewing textual information, popularly known as “post”. Furthermore, additional multimedia data such as images and videos can be easily integrated to a post by directly uploading them or through external entry via URL, when a more elaborated and complex information has to be widespread.

The concept of relationship among users is realized through the “friendship” functionality, which enables people to visualize each other user profile, activities and published or shared posts. Whereas, the interaction among users is based on the sharing each other posts on their profiles, as well as by commenting somebody's post or simply expressing a sense of agreement or disagreement through predefined images, called "emoji", in regard to a specific post and its associated content.

Aside these benefits, Facebook provides an easy communication tool for many criminals or terrorists, who can use the potential of Internet to get in touch with new people around the world quickly, to create communities and discussion groups, to spread propaganda and so on. Indeed, as it is reported on the [42], 837 million pieces of spam and 2.5 million pieces of hate speech and disabled 583 million fake accounts globally in the first quarter of 2018 were taken down, due to broke of policies on coordinated harm and inauthentic behaviour, as well as attacks based on race, gender or sexual orientation.

Due to these reasons Facebook has been considered as reference OSN in order to explain and share the conducted authors' experience in detecting radicalized criminals on social networks.

3 Related Work

The investigation and detection of criminals is an activity that has always existed. Thanks to the advance of IT technology, even the criminal world has moved towards the cyber domain, which makes its monitoring and control more complicated for the Police Forces (PFs). As a consequence, the interest in supporting the discovering process of Internet-based crimes and identify online-criminals is proved by an increasing varieties of research efforts towards this direction. In particular, in Section 3.1 model and methods for supporting crime detection on social networks are discussed, whereas ongoing and past research projects are described in Section 3.2.

3.1 Crime Detection Approaches on Social Networks

A major approach to study social networks is called Social Network Analysis (SNA) [43]. It is a technique for modelling the communication patterns among individuals in a way that illuminates the structure of the network and the importance of individuals within it, in order to understand the human behaviour and its evolution.

Particular attention is devoted to Facebook [44]. Some research efforts have been conducted by employing SNA, such as in [45] where, starting from the monitoring of user's activities of terrorist groups on Facebook, an algorithm, based on a centrality approach, has been proposed to discover the most active users in the group, who were able to recruit the highest numbers of new users. In [46] instead, a system for crawling and analysing organized crime communities on Facebook has been proposed. Specifically, different heuristic algorithms have been implemented in order to extract specific properties (i.e. video, audio and text) of Facebook's social graph, and particular user interactions, in order to highlight specific users' relationships. Whereas, a forensic algorithm on Facebook, by using Natural Language Processing (NLP), has been proposed in [47]. It aimed to support and be beneficial about investigation for legal proceeding as well as to facilitate the police or people who take a part in the operation on law for the identification of users, who are related to computer crimes.

Further models and algorithms have been proposed with the aim to enhance the identification of unseen connections between users, based on other common characteristics [20]. In [19], a model for the collection of data coming from social media and, in particular, regarding jihadists is defined by proposing best practices based on the concept of followers, friends, retweets and mentions. Whereas in [14], a system for the detection of online jihadist has been proposed by combining Natural Language Processing and Machine Learning techniques. A further research contribution is presented in [16], where a machine learning based model, for the identification of online users, who disseminate propaganda via Twitter, is described. In [13], a system for supporting the surveillance and monitoring of cyber-trafficking, whose communication is centred on social media, is proposed. Furthermore, some case studies on social networks, where the criminals exploit Twitter as means to communicate among each other in order to organize their criminal activities, are presented in [22, 53–55].

Further models to discover criminal networks are described in [24]. Some of them (such as GDM, OGDM, SoDM, ComDM and XSDM), work with different data, like criminals' surname and home-town similarity, crime location (GDM and OGDM), or social-cultural data (SoDM). For example, the ComDM (Combined Group Detection Model) is a model, based on OGDM and SoDM, that was developed in order to benefit from maximum similarities of criminal behaviour (e.g. choice of crime location, time and modus operandi) between criminals and use of demographic similarities, for example family relations and related one, such as same home-town circumstances [25, 26]. XSDM (Extended Social Detection Model), in turn, overcome some weaknesses of its predecessor, SODM (Social Detection

Model), by considering also other attributes related to living in the same neighbourhood.

Other research studies are centred on the exploitation of artificial neural networks (ANNs) in order to enhance the public surveillance, as described in [21]. Whereas, in more recent research activities, ANNs are adopted for supporting the crime scene prediction by detecting threatening objects [18]. The crime investigation and analysis with particular attention on homicide cases is discussed in [17], whereas in [15], ANNs are used to model legal system analogous to recent decisions on the basis of past court decisions in case of murders.

3.2 Research Projects

The interest in supporting the detection of criminals in the cyber-space is also proved by the increasing number of research projects that focus on different topics in this field.

CHAMPIONs, for example, is an ongoing research project [48]. Its central action is to establish permanent offline working groups combining FLPs (first-line-practitioners) of different disciplines, professions and institutions or agencies, to jointly develop effective detection and response solutions to counter radicalized criminals, build resilience and protect vulnerable groups in their local communities. The ARMOUR project instead aims to address societal polarization caused by the adoption and spread of extremist ideologies by creating an interdisciplinary model of learning [49]. Another recent project, called TAKEDOWN, that was focused on Organized Crime and Terrorism, aimed at providing resources, tools and services, for First Line Practitioners (FLPs) and public, in order to prevent and counter extremism and crime [37, 38]. CAPER (Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organised crime), was another Security Research Project created in cooperation with European Law Enforcement Agencies (LEAs), that aimed at building a common collaborative and information sharing platform for the detection and prevention of organized crime, through the exploitation of Open Source Intelligence (OSINT) [50].

Another previous project in the field of criminal networks was COPLINK, a system for supporting collaboration, information and data sharing about crimes among LEAs. Based on that, Xu [4, 10] elaborated the so called CrimeNet Explorer [9, 11], by performing entities extraction from police narrative reports, and then using them in order to build links between entities among documents and hidden criminal networks. A clustering algorithm based on mutual proximity, to identify links between offenders, and to detect previously unknown groups was used. Another project, called FLINTS [6], used soft behavioural and hard forensic (e.g. fingerprints, DNA) to enable analysts to create graphical representations of the relationships among crimes and criminals. Whereas, in the context of Internet-based crime, the FinCEN project [5] aimed at identifying money laundering networks by comparing financial transactions.

In addition, as discussed in [6], an analysis of cases of burglaries was documented in the context the OVER project, whereas in [7], a similar work on detention of the clusters to filter the surplus of information on possible terrorist networks was carried on. It aimed to present the police a viable and limited subset of suspects to work on. Other research efforts focused, instead, on the behaviour modelling of sexual offenders, based on clustering methods, as described for example in [2, 3].

The above mentioned models considered, for the discovering of criminal networks, patterns of threatening activity, co-offending feature (e.g., who committed crime with when), crime features (crime locations, crime dates, modus operandi), social features (criminal's surname, home-town information, same neighbourhood). However, other information which might be retrieved from public social network and that could be used for criminal's investigation are neglected as well as the ratio and the way of selecting specific features is not clear. In this direction, this research work wants to contribute by sharing the authors' experience gathered during the identification and exploitation process of additional users' features which can be used during the investigation and detection of criminals on social network by highlighting the occurred issues.

4 Background

In this Section, an overview on the machine learning techniques, which are considered in this work, are presented. Machine learning (ML) is a data analysis approach, based on computational algorithms that are able to learn directly from the data without a predefined model.

In principle, the more examples are available the more precise is the algorithm defined, which can improve its performances by adapting it. Through ML patterns from which to extract information that is used to make better forecasts, prediction and decisions are identified. Thanks to the huge amount of available data, it represents one of the most popular data analysis based approaches. It is adopted in several fields by facing with aspects regarding (i) computational finance for the evaluation of credit risk and algorithmic trading; (ii) image processing and artificial vision for facial recognition, motion detection and object identification; (iii) computational biology for the diagnosis of tumours, pharmaceutical research and DNA sequencing; (iv) energy production for price and load forecasts; (v) automotive, aerospace and manufacturing sectors, for predictive maintenance; (vi) natural language processing for speech recognition applications and so on. Among the different machine learning techniques available in literature and their variants, some of the most popular one that are considered in this paper are: Decision Trees (DTs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), which are below introduced.

In particular, DTs are tools based on a tree-model [32]. A DT is navigated from the root to the leaves, each intermediate node represents a decision point and the ramification represents the property that leads to a particular decision. The predicate that is associated with each internal node, which is used to discriminate among the data, is called split condition. When a leaf is reached by navigating the tree, not only a particular classification is associated to the input instance, but thanks to the path, it is possible to understand the reason of a particular result.

SVMs are linear models for classification and regression problems which are used to solve linear and non-linear problems [31, 39]. The idea of a SVM is based on the definition of a line or a hyperplane which separates the data into classes. So based on given labelled inputs, the algorithm produces in output hyperplane-based model which is able to classify new instances. Given a set of training examples (training set), each of which is labelled with the class to which the two classes belong, an SVM training algorithm constructs a model that assigns new examples to one of the two classes, thus obtaining a non-probabilistic binary linear classifier. A SVM model is a representation of the examples as points in space, mapped in such a way that the examples belonging to the two different categories are clearly separated by the widest possible space. The new examples are then mapped in the same space and the prediction of the category to which they belong is made on the basis of the side in which it falls. In addition to linear classification, it is possible to use SVM to effectively perform non-linear classification using the so called kernel method, by implicitly mapping their inputs into a multi-dimensional feature space.

ANNs are computational models, which are, instead, able to represent knowledge based on massive parallel processing and pattern recognition based on past experience or examples. ANNs are inspired by biological networks in which: (i) the information processing occurs at several simple elements that are called neurons; (ii) signals are passed between neurons over connection links; (iii) each connection link has an associated weight, which, in a typical neural net, multiplies the signal transmitted; (iv) each neuron applies an activation function (usually nonlinear) to its net input (sum of weighted input signals) to determine its output signal. A ANN is defined through an initial layer on the basis of the available inputs, a final layer which represents the output of the computation and a hidden layer which is defined in terms of potential multi-layers through which the inputs undergo various transformations and calculations steps as long as the final layer is reached and the output is generated [10]. Thanks to the flowing of the information through such computational steps the structure of a ANN can change by adapting it as well as by learning, in order to identify relationships between inputs and output and the ability of pattern makes ANNs a good alternative classification and forecasting tool in several application domains such as Forecasting/Business, Image Processing and Character recognition [33].

The above mentioned techniques represent different way to approach a classification problem. In the next Section, the proposed model is presented and then its assessment is conducted by using it to train different classifiers in order to evaluate and compare its performances. Additionally, a customized ANN based on multiple input layers is also proposed and compared respect to the above mentioned techniques.

5 An Analysis Method by Using Machine Learning

In this Section, the adopted research method, centred on machine learning, is presented. In particular, an overview of the exploited dataset is firstly provided, secondly the concepts and the workflow are illustrated, then the identified features are described, and finally the implementation details are shown.

5.1 Dataset Description

Before illustrating the adopted methodology, the reference database, which is called *John Jay & ARTIS Transnational Terrorism Database* and that is publicly available, is introduced [23]. It contains a dataset of 2157 entries consisting of 52 attributes which are used to describe individuals. Different characteristics related to the demographic, organizational, and behavioural of each person are collected. For example:

- *demographic data* contains socio-economic and country-specific information, which might be used for evaluating theories concerning what motivates political violence (for example, poverty might represent a catalyst factor for specific actions or people reactions), as well as for determining the relative representation of countries in the database;
- *organizational data* describes the position that individuals ultimately attained within their respective groups. Although such information is preliminary, it might be useful for developing network maps describing organizational structure in order to understanding to what extent groups structurally differ from each other.
- *behavioural data*, related to terrorist events in which individuals take actively part, such as level of participation, played role and so on.

Personal information is neither available nor derivable from them. Whereas, attributes regarding characteristics such as, age, gender, as well as of cultural nature such as city of birth, religion, and others related to the education level, such school level and type of education are available. More detailed information about the dataset description is available online [23].

5.2 Concepts and Workflow

The overall research method, which is organized in lanes named *Concept* and *Workflow*, is illustrated in Figure 2. In particular, the *Concept* lane aims to describe the conceptual process that has been conceived to conduct such experience by identifying 4 main logical parts: *Reference Domain, Expertise and Based knowledge, Observed Source*, and *Analysis Technique*.

More specifically, the *Reference Domain* part concerns the concepts related to the interpretation, extraction and formalization of data that can be used for the generation and derivation of additional information, which is not a-priori evident. A reference domain can be provided in a textual or verbal (e.g. structure, unstructured or semi-structured) format, in mathematical formalism (through formulas) or in any other representation.

The *Expertise and Base knowledge* part concerns the actions related to the identification, distillation and extraction of hypothetical data that are considered relevant for the purpose of the desired analysis. The choice of what to use can be motivated and justified on the basis of previous experiences, that

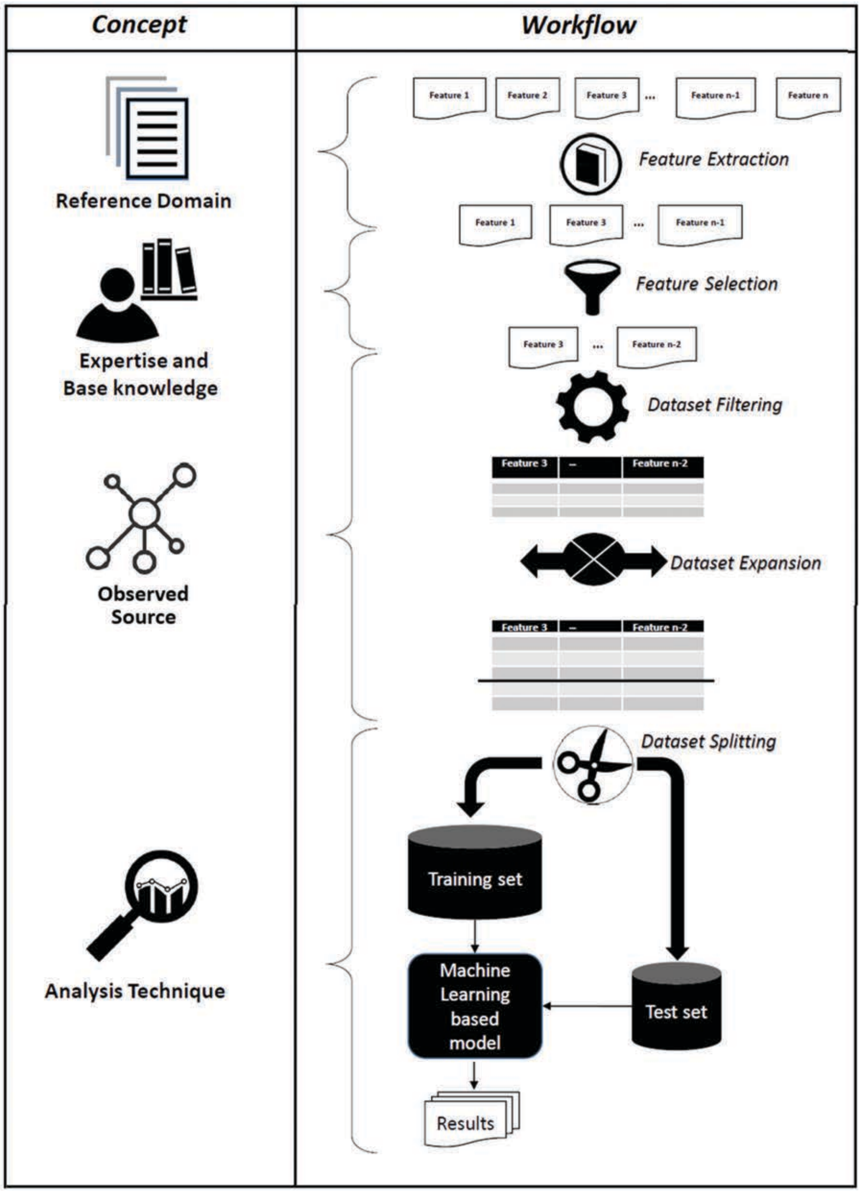


Figure 2 Research method: Conceptual approach and actual workflow.

is to say, guided by the related works, by best practices and/or through the support of domain experts.

The *Observed Source* represents the environment in which a specific study of interest is to be conducted. It has to do with the feasibility of the problem to be analysed on the basis of the available knowledge. In this case, it is important to find the trade-off between (i) the level of availability of the hypothetical data that would be good to have in order to observe a particular phenomenon, and (ii) the actual available data to be able to concretely study the problem in a practical way.

The *Analysis Technique* represents the last part of the conceptual flow, which is related to the choice of the formal tools (such as, statistics, machine learning, as well as key performance indicators and metrics) to be used in order to study and evaluate the case under consideration.

Starting from the above described logic concepts, a *Workflow*, which allows in a concrete way to lead the study, has been defined. It consists of a sequence of 5 main steps: *Feature Extraction, Feature Selection, Dataset Filtering, Dataset Expansion* and *Dataset Splitting*. In particular,

- the *Feature Extraction* step is typically dataset-independent. It aims to identify all potential characteristics (or features), within the *Reference Domain*, that are considered relevant for the study to be conducted. In this case, the ratio of choosing such a data is driven by *Expertise and Base Knowledge*, such as literature review or from domain experts by extracting the *candidate features*;
- the *Feature Selection* step is, instead, typically dataset-dependent. Starting from the overall *candidate features* extracted in the previous step, it aims to select only the characteristics that match those, which are actually available in the considered dataset, but also in dependence with the *Observed Source* (e.g. an OSN).
- after the selection of the actual available features, the *Dataset Filtering* step takes place. This step, guided by the identified characteristics, aims to generate a dataset cleaned from values that are neither useful nor usable for the purposes of analysis in progress. This means that all the available instances are extracted by considering only the selected features in order to generate a database containing only the characteristics which are considered important for the analysis to be conducted;
- the next step, called *Dataset Expansion*, is optional and it depends from the size of the dataset. It aims to enlarge the database (e.g. by duplicating, introducing, sampling instances), as described in [19], in

case the available instances are not enough for the analysis that has to be conducted. This step might very expensive from the temporal point of view;

- at the end of the above mentioned steps, the necessary data for conducting the analysis is available. The *Dataset Splitting* step aims to divide the overall dataset into 2 subsets: Training set and Test set. In particular, the Training set is provided in input to a classifier in order to train appropriately the model, which is built on the basis of the identified features, whereas the Test set is used to validate it.

The above described step-by-step workflow has been applied to the JJATT dataset, described in Section 5.1, in order to illustrate the proposed method, whereas the identified features are presented in the subsequent section.

5.3 Features Identification

As before discussed, the employment of online social networks (OSNs) for conducting illegal activities is becoming more popular due to different advantages for criminals. OSNs allow them to perform their tasks virtually, so as to reducing the risk of being physically exposed, thanks to activities' dematerialization, that are mainly supported by computer techniques by making anonymous their execution. As a consequence, the selection and combination of features that characterize a criminal profile is not trivial, since it can depend not only on the number of the involved variables, but also on their combination.

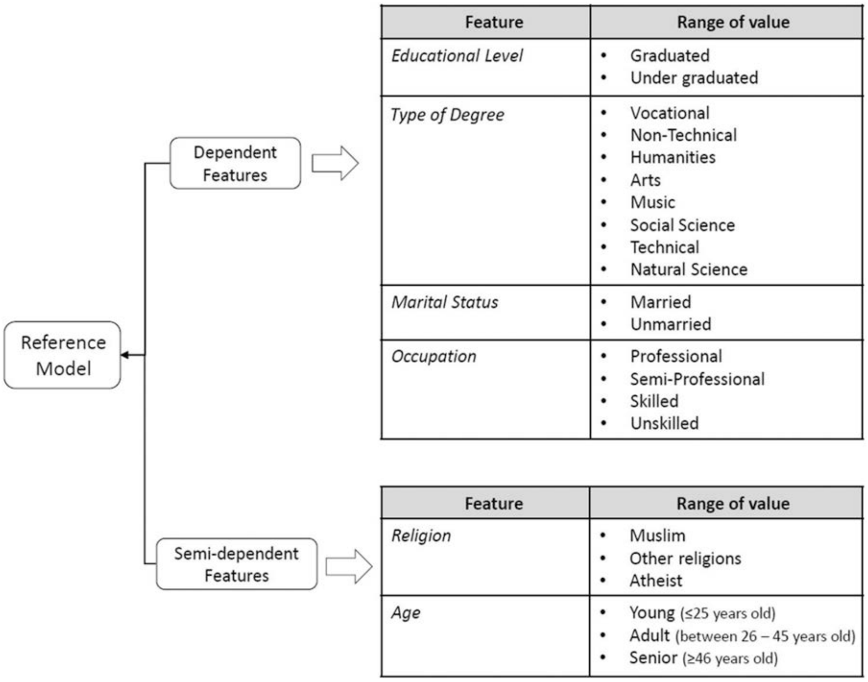
As a consequence, the feature identification process is one of the most important step of such approach by considering the social network perspective. In this case, it has been performed both by applying the above described methodology and through the support and the supervision of the Valencia Local Police (VLP) in Spain [34], as domain experts in the field. In particular, the first step of the methodology has been supported through the best practices provided by the VLP, where a set of candidate characteristics, that are considered relevant, have been selected (e.g. Age, Family, Intimate relationships, Associations, Prison, Religion, Occupations). Due to its popularity (as discussed in Section 2), the reference social network considered in this context is Facebook. According to the second step of the methodology, from the 52 attributes available in the dataset, a reduced set of 6 features has been selected, by crossing these characteristics with both those available in the dataset and subsequently with the metadata available on Facebook. This means that, circa the 88.5% of the data available in the dataset have been discarded, since it was not possible to find a match with those retrievable from the social network. In particular, the remaining and useful identified features are the following one:

- *Educational Level*: this feature provides information regarding the level of education of an individual, because people with a higher level of education have more capacity and freedom of choice; we considered "Under graduated" if one did not achieve at least a bachelor degree and "Graduated" for all persons who got at least a bachelor degree;
- *Type of Degree*: this feature provides information on the type of skill possessed by a person, and consequently on the type of exploitation that can be taken from it;
- *Marital Status*: it provides information on the marital status of an individual, as, typically, married people, especially with kids, are less prone to be involved in illegal activities, in order to avoid risk for the own family;
- *Occupation*: this feature aims to provide knowledge about the occupational status of a person, which can be interpreted as level of job satisfaction by considering the owned type of degree;
- *Religion*: this feature provides information about personal beliefs, because people are often motivated in conducting crimes due to particular beliefs related to specific religions;
- *Age*: it is an important factor, as the way of thinking of the individual and their decisions can vary according to their age. In this case, 3 main categories have been introduced: "up to 25", "between 26 und 45" and "older than 45". The reason of choosing these intervals is related from one site to the education – as most of the people are still studying until 25 years – and from the other side, to marriage status – as most of them tends to create a family within the age of 45 years.

The derived features, as well as specific values or intervals which are identified to discriminate among possible classification, are reported in Figure 3. By analysing them, 2 specific groups emerged. In particular,

(i) *Age* and *Religion* are considered *semi-dependent features*, that is, these are characteristics that have a strong ethnic-cultural or biological root and, as a consequence, they are not strongly connected with the personal lifestyle of a person; whereas

(ii) *Educational Level*, *Type of Degree*, *Occupation* and *Marital Status* can be, instead, considered *dependent features*, that is to say, these are characteristics that might be strongly dependent on the lifestyle of the



Feature	Range of value
Educational Level	• Graduated • Under graduated
Type of Degree	• Vocational • Non-Technical • Humanities • Arts • Music • Social Science • Technical • Natural Science
Marital Status	• Married • Unmarried
Occupation	• Professional • Semi-Professional • Skilled • Unskilled

Feature	Range of value
Religion	• Muslim • Other religions • Atheist
Age	• Young (≤25 years old) • Adult (between 26 – 45 years old) • Senior (≥46 years old)

Figure 3 Reference model, feature categories and range of values.

individuals, the living environment as well as the family standard of living.

Starting from those selected features a database containing only the relevant values has been generated. Then the dataset expansion step has been applied by manually selecting data profile in order to obtain a balanced dataset containing terrorist related profiles and non-terrorist related profiles for a total of 269 entries in total. 85% of the entries have been used as training set and 15% as test set by experimenting different algorithms.

Additional details regarding the technical implementation are explained in the subsequent section.

5.4 Implementation Details

The above described features were tested using 3 types of classifiers in order to assess whether and how the performance varies based on a specific technique rather than another. In particular, a Decision Tree (DT), a

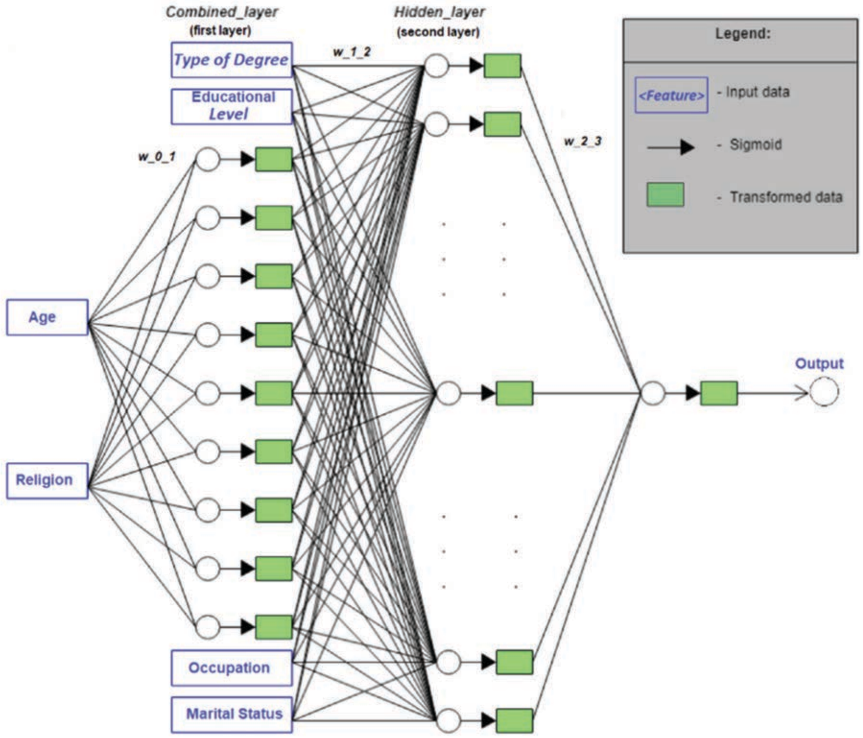


Figure 4 Two-level inputs artificial neural network.

standard Artificial Neural Network (ANN) and a Support Vector Machine (SVM) have been implemented, trained and evaluated based on the identified characteristics. The details on the functioning and characteristics of these algorithms have been already reported and discussed in Section 4. Furthermore, considering the resulting groups of features, as shown in Figure 3, a particular version of ANN has been implemented, considering multiple inputs. Figure 4 shows a graphical representation of the implementation based on two input layers. In particular, *Age* and *Religion* belonging to the *Semi-dependent features* group are used as first input of the network and they generate the input of the *Combined_layer*; whereas *Education Level, Type of Degree*, *Occupation* and *Marital Status*, that belong to the *Dependent features* group, do not play any role at the first input level, but they are combined in the ANN for generating the input of the *Hidden_layer*.

A sigmoid has been adopted as activation function, which introduces non-linear properties in the ANN. In particular, the activation function needs to be differentiable which is important to support the optimization strategy based on backpropagation mechanisms in order to reduce the error during the prediction. For this reason, the sigmoid represented in Equation (1) is adopted. It worth noticing that without it, the proposed ANN would be a simple linear model with a lower level complexity and less power to learn complex functional mappings from data.

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{1}$$

The pseudo-code of the algorithm which has been defined and used to implement the above sketched neural network is reported in Algorithm (1), whereas the DT, SVM and the standard ANN algorithms have been implemented by using standard Python APIs.

The next section provides an overview of the conducted experiments by discussing the obtained results as well as by highlighting the issues that emerged during such analysis.

6 Results Evaluation

In this Section, the assessment of the identified features is reported. It has been conducted by evaluating their performance, in detecting criminal profiles on Facebook, through the trained classifiers described in Section 5. The evaluation criteria are based on the confusion matrix by computing the True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), from which the following three reference indicators are derived.

Accuracy, which elucidates in classification problems the number of correct predictions made by the model over all other kinds of predictions, that is, the degree of closeness of measurements of a quantity respect to the true value (see Equation (2)).

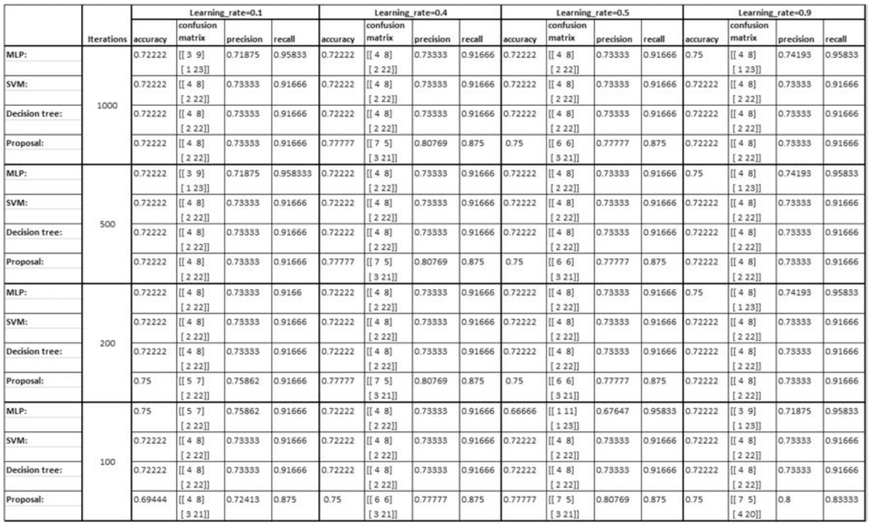
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Recall, which measures the proportion of actual positives that are correctly identified as such (see Equation (3)).

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Algorithm 1: Pseudo-code of the Proposed Multi-level Inputs Artificial Neural Network

```
Input featuresList=list{$f_{1,1}$, $f_{1,2}$, $f_{1,3}$,..., $f_{2,1}$,$f_{2,2}$, $f_{2,3}$,..., $f_{j,i}$,...,$f_{n,m}$}; //$f_{j,i}$ is the feature $i$ belong to the group $j$;
Input trainingSet = list{($e_1$,$v_1$), ($e_2$,$v_2$), ($e_3$,$v_3$),...,($e_{k-1}$,$v_{k-1}$), ($e_k$ $v_k$)}; list of $k$ instances/examples $e$ and related actual value $v$ used to train the Neural Network Model;
NeuralNetworkModel nnm;
ActivationFunction sigmoid=1/(1+np. exp($^{-}$x));
parameter $\lambda$; // learning rate
parameter $\delta$; // error
ListOf Layer loL;
ListOf Weights loW;
ListOf Connection loC;
lastLayer LL;
float eval;
//generation of the hidden layers
for each j in featuresList{
        loL.generateHiddenLayer(j);
}
nnm.add(loL);
//create connection among input and layers and then add further inputs
for each j in featuresList{
    for each i in featuresList{
     //build the connection
     loC.buildConnection(loL(j), i);
     //generate the weight on the generated connection
     loW.generateWeights (j, i);
    }
    //update the hidden layer by the new inputs
    for each i in featuresList{
     loL(j).appendAll (featuresList((j+1, i));
    }
}
nnm.add(loC);
nnm.add(loW);
nnm.add(sigmoid);
for each k in trainingSet {
         integer h = getNumberOf Layers(nnm);
          $eval_k \leftarrow$ evaluate (nnm, $e_k$, $\lambda$);
          if ($|eval_k$-$v_k| > \delta$)
          for each j in loL{
                    loW(h–j).backPropagationUpdate ($\lambda$);
}
return nnm;
```



		Learning_rate=0.1				Learning_rate=0.4				Learning_rate=0.5				Learning_rate=0.9			
	Iterations	accuracy	confusion matrix	precision	recall	accuracy	confusion matrix	precision	recall	accuracy	confusion matrix	precision	recall	accuracy	confusion matrix	precision	recall
MLP:	1000	0.72222	[[3 9] [1 23]]	0.71875	0.95833	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.75	[[4 8] [1 23]]	0.74193	0.95833
SVM:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Decision tree:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Proposal:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.77777	[[7 5] [3 21]]	0.80769	0.875	0.75	[[6 6] [3 21]]	0.77777	0.875	0.72222	[[4 8] [2 22]]	0.73333	0.91666
MLP:	500	0.72222	[[3 9] [1 23]]	0.71875	0.958333	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.75	[[4 8] [1 23]]	0.74193	0.95833
SVM:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Decision tree:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Proposal:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.77777	[[7 5] [3 21]]	0.80769	0.875	0.75	[[6 6] [3 21]]	0.77777	0.875	0.72222	[[4 8] [2 22]]	0.73333	0.91666
MLP:	200	0.72222	[[4 8] [2 22]]	0.73333	0.9166	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.75	[[4 8] [1 23]]	0.74193	0.95833
SVM:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Decision tree:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Proposal:		0.75	[[5 7] [2 22]]	0.75862	0.91666	0.77777	[[7 5] [3 21]]	0.80769	0.875	0.75	[[6 6] [3 21]]	0.77777	0.875	0.72222	[[4 8] [2 22]]	0.73333	0.91666
MLP:	100	0.75	[[5 7] [2 22]]	0.75862	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.66666	[[1 11] [1 23]]	0.67647	0.95833	0.72222	[[3 9] [1 23]]	0.71875	0.95833
SVM:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Decision tree:		0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666	0.72222	[[4 8] [2 22]]	0.73333	0.91666
Proposal:		0.69444	[[4 8] [3 21]]	0.72413	0.875	0.75	[[6 6] [3 21]]	0.77777	0.875	0.77777	[[7 5] [3 21]]	0.80769	0.875	0.75	[[7 5] [4 20]]	0.8	0.83333

Figure 5 Results evaluation of the machine learning algorithms.

Precision, which measures the proportion of actual negatives that are correctly identified as such (see Equation (4)).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Figure 5 summarizes the results in the detection of criminals' profiles on Facebook by showing the accuracy, recall and precision obtained from the experimented classifiers.

In particular, since ANNs are based on two parameters, i.e. learning rate and epochs, during validation process several values have been considered for each of them. So, as learning rate the following values were chosen: 0.005, 0.01, 0.05, 0.1, whereas the epochs' values are the following one: 1000, 5000, 10000, 15000. As a consequence, the performance of each algorithm was evaluated by considering *iteration steps*, *learning rate*, and *epochs* when applicable. The diagrams represented from Figure 6 to Figure 9 show the performance evaluation in terms of average accuracy value, which has been computed for all four experimented classifiers.

Figure 6 presents the results by using a learning rate = 0.005. The best results for this value of learning rate were achieved from the Decision Tree, followed by the MLP, the proposed method and SVM.

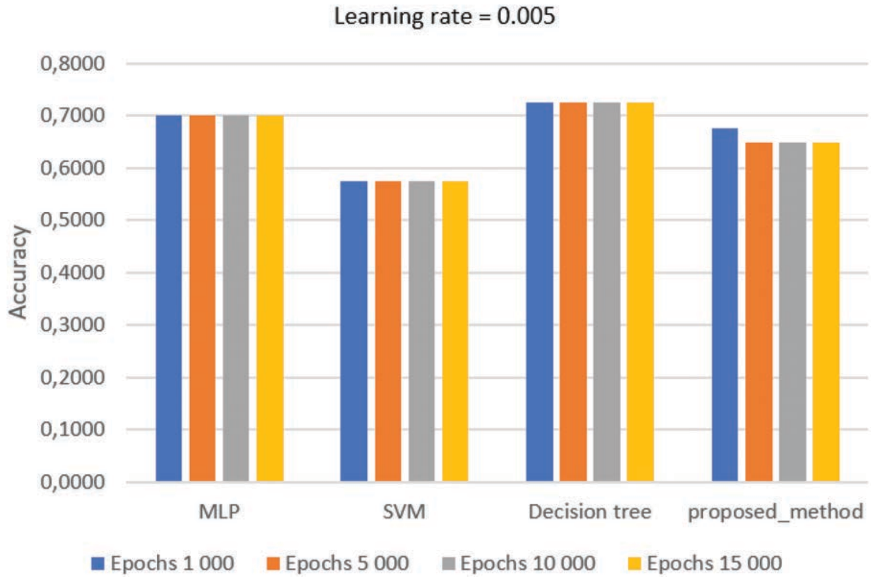


Figure 6 Average accuracy value with learning rate 0.005.

From Figure 7, with a learning rate = 0.05, similar values can be observed from the proposed 2-layer ANN, Decision Tree, SVM and the MLP standard. No crucial improvements of the accuracy emerged even the change of the epochs' values. However, the proposed ANN delivered the better results.

In Figure 8, with a learning rate = 0.01, it is possible to observe the dynamic behaviour of the proposed method, which does not occur in the other algorithms, whose accuracy increases by increasing the epochs.

In Figure 9, with a learning rate = 0.1, the best results are achieved by the Decision Tree, followed by the MLP, the proposed ANN and SVM. However, it is also possible to observe an increase in the accuracy of the proposed method in relation to the epochs' values.

Thus, according to the gathered results by applying the defined model, which is centred on the identified features, it can be seen that any algorithm provided optimal results, as the accuracy varies between 57.5% and 79%. Even if varying the parameters setting in terms of epochs, learning rate and iterations, no particular improvement emerged, which can be interpreted with a modelling problem. Since the model is, in turn, based on the input data, here emerged the gaps due to the lack of proper data related to the phenomenon under study, which are identified and described in the next section.

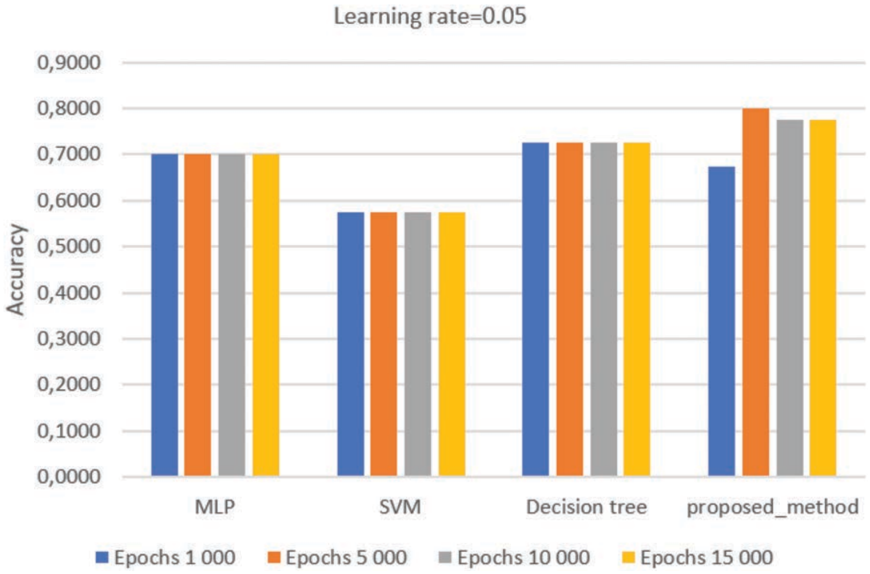


Figure 7 Average accuracy value with learning rate 0.05.

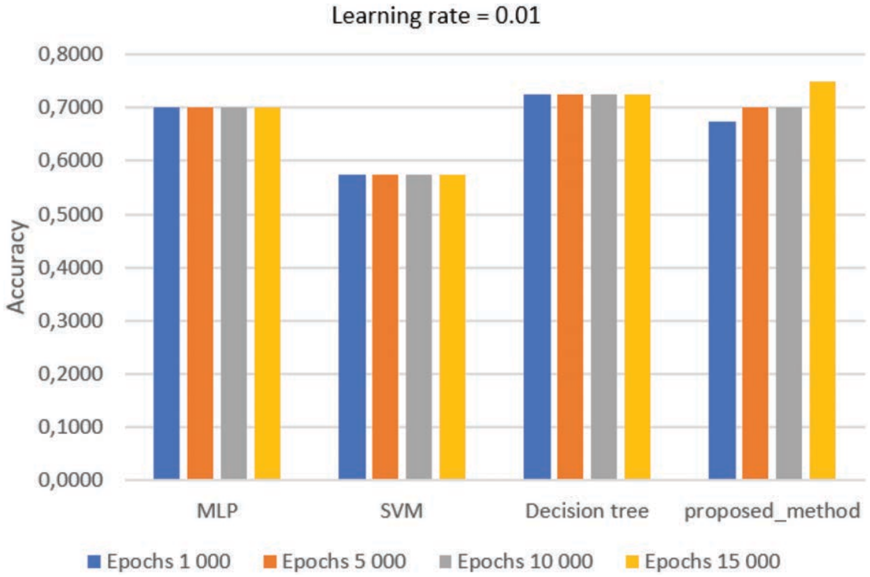


Figure 8 Average accuracy value with learning rate 0.01.

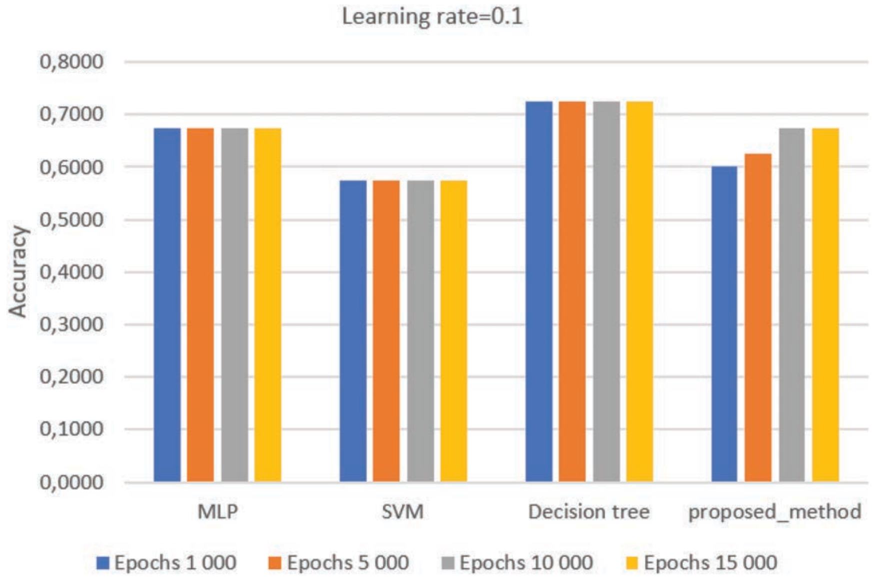


Figure 9 Average accuracy value with learning rate 0.1.

7 Discussion on the Emerged Issues and Limits in the Data

From the conducted evaluation, it is notable that none of the methods excels in the classification. In particular, the lack relies in the currently available data that does not reflect the reality related to online social networks (OSNs) and, consequently, it does not allow it to be properly modelled, but only partially. The main identified issues (see Figure 10), which emerged during the conducted experience, are related to:

(i) *data availability*: it is challenging both to find databases regarding terrorists or criminals that are already publicly and freely usable, as well as to collect data regarding terrorists and criminals, which can be used as the basic to define proper models and define specific algorithms on the top of them, in order to support the online analysis. Most of the scientific papers claim the use of datasets to conduct their analysis, but none of them is freely provided as benchmark;

(ii) *data restriction*: due to the laws and regulations most of the data, which could be exploited to extract or derive additional information, that might be useful in the identification process, is omitted;

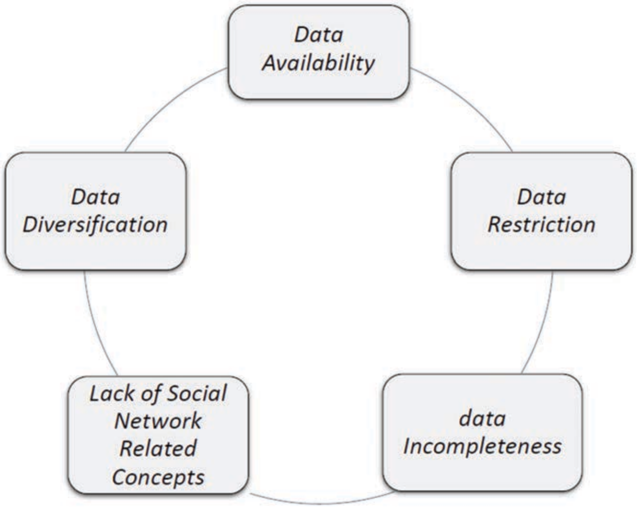


Figure 10 Limits in the data for supporting the detection of criminals on OSNs.

(iii) *data diversification*: the available data is typically related to specific countries; as a consequence, this might bias the criminals' detection model, excluding by default some people, only due to their nationality, religion and so on;

(iv) *data incompleteness*: even if datasets are available, most of the provided entries are incomplete due to the lack of some values and, as a consequence, the available sample is not sufficient enough to be considered significant.

(v) *lack of social network related concepts*: the current datasets do not provide data regarding potential user' relations (which could be associated to the concepts of followers or friendships on social networks), as a consequence, some relevant and characterizing features cannot be applied in reality.

8 Conclusion

The paper discussed the experience conducted on the detection of criminal users on social networks driven by features. In particular, a methodological approach for the features identification related to criminal users has been adopted. The experimentation was based on the *John Jay & ARTIS Transnational Terrorism Database* that is publicly available and freely downloadable. It contains a dataset of 2157 entries consisting of 52 attributes, with data related to the demographic, organizational, and behavioural information, which are used to describe individuals.

By applying the proposed methodology, a set of 6 features have been extracted and grouped in two different categories: *semi-dependent features* and *dependent-features* on the basis of the social-cultural environment around the user. In order to evaluate those features, machine learning techniques have been adopted by implementing 3 standard classifiers by using such features as well as an artificial neural networks algorithm, based on 2 input layers, according to the 2 different groups of features.

Facebook has been considered to show the proposal, as it represents not only one of the most popular online social network, but also because it is highly exploited from criminals as communication means to disseminate criminal information, to spread propaganda and so on.

From the conducted experience, we noticed that most of the data available on current datasets were not able to properly support the social network representation, as it is in reality. Indeed, in our case only 11.5% of the attributes were exploitable in order to define matches on Facebook. This lack reflects why from the performed experiments none of the above mentioned techniques, neither the ad-hoc defined 2-layer ANN excelled particularly in the classification, by obtaining an accuracy varying between 57.5% and 79%.

From such experience, the lack of a proper benchmark dataset emerged, which is specifically conceived for supporting the online social network analysis in criminal context. The identified gap concerns mainly with 5 main issues related to*: data availability, data restriction*, *data diversification, data incompleteness*, as well as *lack of social network related concepts*.

As a consequence of such experience, the definition of a benchmark dataset represents a major starting point which needs to be considered as a future work, both to support a proper analysis as well as to enable the comparison among different research efforts regarding the detection of criminals on online social networks.

Acknowledgments

This work was partially performed in the context of the CHAMPIONs research project, which receives funding from the European Union's Internal Security Fund – Police, grant agreement No: 823705 and by the German Federal Ministry of Education and Research (BMBF) as well as the Hessen State Ministry of Higher Education, Research and the Arts (HMWK) within CRISP. The authors would like to thanks Mr. José L. Diego, Project Manager at Valencia Local Police in Spain, for kindly supporting us in this research activity with his expertise.

References

[1] H. Abadinsky. Organized crime, 10th Ed. Wadsworth, Belmont, USA, 2012.

[2] R. Adderley, A. Badii, and C. Wu. The automatic identification and prioritization of criminal networks from police crime data, EuroISI 2008, LNCS 5376, Springer-Verlag Berlin Heidelberg, pages 5–14, 2008.

[3] R. Adderley and P. B. Mushgrove. Data mining case study: Modeling the behavior of offenders who commit sexual assaults, ACM SIGKDD 2001 International Conference on Knowledge Discovery and Data Mining, New York, pages 215–220, 2001.

[4] M. Chau, J. Xu, and H. Chen. Extracting meaningful entities from police narrative reports. In: National Conference on Digital Government Research, 2001.

[5] H. G. Goldberg and R. W. H. Wong. Restructuring transactional data for link analysis. In: FinCEN AI System, AAAI Fall Symposium, 1998.

[6] G. C. Oatley, J. Zeleznikov, and B. W. Ewart. Matching and predicting crimes. At: AI2004-The 24th SGAI International Conference on Knowledge Based Systems and Applications of Artificial Intelligence, 2004.

[7] D. B. Skillicorn. Clusters within clusters: SVD and counterterrorism, At: Workshop on Data Mining for Counterterrorism and Security, 2003.

[8] A. Tundis, G. Bhatia, A. Jain, and M. Mühlhäuser. "Supporting the Identification and the Assessment of Suspicious Users on Twitter Social Media," 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, 2018, pp. 1–10. doi: 10.1109/NCA.2018.8548321.

[9] J. Xu and H. Chen: CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. In: ACM Transactions on Information Systems, Vol. 23, No. 2, pages 201–226, 2005.

[10] J. Xu and H. Chen: Fighting Organised Crimes: using shortest-path algorithms to identify associations in criminal networks. In: Decision Support Systems, vol. 38, no. 3, pages 473–487, 2003.

[11] J. Xu and H. Chen: The topology of dark networks. In: Communications of the ACM, Vol. 51, No. 10, pages 58–65, 2008.

[12] B. Wellman: The network community: An introduction. In B. Wellman (Ed.). Networks in the Global Community, pages 1–47, Westview Press, Boulder, USA, 1999.

[13] W. Chung, E. Mustaine, and D. Zeng. "Criminal intelligence surveillance and monitoring on social media: Cases of cyber-trafficking," IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2017, pp. 191–193. doi: 10.1109/ISI.2017.8004908.

[14] T. De Smedt, G. De Pauw and P. Van Ostaeyen, 2018. Automatic Detection of Online Jihadist Hate Speech. In Computational Linguistics and Psycholinguistics.

[15] M. M. Janeela Theresa and V. Joseph Raj, 2011. "Analogy making in criminal law with neural network," International Conference on Emerging Trends in Electrical and Computer Technology, Nagercoil, 2011, pp. 772–775. doi: 10.1109/ICETECT.2011.5760222.

[16] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, 2015. "Detecting Multipliers of Jihadism on Twitter," IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 954–960. doi: 10.1109/ICDMW.2015.9.

[17] Q. A. Memon and S. Mehboob, 2003. "Crime investigation and analysis using neural nets," 7th International Multi Topic Conference, Islamabad, pp. 346–350. doi: 10.1109/INMIC.2003.1416748.

[18] M. Nakib, R. T. Khan, M. S. Hasan, and J. Uddin, 2018. "Crime Scene Prediction by Detecting Threatening Objects Using Convolutional Neural Network," Int. Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1–4. doi: 10.1109/IC4ME2.2018.8465583.

[19] D. Parekh, A. Amarasingam, L. Dawson, and D. Ruths, 2018. Studying Jihadists on Social Media: A Critique of Data Collection Methodologies. In: Perspectives on Terrorism, vol. 12(3).

[20] R. R. Petersen. 2013. Criminal network investigation is all about hypertext. SIGWEB Newsl. Autumn, Article 2 (September 2013), 7 pages. doi: 10.1145/2528272.2528274.

[21] M. A. Rashidan, Y. M. Mustafah, S. B. A. Hamid, N. A. Zainuddin, and N. N. A. Aziz, 2014. "Detection of Different Classes Moving Object in Public Surveillance Using Artificial Neural Network (ANN)," International Conference on Computer and Communication Engineering, Kuala Lumpur, 2014, pp. 240–242. doi: 10.1109/ICCCE.2014.75.

[22] S. Savaş and N. Topaloğlu, 2017. "Crime intelligence from social media: A case study," IEEE 14th International Scientific Conference

on Informatics, Poprad, 2017, pp. 313–317. doi: 10.1109/INFORMATICS.2017.8327266.

[23] JJATT 2018 – John Jay & ARTIS Transnational Terrorism Database – online availabe at http://doitapps.jjay.cuny.edu/jjatt/attributes.php.

[24] F. Ozgul and Z. Erdem. 2012. Detecting Criminal Networks Using Social Similarity. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (ASONAM '12). IEEE Computer Society, Washington, DC, USA, 581–585. doi: 10.1109/ASONAM.2012.98.

[25] F. Ozgul, Z. Erdem, C. Bowerman, and J. Bondy (2010). Combined Detection Model for Criminal Network Detection. In: Chen H., Chau M., Li S., Urs S., Srinivasa S., Wang G.A. (eds) Intelligence and Security Informatics. PAISI 2010. Lecture Notes in Computer Science, vol 6122. Springer, Berlin, Heidelberg.

[26] F. Ozgul, M. Gok, Z. Erdem, and Y. Ozal (2012). Detecting criminal networks: SNA models are compared to proprietary models. 156–158. 10.1109/ISI.2012.6284278.

[27] A. Berzinji, F. S. Abdullah, and A. H. Kakei, "Analysis of Terrorist Groups on Facebook," 2013 European Intelligence and Security Informatics Conference, Uppsala, 2013.

[28] A. T. Chatfield, C. G. Reddick, and U. Brajawidagda, "Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks," in Proceedings of the 16th Annual International Conference on Digital Government Research. ACM, 2015.

[29] R. A. Bates, "Dancing with wolves: Today's lone wolf terrorists," The Journal of Public and Professional Sociology, vol. 4, no. 1: p. 1, 2012.

[30] D. Masciandaro, Global financial crime: terrorism, money laundering and offshore centres. Taylor & Francis, 2017.

[31] G. Cheng and X. Tong, "Fuzzy Clustering Multiple Kernel Support Vector Machine," 2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Chengdu, 2018, pp. 7–12.

[32] P. Perner, "How to Compare and Interpret Two Learnt Decision Trees from the Same Domain?" 2013 27th International Conference on Advanced Information Networking and Applications Workshops, Barcelona, 2013, pp. 318–322.

[33] L. Q. Yu and F. S. Rong, "Stock Market Forecasting Research Based on Neural Network and Pattern Matching," 2010 International Conference on E-Business and E-Government, Guangzhou, 2010, pp. 1940–1943.

[34] Valencia Local Police – Online http://www.carismand.eu/valencia-city-council-local-police-plv-spain.html - http://forensor-project.eu/author/josdie/.

[35] R. Borum. Radicalization into Violent Extremism I: A Review of Social Science Theories. Journal of Strategic Security. Vol. 4 Issue 4. (2011) pp. 7–36.

[36] A. P. Schmid (2013-03-27). "Radicalisation, De-Radicalisation, Counter-Radicalisation: A Conceptual Discussion and Literature Review". The International Centre for Counter-Terrorism – The Hague (ICCT).

[37] TAKEDOWN – A EU H2020 Research project – Online available at https://www.takedownproject.eu/.

[38] Vincenzo Ruggiero, "Organized Crime and Terrorist Networks", London, Routledge, 1st Edition, 2019, pp. 220, doi: 10.4324/9780429435102.

[39] S. Kavitha, S. Varuna, and R. Ramya. "A comparative analysis on linear regression and support vector regression". In Proceedings of the 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Kuala Lumpur, Malaysia, 25–27 July 2016; pp. 1–5.

[40] L. Burita, "Information Analysis on Facebook," 2019 Communication and Information Technologies (KIT), Vysoke Tatry, Slovakia, 2019, pp. 1–5. doi: 10.23919/KIT.2019.8883471.

[41] J. Clement, Statista – November 2019. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

[42] A. Moltzau, Towards Data Science – July 2019. https://towardsdatascience.com/artificial-intelligence-and-terrorism-in-social-media-cf166adaf78e.

[43] J. Hu, M. Liu, and J. Zhang, "A semantic model for academic social network analysis," 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, 2014, pp. 310–313. doi: 10.1109/ASONAM.2014.6921602.

[44] N. Akhtar, H. Javed, and G. Sengar, "Analysis of Facebook Social Network," 2013 5th International Conference and Computational Intelligence and Communication Networks, Mathura, 2013, pp. 451–454. doi: 10.1109/CICN.2013.99.

[45] A. Berzinji, F. S. Abdullah, and A. H. Kakei, "Analysis of Terrorist Groups on Facebook," 2013 European Intelligence and

Security Informatics Conference, Uppsala, 2013, pp. 221–221. doi: 10.1109/EISIC.2013.53.

[46] C. Aliprandi, et al. "CAPER: Crawling and analysing Facebook for intelligence purposes," 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), Beijing, 2014, pp. 665–669. doi: 10.1109/ASONAM.2014.6921656.

[47] M. Ketcham, T. Ganokratanaa, and S. Bansin, "The Forensic Algorithm on Facebook Using Natural Language Processing," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, 2016, pp. 624–627. doi: 10.1109/SITIS.2016.103.

[48] CHAMPIONs – A European Union's Internal Security Fund – Police research project – Online available at https://www.championsproject.eu/.

[49] ARMOUR – A European Union's Internal Security Fund – Police research project – Online available at https://www.armourproject.eu/.

[50] CARPER – A Seventh Framework Programme for Research and Technological Development – Online available at http://www.fp7-caper.eu/.

[51] S. Wagstyl, G. Chazan, and T. Buck, 'Merkel Wins Fourth Term but Far-Right Populists Make Gains', Financial Times. Sep 25, 2017. Online available at https://www.ft.com/content/12de72a0-a11c-11e7-9e4f-7f5e6a7c98a2.

[52] G. Delanty (2017). A divided nation in a divided Europe: Emerging cleavages and the crisis of European integration. Brexit: Sociological Responses, p.113.

[53] A. Tundis, G. Bhatia, A. Jain, and M. Mühlhäuser, "Supporting the Identification and the Assessment of Suspicious Users on Twitter Social Media," Proceeding of the IEEE 17th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, 2018, pp. 1–10. doi: 10.1109/NCA.2018.8548321.

[54] A. Tundis, A. Jain, G. Bhatia, and M. Muhlhauser, "Similarity Analysis of Criminals on Social Networks: An Example on Twitter," Proceeding of the 28th International Conference on Computer Communication and Networks (ICCCN), Valencia, Spain, 2019, pp. 1–9. doi: 10.1109/ICCCN.2019.8847028.

[55] A. Tundis and M. Mühlhäuser, "A multi-language approach towards the identification of suspicious users on social networks," Proceeding of the International Carnahan Conference on Security Technology (ICCST), Madrid, 2017, pp. 1–6. doi: 10.1109/CCST.2017.8167794.

[56] A. Tundis and M. Mühlhäuser, "The role of Information and Communication Technology (ICT) in modern criminal organizations." Book Chapter in: Organized Crime and Terrorist Networks, London, Routledge, 2019. doi: 10.4324/9780429435102.

[57] V. Jirovský, A. Pastorek, A. Tundis, and Max Mühlhäuser. "Cybercrime and Organized Crime," Proceeding of the International Conference on Availability, Reliability and Security (ARES 2018), Hamburg, Germany, August 27–30, 2018, doi: 10.1145/3230833.3233288.

[58] A. Tundis, W. Mazurczyk, and M. Mühlhäuser. "A review of network vulnerabilities scanning tools: types, capabilities and functioning." Proceeding of the International Conference on Availability, Reliability and Security (ARES 2018), Hamburg, Germany, August 27–30, 2018, doi: 10.1145/3230833.3233287.

[59] A. Tundis, L. Böck, V. Stanilescu and M. Mühlhäuser. "Limits in the data for detecting criminals on social media," Proceeding of the International Conference on Availability, Reliability and Security (ARES 2019), Kent, Canterbury, UK, August 26–29, 2019. doi: 10.1145/3339252.3341483.

Biographies

Andrea Tundis is a Senior Researcher and his area of expertise are infrastructure protection, Internet organized crime and human safety. In 2014 he got a Ph.D. degree in Systems and Computer Science from the DIMES department at University of Calabria (Italy). He is currently working at Department of Computer Science at Technische Universität Darmstadt (TUDA) in Germany and member of the Telecooperation Division (TK). He is involved in a Horizon 2020 European research project on organized cyber-crime and terrorist networks by investigating on models and methods for the identification, prevention and response of Internet-based crimes.

Leon Böck is a Ph.D. candidate at the Telecooperation Labs at Technische Universität Darmstadt. His research focus is detection, monitoring and prevention of Peer-to-Peer botnets. In addition to the technical aspects of his research, he is interested in the legal and privacy concerns related to fighting botnets and malware. He received his M.Sc. in computer science from TU Darmstadt in 2017 with a master thesis on the topic "On P2P botnet monitoring in adverse conditions".

Victoria Stanilescu is a master student at Technische Universität Darmstadt (TUDA), in Germany. She received her B.Sc. degree in Economics from the Municipal University of Chisinau, Moldova as well as a Bachelor degree in Informatics TUDA, in 2019. She is currently working as MES-Developer at Siemens AG.

Max Mühlhäuser is a full professor at Technische Universität Darmstadt and head of Telecooperation Lab. He holds key positions in several large collaborative research centers and is leading the Doctoral School on Privacy and Trust for Mobile Users. He and his lab members conduct research on The Future Internet, Human Computer Interaction and Cybersecurity, Privacy & Trust. Max founded and managed industrial research centers, and worked as either professor or visiting professor at universities in Germany, the US, Canada, Australia, France, and Austria. He is a member of acatech, the German Academy of the Technical Sciences.