Crime Intelligence from Social Media Using CISMO

Ogerta Elezaj[1], Sule Yildirim Yayilgan[1], Javed Ahmed[1], Edlira Kalemi[2], Brumle Brichfeldt[1] and Claudia Haubold[1]
[1]Department of Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU), Norway
{ogerta.elezaj, sule.yildirim, javed.ahmed}@ntnu.no
{brumle.brichfeldt, caludia.haubold}@stud.ntnu.no
[2]University of Tirana, Albania
edlira.kalemi@unit.edu.al

Abstract. Nowadays, Online Social Networks (OSNs) are being used as a hosting ground for criminal activities, and the legal enforcement agencies (LEAs) are struggling to process and analyse the huge amount of data coming from these sources. OSNs generate a huge massive volume of unstructured data making difficult for the LEAs to 'patrol the facts' and to gather intelligence in order to provide it to the legal domain. There is no ontology model, among those found in literature, that allows to exhaustively describe all the aspects of crime investigation targeting data integration, information sharing, collection and preservation of digital evidences by using biometric features, and query answering. To bridge this gap, this paper presents an extended version of our earlier SMONT ontology, called CISMO as a semantic tool suitable for gathering digital evidence from OSNs helping LEAs to develop new investigative systems to counter the threat of different crimes. The new version introduces the core concepts related to crime cases in the police repositories, biometric data and digital evidences collected by OSNs, making possible for LEAs to classify crimes, investigate hidden crime patterns or predict future crime patterns. CISMO is more concise and has a richer concept knowledge base compared with the previous version SMONT. We prove the effectiveness of CISMO in a case study covering some general aspects in criminal cases in OSNs, demonstrating how this semantic approach can help LEAs to gather knowledge for crime investigation using natural language processing and machine learning to process messages shared in online platforms and also applying reasoning rules, as semantic inferences.

Keywords: Ontology, crime, online social networks, digital evidence, biometrics, security, reasoning.

1 Introduction

OSNs have changed the way how people communicate and connect between them. Social media such as Facebook, Twitter, Instagram, Google+, etc. are being used

by millions of users every day and the data which are freely shared in these networks is like a treasure if it is properly processed, and the knowledge is extracted. The data generated in OSNs contain exabytes of information related to people day-to-day activities and stand as an important source for Big Data [29]. What matter most is not the data itself but rather the information and knowledge that can be extracted in order to be used in decision-making in different domains.

Despite the advantages of OSNs in allowing people to stay connected, there is a darker side to these networks, as criminal activities being committed in these platforms are becoming a central problem in every country. In 2014, the FBI's internet Crime Complain Centre reported that 12% of all logged complaints involved social media, equates to 32,330 complaints received during a year[1]. INTERPOL used social media platforms to find out potential witnesses in different terrorism acts as was the case of London Bridge attack in the UK in 2017[2]. More than 4.7 million counterfeit products were seized in an operation against trafficking of illegal goods, which was carried out by LEAs of 18 countries in collaboration with Europol. In this operation 16 470 social media accounts and 3 400 websites selling counterfeit products were closed[3]. Of major concern to investigators is the fact that social media facilitates the attraction and the recruitment of new members in extremist groups, becoming a topic of major concern for many legal agencies [26]. It is noted that half of teenagers have experienced bullying on OSNs, resulting in low self-esteem and consideration of suicide. Moreover, only 1 out of 10 teenagers tells a parent if they have been a victim, which demonstrate that cyberbullying crimes are not being reported, and thus unpunished[4].

Different type of digital crimes can be collected by OSNs, coming in different forms such as messages, photos, videos, audios, local-based data, etc. In the crime investigation processes, LEAs have to collect and provide reliable and authenticated evidences to ensure their admissibility in the court. OSNs are one of the key driving factors for the evolution of crime arenas bringing new opportunities and challenges to legal agencies to investigate threats from individuals who may or may not be members of groups. Various data generated by users, known as user generate content (UGC) can be used to investigate committed crimes or to predict future

---

[1] https://pdf.ic3.gov/2014_IC3Report.pdf

[2] https://www.interpol.int/en/Crimes/Terrorism/Analysing-social-media

[3] https:// /www.europol.europa.eu/newsroom/news/counterfeit-crackdown-hits-two-organised-criminal-groups-more-30-suspects-arrested

[4] http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html

crime patterns [17]. Digital evidences which contain social media contents has been accepted by different courts to identify suspects, locate witnesses, and convict defendants [16].

However, the characteristics of OSNs data render the existing solutions insufficient to consider the new challenges of LEAs to handle digital evidences in crime investigation and prevention [21]. The exponential growth in the volume, velocity, and variability of OSNs data prevents LEAs to efficiently process and manage the large criminal datasets using traditional methods [20]. Due to the heterogeneity, noise and the massive size of unstructured data generated in social media platforms, LEAs have to take real efforts to face the challenges of collection, processing and analysing the digital evidences in a timely and efficient manner implementing comprehensive solutions.

A well-defined and standardized representation of OSNs data could be achieved using Semantic Web technologies supporting LEAs in structuring and better integrating the crime records and to model formal knowledge in the crime domain. Semantic Web Technologies combine a unique addressing mechanism (Uniform Resource Identifiers: URI) with a formal knowledge representation (RDF and OWL) and a common query language (SPARQL). In this research we use ontology model as one of the major concepts used in Semantic Web applications, to represent a set of concepts and their relationships within a domain into a machine-made form.

CISMO ontology is used to model the environment of different categories of crime happing in OSNs. With the help of this ontology, data can form an interconnected knowledge base of the different evidence objects extracted from social media [15]. The evidence can also be merged with the reports and evidence streams that exist in LEAs repositories to identify new and implicit knowledge using inference engines. To the best of our knowledge, there is not a generic ontology in the crime domain for OSNs to use it as a knowledge-based tool for data mining applications.

Despite this importance, the existing forensics analyses tool are currently limited to face the identified challenges and very few semantic solutions have been developed to help investigators to cope with new technologies [2]. For this reason, in this paper the existing ontologies used for crime investigation through OSNs content are examined with the objective of better understanding the challenges and gaps unique to crime investigation from OSNs, and to provide methods for addressing those challenges and the gaps by developing intelligent systems to sift through massive amounts of online information and to extracts what's useful to the investigators. To the best of our knowledge, there is no ontology model, among those found in literature, that covers each aspect of crime investigation such as data integration, digital evidences, biometric features, and data coming from OSNs. The existing ontologies are not generalized and mostly are platform based, capable to

deal with data coming only from particular social media. The main motivation behind the study proposed in this paper is to develop an ontology, asone of the main components of a knowledge-based graph framework, introduced by authors to gather intelligence from OSNs in order to assist LEAs to detect and prevent criminal activities. CISMO ontology is based on linked concepts like agents composed of persons or organized criminal networks, institutions involved in crime investigation, digital evidence collection processes and biometric modalities. CISMO is important, firstly, to model relationships among user activities in OSNs and to identify suspects related to different crime categories such as ordinary crime and cybercrime. Secondly, SMONT is conceptualized as the main backbone of an intelligent knowledge-based system used in crime detection and prevention. This framework has the capacity to support LEAs in crime investigation activities, starting from data collection to preservation of digital evidence admissible in court. One of the components of this framework is the ontology used to model and structure the social media content in a well-organised structured way. Once all online activities are stored by considering this structure and integrating them with the data coming from LEAs repositories, investigators can though inference rules and reasoning, infer accurate knowledge related to the crime domain. In order to illustrate on how CISMO can be used as a tool in crime investigation, a case study based on a recently leaked dataset collected from Nulled.io[5], an online forum for distributing cracked software, and trade of leaked and stolen credentials, is presented. Specifically, we first manually classify a subset of the private messages, and we train a machine learning model to classify messages belonging to criminal activities or not. All the classified messages and their related data such as IP of the sender, IP of the receiver, timestamp, etc. populate the ontology, and we provide a way for semantically querying the crime ontology.

The remaining part of the paper is organized as follows. Section 2 presents related work on using semantic models and technologies for crime investigation and prevention of OSNs. The methodology used to develop CISMO and its main components are introduced in the section 3. Section 4 describes some crime scenarios that can be solved using our ontology. Finally, the conclusions and future work are given in section 5.

2 Related work

In this section, critically we analyse current literature of knowledge preservation in the crime domain. Based on the meta-analysis and literature review, a summary of

5 https://archive.org/details/nulled.io_database_dump_06052016

existing ontologies used in the crime domain, is provided. The perspective of this review is to find the gaps in the existing semantic solutions in the crime domain. During the review of existing crime ontologies, it has been found out that the existing ontologies in the domain of crime investigation and prevention are sub-domains of various elements that define the crime domain. As the proposed ontology aims to mostly cover all the aspects of the crime investigation in OSNS, including digital evidences by usage of biometrics features, during this research, we have analysed different ontologies capable to handle social media data, to collect and represent biometric data and to maintain defensibly the chain of custody of digital evidences. Based on the depth analyses of the existing ontologies, we have considered some re-use of some existing concepts by providing smooth access to these ontologies and at a second stage based on the identified gaps, we extend the existing ontology developed by authors, SMONT [11] by adding new concepts and by providing advanced support in adapting ontologies to crime domain. The existing ontology, SMONT presented by author in previous research, lack the necessary level of details about the collection and the preservation of digital evidences. The dilemma that we faced was whether to start developing a new ontology from scratch or to examine existing ontologies used to model social media in the crime domain and check if one of them fitted our purposes as it is or in an extended version.

2.1 Ontologies in social media

In literature few ontologies are proposed for semantic social web, such as FOAF (Friend Of A Friend) [8], SIOC (Semantically Interlinked Online Communities) [4] and SCOT (Social Semantic Cloud of Tags) [24], but none of them are detailed enough to be used for knowledge representation and data integration of crime activities in OSNs. Different researchers have extended these ontologies in order to add new concepts to express the interest domain. In [13], it is proposed an extended version of the FOAF ontology, which is evaluated by W3C Consortium as a good ontology capable to model persons and their relationships [9]. In this work, the FOAF ontology is enriched with new classes and properties related to profilers. Still, this ontology is not a unified semantic model for OSN to describe the content of multiple users from different OSNs such as Facebook, LinkedIn, Twitter, etc., that means that it is platform depend and does not offer a generic semantic solution.

SIOC is an open standard otology developed in 2004 aiming to represent social media content in RDF format and is especially designed for modelling user forums [5], missing many of the OSNs concepts. In 2017 authors in [3] enriched this ontology by presenting a new version exSIOCint, adding new classes and relationships in order to model data coming from web forums and to enhance automatic inherences. However, this is a general-purpose ontology focused on modeling web forums and it is not suitable for the intended problem that CISMO is going to solve, assisting LEAs in daily fighting crimes happening in OSNs. Different ontologies are designed and implemented by different authors to semantically represent the knowledge of OSNs (SocIoS) [27], LODE [25]), but all of them are general-purpose ontologies and their focus is not in modelling formal

knowledge for social media forensic.

2.2 Ontologies in forensic analyses

With the widespread of using ontologies as means for knowledge representation, different ontologies have been proposed and applied in the criminal and legal domain but none of them has a focus on analysing crimes in OSNs. However, we have analyses different semantic solution in the legal domain in order to get a better understanding of the domain concepts in order to apply them in CISMO.

Osathitporn et al. [31], present an ontology for the criminal legal domain in Thailand. The objective of this ontology is to represent legal elements in the law domain, and its main artefacts are crime and justification. However, the objective of this ontology is totally different from CISMO as it is not used to collect digital evidences from OSNs. Kastrati in [14] presented SEMCON, a criminal ontology, developed to process semantically and contextually Facebook data of different users and to classify these users as suspects or not. This ontology is platform dependent as it is designed only for Facebook, and its scope is not to model the crime investigation process but only to build up probabilistic predictive models for suspects. The digital evidence collection and preservation are out of the scope of this ontology.

A multilayer semantic framework used to detect crime on OSNs is presented in [2]. This framework is a hybrid solution and its main component is a global ontology derived by mapping different local ontologies for different OSNs. In this paper only some parts of these ontologies are presented, and the global ontology lacks the required level of details of digital evidence gathering. Furthermore, the lack of integration of biometric modalities in digital evidence preservation become a serious problem in the crime investigation.

In [22] authors presented a top-level Cyber Forensics ontology, and its main high-level classes are crime case, criminal, crime type and evidence. This semantic solution does not make an effort to gather digital evidences, but only it suffices with keeping trace of the medium of the digital evidence. Cosic et al. [6,7] presented DEMF, an ontology to represent semantically the digital chain of custody of digital evidences. The developed ontology is a general solution for forensic investigation and its scope it's not related to the social media. The legal concepts used to represent digital evidences in this ontology are considered in our proposal, serving to us as a method to expand on our ontology related to the digital evidences in OSNs. We have used some of the artefacts such as digital evidence integrity methods, chain of custody based on the possibilities to answer to the six interrogatives of police report writing, known in literature as 5W+H investigative model (Who, What, When, Where, Why and How) [10]. Moreover, considering the crucial fact that collecting digital evidences from OSNs is a complex task different from classical forensics, new artefacts are added to the digital evidence class of SMONT.

2.3 Biometric Ontologies on crime investigation

Concentrating on crime investigation in OSNs is very crucial to process multimedia content and to extract biometric features that carries personal

information linked with the person's biometric characteristics for all the suspects of a criminal case. CISMO will be used to process the photographs shared in a social media platform, using a facial recognition biometric system to extract biometric characteristics. As biometrics is considered a strong alternative from crime detection, an automatic system should be in place to identify a person on the basis of his physical characteristics such as finger, face, iris, ears, etc. Unfortunately, in literature there does not exist many attempts focusing in developing biometric ontologies. Authors in [19] proposed a conceptual framework with the core element of human factors ontology for cyber-security, based on socio-cognitive characteristics. In the proposed ontology, there are no evidence of usage of biometric concepts.

The only biometric ontology found is presented in [12], where the authors have developed a biometric ontology and implemented it in a Big Data environment. This otology covers a broad range of biometric aspects related to behavior analyses such as cognitive skills or identification of tacit psychological factors. The ontology is evaluated using the asylum seek and immigrant identification as a real use case. Some concepts of this ontology can be reused to model biometric characteristics of persons based on the multimedia content shared in that persons' social media profile, as our biometric aspects are not related to behavior analyses.

From this review we conclude that no ontology has been developed so far as a complete ontology in order to model all the components of the crime domain capable to gather intelligence from online social networks. Leveraging from existing ontologies (e.g., FOAF, DEMF, SIOC), the objective is to identify the gaps in current semantic solutions and to propose a more generic solution to overcome the identified challenges effectively by representing semantically the crime domain in OSNs.

This research has the following contributions:

1. We proposed a semantical data model for investigation of crimes in OSNs covering the main aspects of the crime investigation process such as digital evidence preservation including biometric artifacts.
2. We implemented the data model in an ontology in OWL using Protégé 5.5.0 [23].
3. We populated CISMO ontology with instances from a hacker forum database. The ontology contains 200 classes, 54 object properties and 18 datatype properties.
4. We semantically query the ontology to find suspects and digital evidences for different crimes using the data coming from an online forum. The resulting ontology can be extended for crime prediction and prevention.

In order to achieve our goal, in the section 3, we present CISMO as a unique ontology in the way it merges important aspects of crime investigation process in OSNs as criminal profile, crime categories, social media content, digital evidence gathering and biometric modalities, enabling a task-driven ontology-developing process.

3 CISMO ontology in OWL

3.1 Methodology

Actually, Figure 1 presents the CISMO methodology composed by six main steps: (i) domain specification; (ii) consideration of reusing existing artifacts; (iii) conceptualization of key concepts; (iv) implementation in OWL; (v) ontology population; and (vi) ontology evaluation..

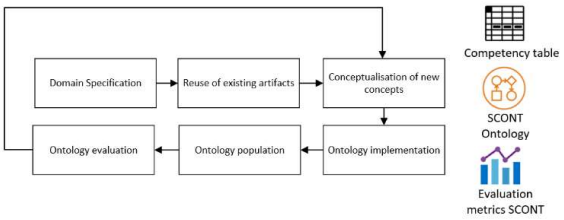

Fig. 1: CISMO Ontology Building Methodology

During the first phase, knowledge about crime domain in OSNs is gained whares in the conceptualisation phase the main classes and subclasses are defined. Bases on the recent literature, the crime domain is analyzed to obtain the required knowledge in order to build CISMO capable to handle digital evidences. Based on the ontology engineering, an integration strategy was applied to import existing knowledge from those fields where concepts are stable and to fit our main goal. The existing ontology, FOAF has been used as a basic to model OSNs artifacts and it is extended with the new concepts. The output obtained by the first step is the competency table containing competency questions which play an important role as it is considered vital to outline and to constrain the scope of knowledge represented by the ontology [18]. Furthermore, the translation of competency questions into SPARQL queries is used to evaluate the ontology including verification and validation [30].

Table 1. Sample of CISMO competency question

	COMPETENCY QUESTIONS
CQ1	What crime categories exist in OSNs?
CQ2	Under what conditions should a person be considered as a suspect?
CQ3	Do difffrent profiles in different OSNs belong to the same person?
CQ4	What biometric features are extracted by facial recognition system accessing photos from social media sites for crime investigation?
CQ5	Can a person of OSNs be considered as a suspect of an *on-line crime* based on online communication with a victim?
CQ6	Can a person be considered as a suspect of an on-site crime based on the geo-content he shared in a social media?
CQ7	Can a crime be *prevented* based on the persons activities in social media such as following crimal profiles, content of comments or statuses shared and likes to suspect pages or persons?
CQ8	Can digital evidenced be collected by using social media artifacts?

CQ9	What are the elements of digital evidences to be admitted in courts?
CQ10	Can the chain of custofy of digital evidenced be maintained?

Specifying competency questions in the specification phase is vital since it allows us to determine the ontology scope. A subset of the CISMO competency questions that covers the main concepts of the crime domain in OSNs is shown in Table 1. Morover, after identifying the goals of CISMO ontology, concepts and the specific classes to the crime investigation in OSNs domain are defined. From the three main methods found in literature to construct ontologiese, Top-Down, Middle-Out and Bottom-Up [28], we have chosen the Top-Down approach to generate this ontology. In a later stage the crime domain ontology is developed using an OWL editor tool developed by Stanford University, named Protégé [23], an open-source and free software. At the end, the solution is tested and evaluated using real data of a hacker forum.

In our proposal, there exist only one global ontology describing the semantics of each OSNs sources, providing a generic solution for LEAs. Additionally, if new data sources are available or it happens that there are done some changes in the existing data sources, the ontology can be adjusted easily.

3.2 Extending SMONT to CISMO

In this section, we present our ontology, developed to store and to arrange all the components of crime evidences in OSNs, providing a common language and a foundation for reasoning. To increase the value of the proposed solution and to generalize its scope, we extended SMONT ontology to a more specific version. The new ontology, named CISMO is a more detailed ontology aiming at supporting LEAs to collect digital evidences and to increase their integrity keeping their chain of custody to stand up in the court.

Initially, the top hierarchy classes of SMONT included the following classes, Person, Crimes and Crime Case Solving. The main extending work lies in the digital evidence collection and keeping its chain of custody in order to prove it in the court. The CISMO ontology generically models key entities relevant for crime investigation and prevention using OSNs and the relations between them. As in Fig. 1, the current ontology, CISMO covers 6 main concepts, two of which are general concepts like *Institution* and *Agent*, whereas the other four concepts are crime related concepts.

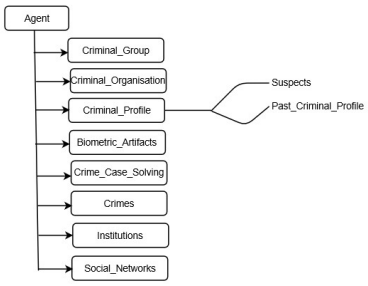

Fig. 2: Top classes of CISMO

The *Agent* and *Institution* concepts define the individuals and institutions who benefit from or are subject to the crime case investigation process. The *Agent* class is a super class of the concepts of *Person, Criminal_Group, Criminal_Organization* and *Criminal_Profile*. Persons are individual agents; an organization is a group of other organizations or persons which acts 'as one' and a criminal group is a group of persons involved in organized crime. A specific individual involved or suspected for a crime is an instance of the class *Criminal_Profile*, which has two sub-classes. The sub-class *Past_Criminal_Profile* stores instances related to previous criminal cases stored in LEAs repositories, which might be useful during criminal investigation for checking and linkage of different profiles. Persons which are declared by the police as main suspects for an open case or individuals that have been considered as suspects after the suspect behavior on the social networking, will belong to the *Suspects* sub-class. The following fifth subsections briefly describe the CISMO concepts.

3.2.1 Social Networks

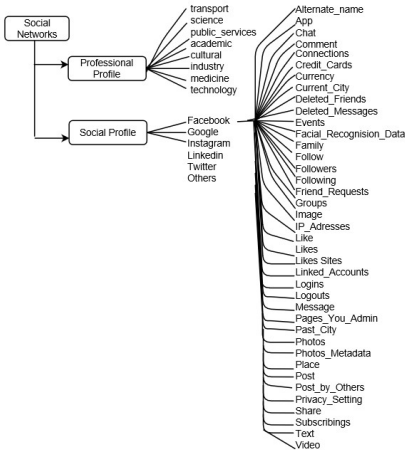

Fig. 3: 'Social Networks' concept of CISMO

The class Social Networks represents information about persons whose data are collected from on-line social networks providers (classes: Facebook, Google, Instagram, etc.) and from their professional profile. As Facebook is the most popular social networking, in fig.3 are illustrated all the sub-classes of the Facebook class.

3.2.2 Crimes

As illustrated in fig. 4, in the *Crime* class we define all the categories of crimes and crime archives. Based on the literature we tried to identify the taxonomy of crime in two classes which are *classical crime* ad *digital crime*. The classical crimes are classified in many categories such as bulgar, corruption, kidnap, sexual crime, etc. The *digital crime* is classified in three main categories, adult crime, child-crimes and malware. This classification is particularly difficult due to the lack of standardized concepts across countries. *Crime archive* contains instances of previous crime investigated by LEAS.

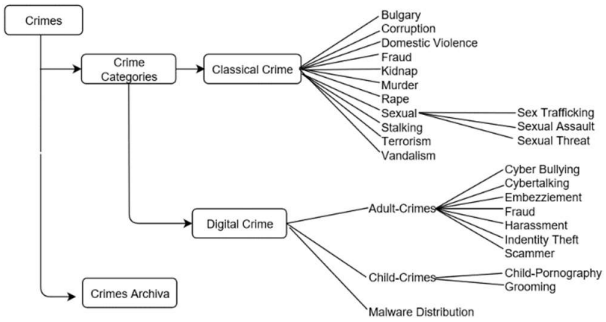

Fig. 4: Underlying classes of 'Crimes' concept

3.2.3 Crime Case Solving

In the past, reacting to crimes was the main purpose of policing, placing focus on crime investigation, but now the focus has changed to finding new ways for the police officers to shift from crime investigation to prevention using intelligent systems. LEAs need to use modern technologies to access and to manage the data coming from different sources including OSNs to automate insights, to create actionable intelligence in order to be focused on investigating the crime on the front line rather than dealing with time-consuming activities. of LEAs. As illustrated in Fig. 5, the concept *Crime_Case_Solving* addresses the wide aspects of collection and preservation of digital evidences. The module of C*rime_Case_Solving* in CISMO is used to model the activities carried out by LEAs in order to collect and manage the digital evidence of the crimes happening in OSN and to maintain an accurate and complete chain of custody, protecting the integrity of the digital evidence itself. Collecting digital evidences from OSNs is a complex task because the evidences are not saved in the hard drive, and the social media artifacts are stores in different places.

We have taken into account different places of evidences found in OSNs, such as browser history, events, hidden and system files, log files, pictures, images, digital photos, videos, system files and temporary files. In most of the OSNs case investigation, LEAs normally collect the digital evidences in context to the specific case and person and manage them is a tough task, and hence an intelligent system is required. Many digital forensic tools are developed to collect digital evidences from computers, but less efforts are made to support the collection of digital evidences from different social media platforms, being that OSNs are new field in digital forensics and there are not commonly accepted standards and guidelines for the forensics investigation based on data coming from OSNs [1].

An investigation starts with analysing a specific social networking web page. For example, a victim got a message in a specific date and time with the content "See you soon" and the icon of a gun, in his account in Facebook. Later, the sender of this message may change it, just deleting the gun icon. An investigator has to analyse the artifacts, and he has to collect immutable elements in order to process with case solving. He has to identify and use different sources for the evidence acquisition, such as suspect's device, victim's device and social network services. All these sources are categorized under the class *Hardware*, belonging to the class *Technology*. The *Hardware* section of CISMO is broken up into three different parts: *Computers, Large Scale Digital Devices* and *Small Scale Digital Devices.*

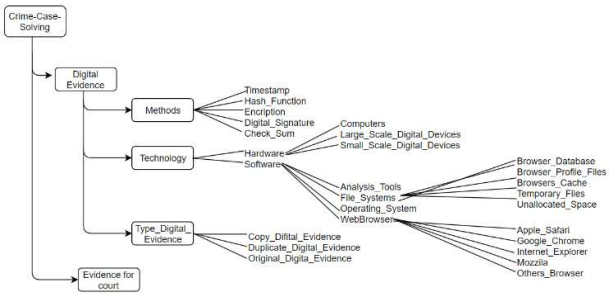

Fig. 4: 'Crime_Case_Solving' concept of CISMO

For the purpose of this ontology, *Small Scale Digital Devices* is one of the most important concepts, broken down into *cell phones* and *PDA*, as people use their phones overwhelmingly to text, share and comment via social networks. In the *Software* section of SMONT are presented the tools used to analyze the evidences, the operation system used by suspects and victims, the web browser used to navigate in the social media, and the file system were different forensic tools try to collect Facebook artifacts of a particular user. The *Method* category focuses on several

methods been adapted from the computer science and information security to the domain of digital forensic to prove the integrity of digital evidence, such as *timestamp, hash functions, encryption, digital signature and check sum.*

In order to examine a system and maintain the chain of custody of a digital evidence defensible, investigators have to freeze it and examine a copy of the original data acquired [32]. So, the digital evidences collected from OSNs using SMONT are categorized in the class *Type_Digital_Evidence*. *Original_Digital_Evidence* refers to the physical items and objects related to these items at the time of seizure, *Duplicate_Digital_Evidence* refers to the duplicate of the evidence on the original physical item, and *Copy_Digital_Evidence* refers a copy of the original evidence independent from the original physical item. When the integrity of a digital evidence is preserved, it goes under the class *Evidence_for_cort.*

3.2.4 Biometric Artifacts

Using biometric artifacts in crime investigation is becoming an important task for LEAs to narrow down the list of persons accused or suspected of committing a crime, and to represent the biometric evidences with strong statistical basis to a court of law. The biometric class is composed of 5 sub-classes. The *Physical_Biometric* class include face, finger, hair, iris, and head. The other category, *Behaviour_biometrics* mainly deals with human behaviors of profiles of social media which are usually extracted by posted videos in their social platforms.

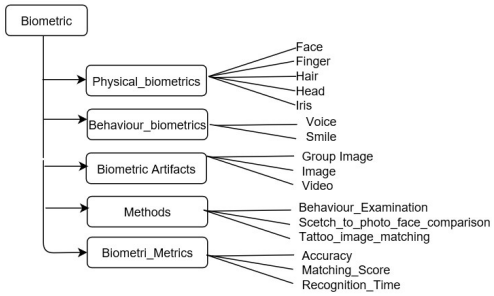

Fig. 5: 'Biometric' concept of SMONT

The biometric artifacts of OSNs are collected by routinely seized cameras and mobile devices of suspects. The images, group images and videos, retrieved from these devices provide key evidences in the crime investigation. In SMONT we have defined the common methods use to extract these features, under the class *Methods*. Biometric processes are evaluated in terms of accuracy, matching score and recognition time, defined under the class *Biometric_metrics*.

3.2.5 Object and Data Properties

CISMO consists of several object and data properties and most of them are owned by Person and Digital Evidence. The data properties are used to link the individual of a class to a data value. For instance, a social media post may have date and time recorded, so each message exchanged in a social media platform will have a timestamp associated with it. In order to relate a message of a social media to its timestamp, data type property *hasTimestamp* was created. On the other hand, object properties are defined to relate individual of two classes. For example, *isSuspected* is object property owns by person and is used to relate him/her to an instance of crime class. *hasbiometric* is another property owns by person and is used to relate a person to an instance of biometric artifact class. Fig. 6 shows the relationships among the concepts of CISMO, briefly explained above.

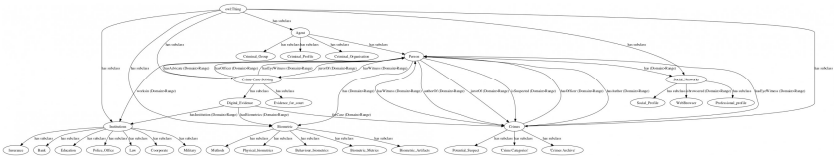

Fig. 6: Class dependencies of CISMO

4 Case Study

The data generated by users in OSN are proprietary of the social media platforms and LEAs can access their non-public data based on signed agreements. Recently, the U.S. and the U.K. have signed a first-of-its-kind agreement to access user-generated data from OSNs companies related to different criminal cases[6]. In this research, we have to evaluate the proposed solution with real data coming from social platforms, but as per confidentiality issue we could not make use of this data. To evaluate and to show the effectiveness of CISMO we used a dataset from Nulled.io, a forum used by cybercriminals to trade and purchase leaked information, stolen credentials, nulled software, hacking tools and cracks. We do not claim this forum represents all different categories of crime happening in OSNs, but this data is a treasure trove for LEAs to investigate criminal activities such as illegal sales. During our research, the data which are publicly available, are processed preserving the sensitive information in order not to allow direct or indirect identification of members. In this dataset it stored the profile information of 599,085 members and their online activities. In our analyses we are based on the 800,593 private

[6] https://www.cnbc.com/2019/10/04/us-uk-sign-agreement-to-access-data-from-tech-companies-like-facebook.html

messages, stored in the table *message_posts*, that has the following fields: *message id, message topic id, message date, message post, message post key, message author id* and *message IP address*. This information is relevant for a criminal case to produce digital evidences to be admitted to the court. During the data pre-processing, the welcome messages send by the system or administrators for the new members are deleted. The remaining records are processed to remove HTML tags using Beautiful Soup, an HTML parser. Next, we also use lemmatizers in NLTK to convert nouns and verbs to their lemma. We also removed all the special characters, stop words and punctuation marks from message contents.

As a case study, we randomly selected 500 messages and labelled manually in normal and criminal activity. This dataset is used to train different machine learning classifiers and after the tuning of the parameters of these classifiers, we classified 2000 messages randomly selected from the dump dataset. These messages are instances to the CISMO ontology. The method called Filtered Classifier is used as it allows a filter to be paired up with a classifier. Even though SVM is the most used in the existing literature language processing, we compared its performance to other classifiers such as C4.5 decision tree, Artificial Neural Networks (ANN) and decision tree, and the classifiers performance is presented in table 2. We applied StringtoWordVector filter to convert string attributes to numeric in and applied the NgramTokenizer that consider the word order in a local context by n-gram features, helping to discover hidden patterns between words which represent a meaningful context. The experiments were conducted using 10-fold cross validation that has been previously explored for text analysis and it is a recommended approach for small datasets.

The explained experiments are implemented in Waikato Environment for Knowledge Analysis (WEKA 3.8) [33] and are executed on a PC with Intel® Core i7 processor,2.1 GHz speed and 8 GB RAM.

Table 2. Classification accuracy of three classifiers

Group	Filter	Learning Methods	Accuracy
Meta	**StringtoWordVector GramMaxSize=3 GramMinSize=1**	**C4.5**	92.7%
		MLP	93.4%
		SVM	95.1%

The results indicate better performance of SVM outperforming C4.5 and MLP. What stands out in these table is that the highest accuracy for classification of messages was 95.1% with the SVM classifier, whereas, the lowest was 92.7% with the C4.5 classifier. Based on this conclusion, we used the SVM classifiers with tuned parameters to classify all 2000 private messages that are used to populate the proposed ontology, as the objective of this research is not to design the best machine learning-based model to classify the messages, but to create a dataset to be used for ontology evaluation. The data instances that populated the ontology are presented in Table 3.

Table 3. Classification accuracy of the three classifiers

Message	%	Number of messages
non-criminal	31.2	624
criminal	68.8	1376

The classified messages are imported into the ontology using a Protégé plugin called Celfie used to map excel spreadsheets based on predefined rules. Since the aim of this paper is to present the performance of CISMO at detecting suspects from OSNs, a simple case study relying on the data of the mentioned forum is produced to represent a case of online crimes and the need of LEAS to obtain information about online activity of persons suspected for a criminal activity. The CISMO is a .owl file that captures the corpus (2000 RDF triples), that was then loaded in an Apache Jena Fuseki server to process dynamically the queries. This server provides a web service framework to support for querying through SPARQ.

Reasoning

The data uploaded in the CISMO ontology present represent facts and based on these facts, new ones can be defined. Rules can be written for categorizing instances in ontology, as a part of their natural belonging or as a part of other categories. Some reasoning examples in the context of crime detection and prevention could be:

- a person has sent a message to another user in Facebook, that might contain criminal content. This person is a possible source of suspect and digital evidences can be collected based on his online activity. If the message sent by a Social_Profile contains criminal content, this person should be categorised under *Criminal_Profile*, and especially in the *Suspects* class.
- a new digital evidence is created if: i) the message has *timestamp* and ii) the message has *msgSentBy* and iii) the message has *msgSendTo* and iv) the message *hascontent* classified as criminal content using NLP. Reasoning rules can be defined to categorize instances in ontology. If the digital evidence has *integrity* and *biometricartifacts*, then the digital evidence is categorized under the *Evidence_for_court* class.

Online Crime

We illustrate the potential of using CISMO in detecting online crime by presenting a case study using a small subset of Nulled.io, analysing the content of private messages leading to find out potential suspects engaged with cybercrime. In this scenario we prove the use of CISMO in analysing the forum data related to online crime and to support LEAs to collect information about persons online activity to better understand the behaviours of offenders and pathways into crime. For the illustration we show a relevant example (Q1) to the online crime case based on the messages exchanged between users of Nulled.io forum. In the second case, using some fake instances about a crime, we show the effectiveness of our solution followed by the query that processed the interested data, and part of the result set.

Q1: Users who have send messages with criminal content

```
PREFIX cismo: <http://www.cismo.org/v1#>
PREFIX rdf: <http://www.w3.org/...rdf-syntax-ns#>
SELECT ?senderProfile WHERE {
?msg cismo:msgSentBy ?senderProfile .
?msg cismo:msgSentTo ?receiverProfile .
?sender hasAccountIn ?receiverProfile.
?msg cismo:hascontent ?content .
VALUES ?content { cismo:Criminal}}
```

This query returns the member IDs of 1376 loaded in the ontology that has criminal content in the messages sent in the forum.

Q2: Which are the persons that checked in a place where a crime has happened and in the same time that the crime happened?

```
PREFIX cismo: <http://www.cismo.org/v1#>
PREFIX rdf: <http://www.w3.org/...rdf-syntax-ns#>
SELECT (concat(?personName," ",?personSurname) as ?Person) ?checkedIn  ?Crime ?cplace
?checkedInDate ?crimeDate
WHERE {
?person cismo:Person.
?person cismo:hasName ?personName.
?person cismo:hasSurname ?personSurname.
?person cismo:owns ?Socia_Profilel.
?person cismo:checkIn ?checkedIn.
?person cismo:checkInOnDate ?checkInDate.
?checkedIn cismo:hasName ?checkInPlace.
?ccase a  cismo:Crime.
?ccase cismo:happenedOn ?cplace.
?ccase rdfs:label ?Krimi.
?ccase cismo:happenedOnDate ?crimeDate. BIND(year(xsd:date(?checkedInDate)) as
?checkedInYear). BIND(month(xsd:date(?checkedInDate)) as ?checkedInMonth).
BIND(day(xsd:date(?checkedInDate)) as ?checkedInDay). BIND(year(xsd:date(?crimeDate)) as
?crimeYear) BIND(month(xsd:date(?crimeDate)) as ?crimeMonth)
BIND(day(xsd:date(?crimeDate)) as ?crimeDay) .
 filter(?checkedIn in (?cplace) && ?checkedInDay=?crimeDay &&
?checkedInMonth=?crimeMonth && ?checkedInDay=?crimeDay )}
```

Based on this query, we find out the persons suspected for a crime based on their *check_in* the place where the crime has happened and considering the time when the crime has happened, illustrated in fig. 7.

	Person	checkedIn	Krimi	cplace	checkedInDate	crimeDate
1	"Miri Xake"	crm:AliDemi	"Sherr ne lagjen Ali Demi"	crm:AliDemi	"2018-12-10T09:00:00"^^xsd:dateTime	"2018-12-10T12:00:00"^^xsd:dateTime

Fig. 7. Query results

5 Conclusions and further work

In this paper a new semantic tool suitable to gather digital evidences from criminal activities happening in OSNs, aiming to guarantee LEAs deeper insights into criminal activities has been introduced. The proposed ontology, called CISMO is an extended version of a previous ontology developed by authors, with new functionalities to model the core concepts related to crime cases in the police repositories, biometric data and digital evidences collected by OSNs, making possible for LEAs to classify crimes, investigate hidden crime patterns or predict future crime patterns.

Conducting some experiments with data coming from a hacker forum, we proved the effectiveness of CISMO, showing how this ontology can help LEAs to gather knowledge for crime investigation using natural language processing and machine learning to process messages shared in online platforms and applying reasoning rules, as semantic inferences. The proposed ontology addresses a wide range of aspects related to crime investigation, including concepts digital evidence collection, biometric data and elaboration of data from OSNs simultaneously. CISMO presented in this research should be accepted as a significant attempt to present a criminal ontology for social media. Moreover, it can be improved and extended with new concepts and relationships and it has to be tested with wider use cases.

Future work will consist in testing this ontology with real use cases obtained from police repositories and real data of OSNs and to evaluate the whole ontology covering a broader range of crimes, which will speed up the crime investigation processes but also made this process more efficient and accurate.

Acknowledgment

This work was carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme.

References

1. Angelopoulou, O., Vidalis, S. (2013) Towards 'crime specific' digital investigation frameworks, in Proceedings of 3rd International conference on Cybercrime, Security and Digital Forensics, 8-9 June, 2013, University of Cardiff, Cardiff, Wales, UK.
2. Arshad, H., Jantan, A., Hoon, G., Butt, A.: A multilayered semantic framework for integrated forensic acquisition on social media. Digital Investigation, Vol. 29, pp.147-158, (2019).
3. Asmae El. Kassiri and Fatima-Zahra Belouadha, 2017. The exSIOCInt Ontology: A SIOC Ontology Extension. Journal of Engineering and Applied Sciences, 12: 8160-8166.
4. Breslin, J., Bojars, U., Passant, A., Fernandez, S., Decker, S., 2009. Sioc: Content exchange and semantic interoperability between social networks. In: W3C Work. Futur. Soc. Netw., pp. 15e16.
5. Breslin, J.G., Harth, A., Bojars, U., Decker, S., 2005. Towards semantically-interlinked online Communities. Semant. Web Res. Appl 500e514.
6. Cosic, J., Cosic, Z., & Baca, M. (2011). An Ontological Approach to Study and Manage Digital Chain of Custody of Digital Evidence.
7. Ćosić,J., and Baća,.M., "Leveraging DEMF to ensure and represent 5ws&1h in digital forensic domain. Int. J. Comput. Sci. Inf. Secur. 13(2) (2015).
8. Goldbeck, J. (2007). The dynamics of Web-based social networks: Membership, relationships, and change. First Monday, 12 (11).
9. Golbeck, J. and Rothstein, M. M., "Linking social networks on the web with FOAF: a semantic web case study," ceeding AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence, vol. 2, pp. 1138-1143, 2008.
10. Hart, G.: The five W's: an old tool for the new task of task analysis Tech. Commun.,43(2), pp. 139-145, (1996).
11. Kalemi, E., Yildirim, S., Domnori, E., Elezaj, O., SMONT: an ontology for crime solving through social media. International Journal of Metadata, Semantics and Ontologies. vol. 12 (2/3), (2017).
12. Kanak, A., «Biometric ontology for semantic biometric-as-a-service (BaaS) applications: a border security use case," in IET Biometrics, vol. 7, no. 6, pp. 510-518, 11 2018.
13. Kassiri, A. E., & Belouadha, F.-Z.: A FOAF ontology 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 3056-3061.
14. Kastrati, Z., Imran, A. S., Yildirim-Yayilgan, S., Dalipi, F.: Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON. Social Computing and Social Media Lecture Notes in Computer Science, 148-157, (2015).
15. Loubiri, A., Obaid, A., & Sadat, F. (2013). An ontology-based system for social networking for health application support. 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, 2013, pp. 1152-1157.
16. Mason, Stephen (et al), Electronic Evidence, 2012, (Reed Elsevier (UK) Ltd.) (Robert J. Currie, Steve Coughlan, Chapter 9: Canada) at p. 293.
17. Mund, B., 2017. Social media searches and the reasonable expectation of privacy. Yale JL Tech 19, 238e238.
18. Noy, N. F., and Hafner,C. D., "The state of the art in ontology design: A survey and comparative review," AI Magazine, Vol. 18, pp. 53-74, 1997.

19. Oltramari, A., Henshel, D.S., Cains, M., et al.: 'Towards a human factors ontology for cyber security'. STIDS, 2015, pp. 26–33.
20. Oludare,A.I., Jantan,A., Omolara,A.E., Singh,M.M., Anbar,M., Zaaba, Z.F. : Forensic DNA proling for identifying an individual crime. International Journal of Civil Engineering and Technology, pp. 755{765, (2018).
21. Oriwoh, E. et al., "Internet of Things Forensics: Challenges and Approaches", 9th IEEE Int'l. Conf. Collaborative Computing: Networking Applications and Worksharing, pp. 608-15, 2013.
22. Park,H., Cho, S., and Kwon, H.-C. "Cyber forensics ontology for cyber criminal investigation," in Forensics in Telecommunications, Information and Multimedia. Springer, 2009, pp. 160–165. [Online].
23. Protégé 5.5.0. https://protege.stanford.edu/ / (Last access: September 2019).
24. SCOT Ontology Specification", Rdfs.org, 2016. [Online]. Available: http://rdfs.org/scot/spec/. [Accessed: 27- Jun- 2019].
25. Shaw, R., Troncy, R., Hardman, L.: LODE: linking open descriptions of events. The Semantic Web, Fourth Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009. Proceedings. Lecture Notes in Computer Science, vol. 5926, pp. 153–167. Springer (2009).
26. Sureka, A., Agarwal, S.: Learning to classify hate and extremism promoting tweets. In: Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint, pp. 320–320 (2014). IEEE.
27. Tserpes, K., Papadakis, G., Kardara, M., Papaoikonomou, A., Aisopos, F., Sardis E. and Varvarigou.T. (2012) "SocIoS: A Social Media Application Ontology", On the Move to Meaningful Internet Systems: OTM 2012 Workshops. Lecture Notes in Computer Science Vol. 7567, pp. 574-584.
28. Uschold, M., Gruninger, M. 1996. Ontologies: principles, methods and applications. Knowledge Engineering Review 11(2), 93-155.
29. Yaqoob I, Hashem IAT, Gani A, Mokhtar S, Ahmed E, Anuar NB, Vasilakos AV (2016) Big data: from beginning to future. Int J Inf Manag 6(6):1231–1247.
30. Zemmouchi-Ghomari, L., Ghomari, A.R.: Translating natural language competency questions into SPARQL queries: a case study. In: First Int. Conf. on Building and Exploring Web Based Environments. pp. 81–86. IARIA (2013).
31. Osathitporn, P., Soonthornphisaj, N., & Vatanawood, W., A scheme of criminal law knowledge acquisition using ontology. 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).
32. Warren G Kruse II and Jay G Heiser. Computer forensics: incident response essentials. Pearson Education, 2001.
33. "Weka 3: Data mining software in java." http://www.cs.waikato.ac.nz/ml/weka/. accessed: 10.05.2019.