Towards Designing a Knowledge Graph-Based Framework for Investigating and Preventing Crime on Online Social Networks

Ogerta Elezaj[1], Sule Yildirim Yayilgan[1], Edlira Kalemi[2], Linda Wendelberg[1], Mohamed Abomhara[1], and Javed Ahmed[1]

[1] Department of Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU), Norway
{ogerta.elezaj, sule.yildirim, mohamed.abomhara, javed.ahmed}@ntnu.no,
{lindawen@stud.ntnu.no}
[2] University of Tirana, Albania {edlira.kalemi@unitir.edu.al}

Abstract. Online Social Networks (OSNs) have fundamentally and permanently altered the arena of digital and classical crime. Recently, law enforcement agencies (LEAs) have been using OSNs as a data source to collect Open Source Intelligence for fighting and preventing crime. However, most existing technological developments for LEAs to fight and prevent crime rely on conventional database technology, which poses problems. As social network usage is increasing rapidly, storing and querying data for information retrieval is critical because of the characteristics of social networks, such as unstructured nature, high volumes, velocity, and data interconnectivity. This paper presents a knowledge graph-based framework, an outline of a framework designed to support crime investigators solve and prevent crime, from data collection to inferring digital evidence admissible in court. The main component of the proposed framework is a hybrid ontology linked to a graph database, which provides LEAs with the possibility to process unstructured data and identify hidden patterns and relationships in the interconnected data of OSNs.

Keywords: Crime, ontology, online social networks, digital evidence, knowledge graph, biometrics, security

1 Introduction

Over the last years, social networking – the most recent innovation in communication – has become an integral part of daily life for many people of all ages. It has changed the behavior and perception people have about the information shared online [22]. The influence of social networking sites, such as Facebook, LinkedIn, Twitter, Google Plus, etc. has exploded in a relatively short amount of time. The number of worldwide users is expected to reach some 3.02 billion by 2021[1], around a third of the Earth's entire population. As of March 31, 2019,

[1] https://www.statista.com/topics/1164/social-networks/

Facebook claims to have 2.38 billion monthly active users worldwide[2]. This expansion in connectedness among people on digital platforms has created a vast repository of information with potential value for LEAs in to use more efficiently for making informed decisions. OSNs have fundamentally and permanently altered the arena of detecting, preventing and solving of digital and classical crimes changing the field of crime investigations [1, 18].

Different types of digital crime evidence can be collected by OSNs. For example, digital evidence may come in the form of public posts, private messages, pictures, videos, tweets, geo-tagged content and, location-based data. Such open source of intelligence can be a reliable mine of evidence that can alter the outcome of a trial. LEAs can data mine social networking sites to identify victims, witnesses and perpetrators. Photographs, videos and other information that witnesses to crime post on social networks intentionally or unintentionally can be used as evidence later in an investigation. In general, a variety of information coming from OSNs can be utilized to solve many types of crime cases. Some instance of crime through open sources include cyberbullying and offenses primarily on Facebook and Twitter [26, 32], terrorism and burglars using social media to find targets [41].

Initially, LEAs started using social media as a communication channel for citizens to offer feedback and to participate in virtual consultations, thus building an online presence on main platforms to increase their legitimacy, transparency and trustworthiness. Nowadays, social media has expanded opportunities for surveillance by providing adequate tools for systematically gathering information from different digital platforms to track people's activities from the prospect of both suspects and victims [33]. Ericson and Haggerty [17] view police officers as "knowledge workers" rather than "crime fighters". LEAs conduct online surveillance to reconstruct events using knowledge management technologies and the corresponding knowledge about crime detection and prevention to assist with identifying crime trends. In legal proceedings, it may be considered good practice for police departments to successfully involve social media as an invaluable tool in their investigations.

The process of criminal investigation involves very high volumes of information that must be handled in a time-critical environment [12]. The success of investigation consequently depends on how the information is turned into evidence [15]. The information and knowledge obtained by OSNs is material that counts as potential evidence. Such material must be properly authenticated in order to be admitted into evidence. There are numerous examples of different courts of law having accepted social media content as digital evidence, leading to convictions and sometimes prison sentences [29, 39, 7].

However, despite the noticeable power of social media as evidence in legal proceedings, LEAs are always scrambling to keep up with new technologies and use intelligent systems to sift through massive amounts of raw social media content. Technically it is challenging handeling the flow of massively voluminous,

[2] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

heterogeneous, unstructured and multimedia content on OSNs. Also, to integrate, process and transform multimedia content on OSNs into intelligence used to identify suspects, locate witnesses and convict defendants. Monitoring social networks and transforming these vast amounts of unstructured data into actionable intelligence can be a daunting task. This overwhelming number of OSNs that serve as potential evidence are in high demand and deplete LEAs. Manual analysis, and keyword-based flagging are impractical, therefore an integrated system is needed to respond to the requirements of OSNs in the crime domain. Such system is required to handle massive volumes of data by supporting simplified automation and interoperability.

Existing solutions to meet requirements in the field of crime investigation are currently limited [7]. There is no exhaustive tool to support crime analysis and prevention on multiple social networks and that is capable of analyzing extremely high volumes of online content relating to chain of custody of digital evidence and their validation using biometric features. The existing approaches are narrow in terms of the integrating different data sources and hence fall short of providing solutions to larger-scale, more complex crime like organized crime and terrorism. Many of the existing digital forensics tools are in the form of individual tools that can only deal with partial aspects of crime investigation and offer limited features for investigations in complex environments [25]. Among the issues to be addressed is the development of automatic tools and techniques for analyzing vast amounts of data that have capability to gather digital evidence. Such solutions should provide visualization features and unified standards [11]. Furthermore, none of the previous frameworks for OSNs take account of the biometric aspect which is significant to the quality of the digital evidence collected. Biometrics technology plays a key role for LEAs during the investigation process stage of narrowing down the persons suspected of a crime. It is an important part of the digital evidence admitted in the court of law.

The contribution of this paper is twofold. In order to identify the research gap, we survey existing frameworks used for crime investigation in the context of OSNs content. Based on the discovered gap, we introduce an intelligent framework, which is a knowledge graph-based framework that is suitable for gathering digital evidence from OSNs, which may help LEAs to increase their analytical capabilities. The main components of the proposed framework are a hybrid ontology for collecting and integrating the unstructured social media content and a graph database used as a storage back-end to store the semantic data and perform efficient querying and storage. This hybrid ontology is an improved version of SMONT [23], which is a semantic-based ontology model originally developed by the authors with the main purpose of enriching an ontology with social media content.

The remaining part of the paper is organized as follows. Section 2 presents a review of literature relevant to existing frameworks used by LEAs to gather intelligence from online social networks to provide to legal bodies. In Section 3 discusses main challenges and solutions in order to build a knowledge graph-based framework. Section 4 introduces the proposed framework architecture and

its main components. The conclusion and future work are presented in Section 5.

2 Related work

Based on the meta-analysis and literature review, a summary of existing frameworks for crime investigation based on OSN content is presented in Table I. We conducted a keyword search in scientific databases of the keywords crime, ontology, biometrics, digital evidence and framework. Articles were filtered by title and abstract while articles that do not cover OSNs were discarded. Focus was on analyzing frameworks capable of handling social media content. The frameworks were evaluated based on the following criteria: - data sources used, type of crime, detection capabilities, prevention capabilities, biometric capacities, support for the collection of digital evidence, support for visual analysis, and methods used for crime detection and prevention.

According to our analysis, few semantic solutions exist that are specifically designed for investigating crime on social media. Moreover, no solution is sufficiently detailed to cover important aspect of crime investigation such as digital evidences and biometrics that have a crucial role in investigations.

Arshad et al. [7] introduced a multilayer semantic framework for OSNs that is based on a methodology of mapping multiple ontologies into a global one capable of integrating unstructured data drawn from different data sources. This framework lacks the required level of details about digital evidence gathering. Furthermore, the framework was not developed for processing biometric data.

The criminal ontology presented by Kastrati [24], SEMCON, is a simple ontology developed to identifying if a Facebook user is a possible suspect or not. The proposed model uses Facebook API to retrieve user' data and exploits it semantically and contextually. However, SEMCON does not cover all perspectives necessary to meet the complexity in criminal cases. To be usable, it needs to be extended to deal with all crime investigation aspects.

In [30] is presented a computational framework that focuses on physical evidence from a crime scene. This framework has three main components: - a physical biometrics ontology, a law enforcement ontology and several supporting stubs. The framework is interesting but does not emphasize crime investigation by OSNs. However, it must be mentioned that this is the only framework that cover biometric aspects of a crime scene in a semantic context. There is a lack of evaluation of the proposed framework.

Nouh et al. [34] presented an outline of a multipurpose cybercrime intelligence framework that helps LEAs investigate criminal cases and prevent crime. The proposed framework is composed of five layers: - data-handling, analysis, front-end, users and data-sources. It integrates different data analyses, such as Social Network Analysis (SNA), time series, content and sentinel analyses. The top-down approach is used for prevention based on various hypotheses of some event incidents and to detect and solve different crime cases. This solution does not

address digital evidence and biometric aspects. The concept of connected data in OSNs and modeling with graph technologies is not addressed either.

In [10] authors built a predictive model pertaining to reactions on Twitter in order to analyze the Woolwich, London terrorist attack that took place in 2013. The model use statistical methods and machine learning to predict the size and survival of information flow related to the terrorist attack. It lacks semantic capability and is limited to predictive capacity.

Cosic et al. (2015) developed an ontology to manage the digital chain of custody of digital evidences. This ontology based on the top-down methodology deals with the management of the chain of custody of digital evidence. Furthermore, it may serve as a method to expand on our ontology related to the digital evidence.

Table 1. Review of existing frameworks in crime investigation.

Authors	Framework description	Data Sources	Type of crimes	Detection	Prevention	Semantic Analyses	Biometric	Digital Evidence	Visual Analyses	Methods	Technology
[27]	Criminal network prediction model	UCINET cocaine smuggling	Drug Trafficking	X	✓	X	X	X	✓	Deep Learning	Graph
[7]	Semantic framework for social media	Social Media	General	✓	X	✓	X	X	✓	Frequency analysis and clustering	Ontology, Graph
[23]	Ontology for crime solving	Social media	General	✓	✓	✓	X	X	X	Ontology Reasoning	Ontology
[24]	Analysis of OSNs Posts to investigate	Facebook	General	✓	X	✓	X	X	X	Semantic and contextual data-mining	Ontology
[30]	Situation-Based Ontologies Focusing on Crime Scenes		Cybercrime	✓	X	✓	✓	✓	X	Situation Management	Ontology
[3]	Analysis and detection of microblogging spam	Twitter	Cybercrime	✓	X	X	X	X	X	Machine learning	Graph
[37]	Criminal Network Analysis Using Big Data	-	General	✓	✓	X	✓	X	X	Machine Learning	Hadoop, Graph
[34]	Cybercrime Intelligence Framework	OSN	General	✓	✓	X	X	✓	✓	SNA, Space and Behavior Analyses	Relational Database
[10]	Modelling the Social Media Reaction	Twitter	Terrorism	X	X	X	X	X	X	Regression Machine learning	-
[2]	Surveillance of Instant Messages	Social Media	General	✓	X	✓	X	X	X	Association Rule Mining	Ontology, knowledge database
[14]	Ontology DEMF	-	General	X	X	✓	X	✓	X	-	Ontology

According to this review and to the best of our knowledge, it is concluded that none of the existing frameworks cover digital evidence collection, biometric data and elaboration of data from OSNs simultaneously. The previous frameworks are mostly not generalized but are platform based, meaning they are capable

of handling data sourced from one specific social network platform. Moreover, among all frameworks analyzed, only one is graph-based, which is a requirement when OSNs are to be used in criminal cases investigation. Utilizing of graph-based technology is important in analyzing the characteristics and behaviors of suspects or criminals as well as the structure of communities or sometimes an overall network. The graph representation helps LEAs understand a criminal network structure, and identify the cliques, groups and key players in a network [31].

3 Identified major challenges and solutions

In this section, we identify the main challenges and proposed solutions that facilitate LEAs investigating crimes happening in OSNs.

Challenge 1: Developing an appropriate model to organize and integrate the massive volume and different data types obtained from OSNs.

OSNs contain massive volumes of content and linkage data, which can be utilized by LEAs in crime investigation. Generally, the data can be divided into structured and unstructured data. Unstructured data is a textual content, known as User Generate Content (UGC) for a particular user. UGC is often in the form of text, but it can contain images, videos or other type of data. On the other hand, structured data are modeled by graph data models, where the entities are presented as vertices (e.g., people or things) and edges (i.e., relationships of vertices). The OSNs data is heterogeneous and also accompanied with different properties, such as the time stamp and the location related to a specific user activity, which means that processing and managing this kind of data is a bigger challenge compared to other data sources such as web pages or blogs. It is foreseen that 80 percent of worldwide data will be unstructured by 2025 [3]. If LEAs are struggling to manage their unstructured data now, they are going to find it difficult to cope with the increasing volume of unstructured data over time, to turn it into a more structured format.

Previous studies observed that Semantic web frameworks provide a graph model (RDF), a query language (SPARQL) and definition systems (OWL) to efficiently represent and manage heterogeneity of OSNs data [40]. Semantic Web technology uses ontology to model an abstract view of a specific real domain. Only few studies have concentrated on using semantic technologies to integrated data coming from OSNs and mostly offer general solutions not related to crime investigation [8].

In 1995, Gruber [19] originally defined the notion of an ontology as knowledge engineering that use explicit specification of conceptualization. The advantages of ontologies are: a) use of common language; b) manage of unstructured data; c) enable reuse of domain knowledge and d) use of inference steps utilizing ontology reasoning [40]. The most significant advantage that ontologies might bring to the domain of crime investigation is the ability to support the integration and processing of unstructured data received from OSNs.

[3] https://www.idc.com/

The Web Ontology Language (OWL) is one of the most commonly used language to formalize ontologies of different domains and to describe their relation by converting to descriptive logic. In the proposed framework, the crime domain ontology is developed using an OWL editor tool developed by Stanford University, named Protégé [38]. This tool is a JAVA-based open source ontology editor which is compatible with different platforms. OWL uses the Resource Description Framework (RDF) of classes and properties. The framework contains a hierarchical description of conceptual things in the crime domain. Individuals are instances of predefined classes, and properties of each class describe attributes of the concepts.

According to literature, there exists different methods used for developing ontologies. These methods are divided into two main groups: evaluation prototype models and experience-based methods [9]. In the proposed solution, we adopted the 101 method which contains seven defined steps to develop an ontology for a specific domain [35]. First, we develop a conceptual ontology – which is an abstract view of classes – defined and arranged in a taxonomic presentation with sub-classes and super-classes, properties and facts of properties. In the second stage, the physical ontology is developed by using Protégé. Based on the data collected in the acquisition layer, we create instances of each class to model the crime domain. The crime ontology covers classes, data properties, object properties, individuals and relationships.

Challenge 2: Developing an appropriate storage technology to handle complex and dynamic relationships in highly connected OSNs to generate new knowledge for LEAs.

While the use of ontologies in the crime domain give big advantages related to knowledge representation and extraction, it also poses problems. The main key challenge is related to the storage of the data and searching for the relevant information in a big data environment [16]. To solve these problems, nowadays, graph databases are being widely and intensively used for storing and querying data for OSNs. The graph databases overcome the limits of traditional databases for storing and managing data represented as graph-like data. As the usage of social networks is increasing rapidly, storing and querying data for information retrieval by LEAS is critical. This due to the characteristics of social networks like dynamic structure, highly volume, velocity and interconnectivity of data. Relational databases struggle to handle the volume and velocity of data generated in big data environment [5]. Also, an other reason that graph databases are used to store OSNs data is the fact that social networks are modeled as graphs.

Moreover, Not only SQL (NoSQL) is one of the main solutions for storing and processing OSNs data [6]. Its main characteristics is that it is schema-free. Different open-sources NoSQL are available as low-cost solutions. NoSQL is more efficient in comparison to relational databases because it ensures efficient big data storage, provide high performance, high volume, high velocity and can handle complex data structure [20]. NoSQL databases are divided into four types: - graph database, key-value store, column store and document database. As our scope is to model the data received from OSNs, we employed graph databases,

as the most popular storage technology used for analyzing data from OSNs [6], assuring natural modeling of their networks.

We employed a NoSQL open source graph database, named Neo4j, released in 2007 [13]. Neo4j is known as "world leading graph database" and is one of the most popular graph databases, characterized by robustness and high performance [42]. It scales billions of nodes and relationship in a network. Therefore, the entire interconnected data obtained from OSNs, needed by LEAs for investigation, can be stored and managed using Neo4j.

The aim of the proposed framework is to extend the SMONT ontology developed by Kalemi et al [23]. The top-level classes identified in the framework ontology are: Agent, Crimes, Crime Case Solving, Social Networks, Biometric Artifacts and Institutions as shown in Figure 1. Agent class represents information about persons, groups or organizations whose data are collected by police reports/evidence streams and from OSNs.

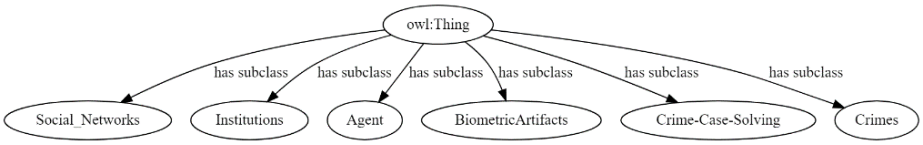

Fig. 1. Top hierarchy classes of crime ontology

Crime class describes the types of crimes based on widely accepted classification by law and jurisprudence. Social Networks represents all data collected by OSNs used for investigation. Biometric artifacts class represents physical biometric features (e.g., fingerprint, iris) about persons that committed a crime or anyone who is suspected. Institutions represent a class that contains information about institutions involve in crime investigations like LEAs, banks, courts, insurance companies etc. Crime-Case-Solving contains information about cases that have not been solved. Using reasoning rules, machine learning and SNAs the information of all the classes of ontology is contributing to provide instances in the Digital Evidence that is a sub-class of Crime-Case-Solving. Class dependencies are as shown in Figure 2.

Challenge 3: Developing an appropriate method to automatically extract and visualize the relevant knowledge.

The increase demand for structured knowledge in investigating crimes has created considerable interest for LEAs in crime analytical techniques in order to provide LEAs better insights into criminal networks. As most of the social networks are big graphs with unpredicted volume, velocity and variety, LEAs face challenges related to effective information search and crime analysis. LEAs require to deploy automatic procedures for analysis related to detection and prediction of crime incidents.

To enhance national security, LEAs and intelligent agencies collect massive amount of data which must be transformed into information, knowledge and

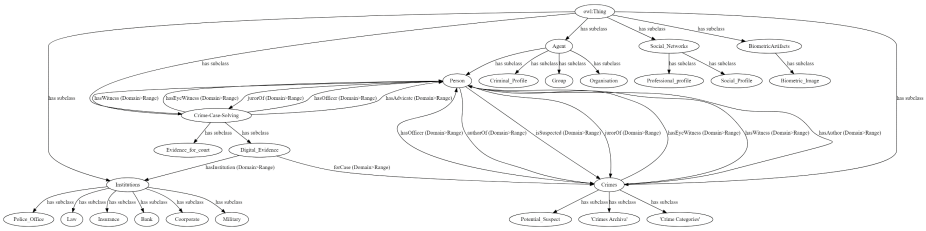

Fig. 2. Class dependencies of crime ontology

intelligence. As LEAs struggle to manage and process this enormous volume of data by traditional methods, the use of machine learning (ML) discipline offers efficient solutions for crime detection and prevention in large crime datasets [36]. Based on ML, LEAs can develop and deploy models to classify crimes, investigate hidden crime patterns or predict future crime patterns. Different classification algorithms are nominated for crime analysis, crime prediction, and to get insight into potential crime hotspot areas [4].

A common problem for LEAs during investigation is to analyze groups of people involved in organized crime and dark networks. In order to analyze the dynamic structure of criminal networks, LEAs must employ quantitative measurements of SNA, which provides models and techniques to analyze OSNs based on graph theory and visualization tools. SNA methods has the ability to find out leaders of the networks, observe network changes over time and model diffusion processes. Also, SNA methods can be used to discover the leaders of a criminal network based on centrality and prestige measures of direct relations. The centrality and prestige are linked with the members that are extensively linked or involved with other members. Finding these leaders in such networks and removing them may defragment the criminal network or disrupt it.

The proposed framework based on machine learning methods, deep learning and SNA can provide information to LEAs in order to predict different crime categories including whether a person or organization will or motivated to commit a crime, crime target, time that a new crime might be committed, the crime location and type of crimes etc. Thus, combining machine learning techniques, SNA analyses and ontologies reasoning inferences, the solution aims to improve prediction performance helping LEAs to predict different types of crime that might occur in a particular geographical location and in a given time period.

Challenge 4: Developing an appropriate method for preservation of digital evidence.

In crime investigation process, LEAs must collect and infer digital evidence that require authenticity and non-repudiation properties to rely on in the court. Digital evidence gathering and analyses requires special procedures and techniques to be used and accepted as evidences in the court. The goals of the systems that produce digital evidence is the maintenance of the chain of custody, which in legal context is the documentation of the order of handled items of evidence during a crime investigation. Since the process of collection, validation, preser-

vation and documentation of digital evidences is complex and dynamic, chain of custody should be kept. Some common problems faced by LEAs in maintaining the integrity of evidence with chain of custody include data integrity, modification of elements of digital evidences and access control for the storage.

The proposed framework aims to ensure the management of the chain of custody from the starting point of a new criminal case. To maintain the chain of custody, in our framework, we are based on the 5W+H investigative model (Who, What, When, Where, Why and How) [21]. In the crime ontology (top-down methodology), we have created sub-classes under class of Digital Evidence. The elements of the sub-class Digital Evidence can respond to the 5W+H aspects of the chain of custody. Combining two classes (Digital Evidence and Biometric), the integrity of chain of custody is strength and the framework can produce biometric digital evidence which are admissible in court for prosecution of offenders and attackers.

Challenge 5: Developing an appropriate method to acquire data from social networks.

In general, there are three different data types that can be retrieved from OSNs. The first, content data includes user profiles, photos, videos etc., which are mostly unstructured data. The second is behavior data which is classified into three different categories: user-user behavior – interaction between two individuals such as following or sending a private message; user-entity behavior – interaction between a user and an OSNs entity such as liking or writing a post in Instagram and user-community behavior – interaction between a user and an online community such as participating in community discussions. The third type is network structure data that consists of explicit hyperlinks between users and their content.

Social network like Facebook provides application program interface (API) and query languages which are key part for researchers in academia and enterprises too [26]. Recently OSNs platform have clearly restricted the access on data that developers can retrieve. For example, Facebook, after the Cambridge Analytica scandal, changed several APIs in order to protect user data. A string called Facebook access token that identifies a user, or an application is used to make API calls and now based on the new rules. All these requests must be reviewed and approved by Facebook. Also, Twitter hands down new strict rules restricting the volume of data that third-party developers can get access to. For example, as for June 2018, the Twitter API can return only the most recent 3200 tweets and during a single request only 200 tweets are return.

To overcome these limitations, the proposed framework combines API and web crawlers to feed real time data and to optimize quality for each data sources. Crawlers are used to extract the information that cannot be collected automatically by API. Also, the framework must address the issue of metadata, which are fundamental important to be collected along with respective data.

4 Knowledge graph-based framework architecture

Aimed at addressing the main challenges in gathering intelligence from OSNs in the crime domain, this paper presents a knowledge graph-based framework incorporating semantic web and graph database. The architecture of the proposed framework is given in Figure 3. It is composed of 4 layers, data acquisition, knowledge creation and semantic enrichment, knowledge extraction and application layer.

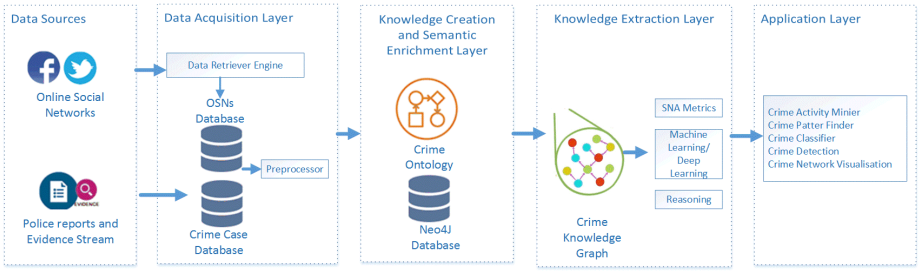

Fig. 3. The architecture of the knowledge graph-based framework

Two data sources have been identified for crime investigation, reports and evidence streams that exist in LEAs repositories and OSNs data. The data acquisition layer contains a data retriever engine capable to feed real time data from OSNs combining API and developed crawlers. For different social networks, different API plugin are included in the engine. To deal with rate limits of API request per hour and per user, crawlers are used. This layer includes data preprocessing focused on removing errors and inconsistency from the data, imputing missing data, and data integration. As data coming from OSNs are user generated data and their quality varies from valuable data to rubbish, data preprocessing is a crucial process. Having good quality of data is more important than having in place efficient machine learning algorithms that run in poor datasets. The framework should be capable to automatically extract information from multimedia content taken from OSNs and format it. This process includes different tasks related to audio and video segmentation and transcriptions,image processing to extract logos, weapons, biometric features etc. The framework can process files of diffrent formats such as audio file, image file, video file, text files, that are stored locally in LEAs repositories or can optionally be extracted by specific OSNs.

The next layer is the knowledge creation and semantic enrichment, where the data are modeled based on the developed crime ontology and are stored in a Neo4j graph database. A mapping schema of the Neo4j data model and the crime ontology that is an RDF directed graph is required in order to preserve the details. The RDF graph triples are composed by a subject, a predicate and an object, where the subject and predictor are resources and the object can be a

resource or a literal. Each resource is identified by a Uniform Resource Identifier (URI). All subjects of triples are mapped to nodes in the Neo4j graph. If the object is literal, predicates are mapped to node properties, otherwise if the object is a resource, predicates are mapped to relationships.

Knowledge extraction layer covers application of machine leaning techniques, SNA and ontology reasoning to analyze the graph data to extract the required knowledge. Using machine learning algorithms, criminal behavior is modeled. Based on anomaly detection, the system alert LEAs for unusual patterns or behaviors. All the methods applied at this layer will be tuned to avoid false positive alarms (FP). If the system has a high FP alarms rate, it makes it very challenging for LEAs due to the high workload required. Thus, it is necessary to train appropriate algorithms in order to obtain high detection rate and to ensure a low FP rate. SNA metrics are used to discover strategic members who belong to criminal networks.

The last layer, the application layer is the dashboard supporting user to customize different processes related to crime investigation and prediction.

5 Conclusions and further work

We have conducted a review of the literature on the use of OSNs data in crime detection and prevention. The aim of this review was to analyze existing intelligent crime solving frameworks and to identify various challenging factors that restrict the use of OSNs by LEAs in the preventive policing of criminal activities. Lack of efficient models to organize and integrate the massive volume and different data types coming from OSNs and the increased demand for structured knowledge in investigating crimes, were found to impede the implementation of efficient intelligent crime frameworks for LEAs.

Existing crime frameworks do not consider digital evidence collection, biometric data and elaboration of data from OSNs simultaneously such that existing solution remain unsuitable to meet LEAs requirements in the process of crime investigation. Solutions that can efficiently collect, store and query the unstructured, high volume, velocity and inter-connected data and guarantee LEAS deeper insight into the criminal activities are demanded. Based on the identified challenges and the requirements of LEAs we have introduced our initial design of a knowledge graph -based framework for investigation and preventing of crime on OSNs. This framework is a hybrid ontology linked to a graph database, which provides LEAs with the possibility to process unstructured data and identify hidden patterns and relationships in the interconnected data of OSNs with the focus on crime investigation and prevention.

Future work will consist in fully implementation of this framework. We aim in developing real use cases obtained from police crime cases and real data of OSNs and to evaluate the whole system and the performance of prediction methods covering a broader range of crimes.

ACKNOWLEDGEMENTS

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

References

1. Abdalla, A., & Yayilgan, S. Y.: A review of using online social networks for investigative activities. Social Computing and Social Media Lecture Notes in Computer Science. pp. 3–12, (2014)
2. Ali, M. M., Mohammed, K. M., Rajamani, L.: Framework for surveillance of instant messages in instant messengers and social networking sites using data mining and ontology. Proceedings of the 2014 IEEE Students Technology Symposium. (2014)
3. Almaatouq,A., Shmueli, E., Nouh,M., Alabdulkareem, A., Singh,V., Alsaleh,M., Alarifi, A., Alfaris,A. and Pentland, A.: If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts. International Journal of Information Security, pp. 1–17, (2016)
4. Almanie, T., Mirza, R., Lor, E.: Crime prediction based on crime types and using spatial and temporal criminal hotspots. International Journal of Data Mining & Knowledge Management Process **5**(4). (2015)
5. Alsubaiee, S., Carey,M.J., Li, C.: LSM-Based storage and indexing: an old idea with timely benefits ACM Workshop on Managing and Mining Enriched Geo-Spatial Data - GeoRich15, pp. 1-6, (2015)
6. Appel, A. P., Moyano, L. G.: Link and graph mining in the big data era. Handbook of Big Data Technologies, pages 583-616. Springer. (2017)
7. 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)
8. Breslin,J., Bojars,U., Passant,A., Fernandez,S., Decker, S.: Sioc: Content exchange and semantic interoperability between social networks. W3C Workshop on the Future of Social Networking, pp. 15-16, (2009)
9. Brusa,G., Caliusco,M. L., Chiotti,O.,: A process for building a domain ontology: An experience in developing a government budgetary ontology. 2nd Australas. Workshop Adv. Ontologies, Hobart, TAS, Australia, 72,pp. 7–15, (2006)
10. Burnap, P., Williams, M.L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., Voss, A.: Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. Social Network Analysis and Mining 4, p. 206, (2014)
11. Caviglione, L., Wendzel, S., Mazurczyk, W.: The Future of Digital Forensics: Challenges and the Road Ahead. IEEE Security and Privacy, (6), pp.12-17, (2017) Investigation, Vol. 29, pp.147-158, (2019)
12. Chen, H., Schroeder, J., Hauck, R. V. et al.: COPLINK Connect: Information and Knowledge Management for Law Enforcement. Decision Support Systems **34**(3): 271–285, (2003)
13. Chen, Y., Chen, Y.: Decomposing DAGs into spanning trees: A new way to compress transitive closures. 2011 IEEE 27th International Conference on Data Engineering, Hannover, pp. 1007-1018, (2011)
14. Cosic, J., Baca,M.: Leveraging DEMF to Ensure and Represent 5ws&1h in Digital Forensic Domain. International Journal of Computer Science and Information Security, **13**(2), (2015)

15. Dean, G., Fahsing, I.A., Gottschalk, P.: Profiling police investigative thinking: a study of police officers in Norway. International Journal of the Sociology of Law, Vol. 34, pp.221–228, (2006)
16. Elbattah, M., Roushdy, M., Aref, M., Salem, A. M.: Large-scale ontology storage and query using graph database-oriented approach: The case of Freebase. IEEE Seventh International Conference on Intelligent Computing and Information Systems,pp 39–43, (2015)
17. Ericson, R.V., Haggerty, K.: Policing the Risk Society, University of Toronto Press. Toronto, (1997)
18. Europol. Crime in the age of technology, Europol unclassified –Basic protection level, The Hague 12.10.207, EDOC#924156v7, (2017)
19. Gruber, T. R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, **43**(5–6), 907–928, (1995)
20. Gupta, A., Tyagi, S., Panwar, N., Sachdeva, S., Saxena, U.: NoSQL databases: Critical analysis and comparison. 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), (2017)
21. Hart, G.: The five W's: an old tool for the new task of task analysis Tech. Commun., **43**(2),pp. 139-145, (1996)
22. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys **47**(4), (2015)
23. 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)
24. 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)
25. Khan, S., Gani, A., Wahab, A. W., Shiraz, M., Ahmad, I.: Network forensics: Review, taxonomy, and open challenges. Journal of Network and Computer Applications, 66, 214-235, (2016)
26. Kokkinos, C. M., Baltzidis, E., Xynogala, D.: Prevalence and personality correlates of Facebook bullying among university undergraduates. Computers in Human Behavior, 55, 840-850, (2016)
27. Lim, M., Abdullah, A., Jhanjhi, N., Supramaniam, M.: Hidden Link Prediction in Criminal Networks Using the Deep Reinforcement Learning Technique. Computers, 8(1), 8, (2019)
28. Lomborg, S., Bechmann, A.: Using APIs for data collection on social media. The Information Society, **30**(4), 256–265, (2014)
29. Mason, S. (et al): Electronic Evidence, Elsevier (UK) Ltd.)(Robert J. Currie, Steve Coughlan, Chapter 9: Canada) at p. 293, (2012)
30. Mcdaniel, M., Sloan, E., Nick, W., Mayes, J., Esterline, A.: Ontologies for situation-based crime scene identities. SoutheastCon 2017 pp. 1-8, (2017)
31. Mena, J.: Investigative Data Mining for Security and Criminal Detection. Butterworth–Heinemann, (2003)
32. Moore, K.: Social media 'at least half' of calls passed to front-line police. BBC Radio 4's Law in Action. (2014) [Online]. Available: http://www.bbc.co.uk/news/uk-27949674.
33. Murphy,J., Fontecilla, A.: Social media evidence in government investigations and criminal proceedings: a frontier of new legal issues Rich. JL Tech., XIX, pp. 1-30, (2013)

34. Nouh, M., Nurse, J. R., Goldsmith, M.: Towards Designing a Multipurpose Cybercrime Intelligence Framework. 2016 European Intelligence and Security Informatics Conference (EISIC), (2016)
35. Noy, N., McGuinness, D. (2000). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, (2000)
36. Oludare,A.I., Jantan,A., Omolara,A.E., Singh,M.M., Anbar,M., Zaaba, Z.F. : Forensic DNA profiling for identifying an individual crime. International Journal of Civil Engineering and Technology, pp. 755–765, (2018)
37. Pramanik,M. I., Zhang,W., Lau,R. Y., Li,C.: A framework for criminal network analysis using big data. e-Business Engineering (ICEBE), pp. 17–23, (2016)
38. Protégé 5.5.0. https://protege.stanford.edu/ / (Last access: June 2019)
39. Recchia, M. : Court of Appeals Declares Facebook"Private Data"and Other Social Media Subject to Discovery, New York Law Journal, www.law.com.
40. Turnbull, B., Randhawa, S.: Automated event and social network extraction from digital evidence sources with ontological mapping. Digital Investigation 13, 94-106, (2015)
41. Weimann, G.: New terrorism and new media. Washington, DC: Commons Lab of the Woodrow Wilson International Center for Scholars, (2014)
42. Williams, D. W., Huan, J., Wang, W.: Graph Database Indexing Using Structured Graph Decomposition. IEEE 23rd International Conference on Data Engineering,pp. 976-985, (2007)