Machine Learning Models on Criminal Networks (MLMoCN): Artificial Intelligence to Disentangle Crime

EDUARDO SALCEDO-ALBARÁN
LUIS JORGE GARAY SALAMANCA
JOSÉ CANO-MELANI

VORTEX

No. 65

Vortex Working Paper No. 65.

Abstract

Artificial intelligence ("AI") is an outstanding technology for analyzing massive volumes of structured and unstructured data. Within the theoretical and methodological framework of Criminal Network Analysis ("CNA"), AI assists in classifying and extracting structured and unstructured data, and generating Criminal Network Graphs ("CNG"). To accomplish these tasks, Machine Learning Models must be trained with real/empirical data that describes the characteristics of the nodes/agents involved in the criminal networks and their interactions. This paper discusses the characteristics of the resulting models, herein defined as Machine Learning Models on Criminal Networks (MLMoCNs), and the prospects and obstacles for applying the most sophisticated AI techniques to CNA.

Introduction

1

Criminal Network Analysis (CNA) is a theoretical and methodological approach based on Social Network Analysis (SNA) to understand the structure and dynamics of complex and systemic illicit phenomena (Basu & Sen, 2021; Cavallaro, et al., 2020; Morselli C., 2008; Morselli C., 2012). In the general theoretical framework of SNA (Degenne & Forsé, 1999; Wasserman & Faust, 1994; Knoke & Yang, 2019; Borgatti, Mehra, Brass, & Labianca, 2009), a node represents any abstract entity; however, when Criminal Network Analysis (CNA) is applied, it is useful to represent social entities with moral agency such as individuals, companies, and public organizations. In this sense, CNA allows identifying micro characteristics of specific nodes or social agents and their interactions, as well as structural macro characteristics, such as the most relevant categories of those social agents and their interactions, characteristics of sub-networks, among other dynamics. Criminal Network Analysis has been applied to analyze several outstanding characteristics of illicit phenomena that includes money laundering, corruption, and various types of criminal markets (Morselli C., 2008; Morselli C., 2012; Basu & Sen, 2021; Garay Salamanca & Salcedo-Albarán, 2012; Garay Salamanca, Salcedo-Albarán, & Macías, 2018d; Garay & Salcedo-Albaran, 2012c; Garay-Salamanca & Salcedo-Albarán, 2012a; Garay-Salamanca & Salcedo-Albaran, 2015; Garay-Salamanca, Salcedo-Albarán, & Duarte, 2017).

Bearing this in mind, the concept of “node/agent” has been used to highlight the moral agency of the social entity analyzed through CNA (Salcedo-Albarán & Garay-Salamanca, 2016). As a complex system (Sayama, 2015) influenced by surrounding stressors (Taleb, 2012), it is essential to understand the dynamics of interaction between the criminal networks' inner and external components, which Criminal Network Analysis particularly facilitates.

This document aims to define some theoretical basis of Machine Learning Models on Criminal Networks (MLMoCN), an essential tool for reinforcing the descriptive and predictive capabilities of Criminal Networks Analysis. This document consists of five parts. The first part is this introduction. The second part discusses how Artificial Intelligence (AI) amplifies the scope and quality of Databases of Interactions, a fundamental element for CNA. The third part discusses how the size and the randomization of training datasets are critical to empower the predictive performance of any AI model. In the fourth part, the distinction between Machine Learning Models (MLM) and Large Language Models (LLM) is analyzed, highlighting the necessity of providing "specific-and-focused-related" data to any AI model used for specialized expert analysis, such as CNA. The conclusions are discussed in the fifth part.

Artificial Intelligence To Generate Databases of Interactions

2

The first step to apply CNA is to elaborate a Database of Interactions (DoI), which is a set of three "elements" or "semantic entities" as it would be referred to in the context of Natural Language Processing (NLP): (i) The emitter node/agent, (ii) the interaction in which a scarce resource is transferred between nodes/agents, and (iii) the receiver node/agents. Most software available for visualizing social network graphs recognizes and uses this type of DoI structure -usually called "edge list"- and an adjacency matrix as typical ways to represent a network for visualization (Armstrong, Johnson, & McCulloh, 2013).

As previously stated, when CNA is applied, nodes often represent individuals and private and public organizations; therefore, the entities -nodes/agents and interactions- that compose a DoI can be extracted from structured and unstructured data. In this case, unstructured data usually refers to Natural Language Text (NLT) from media, judicial, or administrative sources, and structured data refers to qualitative and quantitative data previously categorized and tabulated.

When using unstructured NLT data as the source for CNA, extracting, tagging, and systematizing the semantic entities for building a DoI need careful human scrutiny for analysis; therefore, it is time-consuming. For this reason, AI models trained to identify, extract, and tag semantic entities can exponentially increase the amount of data sources and speed rates for modeling and analyzing criminal networks. These AI models are pre-trained machine learning models that must later be "fine-tuned" with criminal networks based upon specific and empirical "real" data, a process discussed below. The resulting "fine-tuned" model is a "Machine Learning Model on Criminal Networks" (MLMoCN).

MLMoCNs help generate Criminal Network Graphs (CNG) and facilitate CNA. The CNG generation includes a critical subtask in which any unstructured NLT source is "structured" by identifying, extracting, and classifying every semantic entity, a process often referred to in the context of Natural Language Processing (NLP) as Document Information Extraction (DIE) through Named Entity Recognition (NER) task (Marrero, Sánchez-Cuadrado, Morato, & Andreadakis, 2009). Then, the resulting extracted, tagged, and mapped semantic entities are used to arrange each DoI entry, automatically generating a CNG.

Named Entity Recognition (NER) is an essential and initial task in various approaches for text localization and transcription (Carbonell, Fornés, Villegas, & Lladós, 2020) and is implemented in the MLMoCN that sustains VORISOMA, a web application that simplifies tasks related to Criminal Network Analysis (CNA). As described by Cano-Melani, Salcedo-Albarán & Garay-Salamanca (2022), "through an initial knowledge base, VORISOMA AI (…) executes a first-order NER task to identify, extract and label the entities/nodes and the entities/verbs that conform a relationship/interaction. During the first-order NER task, some layers of Entity Linking (EL) based upon syntax tree analysis (Wang & Han, 2015) are executed to categorize the identified entities/nodes and entities/verbs; a process that has also been defined as Named Entity Disambiguation (NED) for entities identification task" (Oliveira, y otros, 2021; Wang & Han, 2015; Trani, Ceccarelli, Lucchese, Orlando, & Perego, 2018; Shehata, 2022).

Large Language Models and Criminal Network Analysis

3

Machine Learning Models on Criminal Networks (MLMoCN) are neural networks trained with "real" datasets that contain specific information describing characteristics of criminal networks' nodes/agents and interactions. In terms of semantic entities, this includes names of locations, individuals, public and private organizations, as well as the interaction verbs.

Three types of datasets can be used for training Machine Learning Models, depending on the type of data: (i) "real," when it consists of empirical data that describes the phenomenon under analysis; (ii) "synthetic," when the data is artificially created, or (iii) "hybrid" when the dataset consists of some "real" empirical elements and others "artificially" created to replicate the randomized characteristics of the "real" elements. As discussed below, the random distribution of elements comprising a "hybrid" dataset is critical for increasing the MLM's performance (Jordon, Wilson, & van der Schaar, 2020).

For instance, if the purpose is to train MLM that identify and detect characteristics of criminal networks operating in country A, but the training dataset only includes information on a criminal network operating in the city A-C1, the result will be an MLMoCN specialized in detecting, extracting, and tagging illegal interactions in the city A-C1, but it will probably have a biased and inefficient performance on detecting, extracting, and tagging criminal interactions in the cities A-C3, ….C*n*. More importantly, since the MLMoCN only contains "real" empirical data about the city A-C1, it lacks randomized data about the cities A-C2, A-C3, …, A-C*n*. In this sense, it is desirable to preserve high levels of randomized data elements for all and each unit of analysis -in this case, cities A-C1, A-C2, …. A-C*n*- in a training dataset.

In the case of datasets consisting of only "real" data, the elements reflect empirical and randomized characteristics of the analyzed phenomena at each unit of analysis; therefore, a higher randomization inherent to empirically descriptive and real data at each unit of analysis, allows a more robust predictive capabilities of the model than if only synthetic and less randomized data is used. In this sense, as stated in Cano-Melani, Salcedo-Albaran, & Garay-Salamanca (2023), "it is expected that approaches for generating synthetic data address the reduction of differences in the distribution of real versus synthetic or expanded [hybrid] elements".

The size of the dataset is another critical factor for improving any MLM's performance. Therefore, considering that larger datasets improve the predictive MLMoCN's capabilities and that "real" datasets informing of criminal network characteristics are scarce, "hybrid" data in which synthetic elements "imitate" randomized elements has proven a good approach to training MLMoCN. Procedures for generating synthetic datasets have been discussed in Cano, Salcedo-Albaran & Garay-Salamanca (2023).

Pre-Trained Large Language Models and Fine-Tuned Machine Learning Models on Criminal Networks

4

In 2022, Large Language Models (LLM), a particular type of MLM trained with vast amounts of data and consisting of large amounts of tuned hyper-parameters, gained attention among the public for their performance in recognizing, interpreting, and generating NLT. Often, LLMs exceed the 10-billion parameter amount, and some reach 100 billion parameters (Simon, 2023). Although this attention mainly relates to the generative performance of Chat GPT-3 and subsequent versions, by 2023, there are several other MLMs with outstanding generative performance.[**1**] Considering the large levels of training data and tuned hyper-parameters, it is expected that LLMoCNs have a better capacity to generate a DoI than previous MLMoCNs with less training data and hyper-parameters.

Although this attention mainly relates to the generative performance of Chat GPT-3 and subsequent versions, by 2023, there are several other MLMs with outstanding generative performance.

1. Generative AI models not only work on NLT but also images.

Considering the large levels of training data and tuned hyper-parameters, it is expected that **LLM**oCNs have a better capacity to generate a DoI than previous **MLM**oCNs with less training data and hyper-parameters.

It must be noted that **MLM**oCNs and **LLM**oCNs are pre-trained AI models, which means that in both cases, the starting point are MLMs or LLMs with an initial NER capacity -to identify, extract, and predict NLT. However, such initial "pre-trained" capacity is *general,* radically different from the case of AI models that aim to identify and extract semantic entities explicitly related to "real" criminal networks' characteristics. For instance, regarding the initial first-order NER capacity before fine-tuning, GPT 4 performs slightly better than GPT 3, although this capability -without additional second-order NER with more complex semantic analysis- performs better with the TensorFlow's Universal Sentence Encoder. [**2**]

Therefore, training the pre-trained model with "real" data on criminal networks' characteristics is essential to obtain MLMoCNs or LLMoCNs. This training stage is usually called "fine-tuning." The difference between the "pre-trained" nature of an MLM and the "fine-tuned" nature of an MLMoCN is critical since it highlights the necessity of training both types of models (MLM or LLM) with "real" data, in this case, specifically related to -characteristics and dynamics- of criminal networks in each social context.

Due to the popularity of Chat-GPT and other LLMs' remarkable capacity to answer virtually any question unrelated to a censured subject, [**3**] the general and not specialized public tends

2. https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder.

3. According to Chat GPT-3.5, by October 2023, the list of banned subjects includes (i) illegal activities, (ii) harmful actions, (iii) hate speech and discrimination, (iv) misinformation, (v) violence and self-harm, (vi) personal or sensitive information, (vii) inappropriate or explicit content, and (viii) unethical or malicious use.

to interpret LLMs as expert-domain systems. However, even the most potent LLMs require a "fine-tuning" process with specific "real" related data to perform specialized language processing tasks correctly. This means that despite its remarkable generative performance, an LLM that is not trained with specialized "real" data does not necessarily have a better performance when compared to a traditional MLM that is "fine-tuned" with technical "real" data related to the phenomena under analysis. Various open-source LLMs -such as (i) TheBloke/Llama-2-7B-Chat-GGM; (ii) OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5; (iii) bigscience/bloom-3b, and (iv) Facebook/opt-350m-, have shown a worse performance against a "fined-tuned" MLMoCN when asked to recognize, extract, label, array a criminal network's DoI, and deliver inference results in specific JSON formats. In this regard, fine-tuned MLMs are more reliable than open-source LLMs. However, it must be noted that recent developments in Graph Neural Networks (GNN) address the capacity of LLMs to follow format instructions during the graph generation; for instance, Tang et al. (2023) have proposed GraphGPT, "*a framework that aligns LLMs with graph structural knowledge with a graph instruction tuning paradigm*".

Conclusions: Opportunities and Limits of MLMoCNs and LLMoCNs

5

As stated above, MLMs can assist in generating Criminal Network Graphs (CNGs) by automatizing the recognition, extraction, and arrangement of semantic entities according to the DoI structure. Therefore, this is one of the AI areas with the most significant current potential within the CNA field. However, a few clarifications and obstacles that hinder the current massive development and adaptation of MLMoCNs and LLMoCNs, are discussed below.

1.1. Requirement of human supervision

Assistance in the subtasks to generate Criminal Network Graphs (CNGs) significantly increases the amount of data that can be analyzed and reduces the time and amount of human analysis. However, this does not mean that CNA can be instanced as an unsupervised and entirely automatized process since the model's inferences during the first training cycles are often misaligned from the expected objectives.

1.2. Pre-trained Machine Learning Models and Large Language Models are not Expert Systems

Due to the remarkable capacity of some LLMs to answer any question correctly -in a grammatical and syntactic sense-, the public has, in a significant way, assumed that the most popular LLMs are Expert Systems (ES) capable of answering any question with the best data and knowledge available. However, not even LLMs developed by Google, possibly trained upon millions of indexed websites, show such performance. Due to the complexity of the neural network used in LLMs, even empirically correct and massive datasets for training sometimes produce grammatically correct inferences but are substantially incorrect in their content, a situation defined as "artificial hallucinations" that have been observed even in LLMs trained with highly specialized datasets (Kumar, Mani, Tripathi, Saalim, & Roy, 2023). This complexity also produces reducible but potential variability in the delivered inferences; for instance, it is common that even after using the same prompt to ask Chat GPT 3.5 and 4.0 to deliver inferences in specific formats, the results change randomly. In this regard, in various tests, Llama2 [**4**] showed a better capacity to follow instructions requesting a specific format to deliver inference outputs; however, its results on the delivery format also varied even when the same instructions were provided.

"Artificial hallucinations" do not disappear in LLMs trained with highly specialized data and ESs, which is a critical reason for permanently conducting human supervision. However, the phenomenon is expected to decrease when more specialized and calibrated datasets are used during training. In any case, fine-tuned

4. https://ai.meta.com/llama/

MLMs and LLMs must produce inferences with high reliability in qualitative terms and statistical confidence in a quantitative sense.

1.3. Scarcity of Empirical Data on Criminal Networks

The inherent opacity of criminal activities implies a lack of empirical data, which is a primary challenge for developing MLMoCNs. As discussed, the initial NLP capabilities of pre-trained MLMs and LLMs are not suited for conducting expert recognition and extracting the semantic entities that define a criminal interaction; therefore, the pre-trained model must be fine-tuned with the categories used to classify types of nodes/agents and interactions.

To address the lack of empirical data, it has been proposed to produce and use hybrid datasets in which synthetic data complements the randomized distribution of real data obtained through empirical analysis (Cano-Melani, Salcedo-Albaran, & Garay-Salamanca, 2023). However, although hybrid and larger datasets can be used to train MLMs and LLMs, producing and using real/empirical data as a starting point is unavoidable since entirely synthetic data does not reflect the empirical variability of the phenomenon.

1.4. Hardware requirements: GPU, CPU, and Massive Adoption of A.I. for Criminal Network Analysis

Due to the large size of the training data and the hyper-parameters that define LLMs -with some models reaching 10-billion parameter amount, and some reach 100 billion parameters-, these models perform better than MLMs in pre-trained semantic tasks such as Name Entity Recognition (NER) and tagging of entities in NLT. In this sense, using LLMs would be desirable to analyze larger and more complex criminal networks in the context of CAN. However, by 2023, fine-tuning LLMs with the best generative performance requires the computing capabilities specifically provided by, at least, A100 Nvidia, a high-cost Graphic Processing Unit (GPU) and not regular Central Processing Units (CPU). In fact, it has been verified that A100 is the only reference that allows using a single GPU to generate a peak of 139 tokens per second with a 7-billion parameter model such as Llama2 (Dell Technologies, 2023). Therefore, although some LLMs with low numbers of hyper-parameters can be fine-tuned with CPU, those models perform similarly to regular MLMs. In this sense, the cost of GPU computing must still decrease to allow independent researchers to fine-tune LLMs for specific CAN projects, or the architecture of those LLMs must be optimized to allow their training and good performance through regular CPUs (Simon, 2023).

Despite these limitations, MLMs and LLMs are expected to be increasingly adopted and used to model and predict social phenomena, and CNA is not an exception. Expectations on the predictive capabilities of MLMs and LLMs will probably increase as for-profit companies have joined a commercial race to develop and promote the massive adoption of chatbots and other A.I. tools; therefore, those companies will probably hide the limitations and exaggerate the expectations of their models' performance. However, defining procedures for disclosing the characteristics and limitations of those models is critical for enabling users to obtain the best possible performance by using each model to analyze and predict the phenomenon that better fits the training data. Artificial Intelligence models have high levels of indescribability. The training data, the human feedback, and the model's architecture constantly change through calibration, making it challenging to define and describe these elements as uniform corpora. However, both for the ethical and social implications resulting from each model's biases and for the practical reason of achieving the best performance according to the training data, disclosing this information is essential for adopting MLMs and LLMs with high levels of qualitative reliability and quantitative confidence to understand and predict the complexity of criminal networks..

Bibliography

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus, 15(2).

Armstrong, H. L., Johnson, A. N., & McCulloh, I. A. (2013). Social Network Analysis with Applications. New Jersey: Wiley.

Basu, K., & Sen, A. (2021). Identifying individuals associated with organized criminal networks: A social network analysis. Social Networks, 42-54.

Borgatti, S. P., Mehra, A., Brass, D., & Labianca, G. (2009). Network Analysis in the Social Sciences. Science, 323, 892-895.

Cano-Melani, J. A., Salcedo-Albaran, E., & Garay-Salamanca, L. G. (2022). A model for evaluating AI generated network graphs. St. Petersburg: SciVortex Corp.

Cano-Melani, J. A., Salcedo-Albaran, E., & Garay-Salamanca, L. J. (2023). Performance of Machine Learning Models on Criminal Networks (MLMoCN) Trained and Evaluated with Rreal, Synthetic, and Hybrid Data. St. Petersburg: SciVortex Corp.

Carbonell, M., Fornés, A., Villegas, M., & Lladós, J. (2020). A neural model for text localization, transcription and named entity recognition in full pages. Pattern Recognition Letters, 136, 219-227.

Casper, S., Davies, X., Shi, C., Krendl Gilbert, T., Scheurer, J., Rando, J., . . . Siththaranjan, A. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217 [cs.AI].

Cavallaro, L., Ficara, A., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, B., . . . Liotta, A. (2020). Disrupting resilient criminal networks through data analysis: The case of Sicilian Mafia. PLoS ONE 15(8):, 15(8), 1-22.

Degenne, A., & Forsé, M. (1999). Introducing Social Networks. London: SAGE Publications.

Dell Technologies. (2023). Llama 2: Inferencing on a Single GPU. Retrieved from Dell Technologies: https://infohub.delltechnologies.com/l/llama-2-inferencing-on-a-single-gpu/overview-5128/

Garay Salamanca, L. J., & Salcedo-Albarán, E. (2012). Narcotráfico, Corrupción y Estados: Cómo las redes ilícitas han reconfigurado las instituciones de Colombia, Guatemala y México. Ciudad de México: Random House Mondadori.

Garay Salamanca, L., Salcedo-Albarán, E., & Macías, G. (2018d). Macro-Corruption and Institutional Co-optation: The "Lava Jato" Criminal Network. Bogotá, Colombia: Fundación Vortex.

Garay, L. J., & Salcedo-Albaran, E. (2012c). Redes ilícitas y reconfiguración de Estados: El caso Colombia. Bogotá: ICTJ - Fundación Vortex.

Garay-Salamanca, L. J., & Salcedo-Albarán, E. (2012a, March). Institutional Impact of Criminal Networks in Colombia and Mexico. Crime, Law and Social Change, 57(2), 177-194.

Garay-Salamanca, L. J., & Salcedo-Albaran, E. (2015). Drug Trafficking, Corruption and States: How Illicit Networks Shaped Institutions in Colombia, Guatemala and Mexico. Bloomington: Vortex Foundation.

Garay-Salamanca, L. J., Salcedo-Albarán, E., & Duarte, N. (2017). Elenopolítica: Reconfiguración cooptada del Estado en Arauca, Colombia. Bogotá: Fundación Vortex.

Jordon, J., Wilson, A., & van der Schaar, M. (2020). Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods. arXiv.

Knoke, D., & Yang, S. (2019). Social Network Analysis. London: Sage.

Kumar, M., Mani, U. A., Tripathi, P., Saalim, M., & Roy, S. (2023). Artificial Hallucinations by Google Bard: Think Before You Leap. Cureus, 15(8), DOI: 10.7759/cureus.43313.

Marrero, M., Sánchez-Cuadrado, S., Morato, J., & Andreadakis, G. (2009). Evaluation of Named Entity Extraction Systems Mónica. Advances in Computational Linguistics, Research in Computing Science, 47-58.

Morselli, C. (2008). Inside Criminal Networks. Montreal: Springer.

Morselli, C. (2012). Assessing network patterns in illegal firearm markets. Crime Law Soc Change, 129-149.

Oliveira, I. I., Fileto, R., Speck, R., Garcia, L. P., Moussallem, D., & Lehmann, J. (2021). Towards Holistic Entity Linking: Survey and Directions. Information Systems, 95.

Salcedo-Albarán, E., & Garay-Salamanca, L. (2016). Macro-Criminalidad: Complejidad y Resiliencia de las Redes Criminales. Bloomington: iUniverse, Vortex Foundation, Small Wars Journal.

Sayama, H. (2015). Introduction to the Modeling and Analysis of Complex Systems. Geneseo: SUNY.

Shehata, D. (2022). Information Retrieval with Entity Linking. UWSpace.

Silva, K., & Silva, T. (2021). A Review on Document Information Extraction Approaches. Proceedings of the Student Research Workshop associated with RANLP-2021, 174–179.

Simon, J. (2023). Smaller is Better: Q8-Chat LLM is an Efficient Generative AI Experience on Intel® Xeon® Processors. Retrieved from Intel.com: https://www.intel.com/content/www/us/en/developer/articles/case-study/q8-chat-efficient-generative-ai-experience-xeon.html

Taleb, N. (2012). Antifragile: Things That Gain from Disorder. New York: Random House.

Tang, J., Yang, Y., Wei, W., Shi, L., Su, L., Cheng, S., . . . Huang, C. (2023, Oct.). GraphGPT: Graph Instruction Tuning for Large Language Models. https://doi.org/10.48550/arXiv.2310.13023 Focus to learn more .

Trani, S., Ceccarelli, D., Lucchese, C., Orlando, S., & Perego, R. (2018). SEL: a Unified Algorithm for Entity Linking and Saliency Detection. Computational Intelligence, 34, 2– 29.

Wang, J., & Han, J. (2015). Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. IEEE Transactions on Knowledge and Data Engineering, 27(2), 443-460.

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.