# A model for evaluating AI generated network graphs

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

VORTEX WORKING PAPER No. 62

VORTEX

Vortex Working Paper No. 62.

# Abstract

The objective of this article is to propose a model for evaluating network graphs based on text and generated through Artificial Intelligence (AI); it is, network graphs generated through the assistance of Natural Language Processing (NLP) procedures for identifying, extracting, and tagging entities to generate a network graph. The evaluation consists of comparing the characteristics of an AI-generated and a human-generated network graph. This article has five sections. The first section is an introduction to the human and AI tools used to generate the compared network graphs. In the second and third sections the procedure for generating the AI network graph and the evaluation process are discussed. In the fourth and fifth sections preliminary conclusions are presented and discussed.

# Introduction: Vorisoma AI 1.0

1

The objective of this article is to propose an evaluation model for comparing network graphs generated through AI tools with network graphs generated by humans. This evaluation consists of comparing and finding matches between entities that compose an AI-generated network graph, and entities that compose a human-generated network graph. Both graphs are evaluated regarding the similarity between the entities extracted and linked by the AI against those extracted and linked by human agents. This means that the evaluation does not consists only of measuring the precision and capacity for extracting and linking entities according to a base knowledge (Wang & Han, 2015), but according to a set of reference entities previously named, extracted, and linked by human agents.

The analyzed network graphs were generated by using VORISOMA, a software suite developed by Vortex Foundation and SciVortex Corp. to model and analyze complex criminal networks. VORISOMA is a web application that simplifies tasks related to Criminal Network Analysis (CNA) (Basu & Sen, 2021; Cavallaro, et al., 2020; Morselli, 2008; Morselli, 2012), by separating analyzed networks into workspaces for which it provides editing, reporting, and management data tools. Each workspace is intended for the analysis of a single independent network, in which nodes and relationships/interactions are labeled according to *types* or *categories*. Since 2011, VORISOMA has evolved along to the CNA methodological set as instanced by Vortex Foundation and SciVortex Corporation (Garay-Salamanca & Salcedo-Albarán, 2012a; Garay Salamanca & Salcedo-Albarán, 2012). VORISOMA has been applied to analyze domestic and transnational criminal networks (Garay-Salamanca & Salcedo-Albaran, 2015; Petrunov, 2013; Goga, Salcedo-Albaran, & Goredema, 2014; Goga, Goredema, & Salcedo-Albarán, 2017).

The network graphs analyzed herein are modeled after text that informs about interactions between two types of entities represented as nodes: individuals, and organizations such as public and private corporations. In the CNA both entities have been described and referenced to as nodes/agents due to their agency in social contexts (Garay Salamanca & Salcedo-Albarán, 2012; Garay-Salamanca & Salcedo-Albaran, 2015). Within the theoretical framework of CNA, the text corpus has been referred to as “empirical evidence” and includes texts with various levels of reliability: judicial, administrative, and media reports. After modeling a criminal network graph, VORISOMA calculates centrality indicators such as degree centrality and betweenness to identify the most relevant entities/nodes in each case (Degenne & Forsé, 1999; Carrington, Scoot, & Wasserman, 2005; Cavallaro, et al., 2020).

The main purpose of VORISOMA’s AI and non-AI tools is to assist during the modeling of relationships/interactions that inform how nodes/agents interact in a criminal network. Each relationship/interaction consists of two entities linked through a verb and mapped to a piece of text in the “empirical evidence”. In this sense, VORISOMA AI 1.0 is a Natural Language Processing (NPL) tool for Document Information Extraction (DIE) (Silva & Silva, 2021) through Named Entity Recognition (NER) task (Marrero, Sánchez-Cuadrado, Morato, & Andreadakis, 2009). Currently, NER is a basic and initial task under approaches for text localization and transcription (Carbonell, Fornés, Villegas, & Lladós, 2020). Through an initial knowledge base, VORISOMA AI 1.0 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).

However, the task for locating and assigning the verb that correctly conforms a relationship/interaction between the initially extracted, named, and linked entities/nodes is a second-order NER task that is not necessarily inherent to the first-order NER executed during the initial identification and labeling of nodes/entities. This second-order NER task results in Modeling a Relationship/Interaction (MoRI) through human analysis or through VORISOMA AI 1.0, by considering the contextual information generated during the first-order NER and by applying predefined syntax rules. This means that after considering the empirical contextual information, the human analyst or VORISOMA AI 1.0 assign an entity/verb to link the previously extracted entities/nodes, therefore modeling a relationship/interaction (MoRI output).

Before VORISOMA AI 1.0, the first and second order NER tasks were executed by human agents: (i) During the first-order NER task by (i.i) reading the text, (i.ii) identifying, and (i.iii) labeling each entity/node as an individual or an organization, and then, (ii) during the second-order NER by (ii.i) identifying and (ii.ii) labeling each entity/verb as an economic, political, violent, or logistic relationship/interaction to, (iii) finally, articulate the entities/nodes and the entity/verb into a relationship/interaction (MoRI output).

Bearing in mind the process executed by human agents, VORISOMA AI 1.0 executes a similar process to generate MoRI outputs like those generated by human agents, which implies that the AI-generated and the human-generated relationships/interactions consist of the same elements: two entities/nodes (individuals or organizations) that interact through an entity/verb.

The network graph can be defined as a set of relationships/interactions related and resulting from pieces of empirical evidence listed in a single VORISOMA workspace. As previously stated, the AI-generated graph is herein compared with a reference human-generated graph in which human analysts identified, extracted, and linked the entities/nodes of the network. It is not discussed or evaluated if the human analysts correctly executed the task for identifying, extracting, and linking the entities.

As expected, the execution of NER tasks within VORISOMA AI 1.0 resulted in various ambiguity challenges that were addressed not only by increasing the knowledge base (Wang & Han, 2015) -which mainly consisted of using one of the most comprehensive and complete NLP public models available, such as Google-, but through algorithmic definitions for improving delimitation, merging, and adjustments of entities that were incorrectly extracted, named, or linked to the knowledge base. Some of those rules are discussed in the second section of this paper. In this sense, the approach adopted in VORISOMA AI 1.0 can be interpreted as “Holistic Entity Linking” because it considers the first-order contextual data during the identification and naming tasks, as well as the algorithmic definitions of syntaxis rules to address specific disambiguation cases (Oliveira, y otros, 2021); which means that the first-order NER task affects the second-order NER, and both tasks are also affected by algorithms previously specified.

VORISOMA AI 1.0 can process any natural language with a minimum of 20 complete words. In this case, VORISOMA AI 1.0 was applied to a data corpus with a size of 6,869.811 bytes of textual evidence collected from criminal cases, including judicial and administrative public records, as well as media.

# AI Generation and Flaws During Identification and Extraction of Entities

2

## 2.1. Comparing the AI-generated and the human-generated network graph

The following process was executed for comparing the two network graphs.

1. Exporting the body of evidence to a table that directly relates the nodes/agents and relationships/interactions to each piece of evidence documented. Mapping the initially unstructured or semi structured text to a table is often defined as the initial step in the extraction of information process (Silva & Silva, 2021).
2. Bulk exporting all "empirical evidence" into a single text file.
3. Using "Google Colab" a script was implemented based upon the public Google Translation Service, to automatically translate all the texts used as empirical evidence in the modeled network. The resulting text was exported to a text file.
4. Executing the algorithm for generating the AI graph with the translated text table.
5. Mapping the entities/nodes and the entities/verbs to their corresponding evidence in the text table.

   a. When VORISOMA AI 1.0 identified and linked entities, an occurrence was registered and counted for each relationship found. Therefore, the task defined as “Occurrence to Evidence Matching” consists of finding occurrences in the original empirical evidence (text file) after alphanumeric normalization of the associated entities.
   b. In this case, normalization consists of trimming the names, applying only lowercase letters, and replacing consecutive blank characters by single spaces.
   c. Additionally, for matching entities during disambiguation, any non-alphanumeric or blank characters were removed from the compared mention strings used for matching potentially redundant entities.
6. Comparing the networks characteristics described in the next section.

## 2.2. Algorithm for Generating the AI Graph: Strengths and Flaws During the Generation of the AI network Graph

1. Submitting the body of text to the google NLP API for executing the **analyseEntities** and **analyseSyntax** services, to obtain an inventory of entities and a syntax tree from the text by using Google's Machine Learning models.
2. Fixing identified flaws in Google's output:
   a. Adjacent Entities: This happens when Google detects multiple entities that should be interpreted as one entity, in a single sequence of adjacent tokens. For example, in the string of text "General Secretary Carlos Muñoz”, this flaw happens if Google detects and extracts two entities -"General Secretary" and "Carlos Muñoz"– although both entities should be identified and extracted as one. To address this flaw, entities were joined in situations when probably wrong separations were expected, through the "**process.py > joinAdjacentEntities( )" function.**

b. Redundant Mentions: This happens when Google incorrectly detects mentions of an actor inside other mentions of other actors. To address this flaw, only the larger mention was considered; therefore, the contained mentions were merged through the **"process.py > removeRedundantMentions( )" function**. This step was implemented but it was not applied during this specific analysis.

c. Redundant Entities: This happens when Google incorrectly returns entities with the same name. To address this situation, entities with the same normalized name were joined if they had at least one syntax token that is a proper noun in one of the mentions through the **"process.py > rawEntityUnique( )" function**.

d. Missed Pronoun Bindings: This happens when Google doesn't link pronouns to its corresponding entity. In these cases, every pronoun was bound to the immediately preceding entity, when the pronoun was identified to be a reference to another entity. This was done through the **"process.py > process ( )" function** (coupled with the general processing algorithm).

3. MoRi task: Finding relationships by navigating the syntax tree, herein defined as any pair of entities that are connected through a verb in a single sentence.
4. Saving the results of the MoRI task in a single graph file that includes the entities/nodes and entities/verbs for defining each relationship/interaction.
5. Mapping entities/nodes and entities/links to nodes and types existing in a target workspace. At this point these pre-existing elements can be understood as a knowledge base. To preserve privacy of the data used during this evaluation and to prevent leaking the expected results to the AI Graph generation process, this step is not used for the evaluation described below.

# Evaluation

3

The evaluation and comparison algorithm included several caches to facilitate restarting the process from critical breakpoints and to reduce calls to the Google API and save re-running execution time during debugging. The scripting language used to oversee the comparison report was PHP, which was also used for starting the graph generation process, handling some caches, and calling a python script for executing NLP operations. Below, some evaluation criteria are described.

- ## Matching Nodes/Agents

This criteria was calculated through the number of entities/nodes that VORISOMA AI 1.0 identified and extracted, that successfully coincided with those extracted and named by the human analyst. These coincidences, defined as "matching actors", were calculated after (i) translating all reference actor names into English, and (ii) comparing the actor's names from the reference workspace with the ones in the AI generated graph, only examining the space of actors bound to the same piece of evidence, and considering both the original language of the actor and the translated name.

The Algorithm for finding matching-actors considers that two actor names A and B are a match if after being normalized: (i) they are identical; (ii) A is contained in B or vice versa; and (iii) the longest matching substring of A is longer than half the average of the length of the two strings. This criterion is defined as "similarity comparison".

- **Mapping to Existing Types and Nodes**

This mapping occurs independent from NLP operations before rendering a graph for importing it into a target workspace. This step includes mapping entity/nodes, actor types and relationship types to pre-existing such elements in the target workspace.

- **Cli Scripts**

In general terms, the processor is designed to be used through the VORISOMA web interface; however, there is also a set of maintenance and operation commands in VORISOMA intended for use through the system console. VORISOMA AI 1.0 includes two such commands: **cli.php** and **cli.py**.

# 4 Results of Evaluating VORISOMA AI 1.0 vs Human Generated Graph

The following parameters were calculated by comparing the AI generated graph and the human generated graph:


```
"AI Links": "The number of links found by the AI",
 "Workspace Evidences": "The Number of evidences found in the reference work-
space" = 13909.
 "Missed Evidences": "The number of evidences that couldn't be mapped to
links found by the AI" = 8014,
 "Matched Evidences": "The number of evidences that were mapped to links
found by the AI" = 5895,
 "Found Occurrences": "The number of occurrences mapped from AI found Links
to reference evidences" = 24074,
 "Total Occurrences": "The number of occurrences mapped from AI found Links
that could not be mapped to reference evidences" = 35736,
 "Matched Evidence Report": {
   "Count": "The number of reference evidences matched to found links" =
5895,
   "Overall": {
     "Total Ref Actors": "Number of actors in the reference workspace" =
15115,
     "Total AI Actors": "Number of actors found by the AI" = 41192,
     "Total Ref Rels": "Number of relationships in the reference workspace" =
10650,
     "Total AI Rels": "Number of relationships found by the AI" = 32845,
     "Total Matched Actors": "The number of actors found by the AI that were
found to map to a corresponding reference actor" = 8829,
     "Avg Ref Actors": "Average Reference Actors per Evidence" =
2.5640373197625106,
     "Avg AI Actors": "Average Actors per Evidence found by the AI" =
6.9876166242578455,
     "Avg Ref Rels": "Average Reference Relationships per Evidence" =
1.806615776081425,
     "Avg AI Rels": "Average Evidence per Evidence found by the AI" =
5.5716709075487705,
     "Avg Avg Matched Actors": "Average Reference Actors per Evidence" =
1.497709923664122
   },
   "Segments": {
     "Average": "The average number of items (additional/missing relation-
ships or actors) in the segment"=1.8841870824053453,
```

```
     "Evidence Count": "Number of pieces of evidence in the segment" = 449,
     "Avg Ref Actors": "Average Reference Actors per Evidence in the Segment" = 4.3162583518930955,
     "Avg AI Actors": "Average Actors per Evidence found by the AI in the Segment" =
2.4320712694877504,
     "Avg Ref Rels": "Average Reference Relationships per Evidence in the Segment" =
.9910913140311806,
     "Avg AI Rels": "Average Evidence per Evidence found by the AI in the Segment" =
1.3496659242761693,
     "Avg Avg Matched Actors": "Average Reference Actors per Evidence in the Segment" =
0.8930957683741648
   },

   "Same Actors": "Segment of evidence pieces where the AI found the same number of actors as the
ones in the reference workspace",
   "Same Rels": "Segment of evidence pieces where the AI found the same number of relationships as
the ones in the reference workspace"
             },
             "Extra Rels": {
             "Average": 6.184972606313592,
             "Evidence Count": 3833,
             "Avg Ref Actors": 2.428645969214714,
             "Avg AI Actors": 9.36081398382468,
             "Avg Ref Rels": 1.5823115053482912,
             "Avg AI Rels": 7.767284111661883,
             "Avg Matched Actors": 1.7644142968953822
             },
   },
   "Missing Actors": "Segment of evidence with missing actors when comparing AI found actors versus
the reference workspace"
   "Missing Rels": "Segment of evidence pieces with missing relationships when comparing AI found
actors versus the reference workspace",
   "Extra Actors": "Segment of evidence pieces where the AI found more actors than the ones in the
reference workspace",
   "Extra Rels": "Segment of evidence pieces where the AI found more relationships than the ones in
the reference workspace",
"Same Actors": "Segment of evidence pieces where the AI found the same number of actors as the ones
in the reference workspace",
{
             "Evidence Count": 1346,
             "Avg Ref Actors": 2.3135215453194653,
             "Avg AI Actors": 2.3135215453194653,
             "Avg Ref Rels": 1.5096582466567607,
             "Avg AI Rels": 1.312778603268945,
             "Avg Matched Actors": 0.9650817236255572
}
"Same Rels": "Segment of evidence pieces where the AI found the same number of relationships as the
ones in the reference workspace"
{
             "Evidence Count": 1325,
             "Avg Ref Actors": 2.239245283018868,
             "Avg AI Actors": 2.4611320754716983,
             "Avg Ref Rels": 1.3962264150943395,
             "Avg AI Rels": 1.3962264150943395,
             "Avg Matched Actors": 0.9916981132075472
             }
```

## 4.1. False positives and false negatives during actors matching

During the execution of the evaluation, it was identified the importance of optimizing the search of false negative (VORISOMA AI 1.0 failing to match two entities/nodes when it should have) and false positive (VORISOMA AI 1.0 matching two entities/nodes when should have not), to reduce and control noise in the matching ratio.

A sample of 30 actor match results were evaluated to identify the sensitivity and specificity of VORISOMA AI 1.0 when matching actors against the reference entities. It was found a higher sensitivity of VORISOMA AI 1.0 to omit potential matches (10%) than to match entities/nodes that should have not been matched (3%), as illustrated with the following results:

```
"False Negative ": 3 (10%)
"False Positive": 1 (3.3%)
"True Negative": 21 (70%)
"True Positive": 5 (16.6%)
```

Therefore, it was implemented a mechanism to identify, tag and filter only “certified evidence”, meaning those pieces of evidence that contained at least 70% of the words of each reference actor name, either translated to English or in Spanish. This allowed excluding cases in which the evidence was not fully or correctly registered into the initial text corpus, making it impossible for VORISOMA AI 1.0 to find and match the entities/nodes. After tagging and filtering as “certificated evidence”, the average actor matching in the “certified” segment increased when compared with the “uncertified” segment, as shown in the following values.

Values with focus on “certified” evidence.

```
1 "AI Links": 29919,
2    "Workspace Evidences": 13909,
3    "Missed Evidences": 7645,
4    "Matched Evidences": 6264,
5    "Found Occurrences": 29224,
6    "Total Occurrences": 42104,
7    "Report Segments": {
8        "Overall": {
9            "Count": 6264,
10           "Total Ref Actors": 16118,
11           "Total AI Actors": 47982,
12           "Total Ref Rels": 41062,
13           "Total AI Rels": 41062,
14           "Total Matched Actors": 9542,
15           "Avg Ref Actors": 2.573116219667944,
16           "Avg AI Actors": 7.659961685823755,
17           "Avg Ref Rels": 6.555236270753512,
18           "Avg AI Rels": 6.555236270753512,
19           "Avg Avg Matched Actors": 1.5233077905491699
20       },
21       "Certified": {
22           "Count": 2380,
23           "Total Ref Actors": 5316,
24           "Total AI Actors": 16702,
25           "Total Ref Rels": 14210,
26           "Total AI Rels": 14210,
27           "Total Matched Actors": 4206,
28           "Avg Ref Actors": 2.2336134453781513,
29           "Avg AI Actors": 7.017647058823529,
30           "Avg Ref Rels": 5.970588235294118,
31           "Avg AI Rels": 5.970588235294118,
32           "Avg Avg Matched Actors": 1.7672268907563025
33       },
34       "Missing Actors": {
35           "Average": 1.1369863013698631,
36           "Evidence Count": 73,
37           "Avg Ref Actors": 3.136986301369863,
38           "Avg AI Actors": 2,
39           "Avg Ref Rels": 2.506849315068493,
40           "Avg AI Rels": 1,
41           "Avg Matched Actors": 1.0547945205479452
42       },
43       "Missing Rels": {
44           "Average": 1.5168539325842696,
45           "Evidence Count": 89,
46           "Avg Ref Actors": 3.292134831460674,
47           "Avg AI Actors": 2.49438202247191,
48           "Avg Ref Rels": 2.831460674157303,
49           "Avg AI Rels": 1.3146067415730338,
50           "Avg Matched Actors": 1.1123595505617978
51       },
52       "Extra Actors": {
53           "Average": 6.1694459386767075,
54           "Evidence Count": 1859,
55           "Avg Ref Actors": 2.235072619688004,
56           "Avg AI Actors": 8.404518558364712,
57           "Avg Ref Rels": 1.3378160301237225,
58           "Avg AI Rels": 7.3442711135018826,
59           "Avg Matched Actors": 1.9483593329747175
60       },
61       "Extra Rels": {
62           "Average": 6.171081677704194,
63           "Evidence Count": 1812,
64           "Avg Ref Actors": 2.2268211920529803,
65           "Avg AI Actors": 8.524834437086092,
66           "Avg Ref Rels": 1.3062913907284768,
67           "Avg AI Rels": 7.477373068432671,
68           "Avg Matched Actors": 1.9641280353200883
69       },
70       "Same Actors": {
71           "Evidence Count": 448,
72           "Avg Ref Actors": 2.080357142857143,
73           "Avg AI Actors": 2.080357142857143,
74           "Avg Ref Rels": 1.1004464285714286,
75           "Avg AI Rels": 1.0803571428571428,
76           "Avg Matched Actors": 1.1316964285714286
77       },
78       "Same Rels": {
79           "Evidence Count": 479,
80           "Avg Ref Actors": 2.0626304801670146,
81           "Avg AI Actors": 2.1565762004175366,
82           "Avg Ref Rels": 1.1356993736951984,
83           "Avg AI Rels": 1.1356993736951984,
84           "Avg Matched Actors": 1.1440501043841336
85       }
86   },
```

Values without focus on “certified” evidence.

This column shows the difference in the results for data segments when the values on such segments were calculated. This includes comparisons on "uncertified" evidences to confirm that the "Avg Matched Actors" is significantly lower in all segments.

The lines corresponding Overall and Certified summaries were omitted because they are exactly the same on both reports.

```
13 ...
14 ...
15 ...
16 ...
17 ...
18 ...
19 ...

20 ...
21 ...
22 ...
23 ...
24 ...
25 ...
26 ...
27 ...
28 ...

29 ...
30 ...
31 ...
32 ...

33 {
34       "Missing Actors": {
35           "Average": 2.108173076923077,
36           "Evidence Count": 416,
37           "Avg Ref Actors": 4.550480769230769,
38           "Avg AI Actors": 2.4423076923076925,
39           "Avg Ref Rels": 4.125,
40           "Avg AI Rels": 1.3341346153846154,
41           "Avg Matched Actors": 0.9206730769230769
42       },
43       "Missing Rels": {
44           "Average": 2.108665749656121,
45           "Evidence Count": 727,
46           "Avg Ref Actors": 3.9876203576341127,
47           "Avg AI Actors": 2.9284731774415405,
48           "Avg Ref Rels": 3.847317744154058,
49           "Avg AI Rels": 1.7386519944979368,
50           "Avg Matched Actors": 1.0687757909215956
51       },
52       "Extra Actors": {
53           "Average": 7.4479071883530485,
54           "Evidence Count": 4396,
55           "Avg Ref Actors": 2.4838489535941766,
56           "Avg AI Actors": 9.931756141947226,
57           "Avg Ref Rels": 1.7145131938125568,
58           "Avg AI Rels": 8.792993630573248,
59           "Avg Matched Actors": 1.7813921747042767
60       },
61       "Extra Rels": {
62           "Average": 7.6722358722358726,
63           "Evidence Count": 4070,
64           "Avg Ref Actors": 2.443243243243243,
65           "Avg AI Actors": 10.371498771498771,
66           "Avg Ref Rels": 1.605159705159705,
67           "Avg AI Rels": 9.277395577395577,
68           "Avg Matched Actors": 1.8098280098280097
69       },
70       "Same Actors": {
71           "Evidence Count": 1452,
72           "Avg Ref Actors": 2.2768595041322315,
73           "Avg AI Actors": 2.2768595041322315,
74           "Avg Ref Rels": 1.4573002754820936,
75           "Avg AI Rels": 1.2761707988980717,
76           "Avg Matched Actors": 0.9146005509641874
77       },
78       "Same Rels": {
79           "Evidence Count": 1467,
80           "Avg Ref Actors": 2.232447171097478,
81           "Avg AI Actors": 2.481935923653715,
82           "Avg Ref Rels": 1.3899113837764145,
83           "Avg AI Rels": 1.3899113837764145,
84           "Avg Matched Actors": 0.9536468984321745
85       }
86   },
```

# Discussion and Conclusions

5

## 5.1. Regarding VORISOMA AI 1.0

The following considerations can be defined as the benchmark of the evaluation and, therefore, are relevant for interpreting the resulting values presented above:

1. **"Avg Matched Actors" should be equal to "Avg Ref Actors"**, which means that VORISOMA AI 1.0 identifies and names the same entities/nodes as the human agent.
2. **"Avg AI Rels" should be more than or equal to "Avg Ref Rels"**, which means that in average VORISOMA AI 1.0 established at least the same number of interactions/relationships as the human agent.
3. **"Avg AI Actors" should be more than or equal to "Avg Ref Actors"**, which means that in average VORISOMA AI 1.0 identifies and extracts as many entities/nodes as the human agent.
4. **"Missed Evidences" should be zero**, which means that VORISOMA AI 1.0 correctly interpreted and analyzed every piece of textual evidence available in the corpus.
5. **"Matched Evidences" should equal "Workspace Evidences"**, which means that VORISOMA AI 1.0 correctly mapped the occurrences of the found relationships/interactions to the specific piece of evidence that was previously used by the human ana-

lyst. This objective was partially achieved by “certifying” every piece of evidence; however, the amount was still high (55%). Therefore, it is important to implement a mechanism between steps 1 to 3 of the "AI graph generation" process described, to improve the separation of each piece of evidence and, therefore, to prevent Google NLP services from combining pieces from different interactions/relationships during the analysis.

Bearing in mind these considerations, and regarding the specific evaluation of VORISOMA AI 1.0, it can be highlighted that "Avg AI Rels" seems promising in most segments since the overall results in this value is higher than "Avg Ref Rels", which implies that VORISOMA AI 1.0 established more interactions/relationships than the human agent. This result does not imply that all the relationships/interactions established by VORISOMA AI 1.0 are critical to understand the analyzed network; therefore, during next stages of VORISOMA 1.0, the AI capacities must be improved to differentiate between relevant and irrelevant relationships/interactions when explaining and analyzing a criminal network.

Additionally, the "Avg AI Actors" value also seems promising in most segments since the overall result is higher than "Avg Ref Actors", which means that VORISOMA AI 1.0 identified and extracted more entities/nodes in average than the human analyst. As stated above, this result does not imply that all the entities found by VORISOMA AI 1.0 are critical to explain a criminal network, it only reflects its capacity to identify, link and name entities/nodes and entities/verbs. However, further research is required to understand why the "Avg Avg Matched Actors" is approximately 1 actor below "Avg Ref Actors", which means that VORISOMA AI 1.0 missed on average 1 of the actors found by the human agent despite finding on average more than twice the actors found by the human agent, and one additional relationship.

The evaluation through the values previously discussed revealed the strengths and weaknesses of VORISOMA AI 1.0 during the NER tasks to produce each MoRI output and, lastly, to generate an entire network graph focused on describing illicit interactions. The main weaknesses revealed during the evaluation relate to the naming of entities during the execution of the first-order NER task. As pointed out, these problems were not only addressed by expanding the knowledge base but by defining additional rules. Therefore, the proposed evaluation was useful for determining the capacities of VORISOMA AI 1.0, and for determining those tasks and modules that require further analysis.

As discussed in the previous section, the evaluation revealed that VORISOMA AI 1.0 has a higher sensitivity to false negatives than to false positives during the matching task. In this regard, the following cases of interest were identified and require further analysis to improve VORISOMA AI 1.0 capabilities.

1. In one case the translation was partially missing from the text.
2. Syntax Mismatches: In two cases, mentions with matching words in the reference actors weren't matched to the corresponding entity/node.
3. Semantic Mismatches: In one case the match could be inferred from the text by examining the name assigned to the entity/node/individual in question.
4. Some entities include unrelated mentions bound to the same entity: This could be addressed by using a safer mechanism for separating each piece of evidence when submitting data to Google NLP services. For instance, the following names were linked to a same entity; therefore, additional research is required to understand whether this is the result of incorrect syntax definitions when differentiating adjacent entities, or of the output of a Google service:

```
Aragua State ~ aragua state ~ venezuelan ~ venezuelans ~ republic ~ bolivarian republic
of venezuela ~ venezuela ~ binational mining turkey mining venezuela ~ the bolivarian
republic of venezuela ~ venezuelan producer ~ venezuelan government ~ state venezuelan ~
venezuelan state oil company ~ venezuelan official ~ venezuelan dictator ~ venezuelan
budget ~ venezuelan president ~ venezuelan businessman ~ venezuelan officer ~ venezuelan
population ~ venezuelan 'businessman' ~ venezuelan state ~ venezuelan companies ~
venezuelan gold ~ venezuelan people ~ venezuelan citizens ~ venezuelan military ~
venezuelan fleet ~ venezuelan regime ~ venezuelan oil giant ~ aerpostal alas de venezuela
company ~ aragua
```

## 5.2. Optimizing the Evaluation Process

The evaluation process herein described can be optimized by integrating it into a single ***cli*** command that receives the workspace id in the VORISOMA database and generates the evaluation report. This would facilitate measuring the matching capacities of VORISOMA AI 1.0 -and similar tools- after each modification or implementation of rules to address ambiguity and naming problems. Tasks that should be automatized for improving the evaluation process are:

1. Generating the evidence table for each workspace, because currently there is a separate ***cli*** process for generating the evidence table for all workspaces.
2. Translating evidence and integrating the result into the evaluation command.
3. Translating actor names and integrating the results into the evaluation command.
4. Detecting “Actor Type” and “Relationship Type” with Python's SpaCy. Currently, this process requires manual mapping that is applied before rendering the generated graph for import into a target workspace. For this automatization, it is critical to define how VORISOMA AI 1.0 will determine the “types” that should be searched and, more important, those defined as relevant. This automatization will require understanding and modeling the specific cognitive tasks followed by human analysts when categorizing and labeling pieces of evidence.
5. Separating pieces of evidence when submitting the information to Google services to avoid that the occurrences of identified links are detected across different pieces.

Other expected improvements for the evaluation are:

1. Including a redundant mention removal script.
2. Executing a report of parameters for various text corpus to secure scalability of the evaluation process.

Considering the verified outstanding capabilities of AI models for detecting entities/nodes and entities/verbs that conform a relationship/interaction (MoRi output), the optimization of VORISIMA AI 1.0 and the evaluation process will be implemented under VORISOMA AI 2.0, in which in-house trained AI models will replace the use of external NLP services. Therefore, it is expected that VORISOMA AI 2.0 includes predictive capabilities trained with datasets focused on criminal networks characteristics and dynamics.The selection of real data cases and the used of synthetic data for training purposes of the criminal networks AI models will be analyzed and discussed in forthcoming papers of this series.

# Bibliography

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

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.

Carrington, P. J., Scoot, J., & Wasserman, S. (2005). *Models and Methods in Social Network Analysis.* New York: Cambridge University Press.

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.

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

Goga, K., Goredema, C., & Salcedo-Albarán, E. (2017). *The Rhino Horn Trafficking Network of the Groenewald Gang.* Bogota: Vortex Foundation.

Goga, K., Salcedo-Albaran, E., & Goredema, C. (2014). *Using social network analysis to understand the interactions of a Cape Town gang.* Cape Town: Institue for Security Studies.

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.*

Petrunov, G. (2013). *Analysis of Social Network Models of Transnational Criminal Networks operating in the Southeastern Border of the European Union.* Sofia: Vortex Foundation.

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.

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.