JOURNAL OF CRIMINALISTICS AND LAW

NBP 2021, Vol. 26, Issue 3, pp. 81–96
https://doi.org/10.5937/nabepo26-33849
Original scientific paper

ANITA's Text Analysis Services for Fighting Online Illegal Trafficking of Drugs, Weapons and False Charity Claims: A Lateral Thinking Approach

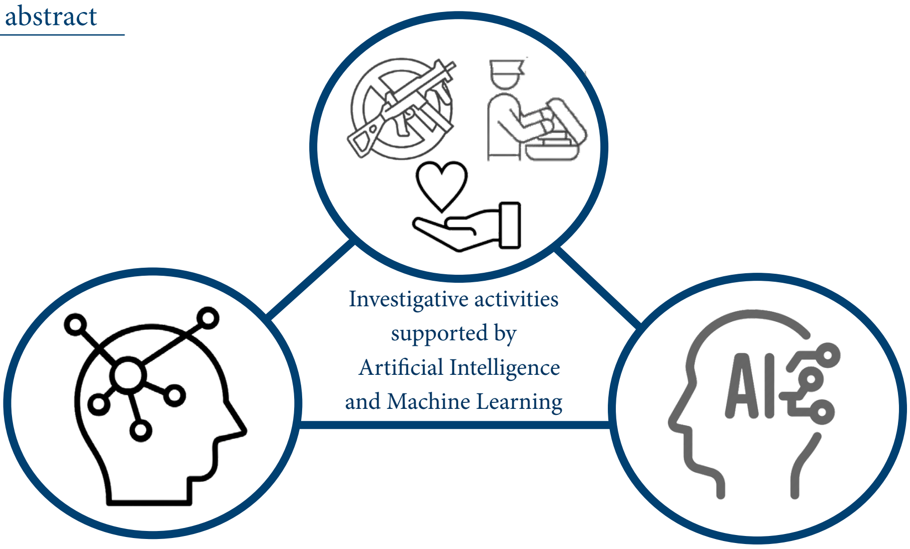
Vincenzo Masucci, Jacopo Colombo, Ciro Caterino, Filippo Nardelli[1]
Expert.ai, Modena, Italy

Submitted: 2021-09-07 • Accepted: 2022-01-19 • Published: 2022-01-25

Abstract: ANITA platform aims at improving investigation capabilities of law enforcement agencies (LEAs) by delivering a set of tools and techniques to efficiently address online illegal trafficking of counterfeit/falsified medicines, novel psychoactive substances (NPS), drugs, and weapons with Artificial Intelligence (AI). Expert.ai has developed a dedicated written content analysis engine to support investigative activity. The market already offers several solutions adopting AI and in particular Machine Learning (ML) to support investigative activities. The particularity of the work carried out in the ANITA project is to adopt a *lateral thinking approach and the supervised generation of knowledge from the analysed contents.* Rather than focusing on searching for specific proper names of substances, weapons, NPS, etc. that are constantly evolving in considerable speed, we propose to concentrate on searching for *precursors*, namely collateral concepts related with the target of the investigation. Is the case of accessories for weapons such as sights for rifles or glass vessels used to produce the drugs themselves? This approach is not compatible with the use of ML, but the use of deep semantic analysis of the text is necessary. This article reports the system's ability to automatically identify precursors and generate stimuli and knowledge which the investigator must validate; that enhances the creative and intuitive approach (*serendipity*) to investigation combining the best of the detective's skills with the power of AI and in particular Natural Language Understanding (NLU).

Keywords: illegal trafficking, drugs, weapons, Natural Language Processing (NLP), Natural Language Understanding (NLU), black market, Artificial Intelligence (AI).

Graphical abstract



[1] Corresponding author: fnardelli@expert.ai • Phone: +39 04 64 44 33 00

ISSN 2620-0406

Citation: Masucci,V., Colombo, J., Caterino, C., & Nardelli, F. (2021). ANITA's text analysis services for fighting online illegal trafficking of drugs, weapons and false charity claims: A lateral thinking approach. *NBP. Nauka, bezbednost, policija*, *26*(3), 81–96. https://doi.org/10.5937/nabepo26-38849

INTRODUCTION

Talking to the investigators involved in the ANITA project, it emerged that over the recent years, online illegal trafficking activities have expanded hugely. For tackling these emerging challenges, a significant part of law enforcement agencies (LEAs)' efforts have been invested in training activities to equip officers and practitioners with the necessary knowledge and skills related to this continuously and rapidly evolving scenery. In order to efficiently understand the organisational structure, the exhibited behaviours/dynamics and their interconnections/interactions, it is of vital importance to collect and analyse all relevant open-source information in near real-time and to combine it with closed-source information provided by the LEAs. The wide range of legitimate AI applications includes systems for crime prevention and detection (Dilek et al., 2015; Li et al., 2010; Lin et al., 2017; McClendon & Meghanathan 2015).

Investigative intelligence is a specialized area of data analytics that serves the needs of those who are hunting for bad actors and typically protecting people, networks, and assets. Thus, it is a data-based approach to investigation. These investigations involve connecting the dots on both structured (well-defined records) and unstructured data (textual and other media) across systems and schemas. At the state of the art, the platforms that manage this type of solution offer:

- The integrated analysis of structured and unstructured databases;
- Automatic generation of graphs, creation of links between analyzed entities;
- Dashboards for real-time navigation of information, and
- The creation of reports or "investigation folders" exportable to other applications.

In this article we focus on a critical element, often underestimated, namely the ability to self-provide new knowledge. The importance of the database in which the knowledge of the investigator is structured is fundamental because it becomes the driver with which the algorithms interpret the big data. The activity of updating this knowledge base is often a "hidden" criticality. ANITA offers a tool to address it with regularity, automation and transparency (Explainable AI). We describe an innovative and divergent approach to the investigation system for analysing heterogeneous (text, audio transcription, video transcription) online (Surface Web, Deep Web, Dark Nets) and offline (LEAs' databases) resources for fighting illegal trafficking activities. The goal of these services is to enrich unstructured text by automatically adding various types of annotations that can help LEAs in their tasks. These annotations will enable end-user functionalities such as filtering, categorization, and document search by entities (for example people, organizations or places) and allow integrating the extracted information with that of other sources or applications (Figure 1). In particular ANITA is focused on designing and evolving an avant-garde semantic-based text analytics engine for automatic content categorization, entities extraction based on a set of semantic analysis features customized for illegal trafficking domain. In this document we will introduce new methodologies and show their capabilities to extract advanced information, like temporal references and relationships among entities, with a new approach oriented to the "lateral thinking" (De Bono, 2016), looking at problems from a variety of angles and offering up solutions that are as ingenious as they are effective, as opposed to the traditional mode that provides concentration on a direct solution to the problem (Maio, 2016).

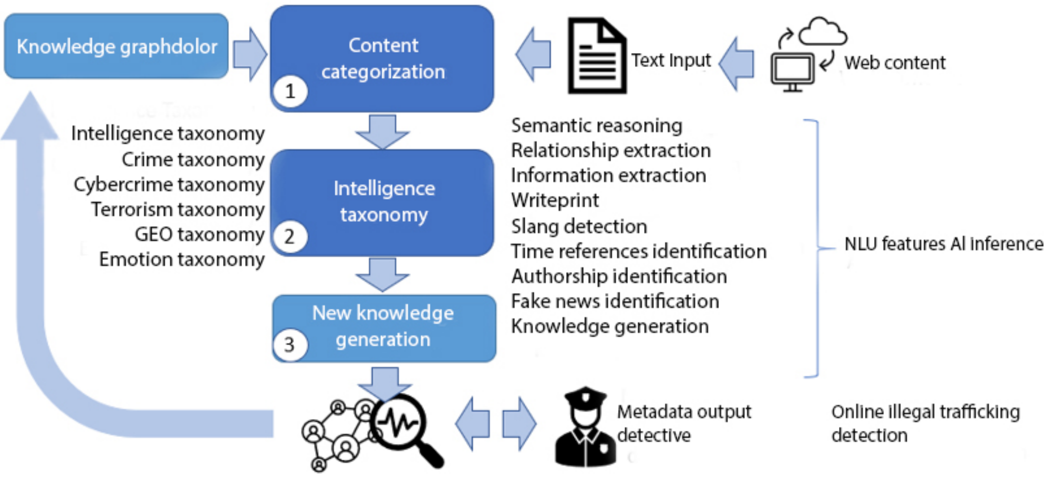


Figure 1. *The Three-Phases Approach of ANITA Lateral Thinking Analysis*

METHODS

Experimental Approach to the Problem

Manual monitoring of the web is impossible even for the largest and most structured organizations. Adoption of AI is required and fortunately there are technical tools that can provide valuable support. The methodology we are proposing involves both semantic Natural Language Processing (NLP) engine and a new approach to problem solving and investigation, through three stages: content categorization, information extraction and new knowledge generation.

First Stage – Content Categorization

This stage can be described as the process in which concepts and objects are recognized, distinguished and finally understood. In particular, this process implies that objects are known and grouped into categories based on their properties. Categories are generated first by formulating their conceptual descriptions and then classifying objects according to the descriptions. This process involves the abstraction of rules that relate the observed object features to category labels. In other words, categorization involves recognizing intrinsic structure in a set of data and grouping objects by using a similarity metric into classes, thus generating a classification structure. It is based on a knowledge graph (according to the knowledge contained in the semantic network called Sensigrafo[1], namely a logical

[1] The Expert.ai Knowledge Graph is an open box, meaning its structure is created by humans and is understandable by humans, the opposite of what happens with pure machine learning systems. Every item in the Knowledge Graph contains a set of attributes (grammar type, semantic link, definition/meaning, domain, frequency) that establish the characteristics of words and concepts. It contains words gathered in groups of words and concepts that express the same meaning and are connected to each another by millions of logical and language-related links describing the relationship between concepts. In the Knowledge Graph, each syncon is a node that is linked to other nodes through semantic relationships in a hierarchical structure. In this way, each node, in addition to its meaning and attributes, is enriched by the characteristics and meaning

and linguistic representation of the lexical database of a language. Unlike traditional dictionaries, here the words are arranged in groups of items expressing identical of similar meaning, called Syncons. Sensigrafo and its Syncons can be customized in order to achieve greater precision and relevancy) and it is performed through the calculation of semantic similarity (namely, their likeliness of meaning or semantic content) (Wang et al., 2018). The categorization engine applies one or more categories to a document when it detects nouns, verbs, or a combination of these which are related to crimes. In this process all the contents and web pages downloaded by the crawlers are filtered on the basis of six domain taxonomies specifically adopted for ANITA: *Intelligence Taxonomy, Crime Taxonomy, Cyber Crime Taxonomy, Terrorism Taxonomy, Geo Taxonomy and Emotions Taxonomy.*

Intelligence Taxonomy

The Intelligence taxonomy is a domain-specific verticalization of the IPTC taxonomy (About IPTC, n.d.), that has been enriched with several categories which mostly pertain to the Intelligence sector's domains such as: illicit trafficking, and arms trafficking in the category "Crime"; criminal organizations and terrorist organizations in the category "Organized Crime"; illicit drugs and controlled substances in the category "Medicine"; paramilitary organizations, defence companies, intelligence agencies and defence contractors in the category "Defence"; cyberwarfare in the category "Act of terror". The proposed examples are just a few. The whole categorization engine developed from the taxonomy is capable of handling 500 categories.

Crime Taxonomy

The Crime taxonomy is based on the standard model as defined under the "Swedish Initiative" (Framework Decision, 2018) ‐ a common protocol and knowledge model which promotes the effective exchange of information and strategic data between the EU States law enforcement authorities. This standard was used as the basis to create a taxonomy structure made up of 32 categories, organized on one single level, where each category represents a crime or a group of crimes. Some of the categories included: "human trafficking", "environmental crimes", "cybercrimes".

Cyber Crime Taxonomy

The Cyber Crime taxonomy consists of 40 categories arranged in three hierarchical levels. The taxonomy and the derived categorization engine focus on two main macro-issues: cyber illegal and cyber security, which are populated with a plethora of sub-domains in order to achieve an accurate categorization of the cyber illegal domain. Examples of second-level (and third-level) sub-domains included in the Cyber Crime Taxonomy are: cyber-attacks (especially DoS attacks and cyber intrusion); content-related computer

of the nodes above or below it. Each of these links identifies a kind of relationship that links the concepts in a language, and the links are principles used to organize the concepts in the Knowledge Graph. For example, concepts are organized starting from less specific to more specific (vehicle/car/SUV), or as potential subjects or objects of a verb, etc.

crime (cyber piracy); cyber security (vulnerabilities, security software and devices, and threats such as malwares and ransomwares); cyber deception (credit card fraud, e-mail spamming, scam, identity theft, phishing); hackers.

Terrorism Taxonomy

The Terrorism Taxonomy features 62 terrorism-related categories arranged in a three-level hierarchical organization. It includes a categorization of terrorism by matrix (religiously inspired terrorism, ethno-nationalist and separatist terrorism, anarchist, left- and right-wing extremism, narco- and eco-terrorism), alongside with terrorist attacks by targets (terrorist attacks on critical infrastructures, NGO's, politicians or institutions, civilians, properties, military and police forces), and critical events and threats potentially related to terrorism (rebellions, demonstrations, disorders, and other extremist activities). Also, the terrorism organizational model is categorized (propaganda and ideology, recruitment and training, terrorism financing), as well as the terrorist tactics (hijacking and skyjacking, kidnapping and hostage taking, assassination, guerrilla, bombing etc.).

Geo Taxonomy

The Geo taxonomy is a taxonomic structure composed of 250 categories, where every category represents a country. The categorization engine, developed from the taxonomy, assigns a category when it detects at least one geographic entity which is associated with a country contained in the semantic network (Sensigrafo).

Emotions Taxonomy

The Emotions taxonomy arranges the main human emotions (Cowie, 2009; Parrott, 2001) and behaviours into a one-level taxonomic structure, where each category represents an emotion or behaviour. The categorization engine, developed from the taxonomy, categorizes documents according to 75 emotion and behaviour related categories. The full path of the categories assigned to the document is returned along with the score and the frequency of each category, in order to define the category's relevance for the analysed document. For ANITA, we have decided to work on the detection of all emotions available with the focus on five key emotions: Anger, Disgust, Fear, Happiness, and Sadness that the LEAs have considered the most relevant for the investigation (Francisco & Gervas, 2013; Sunghwan, 2011; Russell, 1980; Calvo & Kim, 2013).

Second Stage – Information Extraction and Semantic Reasoning

Text mining and named entity recognition (NER) are tasks related to information extraction which aims to locate and classify a text's atomic elements into predefined classes (entity types) such as the names of people, organizations, locations, monetary values, etc. In ANITA, there are many domain-specific entity types such as weapons and firearms (i.e., gun, hunting, rifle, Kalashnikov, taser), counterfeit/falsified medicines (OTFM), drugs (Cocaine, Crystal Meth, Diazepam, Ecstasy), NPS - Novel Psychoactive Substances (Gamma-aminobutyric acid (GABA), Skag), terrorism funding (Islamic Terrorist Organizations, Methods of Money Transfer, etc). Within ANITA we use this functionality for the identification and inference of custom entities and background knowledge facts defined as *precursors*. It is at this stage that the skills and expertise of the investigator are intertwined with the brute force of AI, as it is the domain expert's task to identify in the first phase, the main concepts or objects/substances related to the production or sale of illicit substances but as we will see, the system is able to propose new concepts in autonomy. Other entity types that are often useful like dates, quantities, currencies, time references, etc. are also automatically extracted.

Semantic reasoning is a powerful function which extends and amplifies the Information Extraction feature. In fact, for the entities related to the Intelligence and Security domain, it is possible to automatically infer information not present within the text thanks to the knowledge and the relationships between concepts contained in the Sensigrafo. For example, analysing the text: "Biden is a leader" text analysis usually extracts Joe Biden as a People entity and a World Leader, but Semantic Reasoning based on knowledge graph, goes a few steps beyond to show that Joe Biden is a → president → politician → head of a state → United States of America.

Relations extraction

In addition to extracting entities, ANITA's Text Analysis Service can also extract the "relationships" which may exist between these entities thanks to a dedicated module. Relationship is defined as the conceptual connection between two entities co-occurring within a given scope (sentence). The outcome is a triplet consisting of the connected entities and the type of connection which links them. The module aims to extract the following types of relationships: Communicative (e.g., explain, report, say, etc.); Contact; Criminal action (e.g., hack, etc.); Economic; Existence; Family; Generic (e.g., wage, etc.); Law; Movement (e.g., flee, etc.); Origin; Possession; Social activity.

Writeprint

Ietlt is possible to recognise the author of a text even when it is not handwritten but typed on a computer through studying the so-called stylemes, namely the stylistic features intended as the linguistic choices an author takes during the writing process. They define the personal writing style, which is unique for each author just like his fingerprint.

The Writeprint feature provides semantic, grammatical, structural and statistical text indexes with the purpose of authorship assessment and profiling. Moreover, Writeprint can outline the readability level of a document and predict the grade of education needed to understand it (Figure 2).

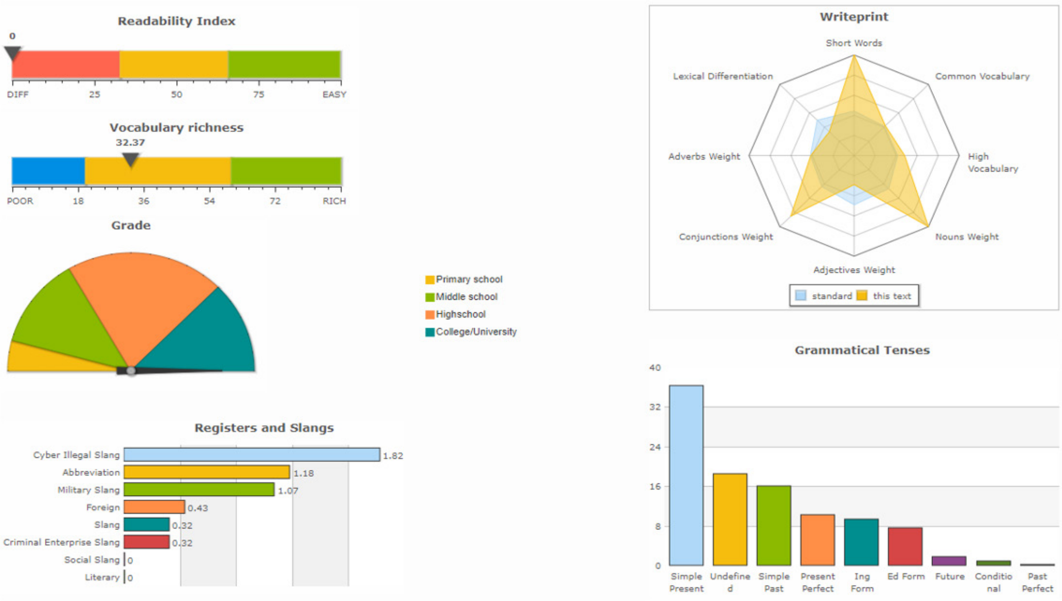
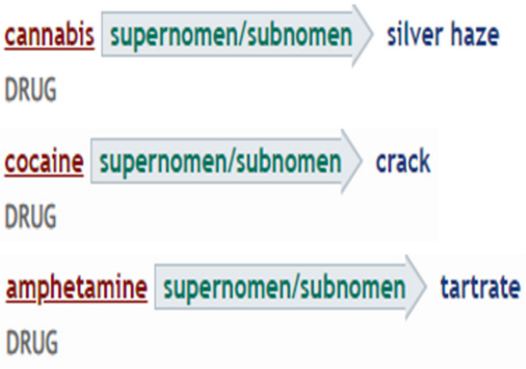


Figure 2. *Writepint Visualization*

Slang Detection

It concerns the identification of non-standard words used by specific groups (Milde, 2013; Elsahar, 2014; Pal & Saha, 2013). In ANITA we focus on the jargons related to weapons and firearms, counterfeit/falsified medicines (OTFM), drugs, Novel Psychoactive Substances (NPS), terrorism funding. Regarding drugs (Figure 3), the 1172 concepts that populate the Sensigrafo coincide with an integration of the complete lists available on UNODC's Multilingual Dictionary of Narcotic Drugs and Psychotropic Substances under International Control and Drugs of Abuse (U.S. Department of Justice, 2017). From the sources analysed, 5911 total new lemmas have been harvested: 586 of which are slang words, while the remaining 5325 are name variants, chemical formulas, mixtures with other substances, and cannabis strains (which are more than 760!).



cannabis supernomen/subnomen silver haze
DRUG

cocaine supernomen/subnomen crack
DRUG

amphetamine supernomen/subnomen tartrate
DRUG

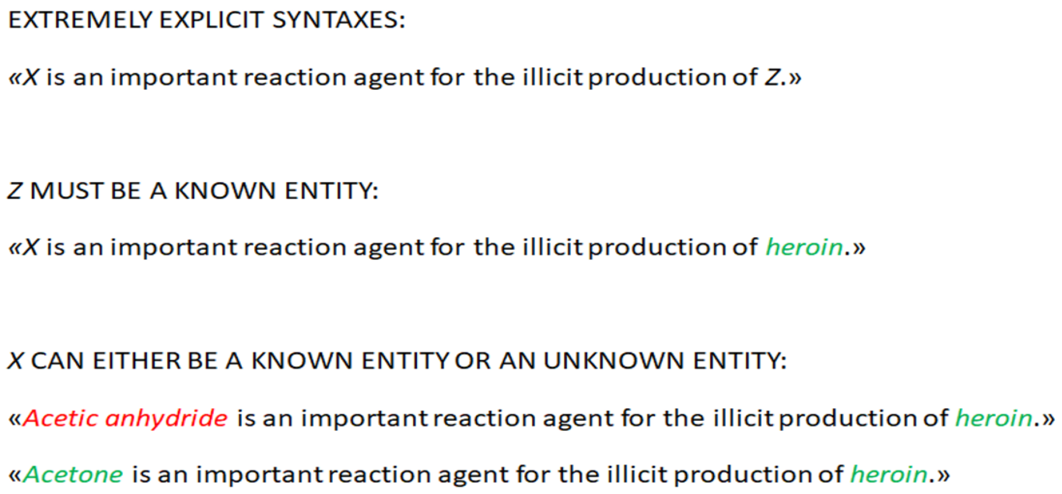
Figure 3. *Example of Slang Detection*

Third Stage – New Knowledge

ANITA's text analysis services have been filled with a huge amount of knowledge and it has to be acknowledged that the domain in which ANITA is immersed evolves to a pace so fast that it would be nearly impossible to keep up with it. Probably, the verticalizations and the algorithms made up to now would not be enough in a time of even one year, due to the uncountable new substances and patterns that could possibly appear in the scene of the illegal trafficking of illicit products, which are led by the needs of the criminal actors to always find new pathways to avoid the law. That is why ANITA adopts a special workflow which integrates algorithms based on NLU with the expertise of the investigators to offer a system capable of detecting, understanding and extracting "New Knowledge" from raw text.

The term "New Knowledge" refers to all the concepts and relations that are not contemplated by the existing Knowledge Base but are nonetheless relevant for the scope of the project, such as new, not-yet-controlled precursors of already known illicit substances or products, or simply precursors that are already known but were not included in the concepts of the verticalization. Normally, an entity cannot be extracted by Extraction/Transformation/Load algorithms if the latter do not possess the instruction for their identification, that is usually given by the Knowledge base or by lists of keywords/lemmas. However, there are special contexts from which new knowledge can be acquired and understood by a human reader, and Expert.ai has found a way to detect and exploit them: innovative algorithms have been designed and developed, to mimic the human ability to grasp new content from certain syntaxes and apply the usual entity extraction on them.

The system is able to recognise new knowledge in the sentences like the one reported as an example and, on the basis of previous knowledge (i.e., information already present in the knowledge graph) can recognize and extract elements already known (in green) or at the moment unknown (in red) or not present in the investigator's database (Figure 4). It is fundamental to highlight how the part of the text from which the system has inferred the particular information is always clearly reported, and it is possible to consult and verify the various concepts by the human expert, according to the paradigms of explainable AI..



EXTREMELY EXPLICIT SYNTAXES:

«*X* is an important reaction agent for the illicit production of *Z*.»

Z MUST BE A KNOWN ENTITY:

«*X* is an important reaction agent for the illicit production of *heroin*.»

X CAN EITHER BE A KNOWN ENTITY OR AN UNKNOWN ENTITY:

«*Acetic anhydride* is an important reaction agent for the illicit production of *heroin*.»

«*Acetone* is an important reaction agent for the illicit production of *heroin*.»

Figure 4. *Process to Recognize New Knowledge*

TECHNICAL ASPECTS

Rules development for the categorization of the documents. In the Expert.ai environment, the term "rules" refers to deterministic algorithms manually written by developers in the proprietary languages of Expert.ai. The proprietary language used for categorization is called "C language", and the categorization rules (algorithms) written in C language are called "C rules" (where C stands for Categorization).

On a conceptual level, to develop a C rule means to identify a set of conditions that, if verified, allow the text in analysis to be assigned with one or more categories in the taxonomy. Of course, speaking of Expert.ai technology, the conditions defined by the algorithms are based on semantics. In fact, C rules do not match simply with mere strings or tokens, since not all the elements that make up a text are relevant to determine what the document is about. They rather focus on the elements that bear the meaning requested by the stated condition.

In more detail, each C rule is univocally associated to a "domain" (intended in Cogito as a category of the taxonomy), and defines a "condition", namely the presence (or absence) of one or more linguistic elements (words, concepts, etc.) in a certain syntactical range. This range, this portion of text, is called "scope". If the defined condition is verified within the given scope, then the rule triggers (or, in Expert.ai terminology, "instantiates"), and assigns a "score" to the domain (taxonomical category) of the rule. Scores are arbitrary numbers according to certain criteria and defined in the rule as well.

Once a text is fully analyzed, and each condition set by all the rules written is verified or not in the text, the text itself is labelled with a set of one or more domains. Each one of these categories possesses a certain score, given by the sum of all the single scores defined in the rules that have instantiated (triggered because their conditions are verified). The scores rank the categories by frequency: the domain/category with the highest score is the most found domain/category, and so on.

Now, a brief recap before providing some examples: it is possible to identify four main elements into a C rule, corresponding to the concepts typed in bold in the paragraph above:

- Scope: the portion of text on which the rule has to be applied;
- Condition: the set of one or more linguistic elements that have to be found in the given scope in order to instantiate (trigger) the rule;
- Domain: the name of one of the categories in the taxonomy, that will be assigned to the text if the condition set by the rule is verified.

Score: an arbitrary number assigned to a domain if its corresponding condition is verified, used to determine the frequency of the domains of the text.

```
    //optional comment describing the rule
SCOPE scope_option
{
    DOMAIN(domain_name:score_option)
    {
        //condition//
    }
}
```

Here the scope_option defines the scope, and it can be, for example, a sentence, a nominal phrase, a relative clause etc.

The domain_name defines the name of the domain, and it can be one of the categories belonging to the taxonomy

The score_option defines the score to assign to the domain if the condition of the rule is verified. It is totally arbitrary, but developers usually assign multiples and sub-multiples of 5 and 10. The choice of the score to assign depends on the specific needs of the project, such as the relevance of the categories to identify (the higher the importance, the higher the score), or on some textual aspects, like syntaxes or other structures that are more representative of a given category than others, condition defines the condition to verify, and it is written in proprietary attributes written in C language.

For a practical example, let us consider the domain of interest of the ANITA project Drugs and Weapon interested in categorizing. The developer receives this text: "cannabis, cocaine and ecstasy are often sold to buy AK-47". The rules he develops are the following:

```
//Rule for the categorization of drugs
SCOPE SENTENCE
 {
DOMAIN(Drugs:10)
   {
        ANCESTOR(1000013AA)//drug
    }
}
//Rule for the categorization of weapons

{

SCOPE SENTENCE
 {
     DOMAIN(Weapons:10)
     {
         ANCESTOR(1000109BB)//weapon
     }
 }
```

These rules use the ANCESTOR attribute, which sets a condition that is verified if the text contains the concept (syncon) indicated in brackets, or any of his hyponyms (Cambridge University Press, n.d.) contained in the Sensigrafo.

The first rule assigns 10 score points to the domain "Drugs" if the concept (syncon) of "drug" (which has the unique identifier 1000013AA inside the Sensigrafo), or any of its hyponyms, is found in a sentence. This rule triggers three times, since three hyponyms of the concept of "drug" are found in the text (cannabis, cocaine and ecstasy). This means that the domain "Drugs" receives a total of 30 points.

The second rule assigns 10 score points to the domain "Weapons" each time the concept (syncon) of "weapon" (which has the unique identifier 1000109BB inside the Sensigrafo), or any of its hyponyms, is found in a sentence. This rule triggers one time, since the one hyponym of the concept of "weapon" is found in the text (AK47). This means that the domain "Weapons" receives a total of 10 points.

The final categorization result will then be the set of the following categories:

- Drug (30 points, frequency 75%);
- Weapon (10 points, frequency 25%).

We can consider a valid set of rules to be a variable number of rules where each is associated to only one domain and each domain can have one or several rules associated to it.

The same rule may "instantiate" (trigger) more than once on different parts of the text, just like different rules may instantiate on the same part of the text. Both cases are contemplated and correct.

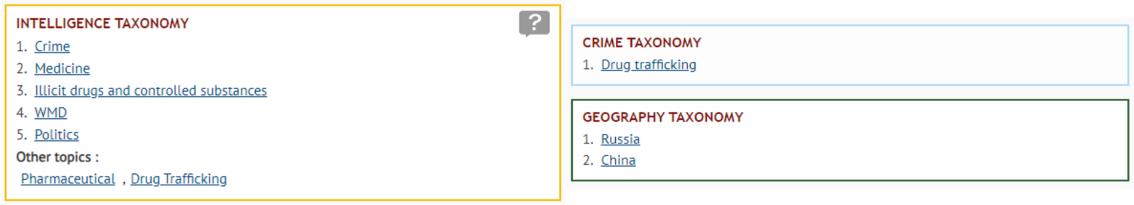
Eventually, errors such as undesired triggers (false positives), or missed triggers (false negatives), are dealt with if they arise in a first phase of test. With the rules it is also possible to handle negations such as "this is not cannabis".

RESULTS

In order to give a clearer idea of the results obtained, examples are given for all three phases included in the process. About content categorization, we explained that the scope is to understand the topic of each text automatically analysed, let us take this text as an example (Kleemans & Soudijn, 2017):

> *"The first measure was aimed at two particular precursor chemicals, i.e., PMK in the case of ecstasy and BMK for amphetamine. These chemicals are the building blocks for certain synthetic drugs and had therefore already been placed under strict international control. By way of smuggling, ecstasy producers were able to get these chemicals from two particular source countries, Russia and China. To counter this problem, high-level discussions with Russian and Chinese authorities (and international pressure from the European Union and the International Narcotics Control Board) resulted in stricter control of legal PMK and BMK producing facilities. As a result, shortages of these precursor chemicals arose, and prices started to rise. As Vijlbrief mentions: "In the language of situational crime prevention, criminals had to increase their efforts, increase their risks, and encounter reduced rewards."*

The result of the automatic Content Categorization Module is presented in Figure 5.



INTELLIGENCE TAXONOMY

1. Crime
2. Medicine
3. Illicit drugs and controlled substances
4. WMD
5. Politics

Other topics :

Pharmaceutical , Drug Trafficking

CRIME TAXONOMY

1. Drug trafficking

GEOGRAPHY TAXONOMY

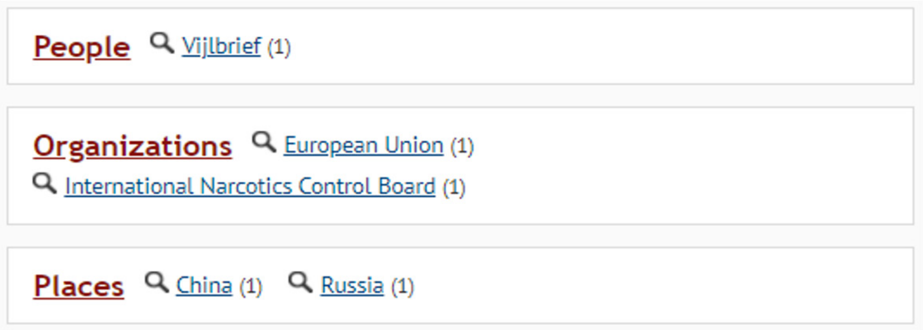
1. Russia
2. China

Figure 5. *Content Categorization Module Output*

These results provide a complete picture of what the selected text contains and thus enables investigators to search for text following a semantic approach and not simply based on keywords. In the case of crawling from the web of thousands of documents whose content is not known a priori, this categorization also allows us to filter the corpus and focus on only those relevant to the investigation.

By analysing about 100 documents manually, it was possible to evaluate the categorisation module and the following results were obtained, as Precision 90%, Recall 80%, and F-Measure 84.7% ("Precision and recall," 2021).

Analyzing the same text with the Information Extraction Module the obtained output is presented in Figure 6.



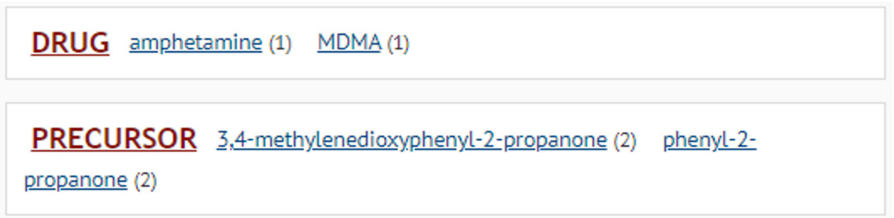
People Vijlbrief (1)

Organizations European Union (1) International Narcotics Control Board (1)

Places China (1) Russia (1)

Figure 6. *Information Extraction Module Output: People, Organizations, and Places*

In addition to the extraction of the generic entities People, Organisations and Places, it is important to note instead this second part of the output (Figure 7).



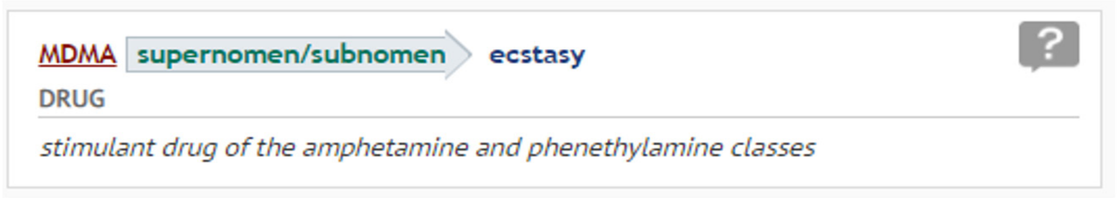
DRUG amphetamine (1) MDMA (1)

PRECURSOR 3,4-methylenedioxyphenyl-2-propanone (2) phenyl-2-propanone (2)

Figure 7. *Information Extraction Module Output: Drug and Precursor*

Note that drugs amphetamine and MDMA are extracted, but in particular note that the word MDMA was never explicitly mentioned in the text analysed. The semantic reasoning functionality has in fact deduced the most distinctive form, namely "MDMA".

Finally, "supernomen/subnomen" stands for the relation that links "MDMA" with the lemma "ecstasy" (Figure 8).



MDMA supernomen/subnomen ecstasy

DRUG

stimulant drug of the amphetamine and phenethylamine classes

Figure 8. *Information Extraction Module Output: Sematic Reasoning*

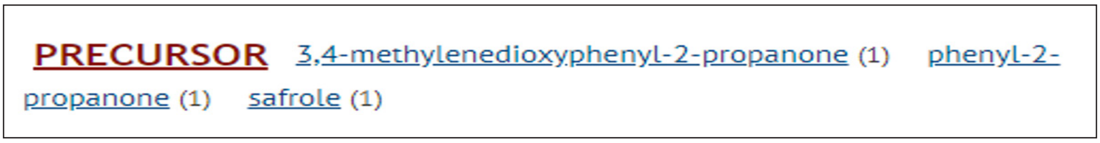
In addition, two precursors have been identified among the substances in the text extract analysed: 3,4-methylenedioxyphenyl-2-propanone (PMK), and phenyl-2-propanone (BMK) automatically recognised again without being explicitly mentioned, but only through their acronyms.

By analysing about 100 documents manually, it was possible to evaluate the Information Extraction module and the following results were obtained, as Precision 85%, Recall 78%, and F-Measure 81.3%.

Very interesting results were also obtained by analysing the module of new knowledge generation. Analysing the following text extracted from (Kleemans & Soudijn, 2017):

> *"Indeed, the almost perfect solution of blocking two important chemical precursors backfired over time. Instead of importing the prohibited but necessary chemicals, offenders started to produce these chemicals themselves, resulting in the use of pre-precursors (such as safrole for PMK and APAAN for BMK)."*

in fact, among the substances listed in the text reported, safrole, 3,4-methylenedioxyphenyl-2-propanone (PMK), and phenyl-2-propanone (BMK) were contained in the Sensigrafo, hence they are contemplated by the Knowledge Base and extracted by ETL algorithms, while APAAN is not (Figure 9).



PRECURSOR 3,4-methylenedioxyphenyl-2-propanone (1) phenyl-2-propanone (1) safrole (1)

Figure 9. *New Knowledge Generation – Precursor*

Furthermore, thanks to the relations contained in the Knowledge Base, it is known that safrole and 3,4-methylenedioxyphenyl-2-propnaone are precursors of MDE, MDA, and MDMA, and that phenyl-2-propanone is a precursor of MDMA.

Nevertheless, it is made clear by the shown context that APAAN is a substance used to synthesize the known controlled substance and precursor of Phenyl-2-propanone. Therefore, APAAN should be considered as a Precursor of Phenyl-2-propanone (and pre-precursor of MDMA as well), and should enter the ANITA's Knowledge Base as such, alongside with the mentioned relations.

As for the NLU approach, New Knowledge has been tested to calculate its performance in terms of precision and recall. Also, here the process was manual with a non-tagged corpus retrieved by Expert.ai's crawler, but it was led before the work of verticalization of the Sensigrafo, in order to find as much new entities and relations as possible. The documents were 18, containing a total of 54 new entities and relations.

The first analysis made by the New Knowledge algorithms for entities retrieved 20 true positives, six false positives, and 16 false negatives. The first 20 new entities have been validated by investigators; a second round of analysis detected other six true positives as pre-precursors.

Finally, the last round of analysis was led with New Knowledge algorithms for relations, which were capable of re-tracing all the relations pertaining to all the new entities retrieved in the first two rounds, for a total of 20 + 6 true positives. The six false positives returned in the first analysis were discarded and so, they would not enter the KB, thus disappearing from further rounds of analysis. The 16 false negatives were instead maintained.

The result is an encouraging and promising score, since it translates in 81 out of 100 useful new entities and relations, Recall 62%, and F-Measure 70%.

DISCUSSION

The methodology proposed in this article, as well as validated by laboratory experiments, has been positively evaluated by the LEAs involved in the ANITA project finding a lot of interest in many stakeholders operating in the world of investigations where intuition and human ability remain fundamental to follow the right track but the amount of data to be analyzed is unsustainable with human resources alone. The ability of AI to automatically suggest new substances or products relevant to the investigation, directly extracted from the analysis of thousands of online contents, enables a fast, efficient and low cost update that can improve the investigative capacity both thanks to new concepts that are introduced in the knowledge base becoming a shared knowledge and thanks to a methodological approach to problem solving focused on elements and features apparently secondary and divergent from the object of the investigation that can give new insights and innovative research keys.

Structuring in the knowledge base the precursors' relations, that is constitutive and essential element of a product or process, represents a lever for the investigator of great potential and interest and certainly does not represent itself an absolute novelty in the methodological approach to investigation. In various projects we have seen the codifying of experts' knowledge in a form suitable for computational processing but, once realized, one of the typical criticalities is its constant updating that can be very expensive. In this article we propose a methodology able to automatically intercept part of the evolutions adopted by the world of crime, reducing the effort (and costs) of updating by the police forces and structuring the knowledge in a shared way, useful to all, not in the heads of a few.

CONCLUSIONS

The use of Artificial Intelligence can impact the modus operandi of investigators, who can adapt their methodologies and procedures based on the results of a new type of "robot-assistant" able to perform tasks impossible for a human, such as reading and analysing millions of texts. Certainly, there are still many limits for the AI in the ability to intuit and follow a track, even by the most sophisticated systems, therefore the leadership of the process must remain in the hands of the expert investigator (supervised process), who however can find great benefit from the robot to remain constantly updated with respect to the continuous evolutions made by criminals in their activities.

The methodology presented here, which uses a *lateral thinking approach* integrated with the capability of the most sophisticated and customized *NLP technology for investigation* support, allows to automatically intercept suspicious activities thanks to the recognition of all the elements that are used to build or sell substances or weapons in the illegal online market. This approach represents an example of how AI can be integrated into investigative processes, fully exploiting its added value. In particular, the ability to identify and propose new elements (*precursors*) that emerge from the analysed texts that can, once validated by the expert, update and enrich the domain knowledge and improve the subsequent analysis and investigation

ACKNOWLEDGMENT

The work presented in this paper was supported by the European Commission under contract H2020-787061 ANITA Advanced tools for fighting oNline Illegal TrAfficking.

REFERENCES

About IPTC. (n.d.). *IPTC*. https://iptc.org/about-iptc/

Calvo, R., & Kim, S. (2013). Emotions in text: Dimensional and categorical models. *Computational Intelligence*, *29*(3), 527‐543. https://doi.org/10.1111/j.1467-8640.2012.00456.x

Cambridge University Press. (n.d.). Hyponym. In Cambridge dictionary. https://dictionary.cambridge.org/it/dizionario/inglese/hyponym

Cowie, R. (2009). Perceiving emotion: Towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*, 3515–3525. https://doi.org/10.1098/rstb.2009.0139

De Bono, E. (2016). Lateral thinking: A textbook of creativity. Penguin.

Dilek, S., Çakir, H., & Aydin, M. (2015). Applications of artificial intelligence techniques to combating cyber crimes: A review. *International Journal of Artificial Intelligence & Applications* (IJAIA), 6(1), 21–39. https://doi.org/10.5121/ijaia.2015.6102

Elsahar, H. (2014). A fully automated approach for Arabic slang lexicon extraction from microblogs. In *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8403 LNCS, pp. 79–91). https://doi.org/10.1007/978-3-642-54906-9_7

Framework Decision 2006/960/JHA (Swedish Initiative) Non-communication of national implementing measures. (2018, November 8). European Union, European Commission. https://ec.europa.eu/home-affairs/whats-new/infringement/framework-decision-2006960jha-swedish-initiative-non-communication-national-implementing-measures_en

Francisco, V., & Gervas, P. (2013). Emotag: An approach to automated mark-up of emotions in texts. *Computational Intelligence*, *29*(4), 680‐721. https://doi.org/10.1111/j.1467-8640.2012.00438.x

Kleemans, E. R., & Soudijn, M. R. J. (2017). Organised crime. In N. Tilley and A. Sidebottom (Eds.), *Handbook of crime prevention and community safety*. Routledge.

Li, S-T., Kuo, S-C., & Tsai, F-C. (2010). An intelligent decision-support model using FSOM and rule extraction for crime prevention. *Expert Systems with Applications: An International Journal*, *37*(10), 7108–7119. https://doi.org/10.1016/j.eswa.2010.03.004

Lin, Y-L., Chen, T-Y., & Yu, L. (2017). Using machine learning to assist crime prevention. *2017 6th IIAI International Congress on Advanced Applied Informatics* (IIAI-AAI) (pp. 1029–1030). https://doi.org/10.1109/IIAI-AAI2017.46

Maio, E. (2016, February 23). Lateral thinking: The key to effective deviation investigation. *IVT Network*. https://www.ivtnetwork.com/article/lateral-thinking-key-effective-deviation-investigation

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal*, *2*(1), 1–12. https://doi.org/10.5121/mlaij.2015.2101

Milde, B. (2013). Crowdsourcing slang identification and transcription in twitter language. *Gesellschaft für Informatik eV* (GI), 51.

Pal, R. A., & Saha, D. (2013). Detection of slang words in e-data using semi-supervised learning. *International Journal of Artificial Intelligence & Applications*, *4*(5), 49-61. https://doi.org/10.5121/ijaia.2013.4504

Parrott, W. G. (2001). *Emotions in social psychology: Essential readings* (Vol. 14). Psychology Press.

Precision and recall. (2021). In *Wikipedia*. https://en.wikipedia.org/wiki/Precision_and_recall

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Sunghwan, M. K. (2011). *Recognising emotions and sentiments in text* [Doctoral dissertation, University of Sydney].

U.S. Department of Justice. Drug Enforcement Administration. (2017). *Drugs of abuse: A DEA resource guide*. https://www.dea.gov/sites/default/files/drug_of_abuse.pdf

Wang, C. (2018). Information Retrieval Technology Based on Knowledge Graph Conference. Proceedings of the 2018 3rd International Conference on Advances in Materials, Mechatronics and Civil Engineering (ICAMMCE 2018). https://doi.org/10.2991/icammce-18.2018.65