Reputation Analysis towards Discovery

Raffaele Palmieri, Vincenzo Orabona, Nadia Cinque, Stefano Tangorra and Donato Cappetta
Eustema Spa, via Carlo Mirabello 7, 00195, Roma, Italy

Keywords: Open Source Intelligence, OSInt, Reputation Analysis, Supplier Risk Assessment, Company and Brand Reputation, Semantic Search, NLP, Crawling, Information Extraction, Information Discovery.

Abstract: This work describes the development and the realization of an OSInt solution conducting a supplier risk assessment, focused on the evaluation of suppliers' reputation starting from publicly available information. The main challenge is represented by the data processing phase that exploits NLP technologies to extract facts, events, and relations from unstructured sources, building the knowledge base for reputational analysis. Several measures have been adopted to provide a satisfactory user experience; however, further integrations are still needed to increase efficiency of the developed solution. Particularly, it is necessary to deepen and improve the analysis over the huge volume of data coming from open sources, enhancing the discovery of all possible relevant information influencing the reputation of the targeted entity.

1 INTRODUCTION

One of the main concerns of governmental organizations and companies is to know and be aware of their own standing on the web. For this reason, as well as for the establishment of their strategies and marketing initiatives, these actors express an increasing need to intercept the constant flow of news and information on the internet.

Furthermore, the increasing massive use of social media generates a conflictual overload of information that consequently reinforces the need for distilled information through an expert analysis.

This is particularly important within a business context, where a company normally conducts a risk assessment on business counterparties, giving strong relevance to reputational risk. Although our experience targets supplier risk assessment, it can be applied to whatever kind of counterparty, such as contractors, agents, customers, partners, etc.

The main objective of the work described in this paper is to implement a solution for counterparty risk assessment, focusing on reputational risk. Through the automation of the analysis on publicly available information, it is possible to facilitate the human activity, offering tools to detect promptly those events that influence reputation.

The overall goal is challenging, because information on the web is heterogeneous and difficult to manage; furthermore, the automation process needs information extraction (IE) tools that have a limited efficacy and it is also important to consider that reputation tends to be a subjective concept.

The approach is based on a consolidated OSInt (Open Source Intelligence) process that foresees a set of phases, during which several tools can be used to support analyst's work.

The considered sources are public databases of companies' data, such as legal name, social capital, revenue, composition of its board of directors, as well as online newspapers, feeds, public judgements, from which it is possible to extract entities, events and relations to build the knowledge base.

This paper is organized as follows. The second section describes the related work, synthetizing the scientific and technological state of art, most of which regards software and tools. Section 3 describes the adopted methodology and the solution architecture, starting from the OSInt process. Section 4 shows results of IE and section 5 describes lesson learned and some functional and design changes that we could apply to the solution. Furthermore, the section highlights some new possible use cases.

2 RELATED WORK

Open Source Intelligence (OSInt) is the gathering of information by accessing public access sources (Stalder and Hirsh, 2002); (Glassman and Min Ju

Palmieri, R., Orabona, V., Cinque, N., Tangorra, S. and Cappetta, D.
Reputation Analysis towards Discovery.
DOI: 10.5220/0006487303210330
In *Proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017)*, pages 321-330
ISBN: 978-989-758-255-4

Kang, 2012).

The strategic importance of OSInt process has been reported by military organizations (NATO, 2002). However, there is a growing interest towards OSInt in the business sector for its implications in business and competitive intelligence (Fleisher, 2008). Social media data are central to processes as Social Media Analytics, at individual product level, in the field of competitive intelligence. (He et al., 2015).

The techniques, algorithms and software applications used are many. These include the natural language processing (NLP) techniques used for acquisition of large volumes of text and automatic classification. Advanced versions of NLP techniques allow the automatic corpus-driven construction of domain ontologies (Navigli and Velardi, 2004), including the more frequent relations to perform event extraction task in financial domain (Basili et al., 2002).

Automatic Content Extraction (ACE) is the research program launched in (Doddington et al., 2004) to develop advanced information extraction techniques in order to extract from human language entities' mentions, relations and events.

(Palmer et al., 2005) proposed Proposition Bank, a semantic representation of clauses of a hand-check corpus of Penn Treebank, annotating verbs and surrounding elements according to predefined templates, called framesets. The work did a comparative study with Framenet (Baker et al., 1998).

In (Mintz et al., 2009) the authors describe a distant supervision approach based on the assumption that each sentence with a pair of entities expresses a relation between them. Starting from Freebase's relations, the work extracts the lexical and syntactic features of sentences containing two involved entities and assigns them to a multiclass logistic regression classifier, learning weights for each noisy feature.

In (Krause et al., 2012), the authors start from distant supervision to extract n-ary specific relations, learning those rules which model the links between relation arguments using dependency parsing.

(Mausam et al., 2012) present a system for learning pattern templates, starting from a corpus of asserted relations. It overcomes two weaknesses of some compared IE systems, such as the mediation of verbs to extract relations, and the non-consideration of context for the analysis.

In (Riedel et al., 2010), the authors propose a modified version of distant supervision, reaching an important error reduction.

In (Liu et al., 2016) the authors showed how a crowdsourced approach for annotation can enhance performance of IE tasks based solely on distant supervision. They propose a Gated Instruction Crowdsourcing Protocol, which includes an interactive tutorial to train and screen knowledge workers. Experimental results showed enhancements of both precision and recall for nationality, place of birth, place of residence, and place of death relations.

In (Vossen et al., 2016) the authors describe a system that processes massive streams of news (millions of articles) in four languages (English, Dutch, Spanish and Italian) to build structured events, related to financial and economic domain, supporting decision-making. Their task is to build event-centric information, identifying the actor and spatial-temporal coordinates of an event, representing it in an interoperable RDF format. They implement the multi-language pipeline referencing different types of data sources embodying specific knowledge: semantic knowledge, such as Wordnet, annotated corpora (e.g. PropBank, Framenet, etc.) and episodic knowledge from open source data (e.g. DBpedia).

Recently, the use of deep neural networks applied to NLP tasks has emerged as one of the the main research trend.

(Zhou et al., 2016) show how the performance of their proposed model outperforms most of the existing approaches, without using lexical resources such as WordNet.

In (Kumar, 2017), there is a comparative study between various kinds of deep and non-deep neural models applied to multi-instance learning for relation extraction task. It is noted that all deep learning models perform significantly better than the non-deep learning models.

About the sentiment analysis (Dave et al., 2003) (Pang and Lillian, 2008), it involves automatic processing of opinions, feelings and subjective features based on the availability of digital resources enriched with opinions, such as blogs and social networks. Therefore, the main interest is directed towards Social Network Analysis (SNA) (Barabási and Frangos, 2002). SNA software tools include the Stanford Network Analysis Platform (Leskovec and Sosic, 2016) and the Social Networks Visualizer (Kalamaras). Other techniques of interest in OSInt solutions are the decision trees (Quinlan, 1986), the real time monitoring and situation detection (Nguyen et al., 2005).

Data mining tools are jointly used, among them the most known are R (R Foundation for Statistical Computing, 2013), Weka (Frank et al., 2016) and Apache Mahout (Apache), a machine learning library well known for its scalability features. (TensorFlow) is an advanced and very scalable machine learning

library, which implement various models for natural language understanding.

Some examples of products, directly developed as OSInt solutions, are Shodan (Shodan), Maltego (Paterva) and Cogito Intelligence Platform (Expert System). Other products not specifically designed for this target, but as Big Data Analytics platforms for fusion and visualization, are IBM i2 Analyst's Notebook (IBM), Palantir Gotham (Palantir) and Lumify from Altamira Corporation (Altamira).

As said, reputation analysis on the web is a particular application of OSInt; reputation computation has always been an active research field. Some research lines consider corporate reputation as an intangible asset, with a strategic relevance because able to gain competitive advantage and create value, reflecting in a better corporate performance. Other consider corporate reputation as a construct that resides in perceptions, attitudes and behaviours developed by various stakeholders' categories towards an organization. In both cases, public facts or opinions can influence corporate image and its development.

Some researchers (Feldman et al., 2014) propose a methodology to compute a consumer reputation index based on the social responsibility of a company, the quality of its products and services, the capacity to innovate, leadership, ethicality, the relationship with customers, the capability to generate positive feelings, and its workplace environment.

In (Santarcangelo et al., 2015), the reputation is an asset also for public administrations and a related use case is presented for the measurement of citizens' satisfaction through an OSInt solution, which implements an opinion mining system. In (Lee and Oh, 2015), the authors propose a new algorithm to compute the reputation of nodes within online social networks, which changes over time and depends on frequency and velocity of online interactions.

In (Tian et al., 2015) the authors depict a relationship between trust and reputation for B2C e-commerce, defining a mathematical model to compute it.

3 METHODOLOGY

In the following section, we describe the adopted methodology for the development of the solution. Starting from the defined intelligence cycle, (Best, 2007) we have adapted it accordingly to a specific application, such as supplier risk assessment.

The intelligence process includes several phases during which some activities need to be carried on using supporting tools. The phases are the following:

- Information source and target planning
- Data collection
- Data processing
- Data analysis and visualization
- Data sharing and collaboration

3.1 Information Source Planning

During information source planning, we have identified "Guida Monaci", that is one of the oldest Italian business and marketing intelligence providers of companies' data source (venues, social capital, employee number, vat number, management, and, upon request, certificates of business entity).

Considering the context of "reputation", we have identified as most relevant data sources the archives of the main Italian online public newspapers, as "Corriere della Sera" and "Repubblica", the regional and local newspapers, and institutional sources, among which judgements of the court of auditors ("Sentenze della Corte dei Conti"), financial LEA's web site ("Guardia di Finanza"), communications authority ("AGCOM"). For the solution only "Guida Monaci", "Corriere" and "Repubblica" have been collected as sources.

In the next section, we give details about modelling of target information, namely suppliers' risk assessment.

3.2 Information Model for Supplier Risk Assessment

For our scope, we have defined our domain ontology with its entities, according to type (companies, persons, locations, etc.), relations and properties.

The relations between entities form the perimeter for suppliers' analysis. Relations' types include acquaintance ones, such as friendship, those related to family, localization, work and commercial environment.

Properties are different according to the various types of entities; as an example, if we consider individual persons, we want to know their profession, their age, birth date and place, fiscal code, email address, phone number, blog's and social network's URLs. For companies and organizations, we need to know their legal denomination and form, social capital, their web site's URLs, VAT number, venues, phone and fax numbers.

To model the supplier risk, we apply a risk aggregation method to conduct analysis related to the entity of interest. We identify several types of risks, each one fed by different factors, and for each of these we define some risk indicators that substantially can be expressed in a checklist form (Figure 1).

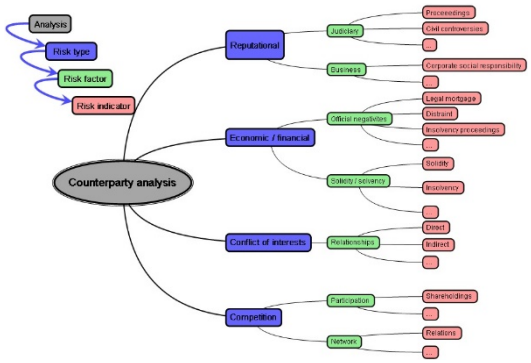

Figure 1: Supplier risk assessment model.

Within the established perimeter, we find evidences, which can switch on a specific alert, related to those entities that are included in the perimeter of the analysed company, up to second level. For example, the system highlights a legal risk, if a friend of targeted company's CEO has been arrested.

Evidences are stored in a knowledge base, showing those events related to entities, and requiring manual verification and validation.

We have started to implement reputational risk analysis considering the relevant judiciary and business risk factors. The following indicators have been considered for judicial risk factor:

- Arrest: is there an arrest?
- Investigations or notifications: is there an investigation?
- Civil or Administrative disputes: ongoing civil or administrative disputes?
- Organized crime's affiliations: proximity to criminal organizations or involvement in organized crime's offenses?
- Statements or plaints: complaints or appeals in its name?
- Administrative or Fiscal illegals: administrative illegality or tax misuse?
- Offences against property: crimes against property?
- Offences against the person; crimes against persons?
- Offences against the public administration: corruption, punishment, bribery, malpractice?
- Offences against the territory: illegal construction or environmental abuse?
- Judgment: is there a judgement?
- Negative sentences: administrative, civil, criminal convictions?
- Confiscations or searches: are there seizures or searches?

As business risk factor, we have identified the following indicators:

- Customer's dissatisfaction: potential dissatisfaction of suppliers' customers?
- Legal risks and job safety: potential risk / impact of non-respect of employee rights by the counterparty?
- Environmental risks related to job safety: potential environmental risk /impact related to the counterparty?
- Violations of competition rules: are there violations of competition rules by companies operating in the same industrial sector?

3.3 Supporting Tools and Solution Architecture

For other phases of OSInt process, we have integrated and developed a set of supporting tools to reach target information.

Solution architecture is illustrated in Figure 2. The **bottom layer** is **data layer** that implements data collection and processing phases and exposes API to access transformed data, which forms Knowledge Base. Due to some specific constraints, this layer hasn't changed and has been integrated in the solution.

Data collection has been performed through a crawler, which is programmable by means of page templates and schedulable using an administration panel. Crawler is implemented in PHP language and can capture textual relevant content of web pages and other relevant metadata (source and category, title, author, language, URL and publication date). The collected pages are stored in Mongo DB and documents are stored in MySQL with their full-text index.

During data processing, the saved contents of the documents are analysed through NLP component, which is composed by pipeline, generic and specialized lexicons on domain, and information extraction rules. The challenging task is to find, among online news, all those evidences of events that

can confirm one or more identified risk indicators: false positives or negatives may occur. Pipeline is implemented in PHP and lexicon and rules are stored using in-memory NoSQL solution Redis.

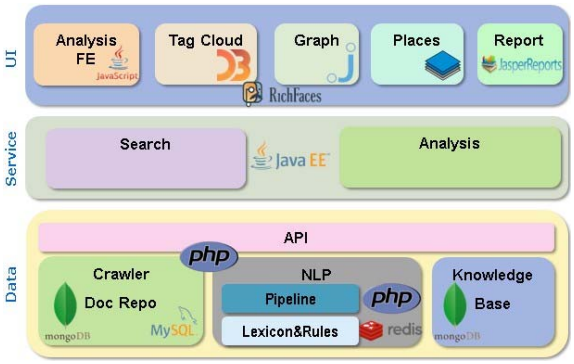

Figure 2: Solution Architecture.

The results of NLP populate the knowledge base, modelled as an entity data warehouse stored in Mongo DB, with recognized named entities, their relations and properties.

Data layer exposes a set of functions to the upper service layer, through a set of API implemented in PHP.

At **service layer**, **search** and **analysis services** implement respectively entities' search and reputational risk analysis. The first one allows a full text search of entities, getting concept cloud and the detail of any entity, as perimeter, properties, retrieving evidences of intercepted events, documents' details, and information on sources.

The analysis service implements the main logic of the solution, namely the risk analysis on supplier companies. To do that, analysis asks for risk types and performs retrieval of all events related to them. The retrieval is done for any entity belonging to the perimeter of targeted entity, up to second level. Therefore, it returns the found risk indicators and evidences, enables updating of confidence scores, saving and closing of the analysis for report creation. Services are implemented as JAX-RS and Enterprise Java Beans Services.

Analysis' service is used by Analysis front end and report components, while Search service is invoked by other UI components.

UI layer hosts interactive components that support activities of **data analysis** and **visualization**, and **data sharing** and **collaboration** phases. These components are implemented as Rich Faces. Analysis front-end makes use of javascript client libraries and performs full-text searches of named entities, retrieval of entities' details, selection of risk indicators, browsing of analysis' results, with analyst's feedbacks to validate or invalidate found evidences and related events.

Tag cloud makes concept clouding of selected entity, namely the retrieval of more recurrent concepts co-occurred with it in a set of documents of a configurable size. The component uses D3.js library.

Graph displays the entity's perimeter with related events and their evidences, activated by apposite functionalities, and allows insertion and removal of nodes from a set of linked nodes, adjusting the visualization accordingly to analyst needs. The graph component is built on top of Jsplumb library (jsPlumb).

Places component uses Open Layer library for geo-mapping and visualizes relations of target entities with places on the map, displaying evidences upon request.

After the validation of evidences and events, the analyst can share the result of his work through a configurable report that synthetizes the findings of the analysis, including evidences and a snapshot of graph. The report is built with JasperReports Library.

3.4 IE Approach

For reputational domain and, particularly, for identified risk indicators, there weren't any relation instances from available knowledge bases (e.g. Freebase, Wikipedia) or labelled data, which could train the bootstrapping phase.

Therefore, we chose to start with implementing **template element** and **relation construction rules**, which are based on generic and domain lexicon, trying to capture syntactic and logical relations of interest between concepts of a sentence.

This approach has been used to extract **entities' events, properties and relations**.

The rules are expressed according to the proprietary format of the NLP component.

Table 1: Logical relations between concepts.

CID1	REL	CID2
S:investigation	AGENT	V:led
V:led	COMP(to)	S:arrest
S:arrest	QUAL	R:Tizio Dei Tizi
R:Tizio Dei Tizi	QUAL	S:director
S:arrest	QUAL	S:accusation#1
S:accusation#1	QUAL	S:corruption
S:corruption	QUAL	S:Public Administration

For example, for the sentence: *The investigation led to the arrest of director Tizio Dei Tizi, with the*

accusation of corruption in the Public Administration", NLP pipeline extracts the concepts and the logic relations between them, in the form of triples. We report in Table 1 the extracted triples.

Within this set of triples, which represent the semantic graph of sentence, we search those ones that can give evidence of facts, events, links and properties that could be of interest.

As an example, for the Arrest event, we get the triples, which satisfy this condition: "*S:investigation->AGENT->V:led->COMP(to)->+S:arrest->QUAL->@1*".

In this case, rule engine triggers the Arrest for person Tizio Dei Tizi. In the same way, we extract the event "Crime against Public Administration" with the rule: "*S:arrest->QUAL->@1->+S:corruption->QUAL->S:Public Administration*".

The formalism of rule expects the sign "+" for co-occurrence of triples within the same sentence.

3.5 Risk Mitigation

The performance of NLP component influences the success of the solution. To mitigate the risk of low performance of information extraction procedures, it has been useful the adoption of a design approach oriented to the validation of extracted information, attributing a confidence score for each singular information.

For the false positives, the solution requires a manual verification and validation of evidences, allowing analyst to select those to be included in the shareable final report.

More work remains to be done for the false negatives. The solution has been designed to show only information with evidences according to identified indicators; thereby, a relevant amount of information could remain undiscovered and unused, although such information could be useful for the analysis.

4 RESULTS

The most important measure of performance is the **user's satisfaction**, which is strongly related to the relevance of the results. Testing approach adopted for analysis of the results is a **black box approach**.
We have followed these three steps:

- Definition of test cases and test data: input selection and definition of expected results
- Test execution
- Output analysis and computation of precision, recall and F-score

In creating the test case, it is necessary to identify documents considered relevant for each type of event (Investigation, Crime, Arrest, etc.), link (works for, administered by, etc.), property (profession, age, revenue, etc.).

In order to build up the set of relevant documents, identified as test cases, we made a selection among those documents used in the solution. We have followed these steps:

a) For those types of information more relevant for the domain and analysis, we have identified concepts related to them: for example, for event "Investigation", we have considered the verbal forms "investigate", "inquiry", etc., its noun forms "investigation", "hearing". We have also considered phrasal verbs used in journalistic lexicon, such as "led to the arrest" for the Arrest event. For CEO relationship, we have considered the forms "chairperson", "executive", "head", other than the normal form "ceo".
b) We have submitted full-text queries with keywords related to these concepts.
c) We have collected and analyzed the results, trying to understand if there would be any evidence of information type that we would want to extract. For each search, we have selected the first most N representative documents among the results. If some documents have been selected previously, these are used contemporarily for more test categories. N depends on the frequency of searched terms and on the presence of logical relations that can be recognized by NLP within the same sentence. Totally, the size of test documents' set is 1% of total documents.
d) We have registered expected results.

Therefore, we submit test cases through a platform-specific API that populates Knowledge Base.

After that, the analyst can manually verify the results, classifying them as TP, FN, FP and TN. Finally, confusion matrix is composed and performance indexes are computed.

Knowledge base has been enriched by new documents in a step-by-step approach and testing has been performed in several moments through specific test sessions (Table 2).

Between each test session, we didn't change IE rules, retaining integration approach for NLP component.

Table 2: Test sessions.

Test Session	Source Type	Sources	Increments of documents
#1	Journals	Home page Corriere 2012-13	~ 8500
		Home page Repubblica 2012-13	~ 9000
#2	Journals	Archive Corriere 2010-14	~100000
	Journals	Archive Repubblica 2010-14	~100000

In the Table 3 we show some results of the first test session. For CEO and director relationships we have obtained the best performance, because of the simplified form with which these relationships are expressed.

Table 3: Results of test session #1.

Information class	Precision #1	Recall #1	F1 #1
Arrest	0.72	0.68	0.7
Investigation	0.71	0.63	0,67
Crime against person	0.6	0.73	0.66
Crime against PA	0.7	0.63	0.66
CEO relationship	0.81	0.85	0.83
Director relationship	0.8	0.83	0.81
Acquaintance relationship	0.79	0.72	0.75

In the Table 4 we report the results of the second test session. From the measurement of performance indexes, we can deduct that the performance of IE engine has declined when the number of documents increased; the degradation in performance is more marked for event extraction.

Table 4: Results of test session #2.

Information class	Precision #2	Recall #2	F1 #2
Arrest	0.64	0.6	0.62
Investigation	0.65	0.59	0,62
Crime against person	0.58	0.64	0.61
Crime against PA	0.68	0.58	0.63
CEO relationship	0.75	0.84	0.79
Director relationship	0.82	0.74	0.78
Acquaintance relationship	0.7	0.62	0.66

Particularly, the recall index has lowered: this could be expected, and has further highlighted how the journalistic lexicon and the expressions or sayings used in newspaper articles complicate the logical analysis of the text, leading the system to false negatives.

The results obtained are not satisfactory and induce to some thoughts that we report as future directions.

5 CONCLUSIONS AND FUTURE DIRECTION

The analysis of the results represents an opportunity to learn some lessons. First, we understood that a complete automatic extraction of events is very difficult to reach. In fact, the lexical variety and richness of a language (in particular for the Italian language) represents a great obstacle to term disambiguation, undermining the system capability to extract the right entity, event and property relations (without falling for false positive or negative). Moreover, the actual project design does not give enough space to the analyst's thinking, who should be able to create his/her own relations, as well as to define new kind of entities and related properties. To reach this scope, the analyst may want to search and analyse documents' content, using more analytics tools, being able to discovery and annotate new relations, verifying at the same time those ones extracted by the system. Crowdsourcing annotation suggested in (Liu et al., 2016) may enhance performances of IE tasks and give feedbacks useful to train machine-learning models. An iterative approach, which is continuously able to gather human feedbacks, seems to be the only way to keep stable the performances of IE. This copes with revisiting the testing approach of NLP, which should occur during its development to anticipate critical issues and verify the non-regression, improving overall accuracy.

These considerations have brought us to review system design: we have decided to implement, in next versions, a new GUI providing such functionalities, namely documents' search (Figure 3) and annotation (Figure 4).

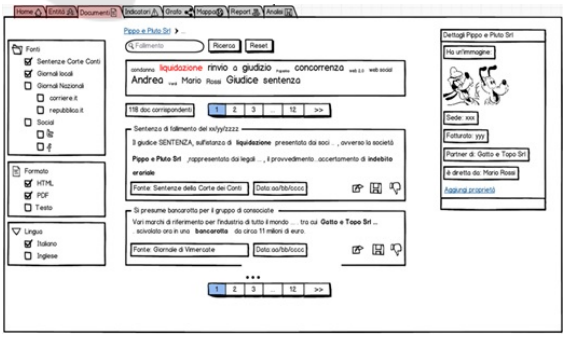

Figure 3: Documents search.

As said, rule-based approach for IE has been initially used due to the lack of labelled data or instances of relations, which could instantiate events of interest, in reputational domain. To overcome this issue, the use of neural models and multi instance

learning for relation extraction seems to be very promising.

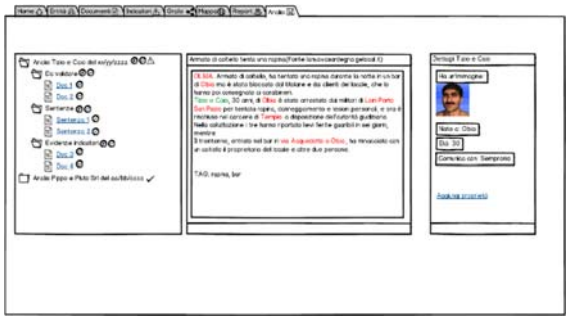

Figure 4: Analyst GUI for annotation.

However, regarding named entity classification tasks and extraction of properties and links (e.g. friend of, knows, parent/child, works for, partner, localization, etc.), the use of open source data (e.g. DBpedia, Freebase, etc.) may help. DBpedia allows the reconciliation of named entities found in news articles, giving more contextual information enabling discovery of hidden relations; moreover, it lays the foundation for building new episodic knowledge (Vossen et al., 2016), potentially regarding the reputational domain. At the same time, the enhanced version of distant supervision approach (Riedel et al., 2010) may help to reach better IE performances for some kinds of relations and properties not specialized on domain.

Another lesson-learned is related to knowledge base modelling. In fact, one of the major concerns is the lack of interoperability with open source data, which have a standard format. For this reason, it is important to formalize a standard language to model entities and its relations. In this field, a winning strategy could thus be the import of Semantic Web technologies and RDF/OWL ontologies, in order to represent the domain. The ontology would represent a "language-agnostic" model to shape the domain; in conjunction with a multi-language lexical database (e.g. multiWordNet), extraction rules can use inductive mechanisms to instantiate ontologies relations, leveraging lexical relations of Wordnet.

Furthermore, through SWRL (Semantic Web Rule Language), it is possible to implement higher-level rules on ontology instances for reasoning over Knowledge Base.

The choice to use open knowledge bases and Linked Open Data for the enrichment of internal knowledge base follows the semantic interoperability model reported in one of our works (Cappetta et al., 2014).

Finally, another important foreseen evolution, with impacts in the design stage, is a "Big Data compliant" architecture implementation. If we would include social channels in our reputation analysis, or manage other kind of sources, surely we will have to deal with "5V" data: volume, variety, velocity, veracity, virality. The last V-characteristic, virality, measures how fastly the information spreads (Mujeeb and Naidu, 2015) and, in the case of reputation, it can have a determining role.

If we consider the opportunity to re-engineer the solution, further applications are possible.

For the antifraud, tools can be useful in supporting the detection and prevention of illegal commerce of counterfeited several types of goods.

For counterterrorism, tools can be useful in the prevention and countering of terroristic activities conducted on the web (propaganda, counter-information, recruitment, cyberattacks, preparation and coordination of terrorist strikes, etc.).

In the Human Resources field, the solution can leverage analytics functions in the research of candidates, analysing their social network profiles, their professional activity on Internet or their scientific publications.

ACKNOWLEDGEMENTS

We would like to thank all company's members that have worked to the development of the solution and contributed to the definition of new use cases for it.

REFERENCES

F. Stalder e J. Hirsh, «Open Source Intelligence,» *First Monday*, vol. 7, n. 6, 2002.

M. Glassman e Min Ju Kang, «Intelligence in the internet age: the emergence and evolution of Open Source Intelligence,» *Computers in Human Behaviour*, vol. 28, n. 2, pp. 673-682, 2012.

NATO, Open Source Handbook, Intelligence Reader, and Intelligence Exploitation of the Internet, 2002.

C. Fleisher, "OSINT: Its Implications for Business/Competitive Intelligence Analysis and Analysts," *Inteligencia y seguridad*, no. 4, pp. 115-141, 2008.

W. He, J. Shen, X. Tian, Y. Li , V. Akula, G. Yan and R. Tao, "Gaining competitive intelligence from social media data," *Industrial Management & Data Systems*, vol. 115, no. 9, pp. 1622 - 1636, 2015.

R. Navigli and P. Velardi, "Learning domain ontologies from document warehouses and dedicated web sites," *Computational Linguistics*, vol. 30, no. 2, 2004.

R. Basili, M. Pazienza and F. Zanzotto, “Learning IE patterns: a terminology extraction perspective,” in *PROCEEDINGS OF THE Workshop of Event Modelling for Multilingual Document Linking at LREC 2002*, 2002.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel e R. Weischedel, «The automatic content extraction (ACE) program–tasks, data, and evaluation,» in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal. European Language Resources Association (ELRA)*, Lisbon, Portugal, 2004.

M. Palmer, D. Gildea e P. Kingsbury, «The Proposition Bank: A Corpus Annotated with Semantic Roles,» *Computational Linguistics Journal*, vol. 31, n. 1, 2005.

C. Baker, C. Fillmore and J. Lowe, “The Berkeley FrameNet Project,” in *COLING-ACL '98: Proceedings of the Conference*, Montreal, Quebec, Canada, 1998.

M. Mintz, S. Bills, R. Snow and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in *ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNL*, Suntec, Singapore, 2009.

S. Krause, H. Li, H. Uszkoreit and F. Xu, “Large-Scale learning of relation-extraction rules with distant supervision from the web,” in *ISWC'12 Proceedings of the 11th international conference on The Semantic Web*, Boston, MA, 2012.

Mausam, M. Schmitz, R. Bart, S. Soderland and O. Etzioni, “Open language learning for information extraction,” in *EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 2012.

S. Riedel, L. Yao e A. McCallum, «Modeling relations and their mentions without labeled text,» in *ECML PKDD'10 Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, Barcelona, Spain, 2010.

A. Liu, S. Soderland, J. Bragg, C. H Lin, X. Ling e D. S. Weld, «Effective Crowd Annotation for Relation Extraction,» in *Proceedings of the NAACL-HLT*, 2016.

P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Aprosio, G. Rigau, M. Rospocher and R. Segers, “NewsReader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news,” *Knowledge-Based Systems*, vol. 110, pp. 60-85, 2016.

P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao and B. Xu, “Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification,” *The 54th Annual Meeting of the Association for Computational Linguistics*, p. 207, 2016.

S. Kumar, “Extraction, A Survey of Deep Learning Methods for Relation Extraction,” *Computation and Language*, 2017.

K. Dave, S. Lawrence and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in *Proceedings of WWW*, 2003.

B. Pang and L. Lillian, “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieva*, vol. 2, no. 1, pp. 1-135, 2008.

A. Barabási and J. Frangos, Linked: The New Science of Networks, Perseus, 2002.

J. Leskovec and R. Sosic, “SNAP: A General-Purpose Network Analysis and Graph-Mining Library,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, p. 1, 2016.

J. R. Quinlan, «Induction of decision trees,» *Machine Learning*, vol. 1, n. 1, p. 81–106, 1986.

Nguyen, T. Manh, J. Schiefer and A. M. Tjoa, “Sense & response service architecture (SARESA): an approach towards a real-time business intelligence solution and its use for a fraud detection application,” in *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP*, 2005.

R Foundation for Statistical Computing, «A language and environment for statistical computing. R,» 2013. [Online]. Available: http://www.R-project.org/.

E. Frank, M. A. Hall and I. H. Witten, “The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques",” Morgan Kaufmann, 2016.

P. Feldman, R. Bahamonde e I. Bellido, «A new approach for measuring corporate reputation,» *Revista de Administração de Empresas*, vol. 54, n. 1, pp. 53-66, 2014.

V. Santarcangelo, G. Oddo, M. Pilato, F. Valenti and C. Fornaro, “An opinion mining application on OSINT for the reputation analysis of public administrations,” in *Choice and preference analysis for quality improvement and seminar on experimentation*, Bari, 2015.

J. Lee and J. C. Oh, “A Node-Centric Reputation Computation Algorithm on Online Social Networks,” *Applications of Social Media and Social Network Analysis*, pp. 1-22, 2015.

B. Tian, K. Liu e Y. Chen, «Dynamical Trust and Reputation Computation Model for B2C E-Commerce,» *Future internet*, 2015.

C. Best, «Open Source Intelligence,» in *NATO Advanced Study Institute (ASI) on Mining Massive Data Sets for Security.*, Ispra, Varese, Italy: IPSC - Joint Research Center - European Commission, 2007.

D. Cappetta, V. Moscato, V. Orabona, R. Palmieri and A. Picariello, “CMS towards semantic interoperability,” in *22nd Italian Symposium on Advanced Database Systems, SEBD*, 2014.

S. Mujeeb e L. Naidu, «A Relative Study on Big Data Applications and Techniques,» *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 4, n. 10, April 2015.

jsPlumb , [Online]. Available: https://jsplumbtoolkit.com/.

Shodan, «Shodan,» [Online]. Available: https://www.shodan.io/.

Paterva, «Paterva,» [Online]. Available: http://www.paterva.com/web6/.
Expert System, «Open Source Intelligence Software and Tools,» [Online]. Available: http://www.expertsystem.com/products/cogito-intelligence-platform/osint-software/.
IBM, “IBM - Data analysis - i2 Analyst's Notebook,” [Online]. Available: http://www-03.ibm.com/software/products/en/analysts-notebook.
Palantir, “Palantir Gotham,” [Online]. Available: https://www.palantir.com/palantir-gotham/.
Altamira, «Lumify | Altamira Technologies,» [Online]. Available: http://www.altamiracorp.com/index.php/lumify/.
D. Kalamaras, “Social Network Visualizer (SocNetV). Social network analysis and visualization software,” [Online]. Available: http://socnetv.org.
Apache, “Apache Mahout: Scalable machine learning and data mining,” [Online]. Available: http://mahout.apache.org/.
TensorFlow, “TensorFlow,” [Online]. Available: https://www.tensorflow.org/.

APPENDIXES

All trademarks, logos and names reported in this paper are property of their respective owners.