AI-based Decision Support Systems for the Management of E-procurement Procedures[*]

Discussion Paper

Pasquale Lops[1,*], Marco Di Ciano[2], Nicola Lopane[3], Lucia Siciliani[1], Vincenzo Taccardi[1], Eleonora Ghizzota[1] and Giovanni Semeraro[1]

[1]*Dip. di Informatica - Università di Bari Aldo Moro, Via E. Orabona 4, 70125 Bari, Italy*
[2]*InnovaPuglia S.p.A., str. prov. per Casamassima km. 3.000 - 70010 Valenzano (Ba), Italy*
[3]*Regione Puglia, Via G. Gentile 52 - 70126 Bari, Italy*

Abstract

Tenders are powerful means of investment of public funds and represent a strategic development resource. Thus, improving the efficiency of procuring entities and developing evaluation models turn out to be essential to facilitate e-procurement procedures. With this contribution, we present our preliminary research to create a supporting system for the decision-making and monitoring process for the entire course of investments and contracts (SIAP). This system employs artificial intelligence techniques based on natural language processing and machine learning, focused on providing instruments for extracting useful information from both structured and unstructured (i.e., text) data.

1. Introduction

Public procurement, especially when the aim is innovation, represents a powerful mean of investment of public funds. Hence, in the area of transparency and monitoring of the entire investment and procurement cycle, it is crucial to improve two main aspects: on one hand, the engagement process of the RUPs[7], procuring entities, administrations, and awarding entities, allowing them to fulfill many of their assigned tasks in a more effective, efficient and sustainable manner, and on the other hand, to develop assessment schemes that correlate particular logical-temporal sequences of facts and contents that can be traced back to specific anomaly indicators. Artificial intelligence technologies and automatic Natural Language Processing (NLP) systems focused on the Italian language represent a new frontier for semantic interpretation, concept extraction, and correlation of texts and documents. This research, leveraging such technologies, aims at developing a system that can interface with existing databases, prepare datasets that are suitable for processing and analysis, execute automatic extraction of relationships between

CEUR Workshop Proceedings (CEUR-WS.org)

textual entities, perform correlation tests between portions of text even of different lengths (paragraphs vs. entire document), then receive queries and return predefined outcomes in web-based format (short report, evidence, reference code, etc.).

2. Methodology

The high-level architecture layout is designed to enable its specialization in the later stages of the project. The architecture consists of four different modules: *Data Collector, Data Pre-Processing, Tender Analyzer, Service Tools.*

The *Data Collector* module gathers tender notices-related data, that can be extracted from several sources: European (TED[1]), national (SIMOG/ANAC[2]) and regional (EmPULIA[3]) databases, both in the form of tabular and textual data. Nevertheless, depending on the use-case requirements, this research may involve sources that do not focus on tendering, such as Feed RSS and other government agencies' databases. Due to the potential heterogeneous nature of the sources, the extraction of data is partitioned into plug-ins, allowing to add new sources or modify the existing ones more simply. A further component, Data Integration, deals with the above-mentioned heterogeneity by combining information gathered from each source, matching the overlapping data, and reporting potential anomalies.

The *Pre-Processing* module transforms the information extracted by Data Collector into entities on which the subsequent analysis tasks can be effectively executed. Thus, it is necessary to establish which features or entities are appropriate for representing a tender notice and subsequently memorize them with the help of databases or suitable indexes.

The *Tender Analyzer* carries out analyses based on the information extracted. Given the nature of such information (structured or unstructured), this module consists of two different components:

- The Data Analyzer deals with the structured information associated with tenders: codes (such as CIG[4], CPV[5], etc), dates, amounts, etc.
- The Content Analyzer elaborates the unstructured information, e.g. any attachments inherent to the notice (determine, specifications, etc.). It performs text analysis using NLP techniques.

The last module of the architecture is designed as a series of *Service Tools*. Given the information and analysis provided by the previous modules, they will be linked to a well-defined set of use cases and perform specific operations to fulfill the requirements.

3. Applications

As discussed in section 2, the proposed framework works on both structured and unstructured data to fully exploit the information associated with the notice and capture all aspects of it.

---

[1] https://ted.europa.eu/

[2] https://simog.anticorruzione.it/

[3] http://www.empulia.it/

[4] Codice Identificativo Gara, *Tender Identification Code*

[5] Common Procurement Vocabulary

3.1. Structured Data analysis

The application on structured data involves deriving metrics that can detect anomalies or deviation from regulatory and normative standards in procurement activities [1]. These can be calculated based on the available public procurement Open Data, we propose the following: *Relative tender value, Variance of bids, Difference between the first and second bids, Concentrated market structure, Static market structure, Cyclical wins, Lack of offers* from a previously active company, *Superfluous bidders, Prevalence of incorrect applications, Prevalence of consortia, Prevalence of subcontracting.*

Such features engineering activity can facilitate the application of machine learning, with the aim of detecting suspicious contracts whose allocation might be the result of collusive agreements among firms participating in the tender or pertaining to that market. The main criticality of the above task is the absence of datasets that record for a given procurement the occurrence of a judicial authority investigation that has proven the collusion among participants. One of the main research areas is the development of unsupervised models, e.g., clustering or anomaly detection. If needed, with records by the relevant law enforcement authorities available in external datasets, it would be possible to make use of supervised learning models at the cost of creating annotated datasets cross-referencing information obtained by data mining on those sources (rulings of the courts, RSS feeds, etc.) and available procurement data (ANAC[3], TED[2], etc).

3.2. Unstructured Data analysis

Unstructured data are handled with the aid of automated tools capable of detecting implicit information such as grammatical and semantic structures present in the text, following approaches based on NLP [2] and Semantic Analysis. Such approach enables various outcomes, for instance, the creation of a search engine capable of receiving queries and returning the most relevant documents based on co-occurrence measures between terms in the query and the documents and their semantic similarity as well [3]. Furthermore, to generate fixed preset outcomes, it is possible to use Natural Language Understanding, and Generation techniques, e.g., the system can generate summaries that condense, with the needed granularity, the information gathered from documents [4]. These functionalities simplify the access to contract information and consequently eases the implementation of innovative business processes.

The analysis of tender documents might allow the extraction of important information, which often is not included in the available metadata. The CPV[6] code, for example, identifies the type of the contract scope, and it is employed for classifying tenders. These codes could be used to train a classifier able to assign a category to a call for tender based on the content collected from its documentation. In addition, it is possible to automatically extract relations between textual entities contained in contract-related documents via Open Information Extraction methods [5], [6]. In particular, our proposed system extracts triples made of a subject, an object, and a predicate that relates them. In this manner, we obtain a machine-readable representation of the information in the documents, namely a fact that can be seen as a truth-bearer, and it can be labeled to be either true or false, relevant or not relevant. Since our study focuses on the

Italian language, our framework OIE4PA[6] leverages an Italian dataset of labeled and unlabeled triples extracted from real invitations to tender acquired from EmPulia. OIE4PA adopts the methodology of a pre-existent system, WikiOIE [6] thus, it adjusts the system to the domain of public tenders by setting up specific tools for extracting text from announcement documents and by creating an *ad hoc* dataset with the resulting facts. We test the self-training approach proposed by WikiOIE on our novel dataset to check whether it enhances the performance of the system or not.

This kind of work proves to be indispensable for resolving several scenarios. For instance, it is possible to cross-check information extracted directly from documents and structured data to detect potential anomalies; moreover, both sources can be exploited for preliminary market analysis, which allows to keep track of specific sectors or contracting authorities. By having enough data, this information can be used to outline profiles that will enable a more accurate identification of irregularities to be further examined by the contracting or competent authorities.

3.3. Contractor information synthetic overview

Considering the extensive information apparatus available, we aim to organize this information focusing on contractors. The objective is, therefore, to contribute to the definition of a "Passport" for companies or economic operators that allows the contracting station and the RUP[7] to access a real-time overview of the available information. For instance, by simulating the audits that the RUP usually runs and/or the economic operator declares, the application connects to the available databases (Registro Imprese, Agenzia Entrate, DURC[8], etc.), extracts and works on the information relevant to each test and, finally, returns them in a synthetic format. Moreover, possible results from the archive of tenders (won contracts, tenders statistics, localization of interventions, KPIs, etc.) are added to the aforementioned information to obtain a thorough and automated informative content about the operator at issue.

4. Conclusions and future works

The research project described so far aims to investigate the use of artificial intelligence technologies to offer the stakeholders engaged in the public procurement process a number of useful tools to facilitate their work in both the engagement and assessment phases. The proposed framework can operate on both tabular and textual data thanks to a modular architecture split into several plug-ins that take care of retrieving information and developing specific applications for each data source used.

The digitization policies are becoming increasingly incisive on a central and peripheral level, and the growing availability of data in a digital format opens up brand new scenarios of usage and integration of applications assisted by artificial intelligence. Therefore, the prototype proposed proves to be cutting-edge within this scenario, exploring the possibilities and, at the same time, being suitable for future extensions and enhancements.

---

[6]Open Information Extraction For the Public Administration
[7]Responsabile Unico del Procedimento, *Head Project Manager*
[8]Documento Unico di Regolarità Contributiva

References

[1] B. Tóth, M. Fazekas, Á. Czibik, I. J. Tóth, Toolkit for detecting collusive bidding in public procurement. with examples from hungary., Report number: CRC-WP/2014:02 (2014).

[2] D. Jurasky, J. H. Martin, Speech and language processing: An introduction to natural language processing, Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey (2000).

[3] P. Basile, A. Caputo, M. D. Ciano, G. Grasso, G. Rossiello, G. Semeraro, Sepir: a semantic and personalised information retrieval tool for the public administration based on distributional semantics, International Journal of Electronic Governance 9 (2017) 132–155.

[4] G. Rossiello, P. Basile, G. Semeraro, Centroid-based text summarization through compositionality of word embeddings, in: G. Giannakopoulos, E. Lloret, J. M. Conroy, J. Steinberger, M. Litvak, P. A. Rankel, B. Favre (Eds.), Proceedings of the Workshop on Summarization and Summary Evaluation Across Source Types and Genres, MultiLing@EACL 2017, Valencia, Spain, April 3, 2017, Association for Computational Linguistics, 2017, pp. 12–21. URL: https://doi.org/10.18653/v1/w17-1003. doi:10.18653/v1/w17-1003.

[5] P. Cassotti, L. Siciliani, P. Basile, M. de Gemmis, P. Lops, Extracting relations from italian wikipedia using unsupervised information extraction, in: V. W. Anelli, T. D. Noia, N. Ferro, F. Narducci (Eds.), Proceedings of the 11th Italian Information Retrieval Workshop 2021, Bari, Italy, September 13-15, 2021, volume 2947 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[6] L. Siciliani, P. Cassotti, P. Basile, M. de Gemmis, P. Lops, G. Semeraro, Extracting relations from italian wikipedia using self-training, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.