

Improving Public Sector Efficiency using Advanced Text Mining in the Procurement Process

Nikola Modrušan¹^a, Kornelije Rabuzin²^b and Leo Mršić³^c

¹Faculty for Information Studies in Novo Mesto, Ljubljanska cesta 31A, Novo Mesto, Slovenia

²Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, Varaždin, Croatia

³Algebra University College, Ilica 242, Zagreb, Croatia

Keywords: Text Mining, Natural Language Processing, Rule Extraction, Automatic Extraction, Data Mining, Knowledge Discovery, Fraud Detection, Corruption Indices, Public Procurement, Big Data.


Abstract: The analysis of the Public Procurement Processes (PPP) and the detection of suspicious or corrupt procedures is an important topic, especially for improving the process's transparency and for protecting public financial interests. Creating a quality model as a foundation to perform a quality analysis largely depends on the quality and volume of data that is analyzed. It is important to find a way to identify anomalies before they occur and to prevent any kind of harm that is of public interest. For this reason, we focused our research on an early phase of the PPP, the preparation of the tender documentation. During this phase, it is important to collect documents, detect and extract quality content from it, and analyze this content for any possible manipulation of the PPP's outcome. Part of the documentation related to defining the rules and restrictions for the PPP is usually within a specific section of the documents, often called "technical and professional ability." In previous studies, the authors extracted and processed these sections and used extracted content in order to develop a prediction model for indicating fraudulent activities. As the criteria and conditions can also be found in other parts of the PPP's documentation, the idea of this research is to detect additional content and to investigate its impact on the outcome of the prediction model. Therefore, our goal was to determine a list of relevant terms and to develop a data science model finding and extracting terms in order to improve the predictions of suspicious tender. An evaluation was conducted based on an initial prediction model trained with the extracted content as additional input parameters. The training results show a significant improvement in the output metrics. This study presents a methodology for detecting the content needed to predict suspicious procurement procedures, for measuring the relevance of extracted terms, and for storing the most important information in a relational structure in a database.


1 INTRODUCTION


Public procurement is a complex process that significantly affects the quality of life of citizens and society in general. A considerable amount of money is spent through this process, all to get the goods, services, and other things needed for the day-to-day operation of the entire public system. As a part of the process, public procurement is defined by the Public Procurement Act, which presents the conditions and steps through which the procedure is conducted (European Commission, 2016).

For the sake of transparency in the spending of the public's money, this process is of great interest to taxpayers. It also raises numerous questions and challenges about the detection, prevention, and protection of citizens' financial interests.

At the EU level, the total amount spent on public procurement is 545.4 billion Euros (DG GROW, 2019). It is also important to mention that the European Union is focused on developing tools and methods that will bring a greater efficiency to this process. For this purpose, systems such as Daisy, Pluto, etc., have been made all under the watchful eye

^a <https://orcid.org/0000-0002-0987-7090>

^b <https://orcid.org/0000-0002-0247-669X>

^c <https://orcid.org/0000-0002-5093-3453>

of OLAF, which is an EU-level anti-corruption agency (OLAF, 2019). On average, corruption accounts for about 5% of the total value of public procurement (Wensink et al., 2013). Therefore, researchers in recent years have been trying to understand any anomalies and detect suspicious activities. If there is a focus on the very early stages of the process, there is a much greater chance to catch these anomalies. The public procurement process consists of the following phases: planning, budgeting and procurement preparation, publishing, gathering information and making an evaluation of the tenders, and engaging in contract bidding and execution (European Commission, 2015).



Figure 1: The PPP phases.

The procedure for preparing the public procurement procedure is, in principle, related to the preparation of the procurement documents, which contain information about the procurement and the required eligibility conditions. In this phase, there is therefore an opportunity for manipulation that gives certain advantages to privileged competitors. Particular sensitive selection criteria and scoring methods are outlined in the bidding documentation. They can be designed to suit particular vendors or service providers and can be formulated so as to highlight the weaknesses of another competitor. This tender specification information can be used to detect procurements that are potentially suspicious (Rabuzin & Modrusan, 2019).

The tender documentation presents a set of unstructured documents of various types (pdf, doc, docx, ppt, etc.) that use an arbitrarily defined structure. However, the contents must include information related to the eligible conditions, excluding criteria, estimated price, technical and professional capacities, and the type of procedure in the documentation. Detection models, or prediction algorithms, are used to process the textual content. Almost all EU member states have increased the amount of their total budgets that they spend using the public procurement procedures. The increase in importance of these processes both for economic activity and from the perspective of the business potential for suppliers, led to adoption of big data techniques and the use of principles in the PPP domain (DG GROW, 2019). It should be emphasized that when one is creating a data model, a large amount of the content comes from clearing away from the

documents the irrelevant content, and thus reducing the amount of data we use in further processing.

The study of public procurement documentation is a specific area in the literature, but so far only one article has been identified from this area (Rabuzin & Modrusan, 2019), so this paper aims to identify the relevant content and to create a model for extracting that content from the tender documentation. This model can also be applied to the documentation downloaded from any public procurement system (we used data from Republic of Croatia), e.g., the Electronic Public Procurement Classifieds. A search was conducted that extracted sections related to the technical and professional capabilities mentioned in all of the tender documents. In this case, a new question is raised: Are we missing some crucial text where other important and suspicious content is hidden? To overcome this problem, we demonstrate our new approach, which extracts only the text related to the special terms defined by PP experts. With this approach, we refined what parts of the text would be extracted in order to find new important content. Moreover, we are trying to emphasize the necessity of determining the boundaries of the content to be extracted, which is a problem given that the documentation is unstructured. This study presents a methodology for detecting the relevant content needed to predict suspicious procurement procedures and to store this information in a structured database. As an example of a rule that detected suspicious behavior, procurement procedures that ended with only one bid were analyzed (Rabuzin & Modrusan, 2019).

This paper is structured as follows: section 2 presents the relevant literature review and is followed by a description of the data. Section 3 explains the experimental setup. Section 4 shows the data preprocessing algorithm, and section 5 explains the experimental setup and results. Finally, in section 6 the conclusions are drawn and future directions for research are envisaged.

2 LITERATURE OVERVIEW

Detecting quality and essential information is a crucial part of the process in any data analysis, especially when it comes to complex models. Bearing in mind that most procurement documents present unstructured data, it is crucial to find mechanisms through which this required data can be extracted from the text, especially if it is to be an automated process. A process in which the data is extracted from the document is an aspect of data mining, i.e., the

extraction of knowledge from the text, which is often used in different domains, especially when it comes to extracting information from web content in a large database (Dragoni et al., 2016; Yi et al., 2019; Tamames & de Lorenzo, 2010; Espejo-Garcia et al., 2019; Ojokoh et al., 2011; Torres-Moreno, 2014). Fissette (2014) looked into extracting information from companies' annual financial reports to find examples of corruption in them. Since quality information is crucial when creating a predictive model, she looked for particular categories of information: specific sections, references that link text, as well as short and straightforward phrases such as the year and company name. The documents in this study were structured and contained particular sections that could be analyzed, which is not the case in our study.

When looking to detect suspicious procurement procedures, a different type of input data is collected in the entire procurement process, namely, indicators or potential rules that can affect the outcome of a prediction model (Fazekas et al., 2016). The best results have been obtained through the use of deep learning methods for detecting corrupt activities (Domingos, 2016; Ferwerda, 2016). It is important to emphasize that such research is based on large volumes of non-filtered content. In particular, all of the information that researchers have been able to gather regarding the subject of public procurement is used as the input data used in the deep learning model. This approach leads to good results because filtering the content can lead to the loss of potentially valuable information. Nevertheless, the disadvantage of not filtering this information is often the problem of processing a large amount of data.

Although very similar, knowledge discovered in the domain of web pages is divided into "1) *generating extraction rules based on the web structure, such as the wrapper method*; 2) *Rule-based text analysis of machine-based learning to extract the relationship for an entity from an open domain*" (Yi et al., 2019). Natural language processing (NLP) is a method that interprets human language from one structure to another and is used in many fields (Dragoni et al., 2016; Geetha et al., 2013; Tamames & de Lorenzo, 2010). For example, one can extract the information and use it to create relational tables from a textual description (Geetha et al., 2013). Specifically, the idea is to use NLP techniques to identify the schema table and its properties so that the primary key attributes are identified based on adjectives using the preferences of the characteristics, rules, and machine learning system.

Because of its document structure, the analysis of legal documents is certainly close to the study of our type of documents. Specifically, they contain permitted, forbidden, or mandatory data in the context of what they regulate. Dragoni et al., using NLP techniques, were able to detect these rules by combining "linguistic information provided by WordNet with syntax-based rule extraction from legal texts, and logic-based dependency extraction between chunks of such texts." It is important to emphasize that their model uses sample extraction using the Stanford Parser and Boxer framework and that their focus was on detecting the conditions rather than the content of interest, which in their case was normative text or the text with bullets. Also, it is essential to note that these two parsers are based on the English language and grammar. What is important for us in this paper is precisely the detection of this content of interest. As they concluded in their work, there is a big difference between handwritten rules and automatic rule extraction precisely because different words are used for the same rule. Moreover, detecting references within the content and detecting their relationships are still significant problems. The issue of knowledge discovery and its extraction from the content is explored in almost all domains, especially when there is a large amount of content.

The growth of data in the area of the Internet of Things (IoT) has become a field of interest for scientists (Yi et al., 2019). They built a model that extracts information about experts from websites that contain this information, and they generated an analysis and lists with their profiles using long short-term memory (LSTM) neural nets. Ojokoh et al. have analyzed deep learning architectures to develop an end-to-end sequence labeler for phytosanitary regulations and have concluded that the best system to use is a neural network that utilizes character embeddings, bidirectional long short-term memory, and Softmax. Hidden Markov Models and Conditional Random Fields (CRFS) are the most commonly used models in terms of the content extraction segment and in terms of neighbor word relation analysis (Ojokoh et al., 2011). An important segment for data mining is a website's metadata, while in the case of structured documents, one can extract all the content one wants. Of course, the amount that can be extracted depends on the domain, especially in the public procurement segment, where areas such as healthcare, agriculture, informatics, or services are intertwined. This is a problem for us because the rules of extraction and boundaries have to be determined in some way.

Rabuzin and Modrusan (2019) used machine learning algorithms for data extracted from the public procurement tender documentation. They extracted and searched for document sections related to the technical and professional capability in order to detect suspicious public procurement tenders. In this article, we want to extend their study by finding other defined terms and check how extracted, “enriched” content can impact the prediction results.

3 UNDERSTANDING THE DATA

The tender documentation represents the basis of any procurement process. The structure of this kind of document includes the quantity of goods, section names, eligible conditions, etc. The particular content is practically left to the entity to choose, and it is challenging to find within these documents the content that is relevant. In the procurement process, the common name for a set of documents describing the quantities, exclusion conditions, aptitude requirements, and other relevant content is referred to as the tender documents. The structure of the tender documentation is not defined by the law. Nevertheless, the Regulation on Procurement Documents and Offers in Public Procurement Procedures in Croatian law defines the information that the contracting authority must provide when announcing its tender offer (MEEC, 2017).

According to this regulation, the tender is divided into several parts, namely the general part in which the description of the subject of the procurement is discussed, the section where the deadlines are given for the beginning and end of the contract, as well as a section with the criteria for selecting an economic entity, i.e., the sections that are important for us to extract any content of interest. In particular, it is important to emphasize that the regulation defines the contents that must be included in the tender documentation, but not where this information is to be located, the name of the sections, its serial number, etc. Therefore, this presents a challenge because it is difficult to define the boundaries and extract the content. The total amount of procurement procedures that included the tender documentation was approximately 15,000, of which 4096 tenders ended with only one bid and 11704 tenders ended with more than one, with the average document size being 200kb. In our previous study, the amount of extracted text from all procedures was 21Mb, and now, after using our new approach, the size of the extracted text for all of the documents was approximately 10 times greater in size (213Mb).

4 DATA PREPROCESSING

In the case of automatic data retrieval, it is possible to create various web scraping scripts that automatically download the documents contained on the pages. This method is difficult to use for the Electronic Public Procurement Classifieds of Croatia, from which the data was retrieved. In particular, the system does not provide the direct links to the documents, but generates a link on the flow because the data are in a database, and not in a file system.

Moreover, the system requires additional authorization for downloading documents. For this reason, business/process owners (Official Gazette) are asked for documentation. They gave us files in several different types of unstructured documents, and we found various documents in .doc, .docx, .pdf, .pwt, and .zip formats. Since our goal is to extract the content from the documents, it is necessary to convert the documents into the desired format and to develop a Python script for this activity (Figure 2).

Specifically, the .docx format is a form of structured document and is easy to read in the Python programming language, so the documents in the .doc type format were converted to docx format using the Python pywin32 for Windows extension or the win32com library. The documents for all of the procurements are stored in one folder, so we created a loop that converts each document to docx format and immediately deletes the old document since we no longer need it. The contents of the .pdf documents were extracted directly by the script.

```
import re
import os
import win32com.client as win32
from win32com.client import constants
fajl = 'C:/Users/admin/Desktop/Eojn_podaci/files-02/'
src_files = os.listdir(fajl)
for file_name in src_files:
    full_file_name = os.path.join(fajl, file_name)
    try:
        if ".docx" not in file_name:
            print(file_name + ' changing document type ')
            #print(full_file_name)
            word = win32.gencache.EnsureDispatch('Word.Application')
            doc = word.Documents.Open(full_file_name)
            doc.Activate()
            # Rename path with .docx
            new_file_abs = os.path.abspath(full_file_name)
            new_file_abs = re.sub(r'\\.\\w+$', '.docx', new_file_abs)
            # Save and Close
            saves=word.ActiveDocument.SaveAs(
                new_file_abs, FileFormat=constants.wdFormatXMLDocument
            )
            doc.Close(False)
            #remove old doc file
            os.remove(full_file_name)
    except:
        #doc.Close(False)
        print("Unexpected error:")
```

Figure 2: Doc to Docx format type converter.

Still, several pdf documents were in an unreadable pdf format, namely those from a documentation scan and are images stored in pdf, so we did not look at these particular documents. Even so, there was a small number of such documents, and it is possible to use the OCR process to convert the images from the file into text content. All zipped documents could not be properly sorted, as different types of unlabeled documents were contained within the zip files. Therefore, it was impossible to detect the bidding documents, and they were excluded from further processing.

5 MODELING AND EVALUATION

After converting all the bidding documents to the docx format, we started extracting the specific content needed for our prediction model. Since the documents were unstructured and their format is not defined by any law or regulation but depends on the choice of the user, it is not possible to automatically find and extract exactly the parts or sections that relate to technical and professional ability or to any eligible conditions. Therefore, the challenge was to define the boundaries. Having examined several hundred examples of the bidding documents, the empirical conclusion was that each of them contained a content related to technical and professional ability or eligibility criteria, and the number of words contained varied from 300 to 10,000 words (and in some cases were even more). In addition to extracting the entire sections, there was also a challenge in finding any content that presents some conditions and may affect the outcome of the procurement process, rather than being directly related to a particular section.

One approach to solving this problem is to find every occurrence of the “technical and professional” trigram, but this approach produces very few results, especially since there may be conditions within the different sections that can significantly affect the outcome of the procurement process. Interviews with experts in the field of the public procurement process revealed that there is a group of words that is very important for our study. Compared to the extraction of the trigram (technical and professional), the number of retrieved contents has increased over three times. For this purpose, the following list of terms have been identified and searched for from the entire corpus:

```
termList = ['minimum', 'maximum', 'requirement',
'expertise', 'certificate', 'qualification', 'minimum',
'highest', 'total', 'evidence', 'technical and expert', ...]
```

It is important to note that the words are taken in such a way that they represent only the root of the word, without prefixes and suffixes, so that all possibilities are taken into account. For example, due to the use of the plural but also the masculine-feminine gender case in the Croatian language, to search for the word expert, experts, etc., the occurrence of the word "expert" was detected to find all of these combinations.

The algorithm goes through each document line or paragraph individually (Figure 3). When the algorithm finds the required content, it stores it in the relational database. When it finds a word, it saves not just the entire paragraph but also the next 10 paragraphs in a row. The reason for this is the fact that after detecting a specific term, the further lines of the text are interconnected or linked and present a descriptive text about the entity that we searched for. It's a similar challenge to what Geetha et al., 2013 has when using a model for creating relational tables from a textual description. That is the reason why we extracted and saved the next ten paragraphs.

```
fajl = 'C:/Users/admin/Desktop/python/2 extraction from documents/dokumenti/'
termList=['minimalno', 'maksimalno', 'uvjet', 'stručnja', 'certifikat',
'kvalifikacij', 'najmanj', 'najviš', 'ukupn', 'popis', 'dokaz',
'tehnicka i stručna']
src_files = os.listdir(fajl)
for file_name in src_files:
    try:
        full_file_name = os.path.join(fajl, file_name)
        document = zipfile.ZipFile(full_file_name)
        xml_content = document.read('word/document.xml').decode("utf8")
        v=0;brojac=0
        tehnicka = ''
        document.close()
        tree = XML(xml_content)
        for paragraph in tree.getiterator(PARA):
            texts = [node.text for node in paragraph.getiterator(TEXT)
                    if node.text]
            stri = ''.join(texts).lower()
            rez = tehnicka.find(stri)
            if any(term in stri for term in termList):
                brojac = 1
            if (v!=10 and brojac==1 and rez<=0):
                tehnicka+=' '+stri
                v+=1
            else:
                v=0
                brojac=0
            cursor.execute("insert into dznTexttest2(id, text) values (?, ?)",
                           file_name[:-9], tehnicka)
            con.commit()
    except: print("Unexpected error:")
```

Figure 3: Python script for extraction of terms.

To remove the redundancies in the extracted content, every detected term from the observed public procurement procedure was compared with the already extracted text, and in the case that it already existed, it was not saved in the relational database.

The whole algorithm was also implemented for the pdf type of documents, although in these documents the search engine examined the content line by line rather than by paragraph, as was the case for the .docx document type. To further process the data, we also needed information about the procedure number or a unique procedure indicator. The title of the document contains the publication number and the title of the procedure itself. The publication number represents the first nine digits of the title, so it was extracted and stored in the database together with the text.

The evaluation process for both extracted options was taking the entire section with the technical and professional content, and taking the content detected by the list of terms for which we used machine learning algorithms (Rabuzin & Modrusan, 2019): naïve Bayes (NB), logistic regression (LR), and the support vector machines algorithm (SVM). The tender documentation describes the procurement description, the technical conditions, the deadlines, and the estimated values, as well as other data. As such, this documentation presents a large set of documents, which in some cases are not adequate for testing and this is a reason why we test only the two mentioned sets of data. The reasons for using this algorithms is that their results overall are easy to understand and because they have already been used in the field of public fraud detection (Wensink et al., 2013). For the purpose of preparing the input data, the process of tokenization and the technique of stemming the word are used. The process for preprocessing the data is the same as in our previous study (Rabuzin & Modrusan, 2019).

Table 1: Prediction results.

Extracted option	Metric	Logistic reg.	SVM	Naive Bayes
Section	Accuracy	0,69	0,69	0,69
	Precision	0,55	0,54	0,59
	Recall	0,31	0,35	0,13
	ROC	0,59	0,60	0,54
Section + Terms	Accuracy	0,76	0,76	0,73
	Precision	0,60	0,61	0,26
	Recall	0,25	0,27	0,01
	ROC	0,60	0,60	0,50

The findings are to be measured through four metric measurements: accuracy, precision, recall, and AUC. The accuracy presents the proportion of the accurate classified examples. The precision is the ratio of the precisely classified examples in a set of positively classified examples, and the recall is precisely the classified examples in the set of all

correct examples. The area under the receiver operating characteristic (ROC) curve, which is called the AUC, provides a general evaluation of the model: a higher AUC suggests the model can better discern between the two classes (Espejo-Garcia, 2019; Fissette, 2017; Rabuzin & Modrusan, 2019).

The results from Table 1 show an improvement in all metrics, especially in the part concerning the accuracy and precision. Even so, there are lower results in case of the recall metrics, and for us it is better to have higher results for precision because then we have a lower rate of false positives. In addition, for the logistic regression we obtained a higher ROC, which means that we developed a better model.

6 CONCLUSION AND FURTHER RESEARCH

Data quality is one of the most important features about which the outcome of an analysis or model development depends. An analysis of the literature showed us that scientists have encountered this problem for many years, especially after big data became a part of everyday life. This analysis found evidence for different challenges, especially in terms of the detection and extraction of the relevant content. A conclusion is that the domain knowledge, which we also used in our case, was indispensable for detecting the necessary terms and sections that we sought within the tender documentation. Numerous applied methods have been detected, of which we will mention the NLP techniques and deep learning.

The approaches to the content extraction from structured and unstructured document types are different. Moreover, it is easier to find content in structured document types than in unstructured documents. In our case, we had a challenge because the content could be found in any part of the document structure. We found that in more than 90% of the documents, there was a section called “technical and professional capability,” but extracting it did not mean that we covered all of the essential content extraction. Because the procurement documentation is not structured, we had to solve this challenge by determining the boundaries and through the process of retrieving all of the content. Through interviews with several important public procurement experts, we came up with a list of important terms/words that we used to extract content. With this approach, the amount of extracted content increased by an average of three times. An evaluation was

performed on both datasets, and the output metrics of the prediction model showed a significant amount of improvement in the case of extracting additional terms.

It is very difficult to find new measures and metrics that can be used as red flags to enhance the detection of suspicious one-bid tenders. Future research should further develop and confirm these initial findings by analyzing the PP process through process mining in order to seek all the connections between the events and the one-bid outcome, or to test the model on a larger dataset, e.g. on the European public procurement portal.

REFERENCES

- DG GROW, 2019. Public Procurement Indicators 2017, Available at: <https://ec.europa.eu/docsroom/documents/38003/attachments/1/translations/en/renditions/native>, (Accessed: 15 November 2019).
- Directorate for the public procurement system, 2017. Statistical report for 2017 year, Available at: http://www.javnabava.hr/userdocsimages/Statisticko_izvjesce_JN-2017.pdf, (Accessed: 15 October 2018).
- Domingos, S.L., Carvalho, R.N., Carvalho, R.S., Ramos, G.N., 2016. Identifying IT purchases anomalies in the Brazilian government procurement system using deep learning. *Machine Learning and Applications (ICMLA)*.
- Dragoni, M., Villata, S., Rizzi, W., Governatori, G., 2016. Combining NLP approaches for rule extraction from legal documents. *1st Workshop on Mining and Reasoning with Legal texts*, Sophia Antipolis.
- Espejo-Garcia, B., Lopez-Pellicer, F. J., Lacasta, J., Moreno, R. P., Zarazaga-Soria, F. J., 2019. End-to-end sequence labeling via deep learning for automatic extraction of agricultural regulations. *Computers and Electronics in Agriculture*.
- European Anti-Fraud Office (OLAF), 2019. The OLAF report 2018, Available at: https://ec.europa.eu/anti-fraud/sites/antifraud/files/olaf_report_2018_en.pdf, (Accessed: 10 November 2019).
- European Commission, 2015., Javna nabava - Smjernice za praktičare, Available at: https://ec.europa.eu/regional_policy/sources/docgener/informat/2014/guidance_public_proc_hr.pdf, (Accessed: 15 January 2020).
- European Commission, Legal rules and implementation, Available at: https://ec.europa.eu/growth/single-market/public-procurement/rules-implementation_en, (Accessed: 15 January 2020).
- Fazekas, M., Kocsis, G., 2017. Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data. *British Journal of Political Science*.
- Fazekas, M., Tóth, I.J., King, L.P., 2016. An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research*.
- Ferwerda, J., Deleanu, I., Unger, B., 2016. Corruption in Public Procurement: Finding the Right Indicators. *European Journal on Criminal Policy and Research*.
- Fissette, M., 2017. *Text mining to detect indications of fraud in annual reports worldwide*. Dissertation, University of Twente.
- Geetha, S., Mala, G. A., 2013. Extraction of key attributes from natural language requirements specification text. *IET Chennai Fourth International Conference on Sustainable Energy and Intelligent Systems*.
- Ministry of economy entrepreneurship and crafts (MEEC), Pravilnik o dokumentaciji o nabavi te ponudi u postupcima javne nabave, Available at: https://narodne-novine.nn.hr/clanci/sluzbeni/2017_07_65_1534.html, (Accessed: 05 January 2020).
- Ojokoh, B., Zhang, M., Tang, J., 2011. A trigram hidden Markov model for metadata extraction from heterogeneous references. *Information Sciences*.
- Rabuzin, K., Modrusan, N., 2019. Prediction of Public Procurement Corruption Indices using Machine Learning Methods. *11th International Conference on Knowledge Management and Information Systems*, Vienna.
- Ratinov, L., Roth, D., 2009. Design challenges and misconceptions in named entity recognition. *In Proceedings of the Thirteenth Conference on Computational Natural Language Learning*.
- Tamames, J., de Lorenzo, V., 2010. EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC bioinformatics*.
- Torres-Moreno, J. M. (Ed.), 2014. *Automatic text summarization*. John Wiley & Sons.
- Wensink, W., Vet, J.M., 2013. Identifying and reducing corruption in public procurement in the EU. *PricewaterhouseCoopers*.
- Yi, L., Yuan, R., Long, S., Xue, L., 2019. Expert Information Automatic Extraction for IOT Knowledge Base. *Procedia computer science*.