

Text Analytics Can Predict Contract Fairness, Transparency and Applicability

Nicola Assolini¹, Adelaide Baronchelli²^a, Matteo Cristani¹^b, Luca Pasetto¹^c,
Francesco Olivieri⁴^d, Roberto Ricciuti²^e and Claudio Tomazzoli³^f

¹Department of Computer Science, University of Verona, Italy

²Department of Economics, University of Verona, Italy

³CITERA Interdepartmental Centre, Sapienza University of Rome, Italy

⁴IIS, Griffith University, Australia

Keywords: Text Analytics, Web Analytics, Web Repository, Economic Analysis.


Abstract: There is a growing attention, in the research communities of political economics, onto the potential of text analytics in classifying documents with economic content. This interest extends the data analytics approach that has been the traditional base for economic theory with scientific perspective. To devise a general method for prediction applicability, we identify some phases of a methodology and perform tests on a large well-structured repository of *resource contracts* containing documents related to resources. The majority of these contracts involve mining resources. In this paper we prove that, by the usage of text analytics measures, we can cluster these documents on three indicators: *fairness* of the contract content, *transparency* of the document themselves, and *applicability* of the clauses of the contract intended to guarantee execution on an international basis. We achieve these results, consistent with a gold-standard test obtained with human experts, using text similarity based on the basic notions of bag of words, the index tf-idf, and three distinct cut-off measures.


1 INTRODUCTION


The exploitation of natural resources has been under the lenses of economic research since the beginning of modern mining industry. On the one hand, the majority of activities concerning natural resources (mining, refinement, transportation, distribution of refined products) is in the hands of a limited number of private companies with headquarters in developed countries, in particular, Europe. On the other hand, the large majority of the resources are to be found on a land of a country that is developing, or emerging economy.


A number of papers regarding the possible interpretations of the underlying data of this enormous, and complex system of resource exploitation has been


written in the mentioned line during the past thirty years; moreover a growing interest has appeared on scholarly journals and well reputed conferences on the topic of *asymmetry* in those contracts. Typically, a major western private or public company, sometimes an international one, is one party of a *contract*, while the other one is a government, in the large majority of cases, from less developed countries. These contracts are often unbalances, to be neatly in favour of the stronger of the contracting parties, often written in a not transparent manner, and inapplicable for those clauses that tend to protect the weaker subject. Although these aspects have been considered as open issues to be confirmed by rejecting null hypotheses asserting the opposites, like “The relationship between the parties in the exploitation of cobalt in the mines on the Congo river is fairly established between European contractors and the Government of the People’s Republic of Congo”. Null hypotheses such the one exemplified above can be quite easily denied and formally rejected on a given confidence interval by means of evaluations on the economic values on the


^a  <https://orcid.org/0000-0001-9950-9592>

^b  <https://orcid.org/0000-0001-5680-0080>

^c  <https://orcid.org/0000-0003-1036-1718>

^d  <https://orcid.org/0000-0003-0838-9850>

^e  <https://orcid.org/0000-0001-8251-0423>

^f  <https://orcid.org/0000-0003-2744-013X>

market and the values agreed, or by more sophisticated tests regarding the cost of single steps in terms of parties involved.

The number of analyses based on *text analytics* is still rather low in the current literature on resource economics. We can mention a few recent papers on the topics that we have considered as starting point, in particular (Wei et al., 2019) poses a problem that is close to those that we deal with here, while (Ambrosino et al., 2018) identifies methodological issues and devise a general approach to the usage of text analytics as a means to reveal economically relevant specific behaviours.

On the above base, we propose to develop an experimental pipeline that takes into consideration the texts of a set of contracts of resource exploitation as a source to look for answers onto three questions:

- **Q1:** Does the following document contain any clause that supports the claim “One of the parties has been leveraging on the weakness of the other one”? Does the document equally treat all the parties? Answer YES (Fair) or NO (Unfair)
- **Q2:** Does the following document contain any clause that supports the claim “It is not clear whether collateral and unidentified claims regarding the agreed matters are known to the parties”? Does the document show transparency? Answer YES(Transparent) or NO (Not transparent)
- **Q3:** Does the following document contain any clause that supports the claim “There is a margin of interpretation that makes it possible to refuse to adhere to the request in an unpredictable way”? Does the document exhibit applicability? Answer YES (Applicable) or NO (Inapplicable)

The foundation of the investigation is the hypothesis that the *content* of contracts could exhibit evidence of the aggregation of them in different groups, where the group formed by one aggregation separates from the group formed by the other aggregation with a low error rate on the separated classes. We perform the above sketched research path on an online, freely accessible web site, resourcecontracts.org where a large number of contracts of the discussed type are gathered. The general idea of the study consists in the following steps:

- **Clustering.** We cluster all the available documents in the repository based on their *text similarity*. The employed approach generates a number of subclasses that cannot be forecasted or fixed a priori.
- **Gold Standard Analysis.** We test the clusters by a *gold standard*. A panel of anonymous ex-

perts on double blinded analysis is asked to answer questions Q1, Q2, and Q3 on a sample subset of the documents in the repository, representative of the clusters obtained in the previous step of the methodology.

- **Confusion Matrix Tests.** We group the clusters in two sets, in view of determining the *best fit model*, by choosing a partition into two classes to be better than another one when the global number of errors is lower than in other models.

The schema of the research methodology is provided in Figure 1.

The rest of the paper is organised as follows. Section 2 describes, with some specific focus on the data/metadata contained in it, the mentioned repository resourcecontracts.org. Section 3 discusses the basics of Text Analytics used in this research, and Section 4 proposes a methodology for constructing the solution in terms of text analytics for the above mentioned problem and describes the specific application of a gold standard measure. Section 5 presents the analysis conducted on the repository introduced before, Section 7 discusses relevant related and previous work, and finally Section 8 takes some conclusions and sketches further work.

2 THE REPOSITORY RESOURCECONTRACTS.ORG

The repository resourcecontracts.org contains an increasing number of documents. We report here the class analysis at date 02/06/2021. The total number of documents in the repository at the date is 2587. Table 1 reports the documents divided by languages. As the reader immediately understands there is not a neat prevalence of one language upon others, but the languages English, French and Spanish are those in which the large majority of the documents are written. In the application of methods as discussed in Section 4 to the repository, we commit only to the analysis of documents written in English.

In figure 2 we report the number of documents in the repository per resource. We grouped in simplified aggregations based on mineral types.

The number of contracts involving developed countries as legal background is 109. The ones involving emerging countries is 212, and the ones involving developing countries is 2266. The number of contracts by Continent is displayed in Figure 3.

The analysis of the documents can be based on the existing annotations, that also involve Company name, Corporate group, Contract type and Key

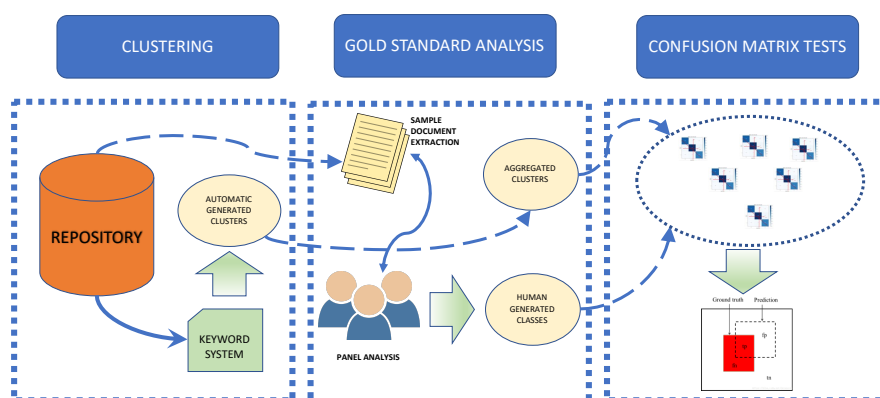


Figure 1: Schema of the experimental path.

Table 1: The 2587 documents by their language contained in resourcecontracts.org.

Language	Number
English	915
French	832
Spanish	752
Portuguese	80
Arabic	5
Chinese	1
Greek	1
Polish	1

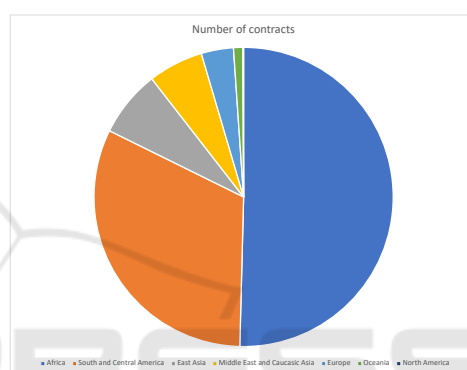


Figure 3: The distribution of documents in the repository by continents.

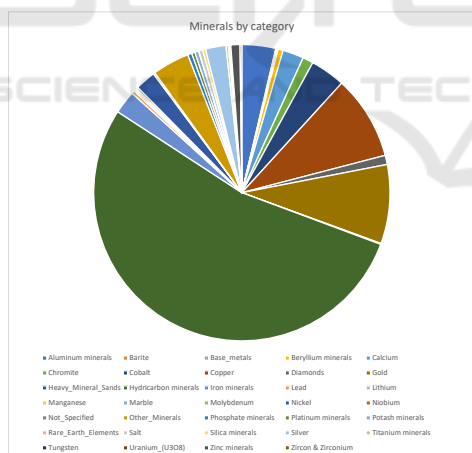


Figure 2: The distribution of documents in the repository by minerals.

3 BASIC DEFINITIONS OF TEXT ANALYTICS

In this section we discuss the basic definitions employed to analyse the content of the documents provided in Section 4, and applied specifically to resourcecontracts.org as discussed in Section 5.

First of all, we treat every document as a sequence of tokens, that we aim at distinguishing in *words* and *non-words*. A word is specifically a token that is lexically correct in a specific language, and a nonword is a token that does not do so. Non-words, in these documents, are mainly strings used in the text as specific elements of the graphic partition of the document itself. Many of these non-words are, in fact, numbers or interpunction symbols.

The first action that we perform on a document to process it analytically is *segmentation*, the operation of separating tokens appearing in the text, in order to represent documents as *sequences of words*, an operation that cannot be performed by just segmenting the text, for we need to distinguish words and non-

clauses, and can be addressed by API, that are made publicly available by the organisation that provides the data.

words. After segmentation, we aim, in this specific context, at devising a methodology that groups documents based on their similarity. To do so we need to get into some general, and then specific details of the analytical process.

First of all, we may be able to generate the so-called *bag-of-world* model, consisting in a transformation of a text into a vector of pairs, one element of each pair being a legitimate token, and the second element being a measure of the *relevance* of the token itself as referred to the document, and in relation with the *corpus* within which we consider it. The reference corpus will be, in the specific case we discuss in this paper, a repository.

We can start with two very basic measures for a token: the *document frequency* (and its variant, the *corpus frequency*) and the *text frequency*. Given a corpus C , we count $D = \|C\|$ the number of documents in C . Elements of the corpus are denoted by d_i with $1 \leq i \leq D$. Length of a text is counted by $l(d_i)$, the number of tokens appearing in d_i . The number of occurrences of a given token w_j in a document d_i is counted by $l(w_{ij})$. The number of documents in which a token appears is denoted by $d(w_i)$. We therefore use the following notations:

$$t_{ij} = \frac{l(w_{ij})}{l(d_i)} \text{ [Text frequency]}$$

$$w_i = \frac{d(w_i)}{D} \text{ [Document frequency]}$$

A more sophisticated measure can consider text and document frequency together. This is the tf-idf measure. It is defined as follows: $\text{tf-idf}_{ij} = -t_{ij} \cdot \log w_i$.

The tf-idf measure is high for those words that *characterise* a document in a corpus, as they are those that are frequent in a specific document but not particularly frequent in the corpus.

On the other hand, words which are very frequent in the corpus (namely those that have a high document frequency) tend not to be very interesting, as they are, usually, common tokens in the language. This brings us to one of the initial steps that improve in a very neat way the notion of similarity on texts. Tokens with a very high document frequency cannot be useful in determining the domain clusters and thus should not be considered. This can be done in two ways: one is the well-known *lexicon-driven* technique of *stop-word elimination*, and by using as a direct selector based on document frequency, known as *direct cutoff*.

Stop-word elimination that is based on the provision of a list of words not useful for document classification: articles, prepositions, common forms of the

auxiliary verbs, pronouns, conjunctions, and, in general, closed parts-of-speech of the language (as opposed to open ones *verbs nouns, adjectives* and *adverbs*).

Direct cutoff consists in eliminating those words that are so frequent in the corpus that cannot be significant to any clustering. Usually the approach is empirical, based on the finding of a threshold of the document frequency that candidates a word to be evicted. Normally the two approaches eliminate a large number of commonly chosen tokens.

After the above mentioned step, the texts are further simplified by using one technique among *stemming* and *lemmatisation*. In rare cases lemmatisation can be performed anyhow after the stemming procedure. Stemming consists indeed in the extraction of the stem of a word, or, as sometimes said, the reverse process of morphological flexions and derivations. Lemmatisation treats each word as referable to a single *lemma* and proceeds mainly as the above mentioned stemming method goes, but can include two further actions: solution to irregularities and association of synonyms.

After the above mentioned processes we go further and compute the bag of words. This can be based on the basic word count, on word frequency or on tf-idf. There is also, sometimes, the computation of the earliest occurrence of the word (stem, lemma) within the text. Once the bags of words of the documents have been generated we can proceed with the computation of the similarity of documents based on the *cosine distance*. The metric co-relates two bags of words w and \bar{w} on the same set of tokens w_i ($1 \leq i \leq n$). The cosine distance between two vectors is defined as follows:

$$\cos(\mathbf{w}, \bar{\mathbf{w}}) = \frac{\sum_{i=1}^n \mathbf{w}_i \bar{\mathbf{w}}_i}{\sqrt{\sum_{i=1}^n (\mathbf{w}_i)^2} \sqrt{\sum_{i=1}^n (\bar{\mathbf{w}}_i)^2}} \quad (1)$$

Based on the above basic measures we can build the methodology illustrated in Section 4.

4 A METHODOLOGY FOR REPOSITORY CLUSTERING

Consider a repository with a large number of documents, and assume that we aim at understanding the documents in terms of their most characterising keywords. This is the premise to any process of similarity analysis. Whenever we have the cosine distance well defined on a chosen measure to devise the bags of words of documents in a repository it makes sense to use any method of selection that can group these documents by means of their closeness. This

is in general, an approach to hierarchical clustering at one level, similarly to what is generally done with k-means as applied to text grouping.

Essentially, the method can be devised as follows:

- **Text Simplification**

- Text cleansing: elimination of unknown characters;
- Tokenisation: based on spacing, it defines the text as a sequence of tokens;
- N-gram generation: tokens are grouped on bigrams and trigram based on pure proximity;
- Stop-word elimination: based on a gazetteer containing a list of very common tokens (articles, prepositions, pronouns,...) the text is shortened;
- Stemming: words in unigrams, bigrams and trigrams are stemmed by means of a well-known Porter Algorithm.

- **Bag of Words Generation**

- For each text we generate the tf-idf index as related to the reference corpus.
- We associate to each n-gram its tf-idf index in each document

- **Cutoff of Word List**

- For each document, we list in decreasing order of document frequency the words appearing in the sequence
- We eliminate those words that belong to the first 99 percentiles (in the Strong Cutoff) or to the first 95 percentiles (in the Weak Cutoff)

- **Computation of Clusters**

- For each pair of documents, on the result of the cutoffs we compute cosine distance among the documents
- For each document d we add the documents that are close in terms of distance to d in the 80th percentile or above to a class;
- We sort classes in decreasing order of size in terms of documents.
- For each document appearing in the class analysed in the order computed in step 3, we remove the document from the smaller classes

- **Keyword Extraction**

- For each n-gram in each cluster we determine the corpus frequency of that n-gram as referred to its cluster (cluster frequency)
- We consider only the 1% most frequent n-grams for cluster as the most useful elements able to characterise that cluster.

The proposed method is an *unsupervised clustering* essentially similar to k-means, with a small but important modification, that makes it finer for the purposes of document classification: it has no fixed a priori number of classes to be generated. The variability of the classes depends upon the method used in this specific. Preliminary results on three different datasets have shown that the method could be adapted to repositories with diverse composition, language and size providing similar performances.

A schema of the working process of text analytics used in this study is presented in Figure 4.

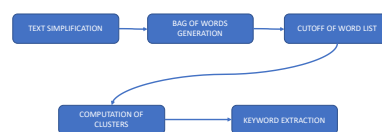


Figure 4: Schema of the clustering method.

5 ANALYSING RESOURCECONTRACTS.ORG

Data Cleaning. While analysing documents in the repository, the number of elements downloaded while selecting the English language is 915. However, a limited number of these documents result null or unreadable when taken by the API. In the preliminary version of the application to the repository we did not examine those data and simply skip the documents. This results in a total of actually utilisable documents of 1102. This is higher than the number of accessible documents in the web site, that is the consequence of the fact that some of the documents in the repository are not equipped with metadata, usually for they have been uploaded recently, or, in some cases, when they are unreadable. The total number of accessible documents is 1281. Therefore there is a set of 179 documents that are unreadable.

At the end of the above process we identify the clusters, whose sizes are omitted for the sake of space.

The Most Frequent N-grams. The twenty most frequent n-grams, after large cutoff, by decreasing document total selectivity are provided in Table 2. Mostly important, readers can note immediately the pertinence of the n-grams and the clear usefulness, especially of the bigrams, for potential match on documents in a repository. The accounts of column three is the *selectivity*, or, in other terms, the document frequency of that n-gram in the corpus.

Empirically, notice that:

- We can determine a number of terms that are related to commercial aspects of the contract: sold, rate exchange, sum paid;
- There is a large usage of terms referring to technical aspects: plant equipment, proportion, acceptable standards, transport storage;
- There are many terms referring the legal background: contractor determination, theft, government presentation, income tax;
- Many terms have specific functions for contracts: contract year, material purchase, event force.

Table 2: The first twenty stemmed n-grams in the 99th percentile.

1	contractor determin	125
2	plant equip	124
3	commensur	122
4	contractor petroleum	121
5	rate exchang	120
6	sold	119
7	domest market	118
8	reason measur	118
9	agreement govern	117
10	duti fee	117
11	connect therewith	116
12	merit	116
13	provis hereof	116
14	govern	115
15	contract year	115
16	own	115
17	sale petroleum	115
18	successor	115
19	transport storag	115
20	contractor subject	114

6 BY-HAND CHECK OF RESULTING CLUSTERS

We submitted a random selection on a single blind process of 25 documents, chosen to be identified as belonging to the clusters proportionally to the corresponding classes in the repository.

Further a panel of three experts has been asked the questions proposed in Section 1. Answers (YES/NO) are reported in a summary table by means of a method of vote known as *unanimity*. The three panelists are asked to agree on the judgments they expressed. The qualitative report they resubmitted qualifies the evidence as *actually agreed* (unanimously reported without discussion) in all cases but 1 for question Q1 e Q2, and in all cases but two for question Q3.

Afterwards, we aggregate the cluster, by using them as direct predictors for each possible behaviour, and use the obvious aggregation method for clustering (for instance that used by k-means in the majority of implementations), based on the selection of the aggregation that best fits the Accuracy, Precision and Recall of the confusion matrices.

For the sake of clarity we report here the definitions required. A *confusion matrix* on a binary classification, that is indeed what we obtain after applying the aggregation method devised above, is a 2×2 matrix in which we have columns representing the *actual classes* and the rows instead the *predicted classes*. In other terms, the cell (1,1) represents the *true positive* [TP] elements of the classification, whilst (1,2) the *false positive* [FP]. On the opposite, the cell (2,1) represents the *false negative* [FN] elements, and finally (2,2) represents the *true negative* [TN]. The sum of the first row is the some of positive elements (P) and the sum of the second row, the total number of negatives (N). The sum of the first column, instead, represents the elements that are *classified* positive, and the second column, instead, the *classified negative* elements.

Accuracy is the total number of correctly classified elements on the total of elements: $a = \frac{(TP+TN)}{TT}$. *Precision*, instead, is the ratio of true positive on positive ones: $p = \frac{TP}{(TP+FP)}$. *Recall*, finally, is the ratio between true positive and the sum of true positive and false negatives: $r = \frac{TP}{(TP+FN)}$.

We are now able to introduce the confusion matrices generated by the method. The first group of matrices is dedicated to the analysis of Question Q1.

- Clusters 1, 2, 3, and 4 form the predictor for “No” and 7,21 form the predictor for “Yes” ($\mathcal{M}_1(Q1)$) and appear in Table 3;
- Clusters 1, 2, 3, 4 and 21 form the predictor for “No” and 7 alone forms the predictor for “Yes” ($\mathcal{M}_2(Q1)$) and appear in Table 4.

Table 3: Measures on the confusion matrix $\mathcal{M}_1(Q1)$.

Accuracy	0,84
Precision	0,86
Recall	0,86

Table 4: Measures on the confusion matrix $\mathcal{M}_2(Q1)$.

Accuracy	0,80
Precision	0,82
Recall	0,90

Analogously, on Question Q2 we have:

- Clusters 1, 2, 3, and 4 form the predictor for “No” and 7, 21 form the predictor for “Yes” ($\mathcal{M}_1(Q2)$) and appear in Table 5;

- Clusters 1, 2, 3, and 21 form the predictor for “No” and 4, 7 form the predictor for “Yes” ($\mathcal{M}_2(Q2)$) and appear in Table 6;
- Clusters 1, 2, 3, 4 and 21 form the predictor for “No” and 7 alone forms the predictor for “Yes” ($\mathcal{M}_3(Q2)$) and appear in Table 7.

Table 5: Measures on the confusion matrix $\mathcal{M}_1(Q2)$.

Accuracy	92%
Precision	95%
Recall	87%

Table 6: Measures on the confusion matrix $\mathcal{M}_2(Q2)$.

Accuracy	80%
Precision	94%
Recall	85%

Table 7: Measures on the confusion matrix $\mathcal{M}_3(Q2)$.

Accuracy	92%
Precision	91%
Recall	91%

The third group of matrices, that indeed is formed by one matrix alone is dedicated to the analysis of Question Q3.

- Clusters 1, and 3 form the predictor for “No” and 2, 4, 7, 21 form the predictor for “Yes” ($\mathcal{M}_1(Q3)$) and appear in Table 8.

Table 8: Measures on the confusion matrix $\mathcal{M}_1(Q3)$.

Accuracy	92%
Precision	95%
Recall	87%

On summary, we can make the following observations, supported by the evidence of the developed experimental trials:

- Contracts are, generally, more frequently established among developing countries on the one part, and private companies of developed countries on the other part, that makes the potential result on unbalancing for the three questions on discussion less accurate on the yes answers, for they are neatly less numerous in the human gold test;
- The results are so good in terms of their credibility that the likelihood of the null hypothesis, if formulated, would be certainly rejected. However, the process sketched in this point of the discussion is *based* on the method we propose here, is not the method per se;
- It is not possible to obtain more accurate results on a limited test, but the confirmation of the qual-

ity of the method is so large that it makes sense to address analogous questions on other document domains.

On top of the above discussion, we can derive that all the questions are answered no with an accuracy, precision and recall of a good level of confidence: contracts can be asserted to be generally unfair, not transparent and not applicable.

7 RELATED WORK

Text analytics is a set of techniques that are also known as *statistical natural language processing*, *text mining* or, sometimes, *automated document analysis*. From the viewpoint of the techniques that we employed here as a base of the methods readers can refer to (Aizawa, 2003).

Most relevantly, we adopted a method that is discussed in (Laxmi Lydia et al., 2018). Research on social effects of usage of digital tools has been prominent in the recent past. On the same pathway we can find studies on how to use these tools, and how the behaviour of these tools can be considered (Yuan et al., 2016).

There is a lack of investigations on the theme of text analytics as applied to economic fields, and in particular, techniques to investigate social or macro-economic issues are neglected as research topics in the recent literature. In the recent past a comprehensive review of the applications of text analytics in finance has been published, (Gupta et al., 2020) that covers the majority of the research topics addressed. As noted in this study, the majority of these investigations focus on the usage of tools of text analysis, but do not lie on web resources.

The majority of studies have focused upon techniques that can be used to understand how text content can be used to support decisions. In particular in (Wang et al., 2020), authors devise a method to employ text content in investment decision, while (Lee et al., 2020) discusses how to identify aspects that are relevant for urban development decision processes by text analysis. In (Basole et al., 2019), authors address the issue of how to use similarity analysis of documents to understand similarity of ventures.

The economic fields involving social aspects have been less focused, in a panorama where the topic is not largely addressed, in general. An exception, fortunately, is indeed the mining industry. The works on which this investigation is based are fundamentally three: the pioneering study on text analytics as means to devise the legal aspects of the Enron case (Eckhaus and Sheaffer, 2018), the study on social responsibility

aspects that can be understood from texts in the specific case of the mining industry (Pons et al., 2021), and finally the study on discourse analysis over time in the extractive industry, where a text analytics perspective is developed on the specific field under the same scope of this paper, but use a text analysis tool that does not take into account corpora (WordStat).

8 CONCLUSIONS

This study has shown that a methodology for text indexing and classification, based in turn on a solid technique of clustering, adapted to guarantee some relevant properties, that are very desirable in a context of usage of document archives as tools for better understanding the social reality, can be used to predict fundamental properties of fairness, transparency and applicability of the contracts themselves. In particular, for this context it would be very interesting to develop techniques that are naturally evolving from basic clustering methods, as we proposed in the study.

From a comparison viewpoint it will be interesting to understand whether convolutional networks and other machine learning methods could be fruitfully employed, without quitting to critical sense (Vincent and Ogier, 2019). We are now adapting the method for other types of repositories. In particular we are making use of the above mentioned approach onto social media, similarly to what has been proposed in (Li et al., 2019).

Further we shall disclose other aspects to be verified by gold standard analysis, that could be contained in documents, and in particular in contract texts, including analysis of textual connections among *named entities* as, in particular, persons and organisations. Economists have long discussed the risk that a large endowment in natural resources may turn into a curse (Van Der Ploeg, 2011). This analysis has shown that the balance of power between companies and states may tilt into one direction, with detrimental effects on fairness.

Similarly on how we did investigate here we also aim at developing novel methods to identify text analytics methods that can be employed with the purpose of *compliance analysis*, within the declarative framework defined in (Olivieri et al., 2015).

REFERENCES

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39(1):45–65.

- Ambrosino, A., Cedrini, M., Davis, J., Fiori, S., Guerzoni, M., and Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics*. *Journal of Economic Methodology*, 25(4):329–348.
- Basole, R., Park, H., and Chao, R. (2019). Visual analysis of venture similarity in entrepreneurial ecosystems. *IEEE Transactions on Engineering Management*, 66(4):568–582.
- Eckhaus, E. and Sheaffer, Z. (2018). Managerial hubris detection: the case of enron. *Risk Management*, 20(4):304–325.
- Gupta, A., Dengre, V., Kheruwala, H., and Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6(1).
- Laxmi Lydia, E., Govindaswamy, P., Lakshmanaprabu, S., and Ramya, D. (2018). Document clustering based on text mining k-means algorithm using euclidean distance similarity. *Journal of Advanced Research in Dynamical and Control Systems*, 10(2 Special Issue):208–214.
- Lee, P., Kleinman, G., and Kuei, C.-H. (2020). Using text analytics to apprehend urban sustainability development. *Sustainable Development*, 28(4):897–921.
- Li, D., Zhang, Y., and Li, C. (2019). Mining public opinion on transportation systems based on social media data. *Sustainability (Switzerland)*, 11(15).
- Olivieri, F., Cristani, M., and Governatori, G. (2015). Compliant business processes with exclusive choices from agent specification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9387:603–612.
- Pons, A., Vintrò, C., Rius, J., and Vilaplana, J. (2021). Impact of corporate social responsibility in mining industries. *Resources Policy*, 72.
- Van Der Ploeg, F. (2011). Natural resources: Curse or blessing? *Journal of Economic Literature*, 49(2):366–420.
- Vincent, N. and Ogier, J.-M. (2019). Shall deep learning be the mandatory future of document analysis problems? *Pattern Recognition*, 86:281–289.
- Wang, W., Chen, W., Zhu, K., and Wang, H. (2020). Emphasizing the entrepreneur or the idea? the impact of text content emphasis on investment decisions in crowdfunding. *Decision Support Systems*, 136.
- Wei, L., Li, G., Zhu, X., Sun, X., and Li, J. (2019). Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures. *Energy Economics*, 80:452–460.
- Yuan, H., Lau, R., and Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, 91:67–76.