What Can Machine Learning Do for the Public Procurement?

Italian Association for Artificial Intelligence

What can Machine Learning do for public procurement?

I. Fatima[PHD_CS], F. Gorgerino[PHD_AL], **R. Meo**[ASS_CS], R. Nai[PHD_CS], P. Pasteris[TEC_CS], **G. M. Racca**[FULL_AL], E. Sulis[RTDB_CS]

Computer Science Department - University of Turin
Management Department – Administrative Law

Agenda

- Introduction
- Case study
- Methodology to reach the goals
- Results
- Conclusion and future work

1.

Introduction

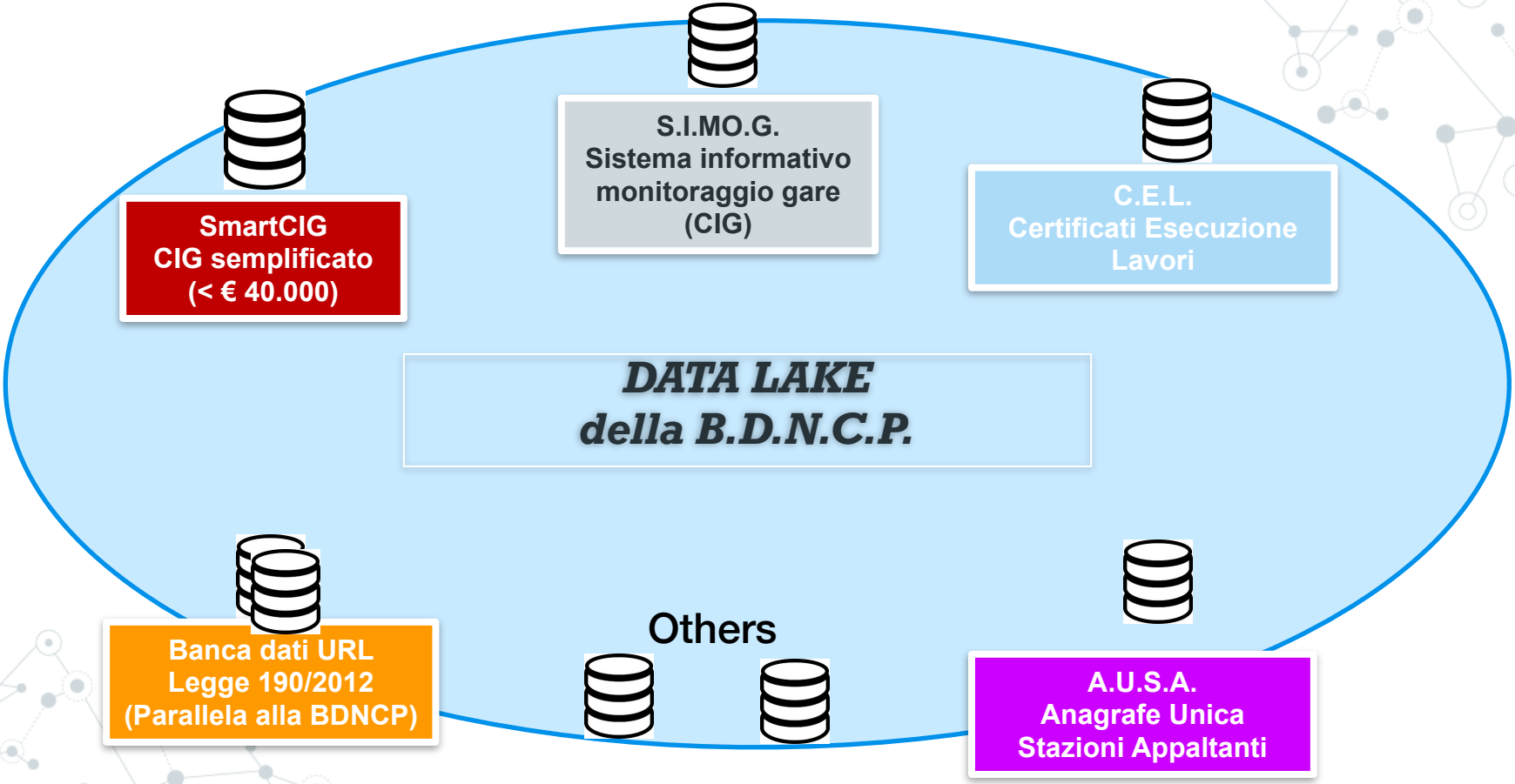
Abstract

- We present the systematic work we conducted on the data about **public procurement** in Italy.
- The **goal** is to clean and integrate various public and open information sources and extract valuable information for the public sector and the companies interested in awarding a contract with the Public Administration.
- Included in the data analysis is the **Regional Administrative Justice** that receives recourses from the involved actors.
- This information coming from recourses is potentially useful for revealing some of the anomalies related to the incorrect behaviour of the partners.
- The obtained results can also make lighter the administrative judges' workload.

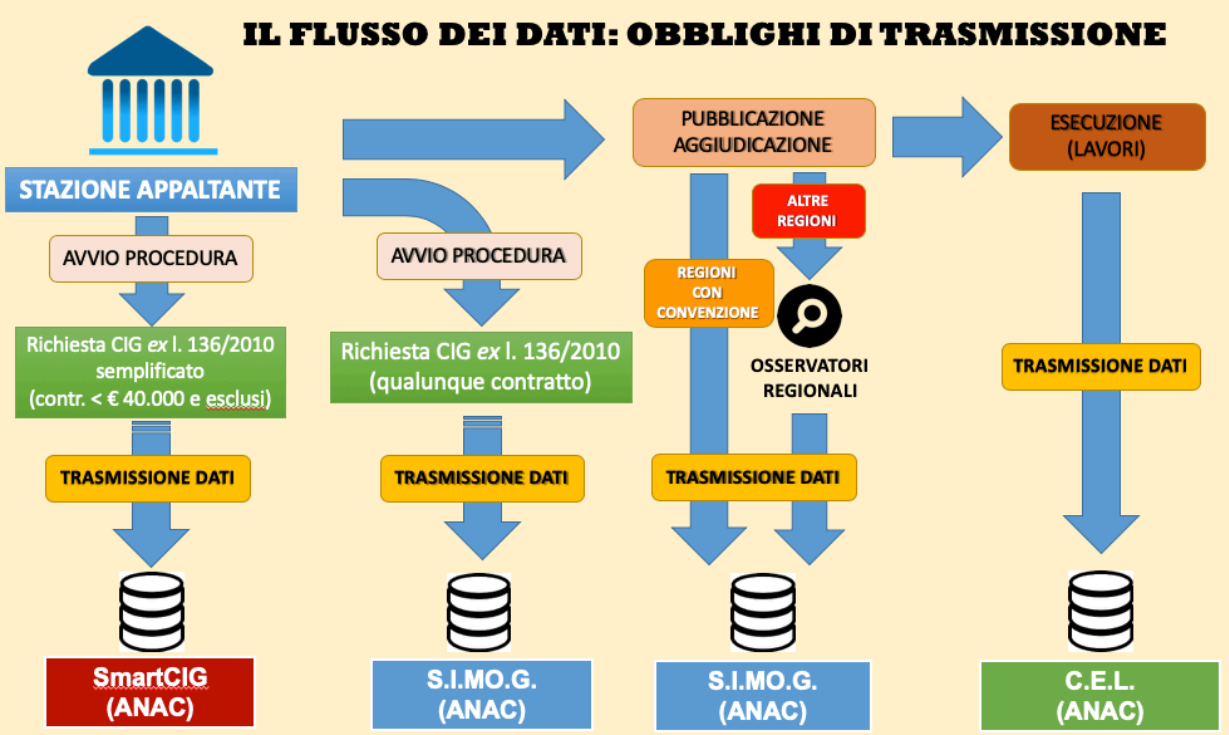
The research project with ANAC

- *ANAC and University of Torino have stipulated a research project in 2019 for data analysis on public contracts*
- *BDNCP* is a ***data lake*** with:
  - *Sistema Informativo Monitoraggio Gare* (SIMOG) with data on tenders
  - *SMART CIG* with data on contracts whose amount is under threshold (<40 K euro)
  - *Anagrafe Unica delle Stazioni Appaltanti* (AUSA)
  - CEL, with data on completion of works (*Certificati di Esecuzione Lavori – CEL)*

Banca Dati Nazionale dei Contratti Pubblici (BDNCP)



The data flow



Goals of the research project

BANCA DATI NAZIONALE DEI CONTRATTI PUBBLICI

DATA ANALYSIS

DATA ELABORATION

DATA COMPARISON

KPI DEFINITION AND USAGE

ENFORCEMENT OF **KNOWLEGDE BASE** FOR THE PUBLIC DECISIONS

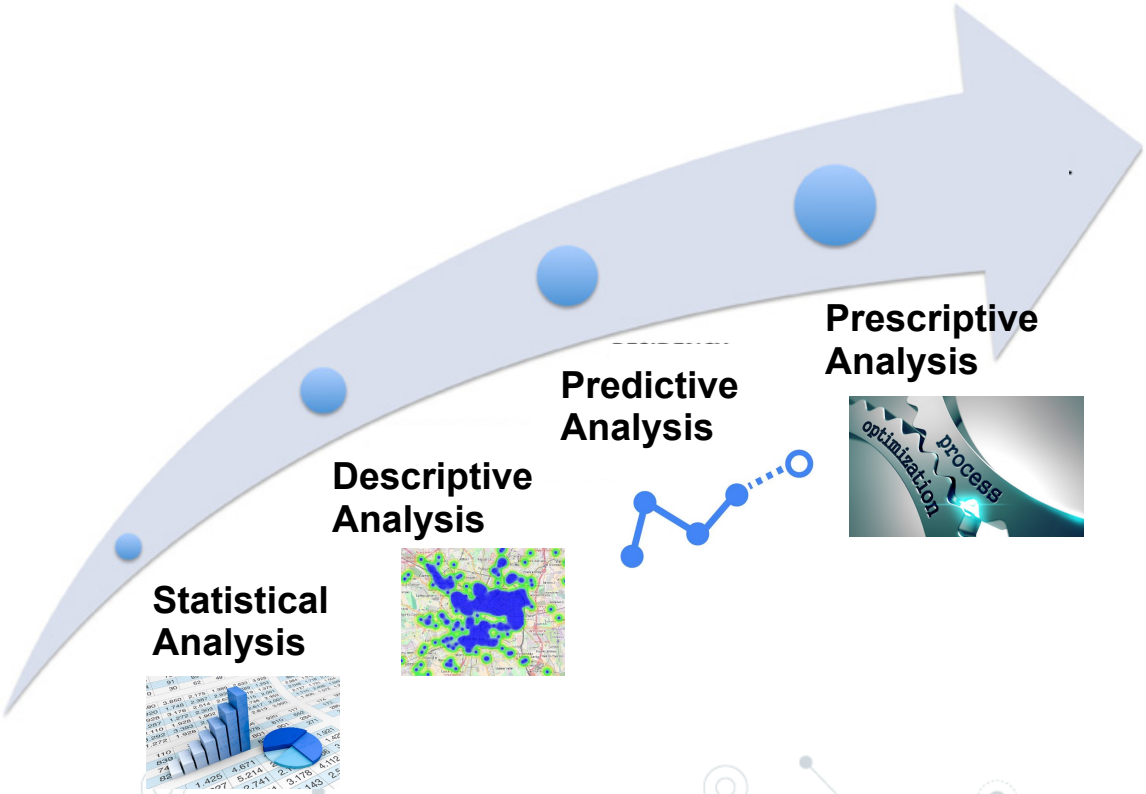
PROMOTION AND DISSEMINATION OF **BEST PRACTICES**

EFFICIENCY AND INTEGRITY OF PUBBLIC ADMINISTRATION

QUALITY AND TRANSPARENCY IMPROVEMENT OF ADMINISTRATIVE SERVICES

PREVENTION OF **CORRUPTION**

The data analysis path



Research questions

🎯 **Research question 1** (RQ1): how can we automatically **extract information from different legal archives**, in order to identify the entities involved in a public procurement?

🎯 **Research question 2** (RQ2): is it possible to set up an experiment to **predict the possible recourses** to administrative courts through the features of a public procurement?

🎯 **Research question 3** (RQ3): is it possible to provide services (**recommender systems**) to help the administrative operators (PA or economic)?

2.

Case study

ANAC Open Data

- The **National Anti-Corruption Authority**, abbreviated to **ANAC**, is an independent Italian administrative authority whose task is to prevent corruption in the Italian public administration, implement transparency and supervise public contracts.
- **ANAC collects data on calls for procurement** from the public contract authority and provides a catalog of Open Data (in CSV, JSON and OCDS format) describing **public procurement**, **contract authority** (public administration), and **contractors** (economic operators winning the procurements).
- https://dati.anticorruzione.it/opendata

ANAC Open Data

2007 - 2022

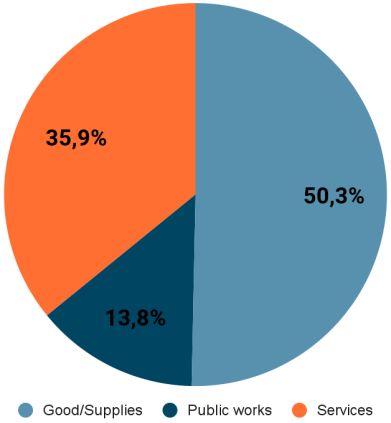
ANAC Procurement

7.2 million procurement

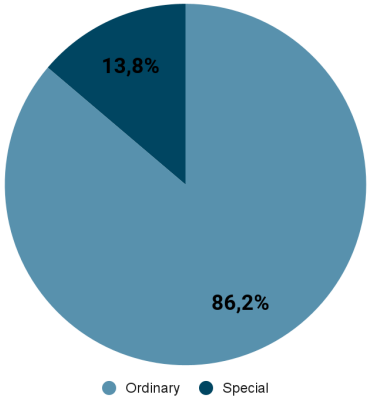
~450.000 per year

Identifier: `CIG` (shared key)

Procurement object



Procurement type



ANAC Open Data

2007 - 2022

ANAC Procurement (SIMOG)

7.2 million procurement

~450.000 per year

Identifier: `CIG` (shared key)

Contractors (PA): 42,393

- Municipalities
- Regional governments
- Universities
- Hospitals
- etc.

Awards: 1,635,609
(~22.71% of Procurement)

- **Private companies** specialized in goods/supplies, public works, services

Economic operators: 265,039

Law 190/2012

- In compliance with Law 190/2012, **ANAC also collects the list of private companies participating in public procurement**.
- The data are transmitted by the individual administrations to ANAC via URL containing XML file.

https://dati.anticorruzione.it/#/l190

Italian Administrative Justice

- On the other side, the **Italian Administrative Justice** (**IAJ**) collects the judges' sentences related to the public procurement appeals.
- Currently, about 80,360 judgments from the regional administrative courts (TAR) are available on the website.
- Every IAJ sentence is a textual file in html (60,284 files available), doc/docx (20,076 files available), or in pdf format (12 files available).
- https://www.giustizia-amministrativa.it

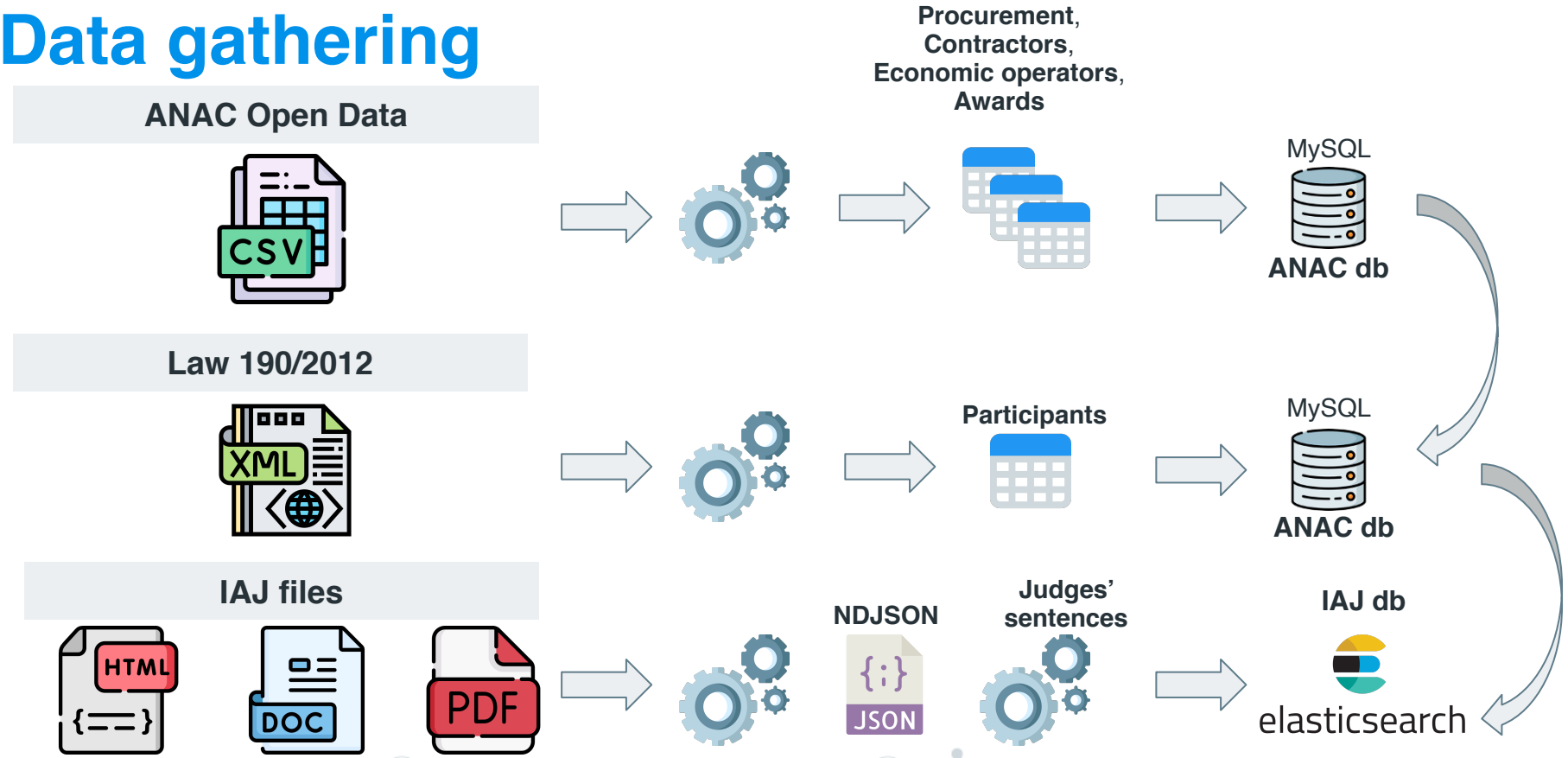
3.

Methodology

Data gathering

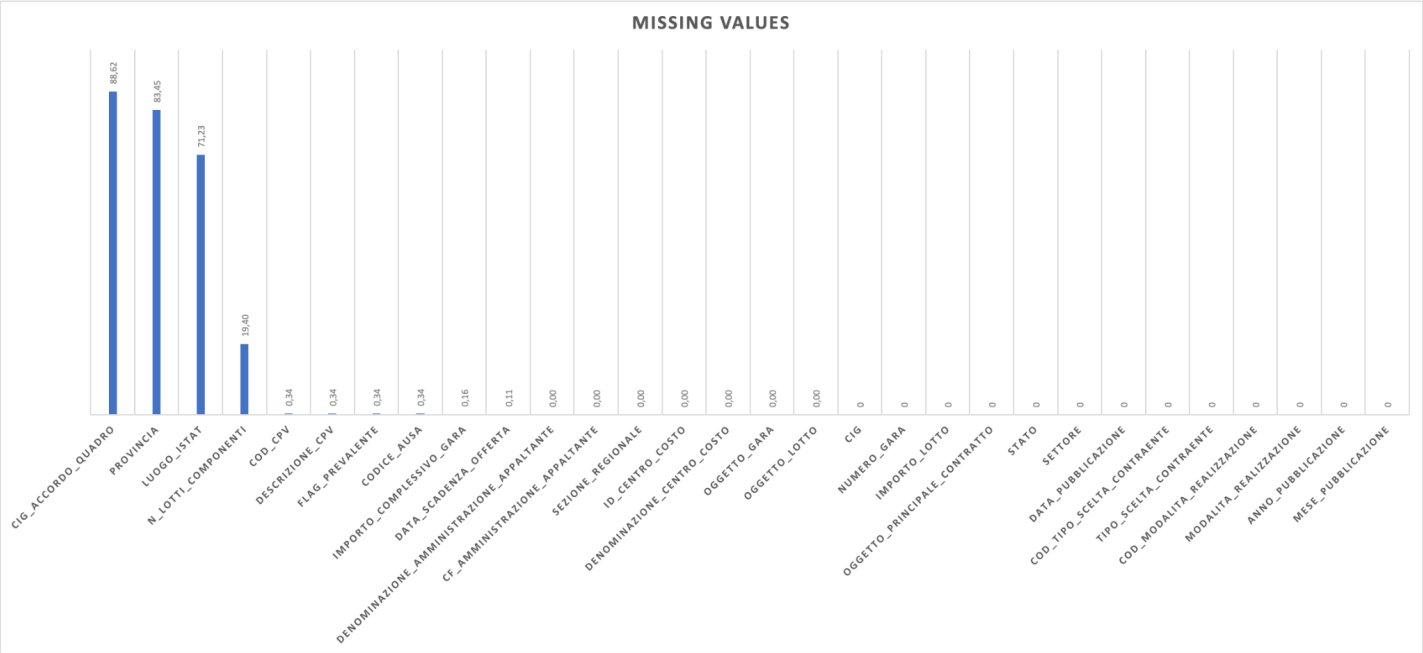
- Regarding **ANAC dataset**, we imported the Open Data of procurement from CSV format into a MySQL database table.
- Regarding **Law 190/2012**, we downloaded and imported the XMLs containing data of participants to procurement into a MySQL database table, with the fiscal code and name of each private company, the CIGs of procurement with its participation.
- We obtained the **IAJ judgments** via web scraping.
  - Since these are text files in html, doc/docx, and pdf format, they were indexed using tools specialized in Big Data processing.
  - According to November, 2022 figures on DB-Engines Ranking of Search Engines, **Elasticsearch** is the most popular search engine software used in Industry (https://db-engines.com/en/ranking_trend/search+engine).

Data gathering



Data quality problems

- **ANAC Open Data**: missing values



Data quality problems

- **Law 190/2012**: missing XMLs

2015 2016 2017 2018 2019 2020 2021 2022

Udine | Ricerca... | Esporta

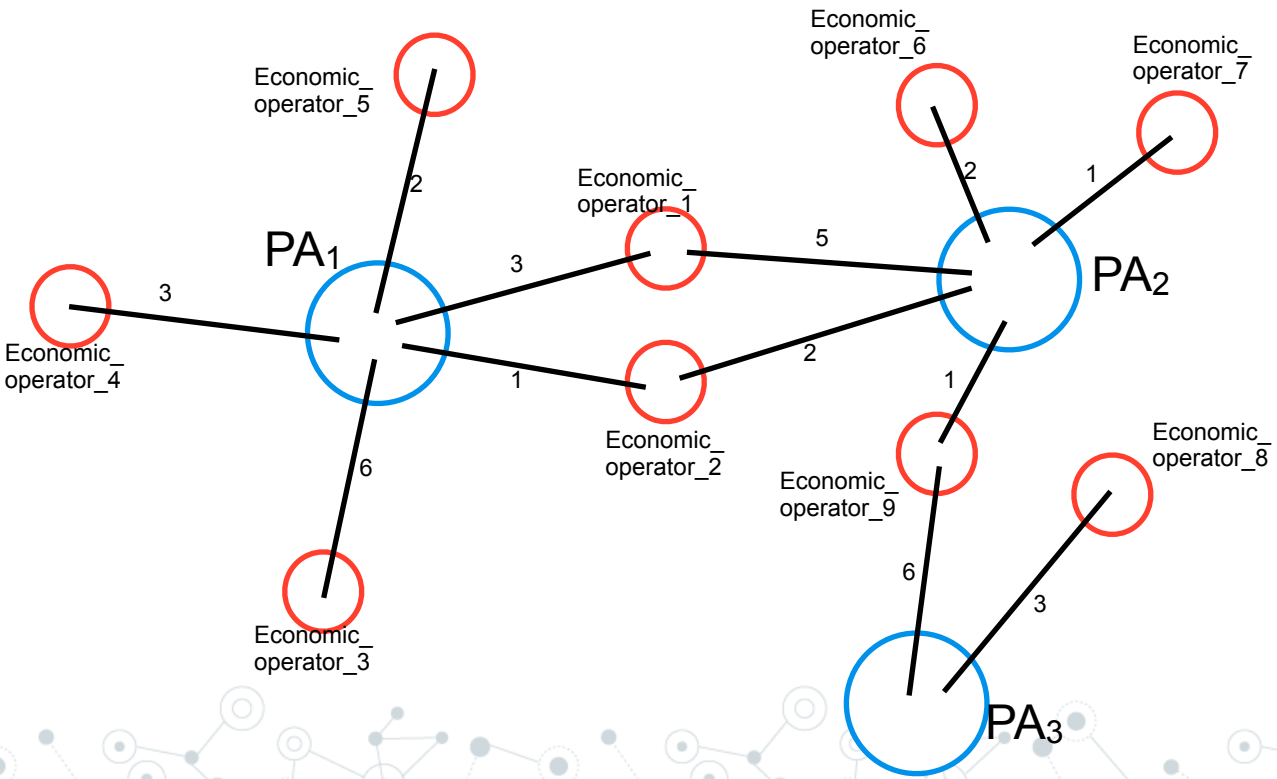
CF Amministr...	Denominazione Amministrazione	Identificativo messaggio PEC	URL	Esito acce...	Data accesso
94127260308	ISTITUTO COMPRENSIVO III DI UDINE	opec296.20220127093747.444221.314.1.24@pe...	XML	successo	16/02/2022 1
80004840304	ARCHIVIO DI STATO DI UDINE	E3317E68-AFAC-E1F5-8A39-C2582BA2B9F8@te...	XML	successo	16/02/2022 1
94127270307	ISTITUTO COMPRENSIVO I DI UDINE	opec296.20220124084121.16790.736.1.23@pec....	XML	successo	16/02/2022 1
00469890305	COMUNE DI PAVIA DI UDINE	1BF8E7F4-213C-E847-5E8F-8017B8B2D1D3@t...	XML	successo	15/02/2022 1
80014550307	UNIVERSITA' DEGLI STUDI DI UDINE	C98D4134.00F2B7FA.B42B3E9E.A2F69463.posta...	XML	fallito	
94150810300	ENTE DI DECENTRAMENTO REGIONALE DI UDINE	716EE483-B6A4-38FD-911C-48DC7D014296@t...	XML	successo	16/02/2022 1
80023240304	LICEO GINNASIO STATALE JACOPO STELLINI UDINE	opec296.20220202121924.378025.452.2.23@pe...	XML	successo	15/02/2022 1
02162990309	FONDAZIONE TEATRO NUOVO GIOVANNI DA UDINE	opec296.20220119083257.01367.848.1.68@pec....	XML	fallito	
00164770307	AZIENDA TERRITORIALE PER L'EDILIZIA RESIDENZIALE DI UDINE	25EF5973.027F352C.90916A4D.BDB32354.posta...	XML	successo	16/02/2022 1
02345670307	SOCIETA' FERROVIE UDINE - CIVIDALE S. A R.L. UNIPERSONALE	opec296.20220128105202.30210.80.1.64@pec.a...	XML	successo	15/02/2022 1
02935190302	CAMERA DI COMMERCIO, INDUSTRIA, ARTIGIANATO E AGRICOLTU...	25E9F05A.02939B28.9F9E6121.548A6726.posta-...	XML	successo	15/02/2022 1
94106210308	ORDINE DEI DOTTORI COMMERCIALISTI E DEGLI ESPERTI CONTA...	25F5397B.029CAA6C.AF107A0C.E576A0A3.posta...	XML	successo	16/02/2022 1
94143540303	CONS.COMUNI DEL BACINO IMBRIFERO MONTANO DELL'ISONZO ...	1FC02A94-9464-C4F9-F4BB-8A863C0250D2@te...	XML	successo	16/02/2022 1

Data quality problems

- **Law 190/2012**: missing XMLs

Year	URLs	Downloaded	NOT downloaded	% NOT downloaded
2022	20,170	18,554	1,616	**8.01**
2021	19,760	17,362	2,398	**12.14**
2020	18,986	15,541	3,445	**18.14**
2019	19,277	13,624	5,653	**29.33**
2018	18,950	11,658	7,292	**38.48**
2017	18,038	10,494	7,544	**41.82**
2016	17,469	9,222	8,247	**47.21**
2015	17,962	8,546	9,416	**52.42**

Descriptive analysis: the graph of public contracts



Analysis per macro-area

◎ **We observe the phenomenon of *segregation***

Big dots:
They award contracts in entire Italy (central in grapth)

Examples:
CNR, roads signals, Tecnositaf (roads security)

4.

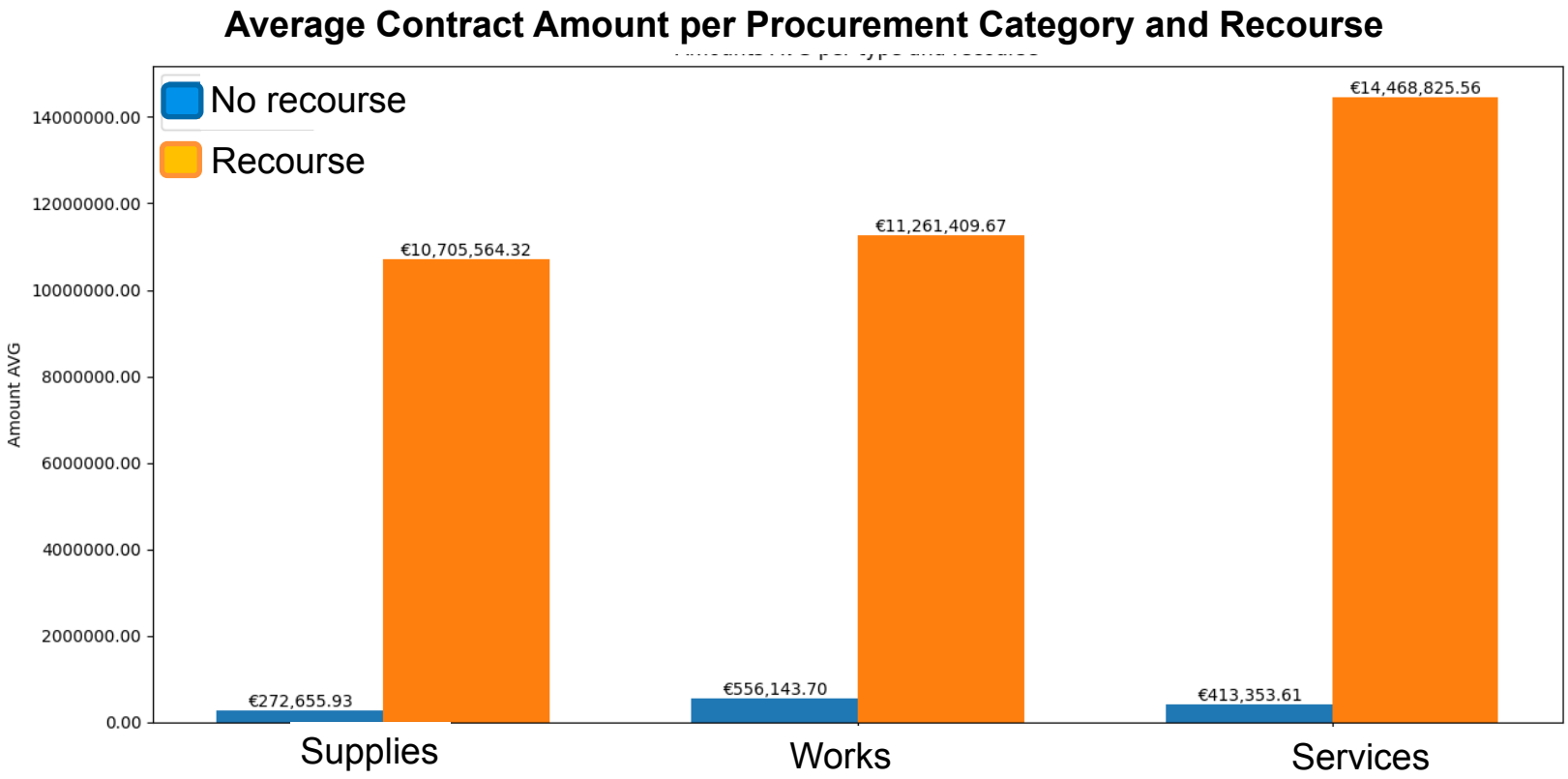
Results

Search by procurement ID (CIG)

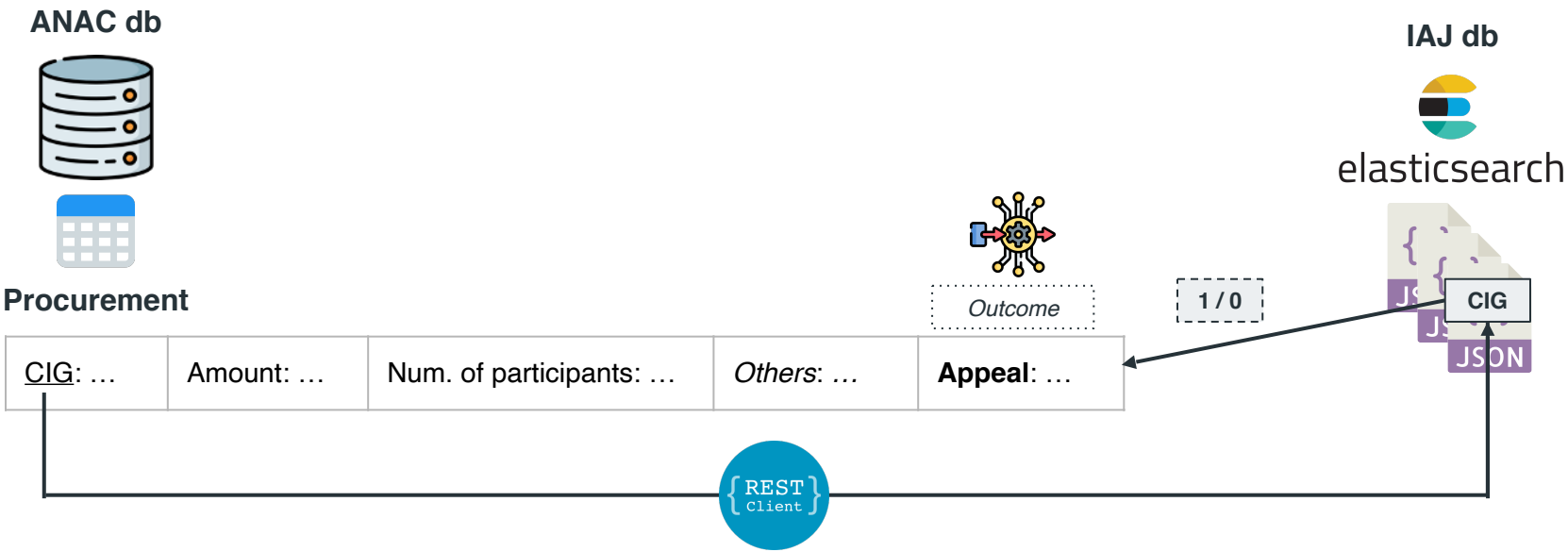
- Connection between the two information sources occurred by:
  - **searching for the procurement CIG in the IAJ sentences archive**;
  - the presence of the CIG in the IAJ dataset were used to label as positive the contract in ANAC procurement table
  - 🎯 This is the supervised input to apply machine learning for the classification of "*irregular*" administrative practice in procurement

  - However, the presence of a recourse could be due to a tendency of economic operators to stop the contract if they were not awarded a *substantious* contract

Correlation with positive labels and contract amount

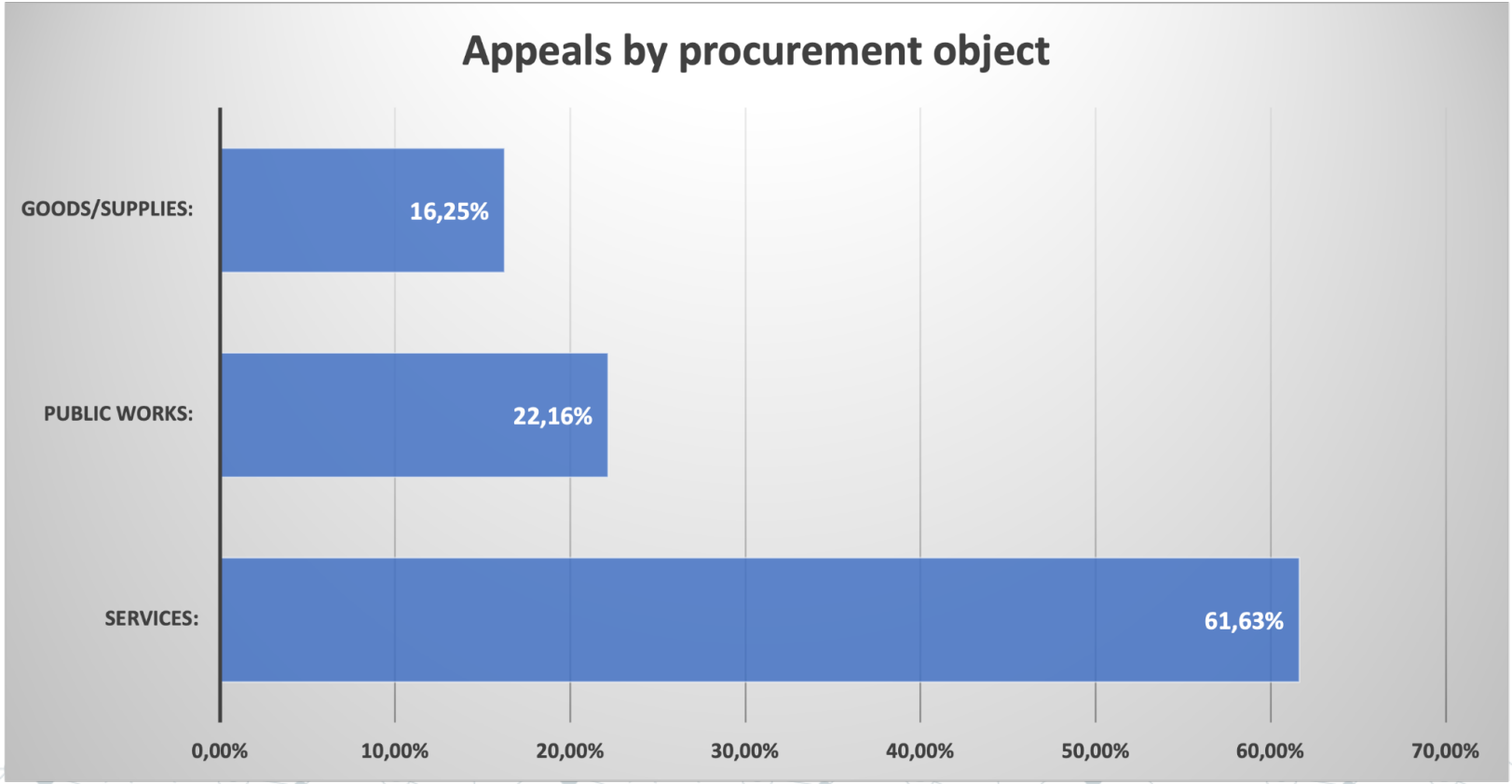


Search by procurement ID (CIG)



The total number of CIG found is **8,062** over 80,360 judgments: this means that the probability that a sentence in our archive refers to a CIG is about 10%.

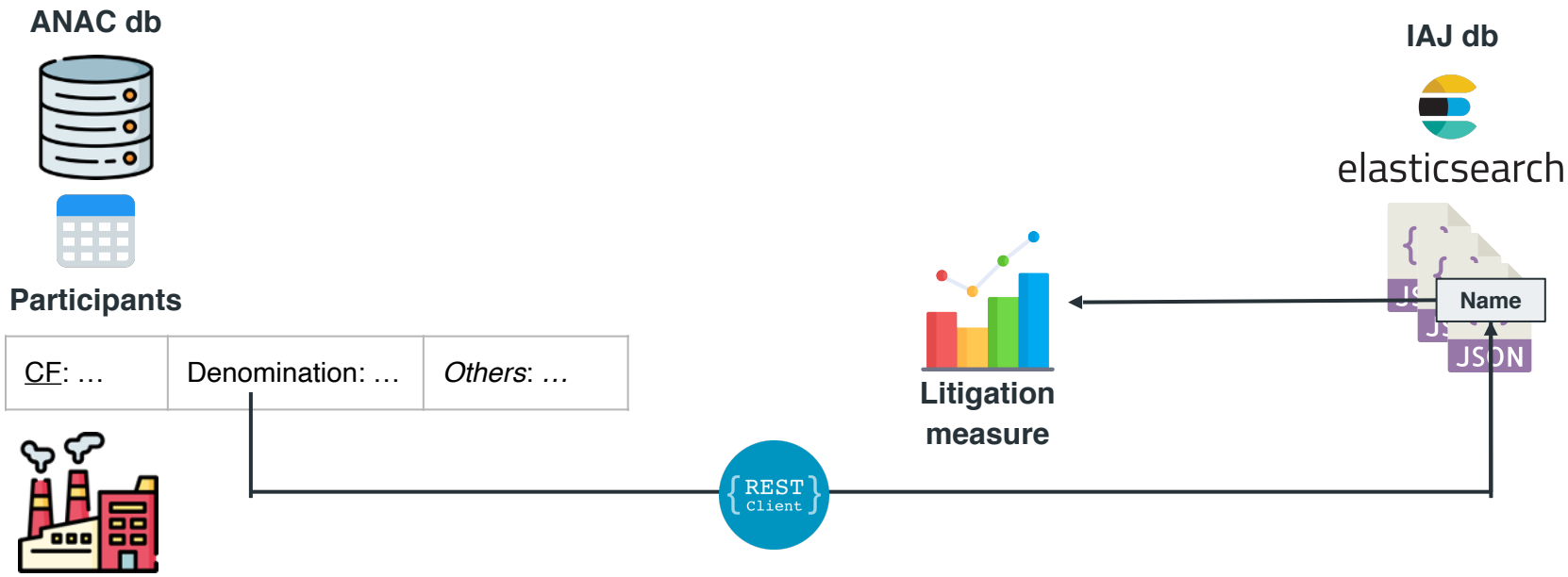
Search by procurement ID (CIG)



Search by participants and contractors

- The connection between the two information sources occurred by:
  - 🎯 **searching for the participants and contractors denominations in the IAJ sentences archive**;
    - the denominations found in the IAJ dataset were useful to find entities inside the legal archive and define a litigation measure between participants (economic operators) and contractors (PA).

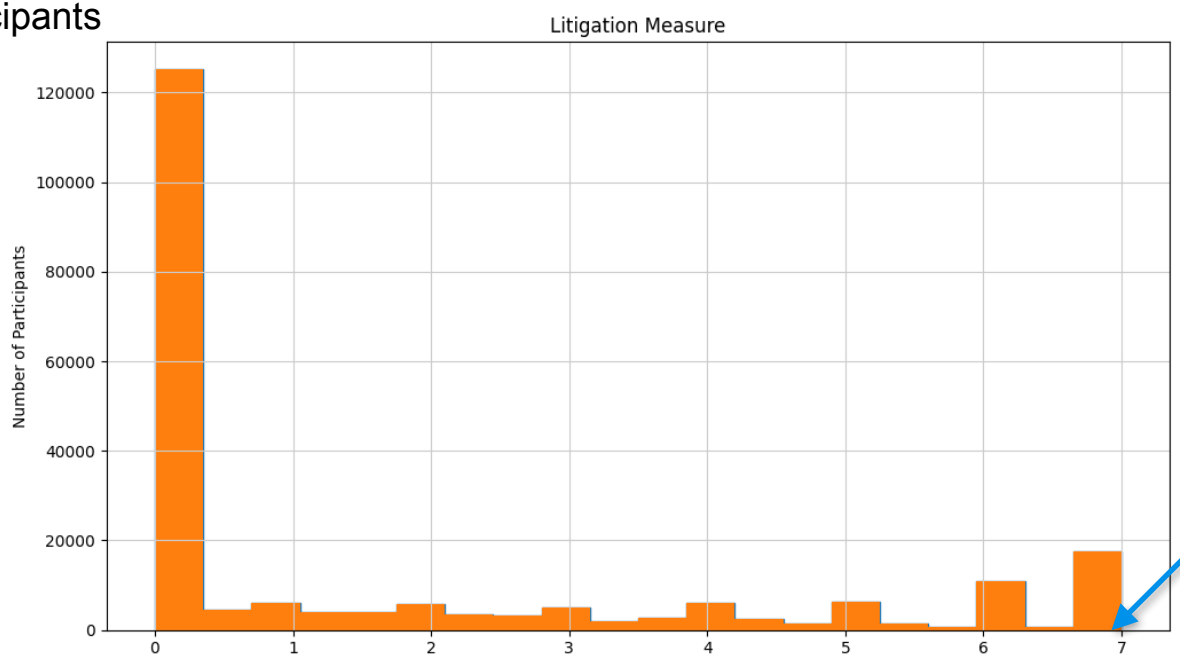
Search by participants' denomination



Results of detection of participants' in IAJ

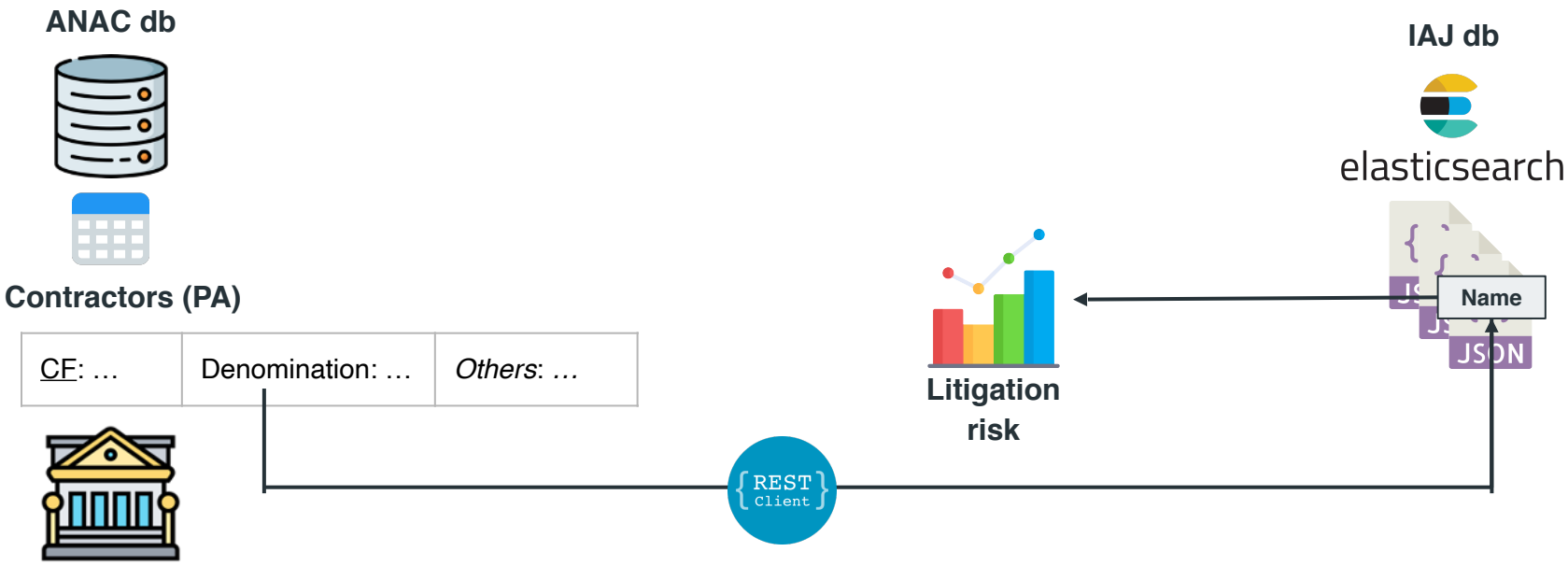
$$\text{Litigation measure} = \frac{Number\ of\ recourses}{Number\ of\ participations}$$

Number of participants



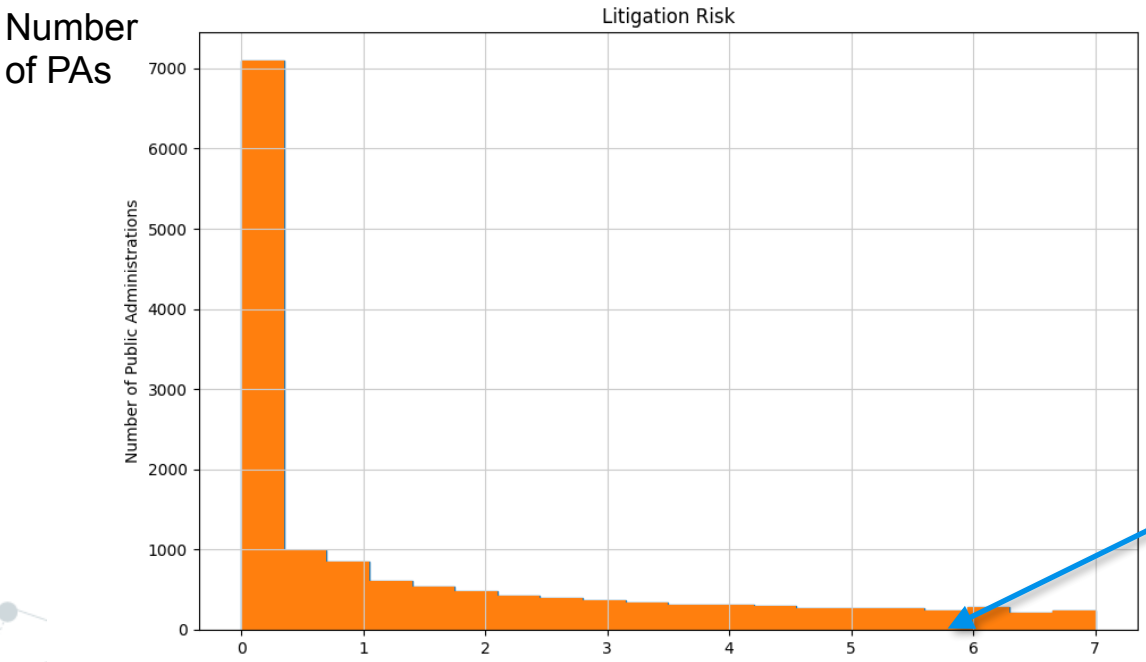
7 is the extreme of the confidence interval at alpha=5%

Search by contractors' denomination



Results of detection of contractors

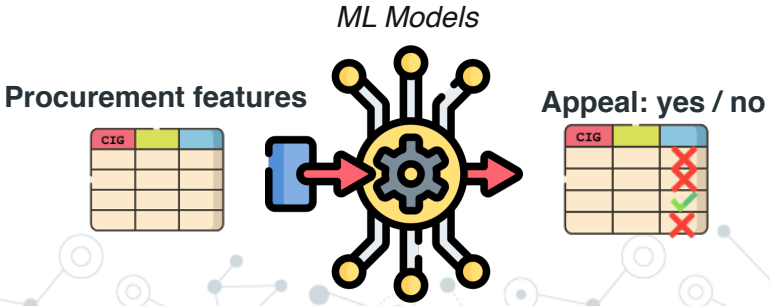
$$\text{Litigation risk} = \frac{Number\ of\ recourses}{Number\ of\ contracts}$$



5.9 is the extreme of the confidence interval at alpha=5%

Machine Learning Models

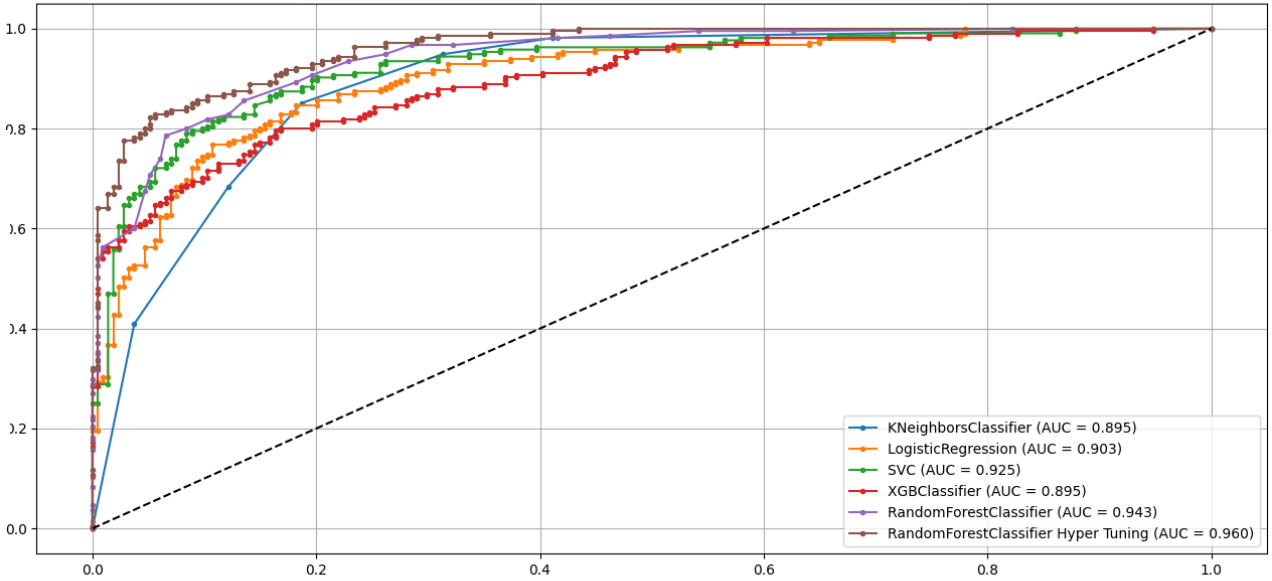
- Following the RQ2, the problem statement is given: find an algorithm such that, given the description of a procurement, it predicts with the highest expected accuracy, the presence of a possible recourse related to that procurement to the Administrative Justice courts.
- We explore consolidated MLAs: it can help to dig into a large amount of data and to discover the predictive patterns based on the procurements features.



Machine Learning Models Results

Procurement object: good/ supplies

TPR

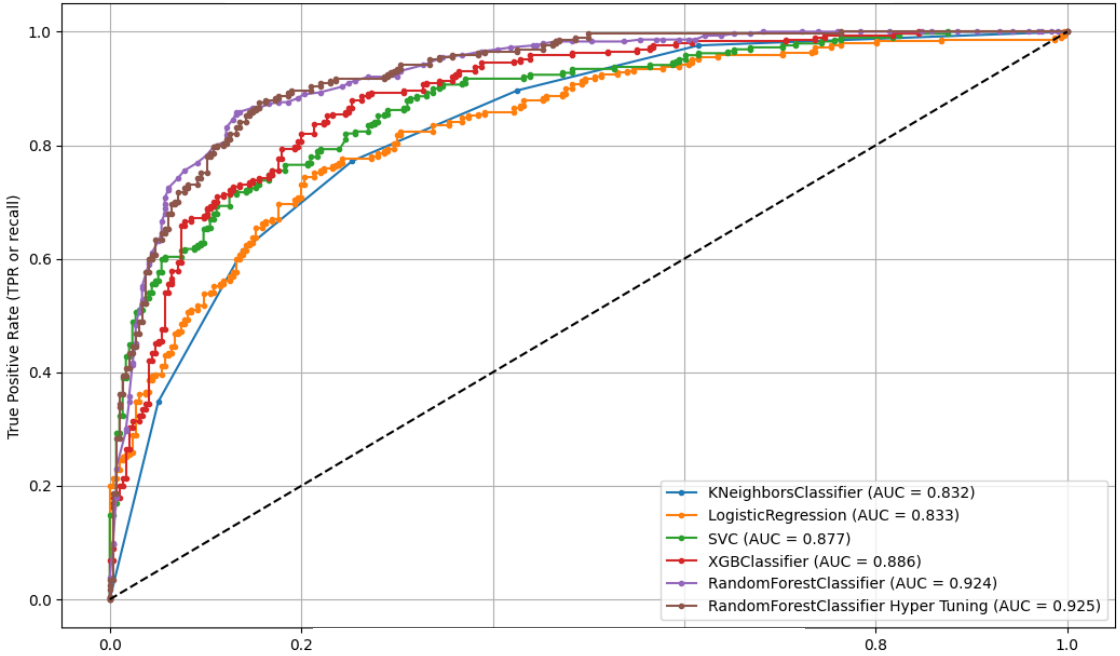


FPR

Machine Learning Models

Procurement object: public works

TPR

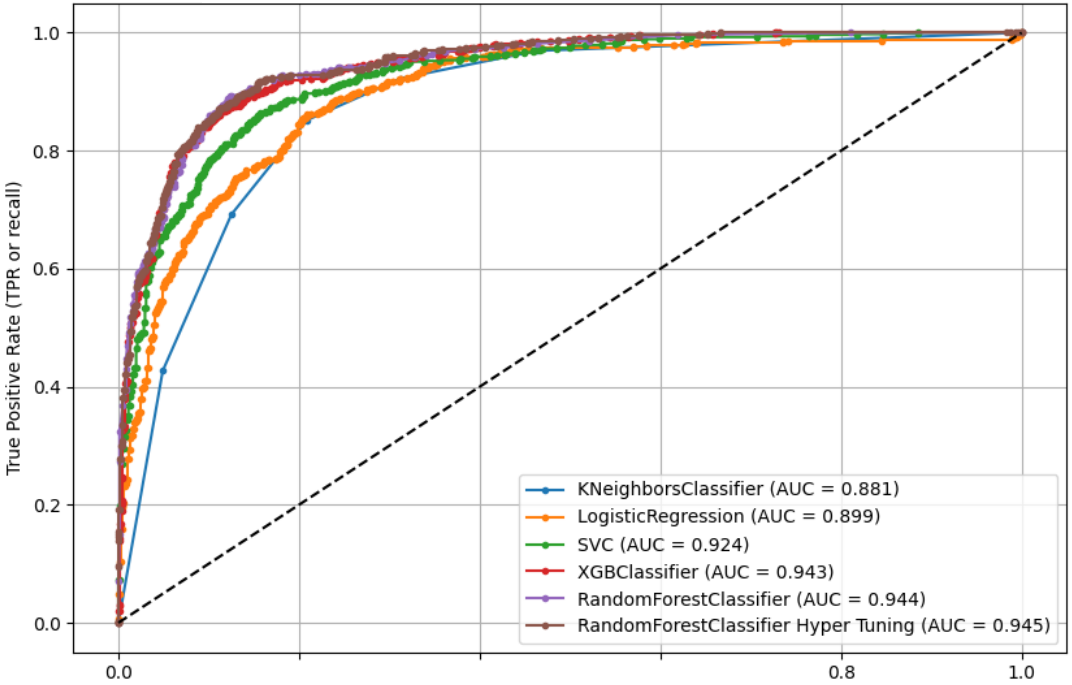


FPR

Machine Learning Models

Procurement object: services

TPR



FPR

Recommender Systems on Contracts (RQ3)

🎯 We want to provide systems to help people (employees from PA or from economic operators searching for the right tenders) working and finding contracts in a large database

- We trained recommender systems on *oggetto del contratto,* a short description that the responsible person of the procurement included in the tender.

- A problem we have in the ANAC data is that the textual descriptions are often ambiguous and carelessly filled by the responsible people of the tenders

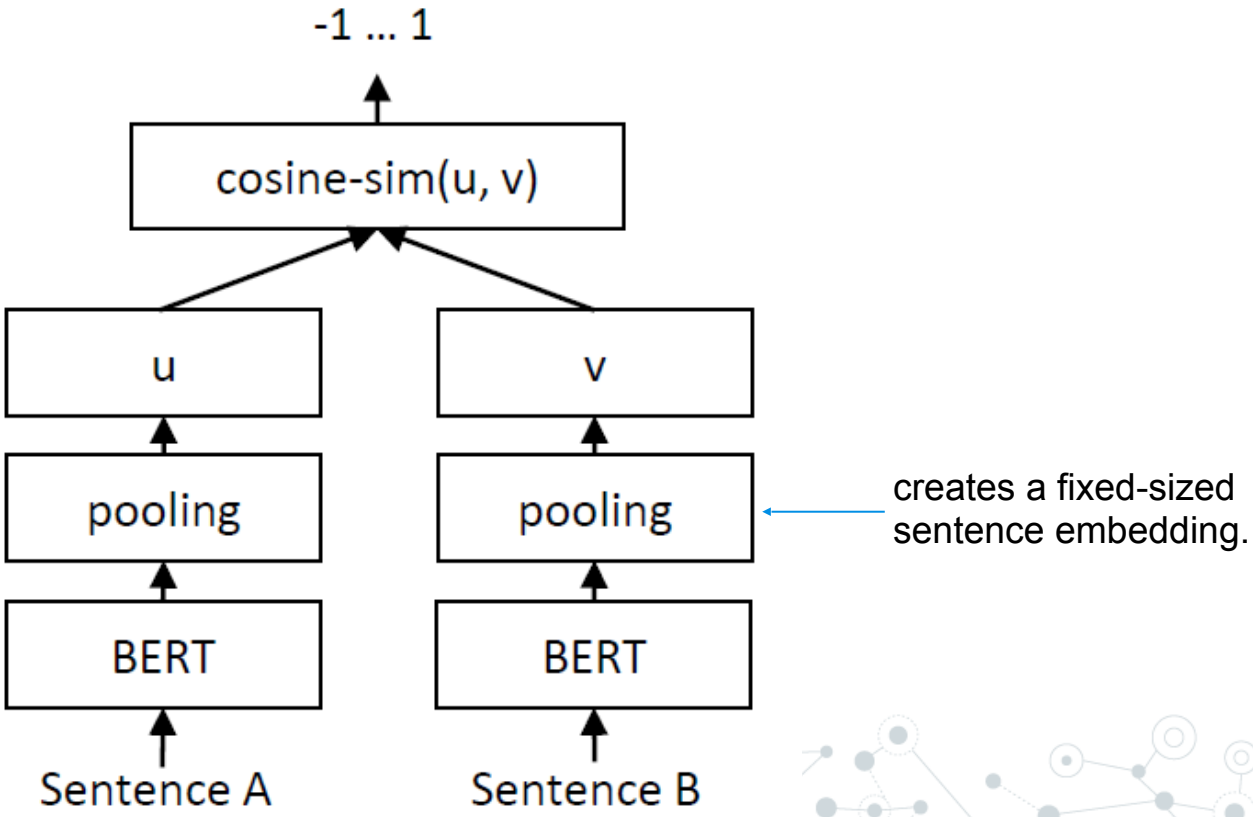
Recommender Systems on Contracts

- We applied S-Bert Deep Neural networks: https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html on the short textual description

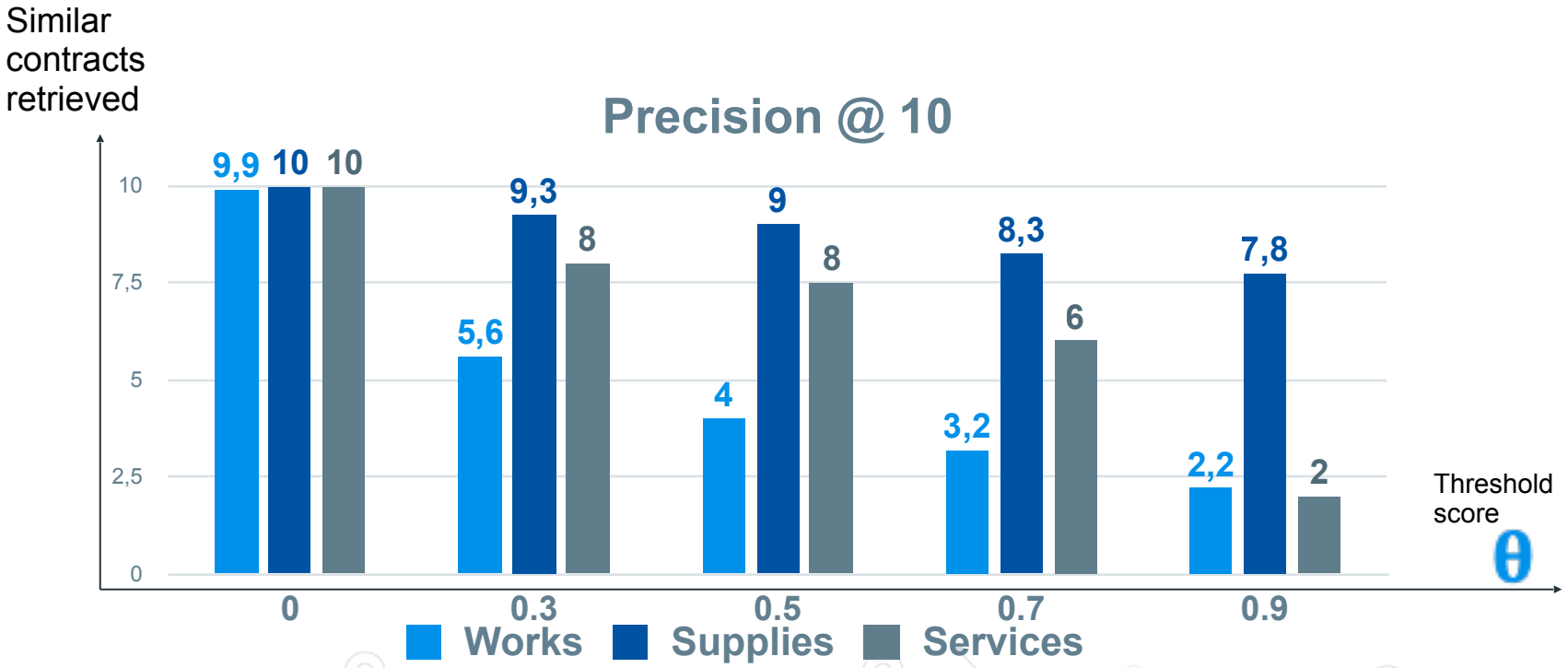
- Transformers outperform both RNN and CNN in translation benchmarks and are trained by ranking on **semantic similarity** a large set of sentences.
- Translation occurs first by encoding the meaning of each word in the sentence and then decoding.

1. We use the encoding vectorial representation of the textual *object of the contract*
2. use *cosine similarity* to retrieve the most similar contracts in the database.

- Transformers, on top of higher translation quality, require less computation, speeding up training by up to an order of magnitude.


-1 ... 1
cosine-sim(u, v)
u
v
pooling
pooling
BERT
BERT
Sentence A
Sentence B
creates a fixed-sized sentence embedding.

Evaluation by a panel of 3 independent experts on a small random training set



Problems with *oggetto del contratto*

- A problem we have in the ANAC data is that the textual descriptions are often ambiguous and carelessly filled by the responsible people of the tenders
- **Examples:**
- Open awarding procedure for assigning a public work contract:
  LAVORI DI 'RESTAURO ARCHITETTONICO DEGLI INFISSI ESTERNI E LA MANUTENZIONE DELLA COPERTURA DEL PALAZZO PRINCIPE DI ARAGONA SEDE DELL'OPERA PIA 'ISTITUTO PRINCIPE DI ARAGONA'
- It is described in a very similar way of the following service whose description is ambigous:
- LAVORI DI RISANAMENTO CONSERVATIVO PER LA MANUTENZIONE EDILE ED IMPIANTISTICA PRESSO IL PALAZZO GADALETA, SEDE DEGLI UFFICI GIUDIZIARI DI TRANI - SERVIZI DI ARCHITETTURA ED INGEGNERIA

5.

Conclusion

and future work

Conclusion

- This work demonstrates the ability to manage a huge juridical dataset from a national public entity to automatically extract meaningful knowledge to address machine learning experiments (RQ1).

Conclusion

- In addition, we explored the results of a predictive experiment to test recourse prediction of the administrative courts on the basis of the features of a public procurement (RQ2).

Future work

- As a future work, we plan to investigate furthermore the explainable AI techniques,
- We need to answer the *causal* questions from the domain experts (law people) on the reasons of a recourse or of an award to an economic operator.

Future work

- As a future work, we plan to expand the usage of sentence embedding and cosine similarity to match the short description of a recourse and of contracts
- We hope this will help us reduce the number of candidate recourses without a corresponding match in the contract database

Future work

- Another direction concerns the adoption of *process mining* techniques for conformance checking and predictive process monitoring applied to a log of temporal events obtained by this database.

Thanks!

Any questions?