




Merging textual and numerical databases: a steppingstone for statistical analyses of illegal events

Maria Francesca Romano¹  · Pasquale Pavone² · Antonella Baldassarini³ · Giuseppe Di Vetta⁴ · Gaetana Morgante⁴

Accepted: 12 December 2024
© The Author(s) 2025

Abstract

This paper aims to define a methodological path—merging judgments and official statistical data—to organize complete, objective, and reliable data in a database, thus simplifying the analysis of illegal social phenomena. Judiciary judgments are a new data source: they deal with illegal events that describe social phenomena—even if they are only the "illegal" ones—and contain objective and reliable data and information. Judiciary judgments are also texts, so the first step is a statistical textual analysis and text mining techniques to discover information and organize it in a statistical database. The final database is obtained by integrating numerical data from other information sources. It therefore has statistical properties such as reliability, completeness and updating. Subsequent statistical analyses or modelling are then possible based on the entire set or subsets of data adequately extracted from the implemented statistical database. We present some results obtained from judgments about corruption in order to demonstrate the advantages of linking textual databases (textual analyses on judgments) and numerical databases (from ISTAT). The proposed methodology can benefit different stakeholders, such as researchers, policymakers, and other enforcement actors. It is independent of the specific software used and remains valid when applied to other illegal activities (e.g., organized crime, tax crime, and money laundering). Furthermore, the results may be even more effective if the institutional actors involved have access to judgments at all levels, thus overcoming potential privacy concerns. The methodology could also be used to support evidence-based policy in the fight against crime and illegal activities.

Keywords Textual database · Merging textual and numerical database · Statistical database · Illegal events · Data science and criminal law

1 Objectives

Judiciary judgments are big data even if their elements are words: the big (texts) data have the same advantages and disadvantages as big data: their use reduces both research costs and time; they permit the design of more ambitious research hypotheses; but they need peculiar attention to their characteristics and require more skills to researchers (or a

Extended author information available on the last page of the article

multidisciplinary research team). Furthermore, it is mandatory always to remember that data and text data are collected, stored, and managed for a precise administrative scope.

Judiciary texts describe sanctioned illegal events in detail. Taking as a whole (by time or by illegal event, for example), a researcher could use judiciary texts to describe social phenomena—even if the only "illegal" ones, but with an undoubtedly pro: from the texts, it is possible to obtain objective and reliable data and information, those being the prerequisite for the final decisions taken by the magistrate.

In this paper, we discuss how it is possible to use judicial texts as a new source of big data, transform them into a textual database, and merge them with other databases to gain more knowledge about illegal events.

Judicial texts are not a new source of data for criminologists, but we think that the automatic textual data process, merging with numerical databases, and database design could be helpful methodologies for increasing knowledge about specific illegal events.

Textual analysis applications are now widespread in the scientific field, and texts, from literary and scientific works to newspaper articles, are becoming increasingly available.

The interest of linguists in the judiciary language has, in Italy, a long tradition of analysing the legal language (Mortara Garavelli 2001; Bellucci 2005; Ondelli 2014) and the legibility of the legal texts (Brunato and Venturi 2014). Furthermore, computational linguistics had expressed interest in the texts of judgments, favouring the construction of ontologies and automatic extraction of the topics dealt with (Dell'Orletta et al. 2008; Lenci et al. 2009; Ceci et al. 2012; Luz de Araujo and de Campos 2020).

Nowadays, the greater availability and computerization of legal documents allow classification analyses on a highly significant number of legal texts (Nese and Troisi 2019; Iezzi and Bertè 2020; Filtzt et al. 2020) and arouse attention to textual analysis methods theses in legal disciplines in several countries. (Bottone 2019; Peruginelli and Faro 2019; Luscombe et al. 2022).

Moreover, there is nowadays increasing attention to a computational approach to judicial texts (Luscombe et al. 2022), and several research groups are exploring judgments with AI tools, the so-called Predictive Justice (Comandè 2019; Viola 2019).

Judgments as a source of information imply some relevant problems:

- (a) The availability of resources to adequately deal with big data textual documents.
- (b) The possibility of obtaining judicial texts.
- (c) Last, above all, all aspects that concern the compliance of privacy.

As for the first problem, computing resources are fine, given the growing availability of hardware and software. In recent years, the availability of proprietary and open-source textual analysis software has increased, and the already cited literature uses judicial or administrative texts.

The research team must have a real multi- and interdisciplinary approach among the necessary resources, however, since the "technical" choices must derive from the researchers' objectives.

The research team's composition also has implications for the second type of problem, as access to legal texts can be facilitated by the presence in the research team of magistrates and/or public administration personnel. As for Italy the transparency of the final acts of the judiciary allows access to the judgments in their final judgment at the Corte di Cassazione (Court of Cassation), and these texts were used to extract financial information like amounts in euros or private companies involved in corruption (Rey 2017). Another

research group led by Zuliani et al. (2009, 2013) has identified in the judgments a source of data for measuring corruption in the Public Administration. They used methods and tools of lexical-textual analysis on the judgments of the Corte dei Conti (Court of Auditors), identifying their temporal dimension and territorial diffusion. As a growing literature shows, data analysis tools play a crucial role in driving policy-decision making concerning crime and illegal activity, and in increasing the policing and law enforcement capacity (Europol 2023; Strikwerda et al. 2024; Porcedda and Wall 2024).

As for ethical aspects, automatic anonymizing of the judicial texts is now affordable (Csányi et al 2021; Romano et al. 2022; Licari et al. 2022). Anonymization surely allows to mitigate the privacy concerns that textual analysis usually raises. However, anonymization procedures could have an impact on statistical properties. The degree of impact also depends on the aims of applying this methodology. Therefore, the impact must be evaluated according to the specific usage scenario. Based on this assessment, optimization procedures can be adopted to minimize information loss. When the proposed methodology is applied to characterize illegal episodes and extract aggregate information, the impact on statistical properties is minor. The processed data retains their usefulness. In contrast, the impact can be much more significant when the methodology is applied as an investigative tool in the context of enforcement activities. In these cases, optimization procedures must ensure a useful balance between privacy protection and application usefulness.

2 Methods

Many analyses concentrate on the extraction of this information (both qualitative and quantitative data) from juridical texts in natural language. Even if these analyses obtain good results (narrative analysis, for instance), these results could be enhanced: in our advice, the challenge is to go beyond the textual analysis. After the first step—obtaining a textual database, we need to merge this database with other numerical databases, with a particular emphasis on the official source of data. The aim is to outline a methodology (or a strategy) to link multiple sources of data/ information. This goal is concrete, with application in multiple contexts, achievable automatically and without limits in the data size.

Integration between information sources is relevant to increasing knowledge of illegal phenomena: data extracted from judgments are valid per se, but they are a stepping stone to merging textual data and data from other statistical sources.

Even though the methodology is general, we choose to analyze corruption, so it is possible that something is more suitable for this illegal event than for others.

From juridical texts, we define and differentiate the elements that make up an illegal event. They are:

- WHO: the natural or legal person who committed it or suffered it
- WHERE: the place or places where the events took place
- WHEN: when the event occurred
- HOW MUCH: the economic quantities (payments of bribes, protection money) or other quantities to be exchanged (such as favours, electoral exchanges)

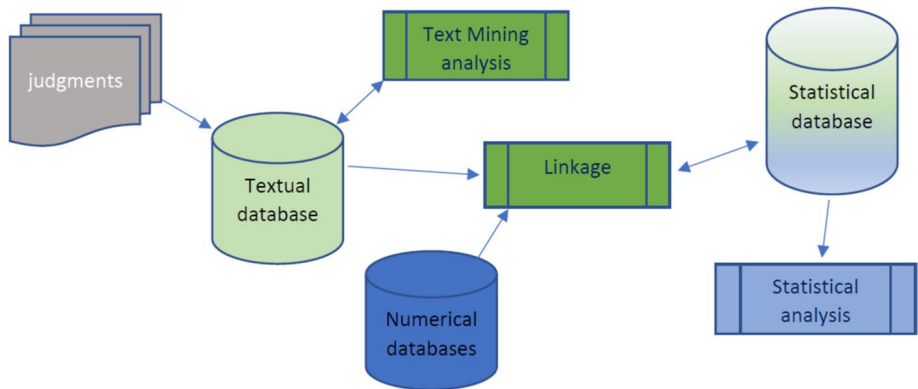


Fig. 1 Scheme of the proposed methodology

- **HOW:** as to how the facts took place (e.g., the possible presence of criminal organizations).¹

The results obtained through the statistical-textual analysis of judicial rulings, suitably integrated with numerical data from other sources of information, allow the design of a database with statistical characteristics (reliability, completeness, updating). Subsequent statistical analyses or modelling are possible based on the entire set or subsets of data adequately extracted from the implemented statistical database.

To know (and then measure) a phenomenon through elements described in judgments starts from transforming texts (judgment) into statistically analyzable data, defining procedures to identify and quantify unambiguously the information considered "useful".

In other words, our work proposes a step ahead by showing the path for "appending" judgments and other legal data to "official" and certified statistical sources to create a more productive environment for social researchers.

This issue cannot be addressed only by computer scientists. However, statisticians have a leading role in its implementation to ensure quality in the processes related to organization and data integration.

It is well known that current statistical similarity measures or probabilistic models are the basis for matching techniques, making the methodology of integrating databases a topic closer to data analysis than archives' simple IT alignment.

We describe the methodology for the relation between judicial sources and other sources, demonstrating its applicability in the specific field of corruption. In contrast, the limits and perspectives of the methodology outlined will be discussed in the Conclusions.

Figure 1 gives a synthetic scheme of the proposed methodology, and Table 1 details the phases and steps. The steps of the methodology indicated in Table 1 are described in Sects. 2.1, 2.3, and 2.3 below, respectively.

¹ these descriptions can be further detailed, for example, considering the role played by each actor (natural person—NP or legal person—LP). For example, an NP may have committed a criminal act by abusing his role (professional or political) within an LP (private company or public institution), and an LP can be involved in the event as a civil party or as an injured party.

Table 1 Phases and steps of the methodology

Phase	Steps
Lexical—textual analysis	Create the data warehouse Text mining Tagging Annotation Defining new variables Export of data
Merging textual and numerical databases	
The design of the statistical database	Database design Implementation Queries to define subsets data

A premise is necessary before describing each methodological step. However, a complete generalization of the method would be difficult. The methodological phases prove to be sensitive to the different software used and to the diverse databases. As such, applying this methodology may need a preprocessing phase and evaluation in terms of comparability.

2.1 Lexical-textual analysis

The first step is to create a textual database (Data or Document Warehouse) and submit judicial texts to textual software.

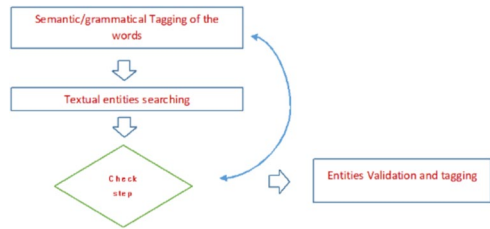
We used TaLTaC (<https://www.taltac.com/>), a software for analysing a collection of texts (Corpus), to describe and interpret its content and some of its properties.² The general structure adopted in the program is known in the literature as a "lexicometric approach" as it allows the direct study of any set of data expressed in natural language, from documents to interviews, from press reviews to messages, according to the principles of "textual statistics" and the Automatic Text Analysis. TaLTaC software (Bolasco 2013; Bolasco and Gasperis 2017) with a peculiar definition of the parsing rules: even if it is possible to normalise some aspects of the text, such as, for example, the reduction of all caps—in order to reduce the number of graphic forms in analysis—we prefer to keep capital letters (except for the forms that follow vital punctuation), to more easily identify the names of legal persons (companies but also consortia or public bodies) which—often (but not always)—are mentioned in the judgments with a capital letter and in a very different way.

In importing the Corpus, another "non-classical" choice—dictated by the same purpose described above—considers the point not a delimiter but a character. Punctuation marks are, in fact, "natural" separators between words, but they can lead to problems in recognising the legal form of a company or the name of cities or places.

Another feature of TaLTaC software is the possibility of respecting the general structure of the judgment, which is organised into ctions: Incipit (list of parties involved), Fact, Law (Fact and Law), and PQM (Disposals).

² The software was used to analyse the corpus in version 2.10 and the recent version 4, officially presented on February 21, 2020, and available for the research group in beta version.

Fig. 2 Scheme of the recursive tagging action



After the Corpus is created, the textual data is already organised in a database (Document Warehouse) that consists of two primary tables: Vocabulary (lexical) and Documents (textual).

The lexico-textual analysis phases add progressively new variables to the Document matrix, along with the presence in the judgment of each identified entity and the number of their occurrences.

After this pre-treatment of the texts, the lexical and textual analysis phases can be moved on.

The object of study in lexical analysis is the lexicon, and the lexical analysis units are the different words recurring in the Corpus. In any case, based on the characteristics of the Corpus under analysis and the objectives of the research, from time to time, they can be considered as lexical analysis units: the single words as reported in the text, the multiword expressions, the headwords, or the roots of the words and the grammatical or semantic categories of each word identified as lexical analysis units.

Lexical analysis occurs by exploring the information structured in the Vocabulary DB, where annotations of various types can be associated with each word (type): grammatical, semantic, and statistical. These annotations are the result of automatic processing at different steps of the analysis.

Each of these properties constitutes an example of meta-information attributed to the type, which can be retrieved by querying the corresponding fields stored in the Vocabulary DB. The extraction/selection of the vocabulary parts "tells" the lexical characteristics of the Corpus, highlighting the significant elements of each "part of the speech" or "illustrating" specific subsets of types and the relationships existing between them.

The object of study in textual processing is the Corpus, a succession of words within a collection of texts to be analyzed, compared, and categorized. The unit of analysis is the unit of context, a fragment (be it a sentence or an entire document). In line with the Lexical Analysis, each context unit constitutes an entry in the Documents DB. The possible modalities of the a priori coded variables and the textual annotations (categorizations) obtained through the explorations performed from time to time are associated. These annotations can be of various kinds: syntactic, obtained by categorizing documents in which specific syntactic structures or groups with variable elements are present; semantics, regarding automatic categorizations based on specific lexicons; quantitative. One of the classic explorations in textual analysis is Regular Expressions (RE). REs search text strings: from time to time, they can be the occurrences of words or complex entities, both of their classes and relationships between classes or single types and classes. The search among the word classes occurs by searching for the grammatical and semantic annotations carried out in the lexical analysis phase. This processing retrieves the fragments that verify the textual query, inventory the list of extracted strings, and, if necessary, notes the Documents (judgments).

To search for such entities of interest, hybrid models of a search for sequences of words and annotated classes of words have been defined from time to time. About the latter, it

was possible to generate classes of annotated forms using the linguistic resources offered by the TaLTaC software, particularly Toponyms (to detect Places) and people's names (to detect natural persons). In addition, the grammatical tagging of the graphic forms in the Vocabulary recognizes the most common multiword expressions.

Each single tagging step results from recursive actions (Fig. 2).

In the end, each judgment is described by variables (qualitative and quantitative) regarding the judicial process (such as the initial judicial seat and the presence of civil parties), but also the presence and number of companies, public institutions, the presence of criminal organizations, and the professional or political role of the natural persons involved.

Along with this phase, the Vocabulary matrix permits the carrying out of queries, and the Documents matrix stores the results. However, they can also constitute the input for subsequent analysis.

2.2 Merging textual and numerical databases

The linkage phase aims to add quantitative and qualitative data on specific entities identified in the Text Mining phases. In the various analyses carried out so far by the research group, these specific values have been the names of the companies or the geographical locations (Italian and foreign).

The choice of databases and the linkage methods are still open to reflection, along with the awareness that the research objectives (and therefore the possible clients/users/stakeholders) and data access guide the process, leaving the empirical applications of the proposed methodology in the background.

The aim of database merging is to organize complete, objective, and reliable data in a statistical database, making every type of analysis about an illegal social phenomenon easy.

Our proposal could extend with minor changes to many textual databases representing many of the so-called Big Data. Moreover, two aspects are crucial to creating a «multipurpose» data set. First, the process should be «automatic» and independent from the research goals, and all the processes must be planned for handling a large amount of data.

We consider official data sources mainly because they respond to criteria of completeness and validation, with certified quality of all data processes, from collection to updating. Official statistical registers are the only ones to have complete (verified and updated) archives of statistical units; the units can be private companies³ or public entities.⁴ Statistical databases use administrative information subjected to a process of normalization and standardization that transforms administrative units and characters into integrated statistical units and variables.

Notice that the statistical database's units and variables depend strictly on the researcher's hypotheses and goals. The starting point is always the Corpus of judicial texts, which is a precious source of information about illegal events.

Suppose the goal is to investigate the economic impact of corruption (how many private companies are involved and in which role, by economic sector, and by number of employees). In that case, the company's name (or its unique identity) is the critical variable linking the illegal event of corruption to other statistical variables.

³ like ASIA Archives by ISTAT.

⁴ Permanent Census of Public Institutions by ISTAT.

Table 2 Corruption database structure

Matrix	Description	Key variables	Source
E	Event \times judgments	ID_event, ID_judgment	Judgments
S	Judgments \times variables	ID_judgment	Judgments
A	Companies \times variables	ID_company, ID_event	Judgments + Orbis + ASIA

Therefore, the crucial point is to define the linkage variable(s): the variable can be the names of private companies, public administration offices, or geographical areas (Regions, Provinces, and/or Municipalities).

We used both an official numerical dataset (ASIA register by ISTAT) and the ORBIS database⁵ in the linkage phase: in the first step, there was the univocal recognition of the most significant number of companies extracted with automatic procedures, the integration of values for missing information in one of the two archives, and last, a validation step. Control strategies were adopted to measure the reliability of the automatic recognition of the company, that is, to test the reliability of the process.

We tested the "false negativity rate" (specificity) of judgments: the failure to identify companies by automatic procedures. First, a random sample (about 5% of the judgments in the Document Table) was carried out to test the procedure: a human reader confirmed the complete absence of false negatives in the sample.

The statistical database, which contains data from textual and numerical tables organized by units, is the basis for subsequent statistical analysis.

The innovative aspect of our approach is the definition of a process of integration between textual and numerical databases. Furthermore, we consider it relevant that all these phases have been carried out automatically to replicate analysis on a substantial amount of data.

2.3 The design of the statistical database

The final data structure is a relational database⁶ with several tables corresponding to the data matrices. For each data matrix, one or more variables are key-variable, making it possible to select units of analysis for the application of statistical analysis, both descriptive (indicators or counts) and inferential (predictive models).

Some variables derive from the lexical-textual analysis phases (source: judgments) or the statistical sources (source: numerical databases).

Note that each variable in the matrices is reliable, deriving from both the source used and the construction method: reliable data are available for each company (such as ATECO classification, number of employees per year, registered office), and objective information (for example, the role in the criminal event) obtained from text mining procedures. However, we can consider all the available data as "statistical."

⁵ The Orbis database is widely used by companies, governments, public sector teams, academics, financial institutions, and professional service firms worldwide. It has information on around 300 million companies worldwide.

⁶ The Corruption Statistical Database was implemented with SAS software.

Table 3 Examples of other available matrices

Matrix	Description	Key variables	Source
AS	Companies × judgments × variables	ID_company, ID_judgment	Judgments + Orbis + ASIA
AES	Companies × events × judgments × variables	ID_company, ID_event, ID_judgment	Judgments + Orbis + ASIA

3 Results: an application of the methodology to the corruption phenomenon in Italy

The methodology was applied to the texts of 684 judgments randomly chosen from the decisions issued by the Court of Cassation between January 2015 and January 2020. All 5,164 decisions contained one or both the words "corruzione" and "conscussione."

The judgments were exported as PDF files and converted into text format through the open-source software Pdftotext (www.pdftotext.org) for import into TaLTaC.

The corpus comprises 684 judgments (Documents), and the Vocabulary consists of 75,436 lexical units for 3,153,050 occurrences.

We followed the phases and steps listed in Table 1.

Table 2 reports the database structure; more details about variables by matrix are in the Appendix.

The database permits the most exhaustive analysis. The researcher interested can choose between different views combining data by key variables. The research hypothesis will guide the records selected from the database. A list of some derived matrices is in Table 3. Each judgment can cite multiple companies. The same company can be present in more than one judgment, even with different roles (civil party, injured party, defendant), and there are also judgments in which no companies are involved.

Furthermore, the settings described above are valid for further linkage: they could be—for example—extended to public bodies involved in the judgments or to tenders managed by public bodies, as well as to possible geographical analyses based on the places involved.

Other analyses can assume companies as the unit of analysis, for example, their presence in one or more judgments (unrelated to each other) or a civil party.

Here, we present some descriptive tables from the implemented Corruption statistical database to demonstrate the increasing knowledge of the corruption phenomenon when using reliable variables on the entities mentioned in the judgments.

Our goal here is to show how the database's statistical structure permits analyses with a different subset of data using variables from textual and numerical databases.

First, it is possible to describe judgments: the unit of analysis is, in this case, the single sentence ($n=684$) for which we can use the variables derived from the text mining phase as a geographical area of events.

Using companies as a unit of analysis ($n=985$), we can describe them by economic sector (Table 4).

Trade, Manufacturing, and Construction Trade are prevalent economic areas, but companies belong to all economic areas.

Table 4 Companies in judgments by economic sector (ATECO Code)

ATECO code	Description	N
G	Wholesale and retail trade; repair of motor vehicles	247
C	Manufacturing	146
F	Construction	146
I	Accommodation and food service activities	79
L	Real estate activities	77
M	Professional, scientific and technical activities	59
E	Water supply; sewerage, waste management	42
N	Administrative and support service activities	38
H	Transportation and storage	35
Q	Human health and social work activities	20
J	Information and communication	15
K	Financial and insurance activities	14
S	Other service activities	14
D	Electricity, gas, steam and air conditioning supply	9
R	Entertainment, sports, and arts	9
B	Mining and quarrying	8
All		958

Table 5 Companies involved by area of crime event and area of registered office

Crime event	Registered office in Italy			All
	Center	North	South	
Center	96	85	29	210
North	57	188	51	296
South	104	169	170	443
Not unique	2	5	2	9
All	259	447	252	958

The comparison between the geographical area of the registered office and the area of the crime event helps describe the same area of involved companies (Table 5).

Less than half (47.4%, 454 out of 958 companies) are involved in judgments in the same area. Furthermore, from the table, companies with registered offices in North Italy were involved in crime events in the same area. Instead, companies with registered offices in South Italy are more involved in crime events in the same area (170 over 252).

A further breakdown by economic sector (Table 6) indicates that the percentage of companies that "commit crimes" in the same area of the registered office is lower for the manufacturing sector (41.10%) and Accommodation and Food Service Activities (37.97%). In comparison, a different percentage value in the same area shows Construction (57.53%) and Trade (46.96%).

The widespread presence of judgments involving companies (about 80%) indicates that the proposed methodology is particularly fruitful. It allows for an accurate and reliable description of these companies' characteristics, using both variables from the textual database (as crime area) and numerical database (economic sector, registered office).

Table 6 Companies involved by area of crime event and area of registered office: a focus on four economic areas

Manufacturing					Trade			
Crime event area	Registered office			All	Registered office			All
	Center	North	South		Center	North	South	
Center	8	17	4	29	21	19	7	47
North	3	37	7	47	16	45	18	79
South	17	38	15	70	24	44	50	118
Not unique						2	1	3
All	28	92	26	146	61	110	76	247
% same area	41.10				46.96			

Accommodation and Food Service					Construction			
Crime event area	Registered office			All	Registered office			All
	Center	North	South		Center	North	South	
Center	8	5	9	22	18	10	4	32
North	2	5	8	15	5	39	5	49
South	9	13	17	39	5	32	27	64
Not unique	1	2		3		1		1
All	20	25	34	79	28	82	36	146
% same area	37.97				57.53			

Table 7 Crime event in judgments without companies involved by geographical area of the event

Crime event	Geographical area of the event							
	North Italy		Centre Italy		South Italy		Total	
	N	%col	N	%col	N	%col	N	%col
Crime against the public administration – corruption	20	52.6	19	61.3	23	50.0	62	53.9
Forgery	1	2.6	6	19.4	6	13.0	13	11.3
Others	7	18.4	3	9.7	4	8.7	14	12.2
Crime against the public administration	3	7.9	1	3.2	6	13.0	10	8.7
Crime against person	3	7.9	1	3.2	2	4.3	6	5.2
Property crime	1	2.6	1	3.2	3	6.5	5	4.3
Mafia-type organized crime offence	3	7.9			2	4.3	5	4.3
Total	38	100.0	31	100.0	46	100.0	115	100.0

Others: Drug offences; Money laundering; Bankruptcy offence; Weapons offence and Tax offence

Furthermore, the statistical database permits the identification of crime events described in judgments, along with other variables of researcher interest. For example, Table 7 shows the 115 crime events by area of the event, only for judgments without companies involved. Moreover, this subset could be the basis for further analysis of the so-called "petty corruption".

The structure of the implemented database permits the detection of other crime events associated with corruption, as listed in the rows of Table 7.

4 Conclusions

The presented work relates data from textual sources with data from statistical sources by bridging databases of different types. The application of text mining to judicial data (to case decisions) has multiple purposes. From a scientific point of view, the method described can provide quantitative and qualitative—statistically significant—information on observed (criminal) phenomena, such as corruption or others. Text mining software represents a tool for enhancing empirical criminal research; it enables the validation of theoretical (or criminological) explanations of crime or phenomena, such as a real correlation between organized crime and corruption. The discussed approach ensures the improvement of the analytical background:

- mapping criminal phenomena in a limited context (national, regional or local)
- elaborating correlations between different dimensions of the same phenomenon
- supporting qualitative statements

The preliminary analyses revealing corruption linked to criminal organizations encourage the continuation of the research, as well as the execution of predictive analyses and the quantification of economic amounts by type of corruption. The ability to extract the sums of money involved could help estimate the economic impact of corruption, which was, in fact, Guido Rey's original objective (Bechi and Rey 1994).

From this perspective, the proposed methodology focused on corruption, where the most important international initiatives to develop data measurement and analysis tools have been undertaken so far (Holmes 2015). In this way, the research ensures high comparability with other methodologies already in use and under development (Lambsdorff 2023).

Regarding corruption, text mining applications will characterize individuals involved in the criminal event, e.g., companies. In this sense, representing the role and (corporate) character of the legal entities involved is very useful in highlighting which types of companies—which type of governance or management order—are more likely to facilitate corruption.

From this perspective, applying text mining to judicial data enables policymakers to adopt an evidence-based approach to monitoring social phenomena such as crime. Furthermore, the method can be applied to empirically verify policies or laws to prevent criminal and other negative phenomena.

Other social phenomena can benefit from the proposed methodology. In this sense, corruption is one of the empirical fields in which the text-mining approach should be tested and validated. Applying this methodology requires the observation of the social phenomenon in statistical sources and databases. Therefore, criminal phenomena represent one of the most significant grounds for testing text mining applications, as described in this article. When we refer to criminal phenomena such as corruption, corporate crime, gender crime, and immigration crime, we are talking about social phenomena inherently equipped with statistical evidence that are judicial data (decisions of judges and courts). The methodology discussed, therefore, takes these textual data (judicial decisions) as a basis; in this

way, the method can provide an objective representation of the phenomenon due to the empirical content of court decisions.

These texts may be useful not only for criminology experts but also for social scientists: although sentences describe ‘illegal’ events, many societal phenomena are part of criminal court documents.

Text mining techniques thus provide social scientists with the tool to extract information from judicial sources to deeply investigate relationships, social features, and manifestations and to deliver explanations and predictions.

Applying textual analysis tools to judicial big data could also help courts at all levels to formulate more predictable decisions and rules and to ensure legal certainty (Xu 2021). We considered the possibility of including predictive analyses but decided against it to avoid producing biased results due to incomplete data across all levels of judgment. Due to privacy concerns, corruption cases resolved before the Supreme Court are missing from our database. Therefore, we have postponed this type of analysis to the desirable continuation of the research in collaboration with institutional partners such as judges and other enforcement agencies.

It is the dream of every social scientist to have complete, reliable, and objective data on the social phenomenon of one’s interest and, moreover, to retrieve the data with a reasonable commitment of time and personnel.

Today, social scientists can use many data sources in the age of digitization and big data (Salganik 2019). Administrative data (e.g. Connelly et al. 2016) are a particular type of big data; among them, text data are a compelling subset of government administrative data.

The full implementation of the *Procedura Penale Telematica* in Italy—currently still in the start-up or first trial phase—will simplify the linking strategy. In the case of unambiguous identification of the parties involved (VAT number or tax code), the connection with ASIA will be immediate, as in the case of the Telematic Civil Procedure. The connection with the Orbis database will enable the identification of companies with registered offices abroad. Further applications will also be required to activate pseudo-anonymization procedures (already tested for another research).

Textual analysis techniques of judicial decisions could improve law enforcement. Text mining tools could be implemented in police investigations to increase detection capabilities, predict criminal behaviour, and obtain valuable evidence (Alcántara Francia et al. 2022; Bifari et al. 2024). Judicial agencies, such as criminal prosecution offices, can use text-mining techniques to target investigations in particular territorial contexts or develop proxy crime indicators. However, further analysis of the statistical database depends on the research hypotheses selected and, ultimately, on the different stakeholders: policymakers or law enforcement actors.

Finally, collaborating with legal and economic language experts (lawyers, economists) will help refine the methodology. The availability of technical vocabulary at the economic level allows for better identification of the company and the availability of terms and expressions typical of legal language, clarifying the substantive and procedural role of the actors involved. Furthermore, linking several data sources is possible thanks to the participation of ISTAT researchers in the research group.

In the future stages of the analysis, the research group plans to involve institutional actors whose roles will allow them to overcome privacy barriers and access judgments at all levels (e.g., courts and financial police). At this stage, we are considering implementing predictive models to quantify the risk of corruption by type and territorial scope.

Appendix

Corruption Statistical DataBase: Data Matrices

E: matrix Event (965 obs)

Variable	Type	Description	Range	Source
ID_judgment	Alfanum	Key variable	Text	Input
ID_event	Alfanum	Key variable	Text	Assigned
Event description	Alfanum	Event type	Text	Results from information extraction

S: matrix Judgment (684 obs)

Variable	Type	Description	Range	Source
ID_Judgment	Alfanum	Key variable	Text	Input
N_Occorrenze	Num	Graphical forms occurrences		Var Taltac
anno_inf	Num	Year of decision	2015–2020	Input
concess	Num	Presence of word “concessione”	0; 1	Results from information extraction and text mining phase
corruz	Num	Presence of word “corruzione”		
Sede_origine	Alfanum	Place of territorial Court	Text	
d_ActivOC	Num	Presence of tag “Attività_OC”	0; 1	
d_Azienda	Num	Presence of companies		
d_DirAz	Num	Presence of top management roles in private companies		
d_Dirigenti	Num	Presence of top management roles		
d_DLG231	Num	Presence of tag “art_25_DLG_231”		
d_ImputDEF	Num	Presence of identified defendants		
d_MetodiOC	Num	Presence of tag “Metodi_OC”		
d_OrbAsia	Num	Presence of companies identified by Orbis/Asia		
d_OrgCrim	Num	Presence of tag “Org_crim”		
d_pagamafia	Num	Presence of tag “money_to_OC”		
d_Partì_civili	Num	Presence of tag “Parte_civile”		
d_PCammCen	Num	Presence of Central PA institutions as civil parties		
d_PCammDEC	Num	Presence of local PA institutions as civil parties		

Variable	Type	Description	Range	Source
d_PCAz	Num	Presence of companies as civil party		
d_PCPersoneFisiche	Num	Presence of natural persons as civil party		
d_PCStkh	Num	Presence of Stakeholder as civil party		
d_persOC	Num	Presence of tag "person_OC"		
d_RuoloPol	Num	Presence of tag "Ruoli_Politici"		
d_STKHolder	Num	Presence of Stakeholders		
n_ActivOC	Num	Number of tag "Attività_OC"	0-max	
n_Azienda	Num	Number of companies		
n_DirAz	Num	Number of top management roles in private companies		
n_Dirigenti	Num	Number of top management roles		
n_DLG231	Num	Number of tag "art_25_DLG_231"		
n_ImputDEF	Num	Number of identified defendants		
n_MetodiOC	Num	Number of tag "Metodi_OC"		
n_OrbAsia	Num	Number of companies identified by Orbis/Asia		
n_OrgCrim	Num	Number of tag "Org_crim"		
n_pagamafia	Num	Number of tag "money_to_OC"		
n_ParticiVili	Num	Number of tag "Parte_civile"		
n_PCammCen	Num	Number of Central PA institutions as civil parties		
n_PCammDEC	Num	Number of local PA institutions as civil parties		
n_PCAz	Num	Number of companies as civil party		
n_PCident	Num	Number of identified "Parte_civile"		
n_PCPersoneFisiche	Num	Number of natural persons as civil party		
n_PCStkh	Num	Number of Stakeholder as civil party		
n_persOC	Num	Number of natural persons with a role in OC		
n_RuoloPol	Num	Number of tag "Political roles"		
n_STKHolder	Num	Number of Stakeholder		
n_PC_non_id	Num	Number of not identified "Parte_civile"		

A: matrix companies (752 obs)

Variable	Type	Description	Range	Source
Company Name	Alfanum			Judgment
Country	Alfanum			Orbis
Score	Alfanum		A to E	
ID_company	Alfanum	Key variable: Vat/ fiscal code	text	
Year	Num	Year of reference		Asia
CAP	Num	Postal code regis- tered office		
Address	Alfanum	Postal address registered office	text	
Addetti_r	Num	Number of employees	0-max	
ATECO	Num	Economic Area (4 digits)	text	
DENOM_ COMUNE	Alfanum	City		
Anno_cost	Num	Year of establish- ment	num	
Cod_f_giuridica	Num	Legal form code		
Status_fonte_CCI- IAA	Alfanum	Status_source_ CCIAA		
Status_fonte_sta- tistica	Alfanum	Status_statistics_ source		
Settore_istituzi- onale	Alfanum	Institutional_ sector		

It is possible to derive any subset of data from these matrices for statistical analysis.

For example, an analysis could focus on the corruption events described in judgments with no company involved; it is easy to detect the companies in more than one judgment, analyze the illegal events in judgments with one or more political roles, and so on.

Author contributions All authors contributed to the study conception and design. The drafting of the manuscript was coordinated by Maria Francesca Romano.

Funding Open access funding provided by Scuola Superiore Sant'Anna within the CRUI-CARE Agreement. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

- Alcántara Francia, O.A., Nunez-del-Prado, M., Alatrística-Salas, H.: Survey of text mining techniques applied to judicial decisions prediction. *Appl. Sci.* **12**, 10200 (2022). <https://doi.org/10.3390/APPI22010200>
- Bechi A, Rey G.: *L'economia criminale*, Editori Laterza, Roma Bari. (1994)
- Bellucci, P.: La redazione delle sentenze: una responsabilità linguistica elevata, Consiglio Superiore della Magistratura, Formazione Decentrata - Distretto di Roma, Il linguaggio e gli stili delle sentenze, Incontro di studio, Roma, 29 novembre 2004. *Dirit. Form.* **5**(3), 447–465 (2005)
- Bifari, E., Basbrain, A., Mirza, R., Bafail, A., Albaradei, S., Alhalabi, W.: Text mining and machine learning for crime classification: using unstructured narrative court documents in police academic. *Cog Eng* (2024). <https://doi.org/10.1080/23311916.2024.2359850>
- Bolasco, S.: *L'analisi automatica dei testi Fare ricerca con il text mining*. Carocci Editore, Roma (2013)
- Bolasco, S., De Gasperis, G.: TaLTaC 3.0. A multi-level web platform for textual big data in the social sciences. In: Lauro, C. et al. (eds) *Data Science and Social Research - Epistemology, Methods, Technology and Applications* pp. 97–103, Springer (2017)
- Bottone L. (2019), *Analisi automatica del testo: prospettive interdisciplinari per il trattamento e la classificazione dei dati testuali*, Tesi di Diploma, Scuola di Specializzazione in Studi sull'Amministrazione Pubblica-SPISA, Università di Bologna, a.a. 2017/18
- Brunato, D., Venturi, G.: Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici. *Inform. Dirit.* **1**, 111–142 (2014)
- Ceci M, Lesmo L, Mazzei A, Palmirani M, and Radicioni DP (2012) Semantic annotation of legal texts through a framenet-based approach, In: Palmirani, M. et al. (eds.) *AICOL Workshops 2011, LNAI 7639*, pp. 245–255, Springer-Verlag Berlin Heidelberg, (2012)
- Comandè, G.: Multilayered (Accountable) liability for artificial intelligence. In: Lohsse, S., Schulze, R., Staudenmayer, D. (eds.) *Liability for artificial intelligence and the internet of things*, pp. 165–187. Art Publishing Nomos, Oxford (2019)
- Connelly, R., Playford, C.J., Gayle, V., Dibbon, C.: The role of administrative data in the big data revolution in social science research. *Soc. Sci. Res.* **59**, 1–12 (2016)
- Csányi, G.M., Nagy, D., Vági, R., Vadász, J.P., Orosz, T.: Challenges and open problems of legal document anonymization. *Symmetry* **13**, 1490 (2021). <https://doi.org/10.3390/sym13081490>
- Dell'Orletta F., Lenci A., Montemagni S., Marchi S., Pirrelli V., Venturi G.: Acquiring legal ontologies from domain-specific texts. In *Proceeding of LangTech 2008*, Rome (2008)
- Europol (2023) *The Second Quantum Revolution—the impact of quantum computing and quantum technologies on law enforcement*, Europol Innovation Lab observatory report, Publications Office of the European Union, Luxembourg.
- Filtz, E., Navas-Loro, M., Santos, C., Polleres, A., Kirrane, S.: Events matter: extraction of events from court decisions. In: Villata, S., et al. (eds.) *Legal Knowledge and information systems faculty of law*, pp. 33–42. Masaryk University IOS Press (2020)
- Holmes L. (2015) *Corruption: a very short introduction*, Oxford, online edn, Oxford Academic
- Iezzi, D.F., Bertè, R.: Big corpora and text clustering: the Italian accounting jurisdiction case. In: Iezzi, D.F., Mayaffre, D., Misuraca, M. (eds.) *Text analytics: advances and challenges (Studies in classification data analysis, and knowledge organization)*, 1st edn., pp. 77–90. Springer, Cham (2020)
- Lambsdorff, J.G.: Measuring corruption across countries. In: Pieth, M., Søreide, T. (eds.) *Elgar concise encyclopedia of corruption law*, pp. 333–347. Edward Elgar, Cheltenham (2023)
- Lenci, A., Montemagni, S., Pirrelli, V., Venturi, G.: Ontology learning from Italian legal texts. In: Breuker, J., Casanovas, P., Klein, M.C.A., Francesconi, E. (eds.) *Frontiers in artificial intelligence and applications (Online)*, pp. 75–94. IOS Press, Amsterdam (2009)
- Licari D., Romano M.F., Comandè G.: Automatic anonymization of Italian legal textual documents using deep learning. In: Misuraca M, Scepti G, Spano M (eds.) *JADT 2022 Proceedings of the 16th International Conference on Textual Analysis of Textual Data*, ISBN 9791280153319, pp. 552–7 (2022)
- Luscombe, A., Duncan, J., Walby, K.: Jumpstarting the justice disciplines: a computational-qualitative approach to collecting an analyzing text and image data in criminology and criminal justice studies. *J. Crim. Justice Edu.* (2022). <https://doi.org/10.1080/10511253.2022.2027477>
- Luz de Araujo, P.H., De Campos, T.: Topic modelling Brazilian supreme court lawsuits. In: Villata, S., et al. (eds.) *Legal Knowledge and Information Systems Faculty of Law*, pp. 113–122. Masaryk University and IOS Press (2020)
- Mortara Garavelli B. (2001), *Le parole e la giustizia. Divagazioni grammaticali e retoriche su testi giuridici italiani*, Torino, Einaudi.
- Nese, A., Troisi, R.: Corruption among mayors: evidence from Italian court of cassation judgments. *Trends Org Crime* **22**, 298–323 (2019). <https://doi.org/10.1007/s12117-018-9349-4>

- Ondelli, S.: Ordine delle parole nell'italiano delle sentenze: alcune misurazioni su corpora elettronici. *Inform. Dirit.* XXIII **1**, 13–39 (2014)
- Peruginelli, G., Faro, S. (eds.): Knowledge of the law in the big data age. IOS Press (2019)
- Porcedda, M.G., Wall, D.S.: Data science, data crime and the law. In: Mak, V., Taj, E.T.T., Berlee, A. (eds.) *Research Handbook Data Science and Law*, pp. 333–359. Edward Elgar, Cheltenham (2024)
- Rey, G.M. (ed.): La mafia come impresa Analisi del sistema economico criminale e delle politiche di contrasto. Franco Angeli, Milano (2017)
- Romano M.F., De Gasperis G., Pavone P., Bolasco S.: Potenzialità di TaLTaC nella anonimizzazione di sentenze della magistratura, in: Misuraca M, Scepti G, Spano M (eds) JADT 2022 Proceedings of the 16th International Conference on Statistical Analysis of Textual Data, ISBN 9791280153319, pp. 758–63 (2022)
- Salganik, M.: Bit by Bit Social Research in the Digital Age. Princeton University Press, Princeton (2019)
- Strikwerda L, Mensik J, Timmers R.: Data Science and criminal law (2024), In: Mak V., Taj E.T.T., Berlee A. (eds.) *Research Handbook Data Science and Law*, Edward Elgar, pp. 227–249
- Viola L (ed) (2019), Giustizia predittiva e interpretazione della legge con modelli matematici, Atti del Convegno tenutosi presso l'Istituto dell'Enciclopedia Italiana Treccani. ISBN: 9788835348115
- Xu, J.: Research on judicial big data text mining and sentencing prediction model. *J. Phys. Conf. Ser.* (2021). <https://doi.org/10.1088/1742-6596/1883/1/012158>
- Zuliani, A., Aurisicchio, G., Canzonetti, A.: Un'analisi statistica delle sentenze della Corte dei Conti: prime evidenze. *Riv. Trimest. di Dirit. Pubblico* **3**, 673–706 (2009)
- Zuliani, A., Aurisicchio, G., De Benedetto, M., Canzonetti, A., Guagnano, G., Liverani, A., Menichino, P., Rispoli, L., Salvi, S.: La responsabilità per danno erariale alla prova del contenzioso. *Riv. Trimest. di Dirit. Pubblico* **2**, 489–518 (2013)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Maria Francesca Romano¹  · Pasquale Pavone² · Antonella Baldassarini³ · Giuseppe Di Vetta⁴ · Gaetana Morgante⁴

✉ Maria Francesca Romano
mariafrancesca.romano@santannapisa.it

Pasquale Pavone
pasquale.pavone@unipegaso.it

Antonella Baldassarini
anbaldas@istat.it

Giuseppe Di Vetta
giuseppe.divetta@santannapisa.it

Gaetana Morgante
gaetana.morgante@santannapisa.it

¹ Scuola Superiore Sant'Anna, Institute of Economics & L'EmbeDS, Piazza Martiri Della Libertà, 56127 Pisa, Italy

² Università Pegaso, Naples, Italy

³ ISTAT & L'EmbeDS, Rome, Italy

⁴ Scuola Superiore Sant'Anna, DirPolis Institute & L'EmbeDS, Pisa, Italy