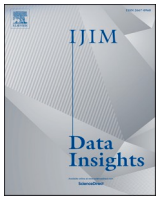
International Journal of Information Management Data Insights 4 (2024) 100247

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei



Textual similarity for legal precedents discovery: Assessing the performance of machine learning techniques in an administrative court

Hugo Mentzingen [a,*], Nuno António [a], Fernando Bacao [a], Marcio Cunha [b]

[a] NOVA Information Management School, Lisbon, Portugal
[b] Ministério Público do Rio Grande do Sul, Rio Grande do Sul, Brazil

ARTICLE INFO

Keywords:
Language processing
Court automation
Case similarity
Imbalanced data

ABSTRACT

The importance of legal precedents in ensuring consistent jurisprudence is undisputed. Particularly in jurisdictions following the Common law, but even in Civil law systems, uniformity in case law requires adherence to precedents. However, with the growing volume of cases, manual identification becomes a bottleneck, prompting the need for automation. Leveraging the capabilities of natural language processing (NLP) and machine learning (ML), our study delves into the potential of automation in identifying similar cases indicative of precedents. Drawing from a unique, substantial dataset of legal cases from an administrative court in Brazil, we extensively evaluated over one hundred combinations of document representations and text vectorizations. Contrary to earlier studies that relied on minimal validation samples, ours employed a statistically significant sample vetted by legal experts. Our findings reveal that models focusing on granular text representations perform optimally, especially when extracting concepts and relations. Notably, while intricate models may not always guarantee superior outcomes, the importance of refining textual features cannot be understated. These findings pave the way for creating efficient decision support systems in judicial contexts and set a direction for future research aiming to integrate technology in legal decision-making.

1. Introduction

Administrative courts specialize in administrative law, a branch of public law focused on public administration (Amaral-Garcia, 2021). It encompasses the set of statutes and legal principles ruling the administration and regulation of government agencies (Cornell University Law School, 2022). In many countries, these courts deal with more cases than criminal or private civil justice, as their role is closely linked to providing public services such as licensing, residence permits, or granting social benefits (Nason, 2018).

The administrative courts' actions positively impact the quality and efficiency of public administration (Batalli & Pepaj, 2022). Many aspects, such as legislative changes, migratory flows, and economic activity, influence their workload. In any way, their intervention must result in prompt and consistent judgments (Gomez, 2021; Rhode, 2004). These institutions, however, deal with limited resources and strive to keep up with the caseload (Popova et al., 2021; Susskind, 2020).

The relevance of data-driven decision-making is burgeoning within all management spheres (Kushwaha et al., 2021). Administrative law courts are no exception, as they are integrating technological advancements to enhance efficiency. Henkel et al. (2017) examined the potential of language technologies in public organizations, concluding that these technologies support the notion that automation, including AI-driven systems, can streamline case processing and decision-making in judicial settings. In the context of vast volumes of data, automation is a crucial factor in increasing the efficiency of a judicial court. It may be characterized as using technology to facilitate or minimize human involvement in case processing. When effectively implemented, it can significantly reduce the duration of lawsuits (Henkel et al., 2017; Velicogna, 2007). Equally important is that courts produce coherent decisions founded on a body of jurisprudence, demonstrating a commitment to certainty and predictability in the law (Mcintyre, 2020).

In a court context, administrative courts included, consistency is typically accomplished through precedents. They serve as the foundation for judges' reasoning. Similar past cases are considered precedents in the Common law system, compelling the outcome of new issues (Rigoni, 2014). Even in jurisdictions that adopt Civil law, courts must consider prior decisions when there is enough uniformity in case law. Typically, when consistent jurisprudence is established, precedents become "soft" law, considered by courts when making decisions (Fon &

* Corresponding author.
E-mail address: hsilva@novaims.unl.pt (H. Mentzingen).

https://doi.org/10.1016/j.jjimei.2024.100247

Available online 15 May 2024

Parisi, 2006).

A judge must know the Court's previous decisions to follow and apply the judicial precedents. Case reports, or "law reports", were established as a system for reproducing judgments of superior courts, and many different series of law reports have been published (Martin, 2008). However, new judgments may take time to be reported. In addition, law reports are usually restricted to the judgments of superior courts, limiting the identification of jurisprudence from same-court decisions. Automating the identification of similar previous cases can fill this gap and improve efficiency and consistency in law courts.

Many methods based on knowledge engineering (KE) have been tested to retrieve similar cases. However, they had scalability as a common constraint as techniques required domain expertise and manual screening. When the number of cases increased, it became unfeasible to identify similar documents. Following the broad utility and relevance of text mining and NLP techniques across various domains (Kumar et al., 2021; Shahade et al., 2023; Zarindast et al., 2021), automated legal precedents retrieval has been progressively associated with NLP and ML. Nevertheless, such applications are barely explored in the literature, and there is still no prominent approach to developing models capable of recognizing precedents (Mentzingen et al., 2023).

This study is rooted in the field of Information Retrieval (IR) and its application to the legal domain, intertwining with the burgeoning field of AI as defined by Dwivedi et al. (2021). AI systems, which emulate cognitive functions typically associated with human intelligence, such as learning, speech, and problem-solving, present a novel dimension to the principles of IR. The theoretical foundation lies in these principles, which involve systematically retrieving relevant information from extensive document collections based on user queries. However, the unique characteristics of legal documents, replete with complex semantics, domain-specific terminology, and intertextual relationships, necessitate a special consideration of IR principles, mainly two: relevance, as it is a subjective concept that can vary from user to user and from context to context, and query formulation, the way users express their information needs through keywords or phrases to retrieve relevant documents.

Under this theoretical lens and leveraging AI capabilities, we evaluate to what extent textual similarity can mimic legal experts' identification of precedents. We employ different ML models to identify analogous cases and compare their results with a statistically significant gold standard. Furthermore, this study recognizes that relevance in the legal domain transcends mere textual similarity. Legal precedents are often considered relevant if they share textual similarities and align with the legal context and principles. As such, this study's unique evaluation set, curated by legal experts, considers the nuanced notion of "relevance" within the legal domain, which is influenced by textual and contextual factors.

The significance of this work lies in its response to pressing challenges facing courts in general, in the context of an escalating caseload, resource constraints, and the need for consistency and efficiency. With administrative courts handling a wide range of cases critical to public services, the timely implementation of court automation is vital to expedite case processing and maintain the quality of public administration. Furthermore, the identification of legal precedents, a cornerstone for judicial decision-making, has traditionally relied on manual curation, leading to delays and limitations in capturing relevant cases. Leveraging recent advancements in NLP and ML, this study bridges the gap in the literature by rigorously exploring automated methods for identifying legal precedents, ultimately recommending a baseline solution to address this contemporary challenge. Consequently, this research is of paramount importance in the current legal landscape.

We used data from the Superintendency of Private Insurance (SUSEP), an independent agency under Brazil's Ministry of Finance with the authority to license and monitor insurance brokers and companies. The agency's duties include prosecuting such economic agents when they violate the rules. SUSEP may start a sanctioning proceeding after an inspection or a complaint and enforce penalties if an infraction has existed. For this purpose, SUSEP has an internal structure to prosecute and judge the supervised agents of the insurance market, acting as an administrative court.

The supervisory jurisdiction of administrative courts varies according to countries and justice systems. In some cases, not only are the acts or omissions of government agencies and authorities subject to review by an administrative court, but the conduct of economic agents under State supervision is subject to scrutiny by such courts. In Brazil, when the interest of administrative authorities is at stake, administrative jurisdiction is given rise (Perlingeiro, 2014).

SUSEP initiated, on average, 822 infraction proceedings per year between 2016 and 2019. Of these, 2471 cases remained undecided in September 2020, when the data was collected. It took, on average, 1113 days to decide. In this context, an immense opportunity exists to improve efficiency by taking advantage of ML. Many documents can be reused, cited, or transformed into templates, providing efficiency to decision-making. It also makes it possible to analyze the arguments and evidence considered in similar infractions, reinforcing the use of jurisprudence and playing a crucial role in favor of consistency in applying laws and regulations.

This study aimed to gauge to what extent ML models can mimic the human notion of similarity in legal practice and be implemented in real-world settings. A second objective was understanding how such models' architecture and text representation influence their results. To this purpose, we tested different combinations of document representation and text embedding methods to identify pairs of similar cases. The similarity scores obtained with the various assemblies were compared with a large set of sample pairs evaluated by legal experts from SUSEP, which was assumed to be the gold standard. This approach also filled a gap observed by Mentzingen et al. (2023): previous studies that used validation samples provided by experts employed a small number of document pairs, insufficient for statistical confirmation.

As an additional advantage, this study evaluates the success of similarity assessment using the infraction's initial document rather than comparing decisions. This method provides the likelihood of similarity in the earliest step of the infraction analysis, allowing the grouping of similar cases to be jointly decided or processed by specialized judges. In the subsequent sections, we will delve into the related literature, methodology, data, and experiments, aligning with this study's theoretical framework.

2. Related work

In assessing judicial precedents based on textual similarity, a challenge arises in balancing legal texts' intricate semantics and domain-specific nuances. Complex semantics, domain-specific terminology, cross-references, and citations to prior case law are some of the features that make similarity learning even more difficult in the legal field (Hu et al., 2022; Mandal et al., 2021; Mentzingen et al., 2023). While the application of computational methods to analyze document similarity in legal contexts has evolved, particularly with neural network-based text embeddings post-2010, the field's momentum has not yet produced a definitive approach (Cho et al., 2014; Mentzingen et al., 2023; Mikolov et al., 2013; Vaswani et al., 2017).

Kumar et al. (2011) underscored the legal terminology's influence on identifying precedents through textual similarity, suggesting that all-term-based methods might not fully grasp the legal concept of similarity. They also noted the potential of citation patterns to reflect the jurisprudential context of judgments. Kulkarni et al. (2017) advanced to more nuanced text embeddings like Doc2Vec but with limited success, indicating the challenges of adapting these models to legal texts.

The experiments on textual similarity are commonly assembled by varying three main components: document representation, text embedding (or text vectorization), and a similarity measurement technique. The latter usually employs a vector distance metric. As an

example, Kumar et al. (2011) used judgments from the Supreme Court of India to evaluate four different models by varying document representations and similarity measures: cosine similarity of all documents' terms, cosine similarity of legal terms identified in the verdicts, a similarity score derived from the number of equal citations in a pair of documents, and a similarity score represented by the number of documents in which the pair under analysis has been cited together. The text embedding technique was the term frequency-inverse document frequency (TF-IDF) (Luhn, 1957; Spärck Jones, 1972). TF-IDF normalizes word occurrence frequency based on the number of cases where each term is present.

The authors compared the models' results with similarity scores given by five experts to 20 pairs of judgments. The second and third models best matched the experts' scores. The authors concluded that "since a legal concept can be explained using various combinations of words, text-based similarity methods fail to satisfy the human notion of similarity." Regarding the third model, the authors inferred that "if two judgments cite the same judgments, both agree to the context of the judgment."

After a six-year gap, Kulkarni et al. (2017) advanced to more nuanced text embeddings like Doc2Vec (Le & Mikolov, 2014) but with limited success, indicating the challenges of adapting these models to legal texts. Zhang et al. (2017) showed that tuning the text processing pipeline and incorporating new algorithms could overcome such limitations. The authors integrated genetic algorithms (GAs) to detect similar cases using k-nearest neighbors (KNNs). They improved document clustering by using GAs to compute the weight coefficients of documents before applying the KNN method.

In their turn, Mandal et al. (2017) reviewed approaches to identify similar legal documents. The authors categorized these works into network-based methods (which utilize the citation network), text-based methods (which utilize the textual content of legal documents), and hybrid methods (based both on the text and the citation network). The authors mention the usual sparsity of citation networks as a limitation of the network-based approach, limiting the method's viability. Regarding the text-based approaches, the authors argued that previous works' methods were predominantly primitive, based mainly on term-frequency inverse document frequency (TF-IDF) for text vectorization.

The study explored various document representations (the whole document, the set of paragraphs, summaries of the document, and the text surrounding a citation). Likewise, other vectorization techniques were evaluated, including topic modeling (LDA) and the neural-network-based approaches Word2Vec (Mikolov et al., 2013) and Doc2Vec. The cosine between two vectors calculated the similarity score, hypothesizing that these methods could effectively capture semantics in legal documents. Mandal et al. (2017) used 47 pairs of cases graded by legal experts on similarity at a scale of 0 to 10, considered the gold standard for assessing such methods. Similarity measures obtained with Doc2Vec embeddings using the whole document correlated most with the expert judgments.

Continuing this trend, Ranera et al. (2019) demonstrated the practical effectiveness of Doc2Vec in a large corpus of Philippine Supreme Court decisions, while Di Nunzio (2020) explored dimensionality reduction in text embeddings to improve precedent retrieval. The novel text embedding technique Top2Vec (Angelov, 2020) was also used to retrieve precedents in combination with BM25 (Robertson et al., 2009) by Arora et al. (2020). These embeddings outperformed BM25-only results.

In parallel, Thenmozhi et al.'s (2017) integration of key elements' extraction in document representation presented a novel avenue for capturing the essence of legal documents. The authors represented documents by extracting their concepts (nouns) or concepts and relations (nouns and verbs). This approach focuses on the theoretical perspective that relevance in legal IR transcends mere textual overlap and may be better represented by extracting central terms.

The Forum for Information Retrieval Evaluation (FIRE) track on Artificial Intelligence for Legal Assistance (AILA) from 2019 to 2020 and the Competition on Legal Information Extraction and Entailment (COLIEE) from 2020 to 2023 offered a showcase for the practical application of precedent retrieval methods. In FIRE AILA (Bhattacharya et al., 2019, 2020a), relevance meant similarity between a legal situation described by the user (query) and the legal database, with successful approaches employing BM25 (Gao et al., 2019; Liu et al., 2020; Zhao et al., 2019) and TextRank (Gao et al., 2019; Mihalcea & Tarau, 2004) for keyword extraction and document retrieval. This competition also introduced techniques like query expansion based on Divergence from Randomness (Leburu-Dingalo et al., 2020).

COLIEE (Goebel et al., 2023; Kim et al., 2023; Rabelo et al., 2021, 2022) competitions included a legal case retrieval task that involved reading a new case Q and extracting supporting cases from the provided case law corpus, hypothesized to support the decision for Q. It is similar to the task studied in this paper, as the comparison occurs between cases. However, the dataset for COLIEE's task was automatically produced by extracting cross-citations between cases, limiting the "relevant" cases to those that have been previously cited. Under this approach, precedents may be disregarded for not being cited due to multiple assessment constraints, ultimately being considered wrong candidates. Despite this limitation, COLIEE findings resonate with our work, revealing the efficacy of traditional models like BM25 and leveraging Transformer-based language models like the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019a) in the legal case retrieval task.

The COLIEE 2020 winner (Westermann et al., 2021) utilized a universal sentence encoder for initial candidate case selection and an SVM for final judgments, emphasizing the utility of combining machine learning classifiers with vector space models. COLIEE 2021 illustrated that traditional IR methods could still trump sophisticated neural approaches, with a team achieving top results using a language model for IR that leveraged strategic text fragment identification (Ma et al., 2021). In contrast, transformer-based methods did not yield superior results despite their sophistication.

In subsequent years, COLIEE highlighted the effectiveness of gradient-boosting classifiers combined with BERT-based embeddings and pre-and post-processing heuristics (Li et al., 2023; Rabelo et al., 2023). The substantially larger dataset of COLIEE differs from the small-sized data employed by most studies. This characteristic entails testing whether the efficacy of Transformers in legal precedent retrieval is contingent upon the volume and specificity of training data available. Our study, grounded in the administrative legal domain of Brazil, contributes to answering this question by evaluating a BERT-based model pre-trained in legal documents with no further training in our dataset, assessing the performance of such a model in a context with limited training data.

Lastly, an investigation of 56 different assemblies, crossing eight document representations and seven text embedding techniques, was performed by Mandal et al. (2021). The methods included representing documents by their parts, such as sentences or paragraphs. At the same time, BERT, Word2Vec, Doc2Vec, Law2Vec (Chalkidis, 2018), and TF-IDF were used for text embedding. The authors reported similar performance between neural network-based embeddings and conventional embeddings. In contrast, BERT produced unsatisfactory results when trained on the dataset of 33,500 Indian Supreme Court documents. Also, conventional vectorization techniques that represent text using bag-of-words, such as TF-IDF, outperformed the more sophisticated context-aware methods, such as Law2Vec and BERT.

A critical analysis of these previous works reveals that there is still no consensus on the predominance of neural network-based models over less sophisticated text embeddings. There is a trend towards using new variants of neural networks, although models that do not consider context may offer better results in specific contexts. Furthermore, document representation techniques, such as extracting concepts and

relations, were loosely explored. In general, little relevance was given to the metrics used and their application to real-world problems when validating the results. Finally, the validation samples, when available, contained few observations, inhibiting the extrapolation of results to the entire data set. Our study fills these gaps and builds upon the literature, using a statistically significant validation sample of expert-evaluated cases from SUSEP to propose a baseline model for precedent identification that can enhance the decision-making process in administrative courts.

3. Material and methods

In the infraction examination process represented in Appendix 1, the proceedings are initiated by inspectors and distributed among legal analysts. These produce an assessment of the case, subsiding the decision-making. Their role is crucial to the trial process since legal analysts are the ones who identify previous similar cases, conduct the formal review, and prepare the case for trial, acting as assistants to the judge. The competence for judging each case depends on the infringed rule: from an administrative judge to the agency's board of directors, composed of five members. The responsibilities in the infraction examination process are summarized in Table 1.

The number of participants in this process makes it even more challenging to consider the existing jurisprudence. Furthermore, analysts can consider previous cases differently, depending on the final decision makers to maintain coherence. A tool based on this study can help stakeholders obtain visibility on the precedents of court judgments.

Compared to other justice courts' cases, these administrative proceedings are more consistent regarding wording and document structure. Although complaints may initiate the proceedings, most of the potential infractions are identified by inspectors during their routine activities, leading to infraction notices of similar terminology. They commonly refer exclusively to infringed legislation and do not contain many references to decisions taken in similar cases. In cases where the proceeding begins with a complaint, the inspectors assess whether there is any potential infraction. Still, an infraction notice is not generated, and the complaint continues in its original state.

3.1. Proposed framework

Language plays a central role in legal practice, where words' precise meanings and contextual application can determine the resemblance between cases and the outcome of decisions (Biel & Kockaert, 2023; Vogel et al., 2017). Although ML models have been extensively employed to measure document similarity, when legal ontology is incorporated, ML models must navigate the complexity of legal language, characterized by its unique terminology, syntax, and semantic nuances. Therefore, ML models may not substantially benefit from improvements made in general language understanding (Frankenreiter & Nyarko, 2022).

This study assumes that by transforming the infraction's primary documents into vectors and applying similarity measures between these text representations, it is possible to identify pairs of cases that satisfy legal experts' notion of similarity. Ultimately, the influence of machine learning models' training and internal construction in accurately capturing the subtleties of language as understood by experts is also assessed. To validate this hypothesis, we compare the models' outputs with a statistically significant gold standard and a baseline model. We utilized a representative dataset of administrative infractions to investigate combinations resulting from five document representations, five text embedding models, and a similarity measurement utilizing cosine similarity plus the BM25 ranking function.

Therefore, the experimental design in Fig. 1 combines document representations, vectorization techniques, and cosine similarity measurement. An exception was made to the BM25 scoring function that, due to its characteristics, only allowed testing different document representations. The source code used to run the experiments is shared on GitHub through the following link: https://github.com/hugosaisse/textualSimilarityPrecedents.

The choice of techniques followed what was identified in the literature review, which, in summary, indicated promising results with BM25 and Top2Vec and less encouraging results with BERT. Furthermore, the related works suggest that the comparison between text embeddings based on the frequency of terms and neural networks should be further investigated.

Table 1
Roles and responsibilities in the infraction examination process.

Role	Responsibilities
Inspector	Assess potential infractions during their routine activities and file an infraction notice as appropriate, check whether a complaint corresponds to a possible infraction, and register the infraction notice or complaint in the penalties system.
Defendant	Present defense.
Legal Analyst	Conduct a formal review of cases, notify the Defendant to present a defense, and prepare the case for trial.
Administrative Judge	Issues the final decision in cases whose infraction is within their competence.
Board of Directors	Issues the final decision in cases whose infraction is within its competence.

3.2. Data collection

The sections below detail how data was collected from SUSEP's information system (IS), how interviews with legal experts were conducted, the choice of the sample size, and the evaluation metrics employed in this study.

3.2.1. Legal documents

The legal documents utilized in this study were obtained from SUSEP's administrative process computer system, with decision dates ranging from June 2016 to August 2020. This electronic document management system (DMS) stores the text documents of the proceedings in chronological sequence. We retrieved all born-digital administrative proceedings initiated after 2016, when the DMS was initially implemented so that documents in HTML or PDF format could be appropriately parsed. To avoid comparing recent cases to those where the law is not applicable, the infraction described in the document must be within the rules in force during the dataset date range.

The resulting dataset contained 1109 infractions, including their primary document's text extract, from which 155 (14.0 %) originated from a customer complaint and 954 (86.0 %) originated from infraction notices.

Each proceeding in the DMS may have had one or more potential infractions. Regularly, all infractions are described in a single complaint or infraction notice. From now on, this is referred to as the primary document. Since the infraction was the unit of analysis, it was necessary to split the text related to each. Hence, the corresponding part of the primary document was extracted. Some proceedings also contained drafts and rectifications. Consequently, these cases had two or more copies of the same documents. As a rule, the latest versions were believed to be final.

BeautifulSoup (Richardson, 2007) extracted text from the HTML files and pdfminer.six (Shinyama et al., 2019) served the same purpose for PDF files. Different parsers were built for complaints and infraction notices, in which regular expressions helped remove prologues and epilogues and extract the core text for each infraction. Fig. 2 shows an example of an infraction notice stored as a PDF file in its original idiom, and Fig. 3 contains a free English translation of the text.

3.2.2. Legal expert evaluation and the gold standard

Each assembly, a combination of document representation and text embedding, had as output a list of documents sorted by similarity.

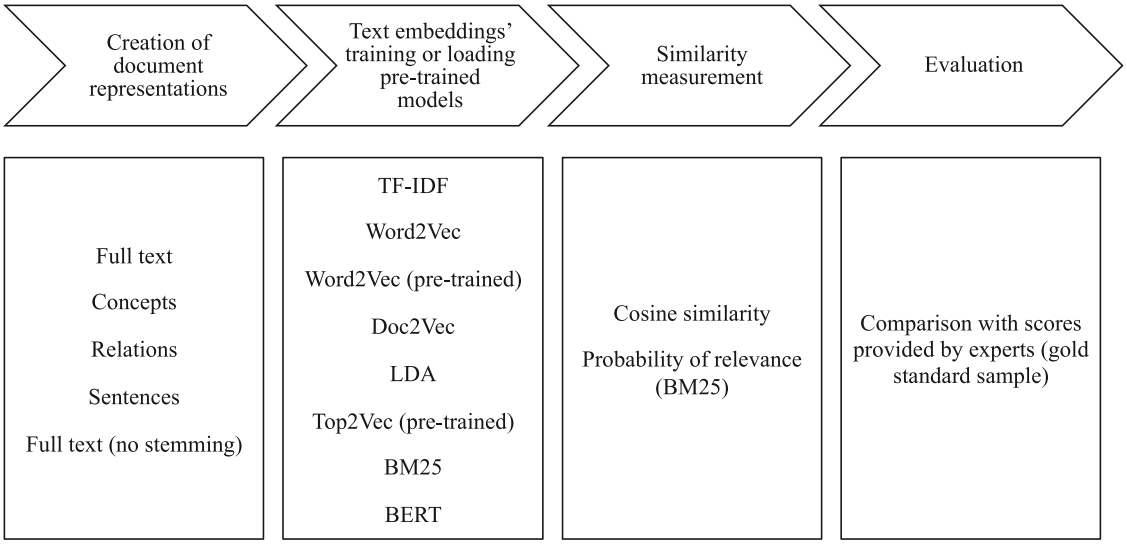


Fig. 1. Experimental setup employed in this study.

Effectively evaluating each assembly's results regarding document relevance demanded comparing the similarity scores the models gave to an assessment made by legal experts. For this reason, a sample of infractions was randomly selected from the 1109 dataset records. These samples were combined into pairs of infractions, and three legal experts from SUSEP evaluated and scored the pair in terms of relevance.

We used Cochran's equation for large populations to ensure the sample was representative of the dataset (Cochran, 1977). In our case, the population comprises 614,386 possible combinations of the 1109 infractions. Since we did not know the population proportion, i.e., the percentage of pairs with a level of similarity different from zero, we assumed 50 % as a rule of thumb. For a confidence level of 95 % and a confidence interval of 3 %, the needed sample size was 1065 pairs. By selecting 50 random cases, we formed 1225 combinations (C_2^{50}). Later, we observed from the experts' evaluations that the population proportion is about 10 %, confirming the utility of the results, with a confidence level of 99 % and an error margin of 2.21 %.

The legal experts were given a spreadsheet containing the 1225 pairs of cases with their text as extracted from SUSEP's DMS. They worked independently and could not access the other experts' evaluations. Each infraction pair received a score from 0 to 5, where 0 represented no similarity, and 5 represented the highest level of similarity. Given a pair (A, B) of cases, the experts were asked to score to what extent case A was relevant to the analysis of case B. Therefore, the concept under evaluation was the utility of infraction B as a legal precedent during the analysis of infraction A.

Moreover, the similarity was understood as reciprocal. Each expert evaluated all pairs, and the mean of the expert's scores was considered the gold standard. Table 2 presents the number of pairs comprised in each similarity level, the mean, and the standard deviation of the scores for each expert, and Table 3 presents the number of pairs comprised in each similarity interval when considering the experts' scores mean.

We observe that the experts' scores distribution has very low variability (most results are equal to zero). Considering a binary classification problem, in which the positive case represents a document pair having a level of similarity different from zero, we observe a very low Prevalence (Marshall, 2005) of the positive class. As a result, pairs with similarities different from zero can be seen as outliers. This situation is mainly derived from the fact that there are 94 different infringed regulations in a relatively small dataset. Therefore, it is expected that many cases will not be similar.

The variation identified among the scores' means and standard deviations denotes the high variability in humans' perception of similarity. This aspect is also recognizable in Fig. 4, comparing the scores of all experts for the same case pairs. Assessing the human cognition of similarity and understanding the parameters that influence this cognition is not among the objectives of this work. Meanwhile, this finding reveals the importance of comparing the various experimental setups' results with the individual experts' assessments, not only the scores' mean. This approach made it possible to assess the ability to mimic the notion of similarity for each individual, i.e., simulating the case of an individual receiving a suggestion for a document similar to a case under analysis.

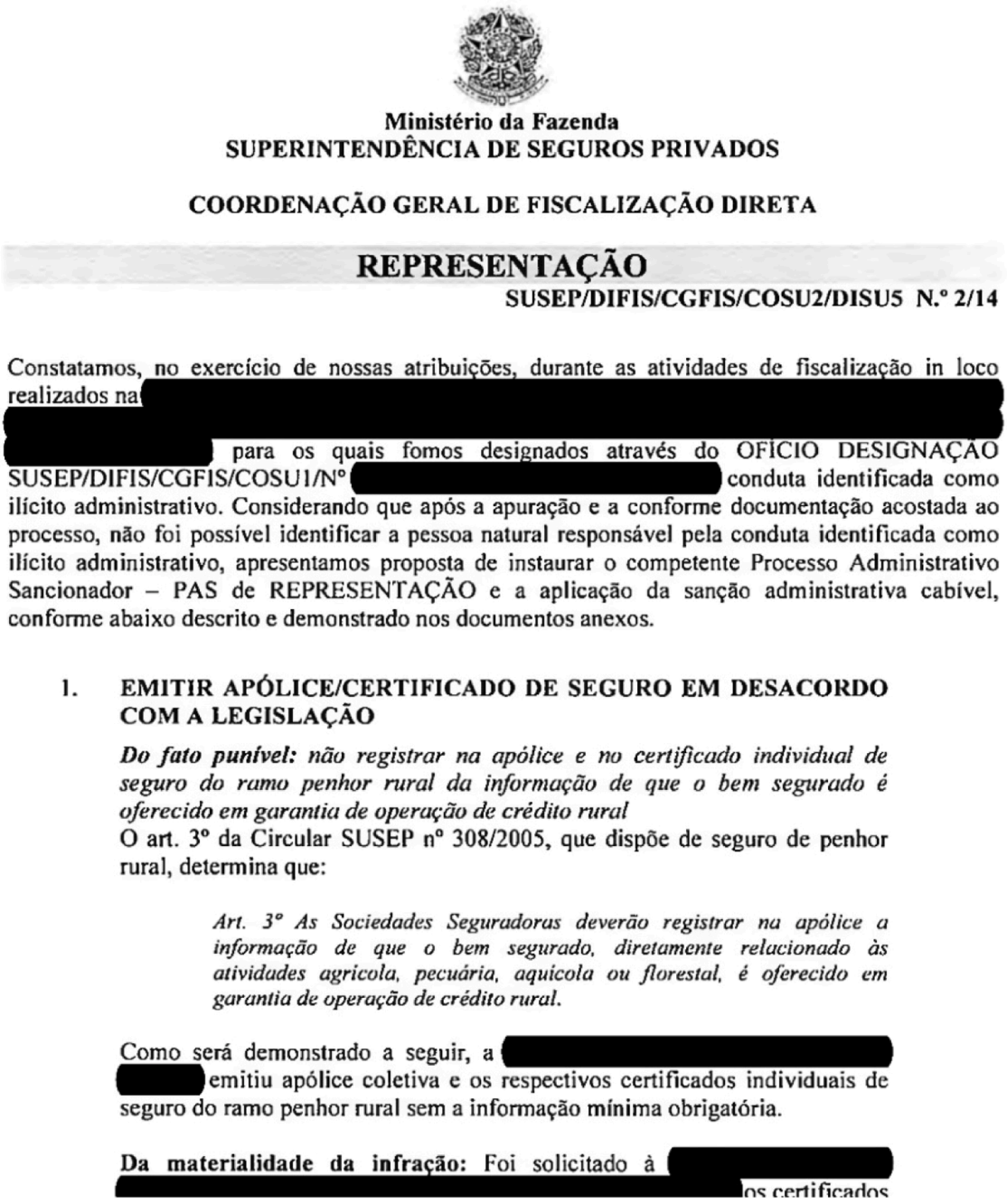
Furthermore, experts' scores being discrete implies that they can only take on a limited number of integer values. This limitation leads to a non-normal distribution of scores where multiple items receive the same score. Notably, as many cases are deemed irrelevant (scored as zero), it leads to a non-uniform variance across the range of machine scores. On the other hand, model-generated scores being continuous means that they can take on any value within the similarity range. These scores will likely have a different distribution and potentially a broader variance than the expert scores.

It denotes heteroskedasticity in the experts' scores, which challenges using certain statistical measures, such as the correlation coefficient (Greene, 2017; Wilcox, 2015). Therefore, the performance metrics must consider this dataset's characteristics to ensure the study's validity. As explained in the next Section, in the operational context of identifying legal precedents, Recall emerges as a more appropriate metric.

3.2.3. *Performance evaluation metrics*

From a business perspective, ensuring that the most relevant documents are retrieved when a new case is under analysis is essential. To this purpose, first, the experiment was transformed into a two-class classification problem, with the positive class meaning that a document pair is 'similar' and the negative meaning 'not similar'. A document pair with a non-zero similarity in the gold standard was classified as 'similar'. Otherwise, it was labeled as 'not similar'. The experiment objective is to retrieve the highest possible number of 'similar' documents for a legal case document under analysis. Hence, Recall is the most critical metric in this business problem, i.e., the fraction of relevant documents that are successfully retrieved or the number of appropriately retrieved infractions divided by the number of results that should have been returned.

Also, the Recall sidesteps the non-uniform variance (heteroskedasticity) issue by concentrating solely on the model's capacity to identify all relevant precedents (Manning et al., 2008). This focus on Recall aligns with the imperative need in legal analysis to ensure no pertinent case is overlooked, which could be critical for the outcome of a



Ministério da Fazenda
SUPERINTENDÊNCIA DE SEGUROS PRIVADOS

COORDENAÇÃO GERAL DE FISCALIZAÇÃO DIRETA

REPRESENTAÇÃO

SUSEP/DIFIS/CGFIS/COSU2/DISU5 N.º 2/14

Constatamos, no exercício de nossas atribuições, durante as atividades de fiscalização in loco realizados na [redacted] para os quais fomos designados através do OFÍCIO DESIGNAÇÃO SUSEP/DIFIS/CGFIS/COSU1/Nº [redacted] conduta identificada como ilícito administrativo. Considerando que após a apuração e a conforme documentação acostada ao processo, não foi possível identificar a pessoa natural responsável pela conduta identificada como ilícito administrativo, apresentamos proposta de instaurar o competente Processo Administrativo Sancionador – PAS de REPRESENTAÇÃO e a aplicação da sanção administrativa cabível, conforme abaixo descrito e demonstrado nos documentos anexos.

1. **EMITIR APÓLICE/CERTIFICADO DE SEGURO EM DESACORDO COM A LEGISLAÇÃO**

Do fato punível: *não registrar na apólice e no certificado individual de seguro do ramo penhor rural da informação de que o bem segurado é oferecido em garantia de operação de crédito rural*

O art. 3º da Circular SUSEP nº 308/2005, que dispõe de seguro de penhor rural, determina que:

> *Art. 3º As Sociedades Seguradoras deverão registrar na apólice a informação de que o bem segurado, diretamente relacionado às atividades agrícola, pecuária, aquícola ou florestal, é oferecido em garantia de operação de crédito rural.*

Como será demonstrado a seguir, a [redacted] emitiu apólice coletiva e os respectivos certificados individuais de seguro do ramo penhor rural sem a informação mínima obrigatória.

Da materialidade da infração: Foi solicitado à [redacted] os certificados

Fig. 2. PDF Infraction notice sample. This document, written in Portuguese, describes the findings of an inspection team and may contain one or more potential infractions.

new case under consideration (Roitblat et al., 2010). Therefore, the study will adopt Recall, specifically the mean Average Recall (mAR), to evaluate the models' performance, ensuring the most relevant documents are retrieved without the misleading influence of heteroskedasticity. In such situations, using the correlation coefficient to measure the strength of the relationship between the gold standard and the models' outputs can be misleading.

Precision, or the Positive Predictive Value (PPV), could be a secondary metric since it measures the fraction of relevant documents among the retrieved documents. However, it is worth considering that Prevalence impacts the PPV of experiments. As the Prevalence increases, the PPV also increases. Likewise, as the Prevalence decreases, the PPV decreases. In our dataset, we observe a very low Prevalence of the positive class, which can distort the evaluation of an assembly. For the same reason, accuracy is also not a good metric for evaluating the models' performance.

Consequently, this study adopts the mean Average Recall (mAR) as the primary evaluation metric, defined as the Average Recall (AR) mean. mAR is defined similarly to the definition of Schröder et al. (2011) for the mean Average Precision (mAP). In this study, however, the number of users was substituted by the number of cases.

For a real-world situation, the recommendation threshold is the number of predicted-similar suggestions presented to a legal analyst when a case is under analysis. Hence, we define the AR @ k (AR 'at' k) as the Recall averaged over all fifty documents in the gold standard when the number of recommendations is set to k. The mAR and AR @ k used in this study are presented in Eq. (1) and Eq. (2), where T corresponds to the number of cases in the evaluation dataset, and N corresponds to the maximum number of recommendations.

Ministry of Finance
SUPERINTENDENCY OF PRIVATE INSURANCE
GENERAL COORDINATION OF DIRECT OVERSIGHT
INFRACTION NOTICE
SUSEP/DIFIS/CGFIS/COSU2/DISU5 NO. 2/14

In the exercise of our attributions, we verified during the on-site inspection activities carried out at (...) for which we were assigned through the DESIGNATION TERM SUSEP/DIFIS/CGFIS/COSU1/ NO. (…) conduct identified as an administrative offense. Considering that after verification and according to the documentation attached to the process, it was not possible to identify the natural person responsible for the conduct identified as an administrative offense, we present a proposal to file the competent Sanctioning Administrative Proceeding - REPRESENTATION PAS - and the application of the appropriate administrative sanction, as described below and shown in the attached documents.

1. **ISSUING A POLICY/INSURANCE CERTIFICATE IN DISAGREEMENT WITH THE LEGISLATION**

The punishable fact: *not registering in the policy and the individual insurance certificate of rural pledge type, the information that the insured property is offered as a guarantee of a rural credit operation.*

The article no. 3 of Circular SUSEP No. 308/2005, which addresses the rural pledge insurance, determines that:

> *Art. 3 The Insurance Companies must register in the policy that the insured property, directly related to agricultural, livestock, aquaculture, or forestry activities, is offered in a guarantee of rural credit operation.*

As will be shown below, the (…) issued a collective policy and the respective individual certificates of insurance of the rural pledge type without the mandatory minimum information. On the materiality of the infraction: It was requested to (…) the certificates (…).

Fig. 3. Infraction notice (English translation).

$$AR\ @\ k = \frac{\sum_{i=1}^{T}(Recall\ @\ k)_i}{T},\ T = 50,\ k \leq T \tag{1}$$

Eq. (1) Average Recall @ k

$$mAR = \frac{\sum_{k=1}^{N} AR\ @\ k}{N},\ N = 50 \tag{2}$$

Eq. (2) Mean Average Recall

Due to the limited human capacity to analyze presented recommendations, Recall@5 is used as a second evaluation metric. It is defined as the proportion of similar documents in the top 5 recommendations (Aggarwal, 2016). We then calculate the average Recall@5 over all documents in the gold standard (Eq. (3)). In some cases, the calculation of Recall can cause a division by zero, happening if there are no relevant documents for an infraction in the gold standard sample. We set Recall to 1 for these cases since no relevant item is left unidentified in the top results.

$$Recall@5 = \frac{\sum_{i=1}^{T}\left(\frac{\#\ of\ relevant\ documents\ in\ top5}{\#\ of\ relevant\ documents}\right)_i}{T},\ T = 50 \tag{3}$$

Eq. (3) Recall@5

Table 2
Similarity score distribution according to legal experts.

	Number of samples		
Similarity score	Expert 1	Expert 2	Expert 3
0	1124	1094	1174
1	36	29	8
2	12	21	7
3	12	23	9
4	39	17	12
5	2	41	15
Mean	0.337	0.140	0.214
Standard Deviation	1.099	0.734	0.813

Table 3
Similarity score distribution when considering the experts' scores mean.

Similarity score	Number of samples experts mean
0	1058
$0 < \text{score} \leq 1$	97
$1 < \text{score} \leq 2$	21
$2 < \text{score} \leq 3$	9
$3 < \text{score} \leq 4$	17
$4 < \text{score} \leq 5$	23
Mean	0.230
Standard Deviation	0.815

3.3. Document representation

As mentioned in Section 3.1, five variants for document representation were used in this study. The first document representation was also the most basic, composed of the full stemmed text. From the study of Thenmozhi et al. (2017), which described encouraging results by extracting concepts and relations from the text, we used POS tagging to generate the second and the third representations by extracting, respectively, the nouns (concepts) and nouns and verbs (concepts and relations) from the whole document. In both cases, the resulting text was later stemmed.

This work also considered the results Mandal et al. (2021) achieved by seeking legal issues into different levels of granularity, i.e., splitting the documents into paragraphs and sentences, obtaining document summaries, and extracting catchphrases from the text. Nevertheless, because part of our dataset was extracted from PDF files, we lacked the paragraph markers to partition all documents into their paragraphs. On the other hand, it was possible to break documents into sentences using the existing dots. To avoid the improper break of sentences when abbreviation dots were found, we ignored dots in a set of standard abbreviations for the Portuguese language (art., arts., s.a., sa., doc., docs., fl., fls., dr., dra., drs., cia., cias.) as well as consecutive dots that could represent suspension points. As in the other representations, the final text was stemmed.

Finally, the fifth document representation was the full text without stemming, employed for evaluating pre-trained Word2Vec or Top2Vec models. In Mandal et al. (2021), the representations based on document summaries or the extraction of catchphrases showed the best results when associated with the TF-IDF vectorization method but still had lower performance than the full-text representation. For this reason, these specific representations were not implemented in this work.

3.4. Text vectorization

After obtaining one corpus for each document representation, different methods were applied to convert text into vectors. The first was the TF-IDF, in which hyperparameter configurations were tested with unigrams plus bigrams and bigrams plus trigrams. We ignored terms with a document frequency higher than 80 % of the corpus or appearing in only one document. Besides the mentioned n-gram intervals, one TF-IDF vectorizer was trained on each of the four stemmed corpora. The vectorizer trained on full documents was evaluated on all corpora. The other three vectorizers were used to obtain document embeddings from

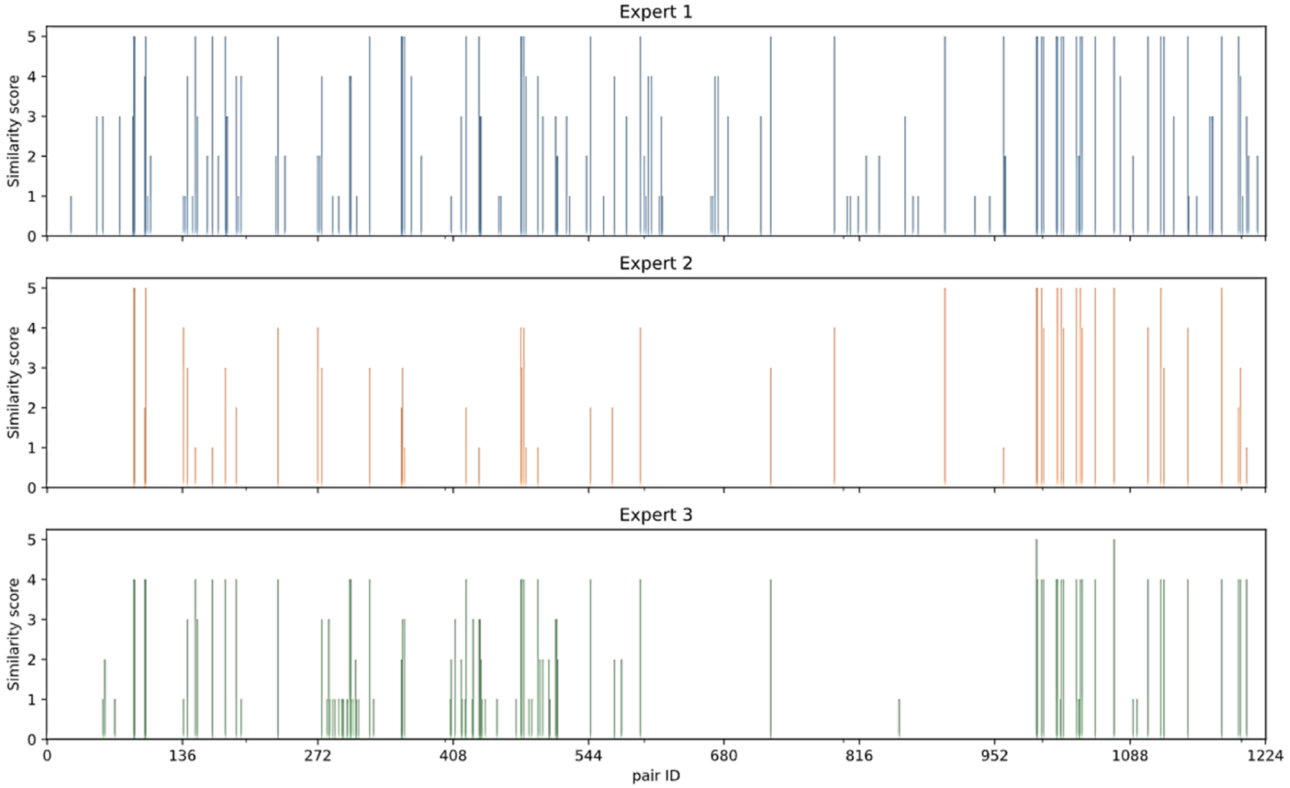


Fig. 4. Expert scores distribution.

their respective corpus. Thus, the fourteen configurations in Fig. 5 were experimented with using TF-IDF vectorizers. For the sake of comparison to the prevalent method for document similarity retrieval, this study assumes the TF-IDF model created with the full stemmed documents using unigrams and bigrams as the baseline.

The Word2Vec technique, based on word embeddings, was the second method used to create document representation vectors. Considering that Word2Vec generates embeddings for each word, the vectors for the text segments were obtained by the mean of the constituent word vectors, weighted by the TF-IDF score of the word. First, pre-trained Word2Vec models for the Portuguese language were obtained from the repository maintained by Hartmann et al. (2017) on the Interinstitutional Nucleus of Computational Linguistics (NILC-USP) of São Paulo University. The continuous bag-of-words (CBOW) and Skip-gram methods with the available embedding sizes of 100 and 300 were employed. Then, following the same methods and embedding sizes, other Word2Vec models were trained on each of the four corpora. Again, the models trained on the whole corpus were applied to all corpora. The other models were applied to the respective corpora, resulting in thirty-two assemblies in Fig. 6.

Next, the Doc2Vec technique was implemented to transform each document into vectors of 100, 200, and 300 dimensions. By representing the whole text into a vector, unlike the Word2Vec models, Doc2Vec does not require the combination of embeddings into a final vector. Following the previous model families, twenty-one assemblies resulted from training the model on all corpora, applying the full-corpus-trained model to all corpora and the other models to their respective corpora (Fig. 7).

LDA was the fourth model family tested in this study. Each corpus document is modeled as a finite combination of an underlying set of topics in this model family, where a distribution across words characterizes each topic. The probabilities of belonging to topics explicitly represent a document in text modeling. Following the previous assemblies, we trained LDA over the full text stemmed corpus with 10, 20, 40, and 80 topics and retrieved vectors with these dimensions for the different document representations. Later, we trained LDA models with the same topic range over the three remaining corpora and retrieved vectors for documents in the same corpora. The twenty-eight LDA assemblies are represented in Fig. 8.

Following the promising results by Bhattacharya et al. (2020), we experimented with the BM25 model family by employing the BM25+ algorithm. This implementation addresses one deficiency of the standard BM25: the component of term frequency normalization by document length is not appropriately lower-bounded. As a result, long documents that match the query term can often be unfairly scored by BM25 as having similar relevancy to shorter documents that do not contain the query term at all (Lv & Zhai, 2011). Fig. 9 illustrates the experimental setup for this model's family.

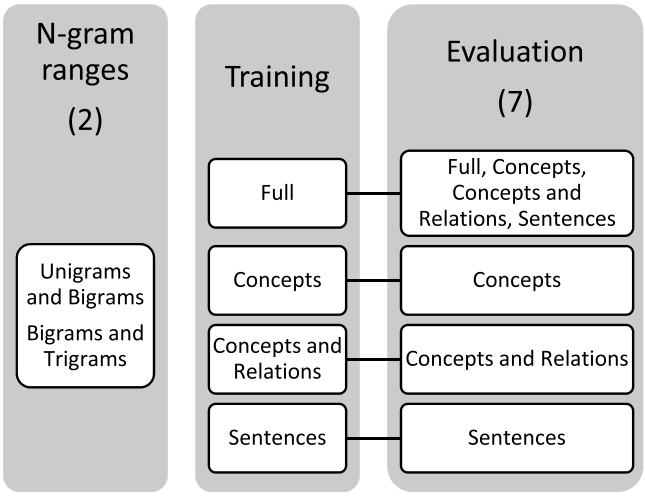


Fig. 5. The fourteen TF-IDF experimental configurations.

The sixth model used was Top2Vec, adopting the Universal Sentence Encoder Multilingual (Yang et al., 2019) as the vectorization model. This encoder is an extension of the Universal Sentence Encoder pre-trained on multiple tasks across languages, including Portuguese. After text vectorization, Top2Vec performs dimensionality reduction using UMAP (McInnes et al., 2018). Later, it creates clusters of documents with HDBSCAN (McInnes et al., 2017), also called topics. Keywords for each document are determined based on the underlying topic. The similarity to the keyword vectors determines the similarity between a query and existing documents. Fig. 10 shows we evaluated the model using the full corpus without stemming.

In addition to the traditional text vectorization methodologies, we explored the capabilities of a BERT-based model specialized in the legal domain of the Portuguese language, known as 'legal-bert-base-cased-ptbr' (Domingues, 2022). This model, rooted in the architecture proposed by Devlin et al. (2019b), has been pre-trained with diverse legal documents from the Brazilian Supreme Court. With its 126 million parameters and training on roughly 65,000 examples of varied legal documents written in Brazilian Portuguese, this model has shown a promising evaluation loss of 0.473 for predicting masked words, indicating its potential for high accuracy in identifying relevant information within the legal corpus. This model's extensive training in Brazilian legal documents makes it a prime candidate for our experimental setup, aiming to discover its efficacy compared to the other vectorization methods we have tested. Fig. 11 will detail the experimental setup for this BERT-based model, showcasing its integration into our methodological framework.

3.5. *Document similarity measurement*

After transforming the document representations into vectors through each previously mentioned method, we applied cosine distance to evaluate document similarity. It is formally defined as the dot product between two vectors divided by the product of their magnitudes. It measures the cosine of the angle between these vectors. When applied to document similarity, cosine similarity is typically preferred over Euclidean distance because even if two similar documents are far apart when compared by the Euclidean distance (due to the difference in the size of the documents), they may still have similar orientations. In this sense, the smaller the angle, the higher the cosine similarity.

As mentioned in Section 3.1, BM25 is a ranking function whose score depends on each query and document. We used the latter as the query under analysis to rank documents in terms of similarity to a single document. Therefore, computing cosine similarity between vectors was not the case for BM25, and the output score could be directly used to retrieve the most similar cases.

As previously explained, this study evaluated the performance of models trained on representation-specific corpora, resulting in 106 assemblies to identify similarities among documents using the matching representation. This approach builds on the assemblies used by Mandal et al. (2021). Using the full corpus, the authors trained the model once for each vectorization technique in their study. In this study, we evaluate the performance of models trained both on the full-text corpus and the representation-specific corpora.

4. Results

As previously stated in Section 3.3, the similarity measures obtained for the 1225 pairs formed from a sample of 50 randomly selected cases were compared with a gold standard composed of similarity scores awarded by legal experts. The models were compared by employing the metrics to the context, i.e., the mean Average Recall (mAR) as the

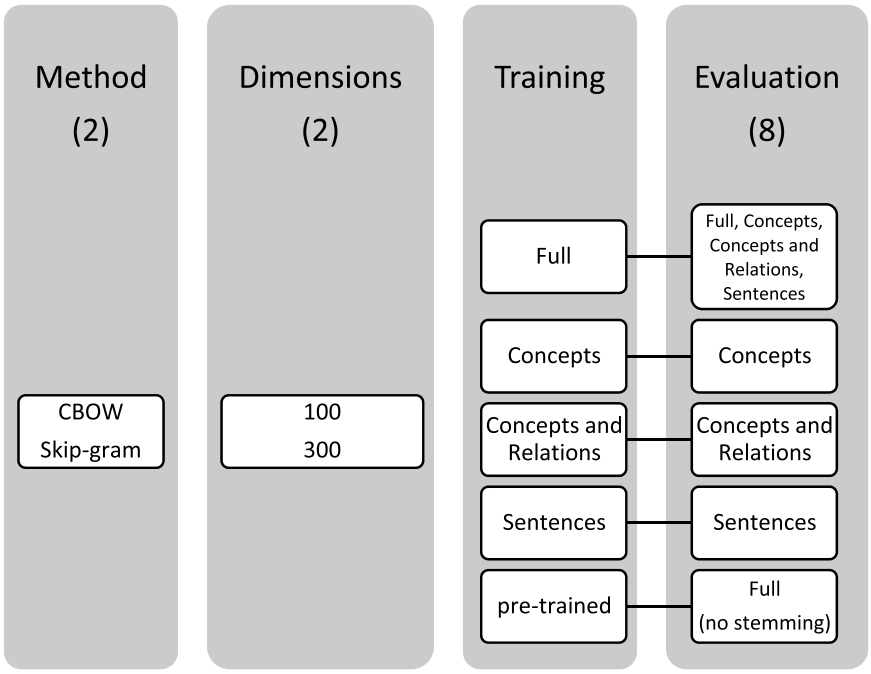


Fig. 6. The thirty-two Word2Vec experimental configurations.

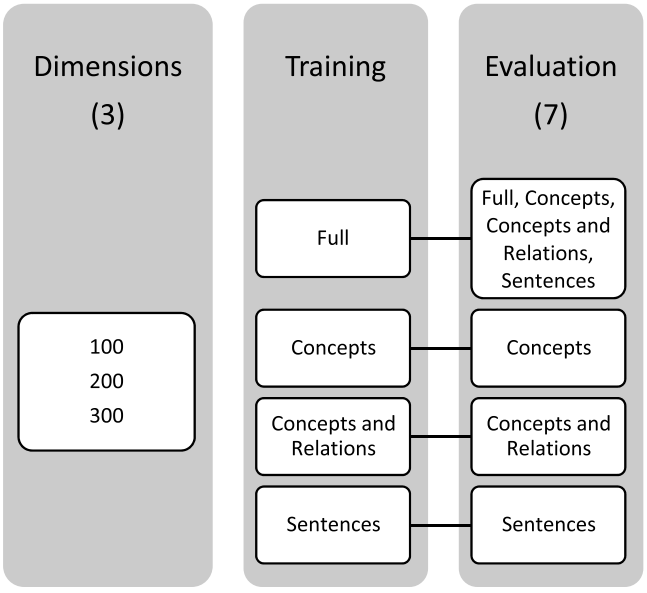


Fig. 7. The twenty-one Doc2Vec experimental configurations.

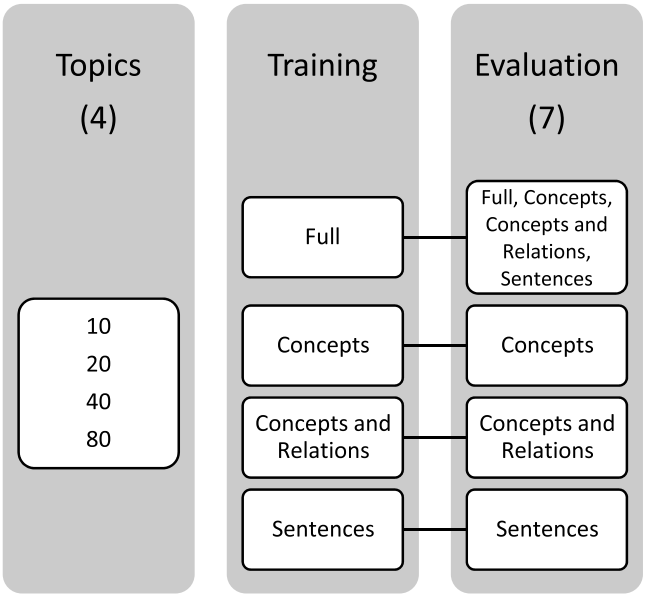


Fig. 8. The twenty-eight LDA experimental configurations.

primary metric and Recall@5 as a secondary metric. Table 4 presents the models that returned one of the three best results for any combination between metric (mAR and Recall@5) and assessment set (experts one, two, and three, and the experts' mean) .[1] The complete results set is presented in Appendix 2.

Looking at the scores, the Experts attributed to pairs of documents, similarities greater than zero become more numerous when the average from individual evaluations is considered. The Recall, in turn, represents the fraction of relevant documents successfully retrieved, while the documents considered similar are those with a score higher than one. Hence, the mAR and Recall@5 results obtained on the experts' mean assessment tend to be lower than those observed in the experts' individual evaluations.

5. Discussion

This study examines the effectiveness of ML techniques in detecting legal precedents. Within the landscape of existing research on ML applications in legal document analysis, studies such as those by Kulkarni et al. (2017), Ranera et al. (2019), and Mandal et al. (2017, 2021) have paved the way for understanding the performance of various text embedding models in legal contexts. This study was conducted through

[1] The three best results for each combination are marked in bold. The Full, Concepts, Concepts & Relations and Sentences corpora are represented by the F, C, C&R, and S labels. For continuous bag-of-words and skip-gram methods, we used the CBOW and SG tags.

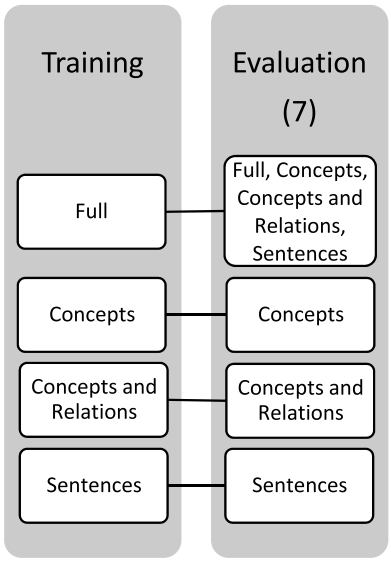


Fig. 9. The seven BM25 experimental configurations.

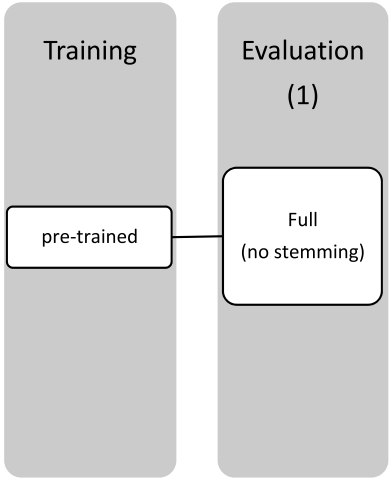


Fig. 10. The Top2Vec experimental setup.

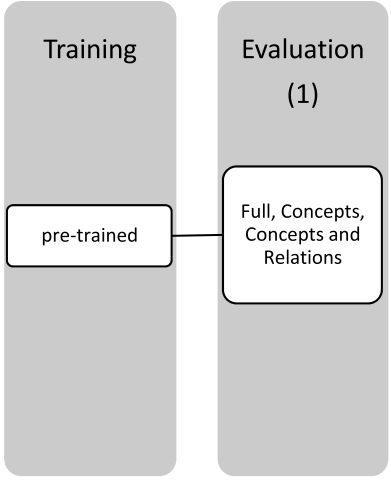


Fig. 11. The BERT-based model experimental setup.

an empirical evaluation using data from Brazilian administrative law and adds new dimensions by utilizing a significantly large and expert-evaluated dataset, thus providing a robust validation of these text embedding techniques. Additionally, our findings enrich the discourse on the effectiveness of neural network-based models versus less sophisticated embeddings, a topic that lacks consensus, particularly in specialized domains like law.

The findings support the initial hypothesis that ML techniques can accurately reflect the human concept of similarity when subject to legal ontology. Several models consistently achieve a mean Average Recall (mAR) greater than 70 %. Beyond this initial success, it is vital to explore the behavior of specific models and the reasons behind it. Our results are juxtaposed with the findings of Mandal et al. (2021), who highlighted the challenges of using Doc2Vec and BERT in legal texts. While they observed limited success, our application of similar methodologies under different conditions provides further insights into the variables that affect performance. This analysis is essential to highlight these models' significance and potential impacts when integrated into IS within the socio-technical context of legal informatics (Kar et al., 2023; Struijk et al., 2022).

Considering the mAR obtained from the experts' mean scores, the Word2Vec models outperformed the baseline and, together with the TF-IDF models, obtained the best Recall scores overall. These models use the most granular text representation (words) among the tested setups. It suggests that combining a relatively small corpus, business-specific words, and ordinary words with particular meanings for the context can favour models that rely on more granular representations. Satisfactory results seem achievable using bag-of-words, such as in TF-IDF, or adopting Word2Vec's distributional hypothesis. The distributional hypothesis suggests that the more semantically similar two words are, the more distributionally similar they will be in turn, and thus, the more they will tend to occur in similar linguistic contexts.

Our findings also indicate that while BERT is a powerful tool in NLP, it does not necessarily outperform traditional models in the specific context of legal document relevance assessments. This finding suggests that the complexity and advanced capabilities of BERT do not automatically equate to higher performance in specialized domains such as law, although the model was pre-trained on a legal corpus. This fact implies that BERT may benefit from additional fine-tuning tailored to specific domains to enhance its performance. The extraction of Concepts or Concepts and Relations also contributed to the degradation of BERT's performance (i.e., BERT achieved its best performance when the full text was employed).

It is also significant that the pre-trained Word2Vec or Top2Vec models evaluated using the complete corpus without stemming did not obtain good results, which is why they are not presented in Table 4. Two factors can explain this phenomenon: the first is that stemming considerably reduces the word vector space, making it easier to identify similar word stems. The second factor is that as the words in the pre-trained versions are analyzed in context, pre-training a model without financial-sector-related texts may not capture the use of certain words with context-specific meanings. Assuming a general-purpose corpus that does not match the domain's vocabulary (the exact words and words in the same senses) cannot be overcome by presenting more data, which would move the word vectors towards standard rather than domain-specific meanings.

The Word2Vec models trained with 100 dimensions stood out when considering the similarity average among experts. When considering the individual evaluations, various Word2Vec configurations obtained good results, followed by BERT, which achieved outstanding performance for Expert 3. The Word2Vec results suggest that the choice of this model prevails over the configuration concerning the corpus, the number of dimensions, and the algorithm. The noticeable variation in BERT's performance across different experts could be due to the subjective nature of relevance, through which Expert 3's understanding of relevance might align better with the way BERT's deep neural network captured

Table 4
Experimental setups with top three scores by metric and expert scores set. The baseline setup is highlighted.

							Experts Mean		Expert 1		Expert 2		Expert 3	
Model	**Training Corpus**	**Evaluation Corpus**	**Topics**	**Dimensions**	**Algorithm**	**N–Grams**	**mAR**	**Recall@5**	**mAR**	**Recall@5**	**mAR**	**Recall@5**	**mAR**	**Recall@5**
W2V	F	C & R	–	100	CBOW	–	**0.705**	0.389	0.746	0.456	0.867	0.764	0.750	0.523
W2V	C	C	–	100	CBOW	–	**0.705**	0.357	0.743	0.431	0.856	0.710	**0.772**	0.526
W2V	C	C	–	100	SG	–	**0.701**	**0.393**	0.747	**0.486**	0.862	0.721	0.753	0.511
W2V	F	F	–	100	CBOW	–	0.697	0.330	**0.761**	0.420	0.876	0.684	0.725	0.481
W2V	C	C	–	300	SG	–	0.696	0.391	0.741	**0.480**	0.858	**0.770**	0.761	0.533
W2V	F	F	–	300	CBOW	–	0.696	0.357	**0.755**	0.438	**0.878**	0.721	0.724	0.509
TF–IDF	F	F	–	–	–	Uni & Bi	0.695	0.385	**0.759**	0.457	**0.884**	0.752	0.733	0.493
W2V	C & R	C & R	–	300	SG	–	0.681	0.381	0.730	**0.468**	0.873	**0.768**	0.741	0.528
TF–IDF	F	F	–	–	–	Bi & Tri	0.655	**0.395**	0.710	0.462	0.874	**0.769**	0.706	0.498
TF–IDF	F	S	–	–	–	Uni & Bi	0.650	0.334	0.712	0.411	**0.880**	0.683	0.702	0.452
BERT	Pre–trained	F	–	–	–	–	0,645	0,290	0,667	0,347	0,685	0,346	**0,763**	**0,541**
TF–IDF	C	C	–	–	–	Bi & Tri	0.638	**0.395**	0.698	0.462	0.856	0.748	0.691	0.507
BERT	Pre–trained	C	–	–	–	–	0,589	0,285	0,682	0,431	0,626	0,328	**0,768**	**0,562**
BERT	Pre–trained	C & R	–	–	–	–	0,589	0,269	0,682	0,409	0,616	0,306	0,754	**0,570**

the vast array of legal documents.

Using the sentence-based corpus did not yield Recall scores comparable to other representations. This fact represents good news for applying document similarity models in real-world scenarios since sentence-level training and evaluation substantially increase computational cost compared to document-level training (about 100 times higher, in our dataset's case).

Finally, the performance of our experiments' Doc2Vec and BM25 models was noticeably low. Although models based on neural networks such as Doc2Vec and Word2Vec require large volumes of training data, the dataset used in this study seems sufficient to obtain good results using Word2Vec, but not with Doc2Vec models. It reinforces the granularity hypothesis mentioned before. Considering that it was impossible to extract paragraphs or sections from the documents, which could contribute to document-level contextual information for the Doc2Vec models, the Word2Vec assemblies benefited from using local contextual information based on the word neighborhood.

Among the pre-trained models, using the full text without stemming and mAR as the evaluation metric, Top2Vec obtained significantly better results than Word2Vec. The Convolutional Neural Network (CNN) on which the Top2Vec vectorization is based seems to have better captured the semantic relationships in the corpus.

5.1. Implications for practice

The first implication for an IS focusing on capturing legal precedents is the need for fine-tuning to accommodate the unique semantics of legal texts. The granular text representation of the fine-tuned Word2Vec models was exceptionally effective due to their alignment with the intricate nature of legal language, which often encompasses context-specific terminologies. In contrast, the advanced capabilities of BERT, even though pre-trained on legal corpora, did not lead to superior outcomes in the Administrative Court under analysis. This fact underscores the challenge of integrating cutting-edge technology with the context-specific demands of legal practice.

Overall, the uninspiring performance of pre-trained models illuminates the critical role of context-specific training and reinforces the importance of tailored approaches within legal IS.

For applications with limitations to fine-tuning the models in a context-specific corpus, Top2Vec would be the best option among the tested assemblies, followed by BERT. When fine-tuning is possible and for applications limiting the number of predicted similar cases presented to the user, Word2Vec and TF-IDF perform similarly, with a slight advantage for Word2Vec for achieving better results in expert individual evaluations. Finally, when achieving good performance with any number of predicted similar cases is desirable, Word2Vec was the best option.

Another implication derived from the results is the need to tune even models that rely on granular text representations, such as Word2Vec and TF-IDF. Practitioners should also expect the need for periodic retraining to adapt to new cases in the dataset and keep good performance.

Expert evaluations further revealed that individual interpretations of relevance could influence the effectiveness of models, highlighting the subjective dimensions of legal decision-making and reinforcing the importance of human judgment for not overruling AI-retrieved precedents (Fagan & Levmore, 2019).

5.2. Contributions to literature

This work contributes to the discourse on legal text relevance analysis by empirically demonstrating the efficacy of multiple neural network-based models against traditional text embeddings within the specialized domain of Brazilian administrative law. It rigorously assessed multiple methods against a gold standard provided by legal experts to evaluate the models' capacity to imitate human cognition to detect legal precedents.

It also recommends baseline solutions for administrative courts, depending on real-world constraints. By examining a BERT-based model pre-trained on legal documents and comparing it with Doc2Vec, Word2Vec, and TF-IDF, this study extends the findings of Kulkarni et al. (2017); Ranera et al. (2019), and Mandal et al. (2017, 2021), confirming that Doc2Vec might be impracticable for small legal datasets, and demonstrating that the even with domain pre-training, BERT may not always translate to superior performance in context-specific tasks, demanding to fine-tune.

Unlike the extensive datasets used in COLIEE and FIRE AILA that leveraged cross-citations to create the evaluation dataset, our work provides evidence based on human knowledge that smaller, specialized datasets can still discern precedents when analyzed with granular text representations. Our study also confirms AI's cognitive mimicking capabilities, as discussed by Dwivedi et al. (2021), showcasing the practical implementation of AI in legal precedent identification and evaluation. Furthermore, it resonates with the emerging management applications supported by big data, as highlighted by Kushwaha et al. (2021), emphasizing the growing relevance of data-driven decision--making in legal administration and reinforcing automated precedents retrieval as an actual possibility.

Also, we produced theoretical insights under a socio-technical lens that deserve further discussion (Berente et al., 2019). First, what is the effect of fine-tuning on the effectiveness of legal precedents retrieval and the adoption of such IS by legal practitioners? Additionally, what should the frequency of such updates be to maintain acceptable performance? Finally, it is crucial to enhance comprehension of how individual interpretation grounds the most relevant legal literature.

Ultimately, our research aligns with Henkel et al. (2017), who discuss the transformative potential of language technologies in public organizations. By empirically testing language models, our work underscores the practicality of AI in legal precedent identification and evaluation, advocating for data-driven methodologies that Henkel et al. (2017) suggest can optimize decision-making in public administration.

6. Conclusions

This study demonstrates the feasibility of extracting relevant legal cases in Brazilian administrative law through textual similarity analysis. Various machine learning assemblies, particularly Word2Vec, proved effective in identifying similar cases, resembling the evaluation of human experts. Using vectorization methods based on granular pieces of text, such as individual words, emerged as advantageous in this context. To this conclusion, it was crucial to consider the appropriate performance metrics, especially the ratio of identified documents that proved relevant to the total number of relevant documents (Recall).

Also, using a significant sample of cases ensured that the comparative results did not happen by chance. Results were measured in a large sample of 1225 case pairs, independently evaluated by three legal experts. To the authors' knowledge, this is the most significant human-curated sample used in studies of this nature.

This study's findings provide valuable insights and hold significant practical implications:

1. Effectiveness of text embedding models: the results demonstrate that specific text embedding models, particularly Word2Vec, outperform traditional models like TF-IDF and even advanced models like BERT in legal document relevance assessment. Word2Vec models consistently achieved the highest mean Average Recall (mAR) scores, indicating their efficacy in identifying similar cases and potential legal precedents. This result suggests that leveraging granular text representations, such as individual words, can be advantageous in this context.
2. Efficiency and consistency: the Word2Vec models not only demonstrated high mAR scores, above 70 %, but also yielded competitive Recall@5 scores, indicating their potential to identify legal precedents relevant to multiple experts efficiently and when a limited number of samples is retrieved. This conclusion can contribute to expedited analysis and decision-making in law courts while maintaining consistency in jurisprudence.
3. Techniques for improving results: text summarization through concepts (nouns) and relationships (verbs) improved the results. The vectorization of concepts was the best option for an application that suggested five similar cases. Therefore, further exploring textual feature extraction techniques may improve performance.
4. Baseline models for real-world constraints: we concluded that optimal baseline solutions for administrative courts depend on constraints such as limited fine-tuning capacity or the number of candidate precedents to be presented to a user. Furthermore, we suggest such baseline models in the context of an administrative court.
5. Need for fine-tuning: the results underscore the necessity for domain-specific fine-tuning. To effectively capture the nuanced legal language, models must be tailored to the legal domain's unique terminology, syntax, and semantic structures. This process involves adjusting model parameters to better align with legal data, a crucial step for enhancing relevance in precedents discovery.
6. Frequent retraining: as a corollary of the previous conclusion, the dynamic nature of law, with evolving legal norms and terminologies, necessitates the frequent retraining of ML models. Keeping models updated with the latest legal documents and decisions ensures they remain effective in identifying relevant cases and legal precedents. Regular updates can help mitigate the risk of obsolescence due to changes in legal ontology.

Finally, the adoption of ML solutions in legal IS must consider the elements highlighted above, which are essential for maintaining the accuracy and relevance of such systems in legal practice, where the precision of language and the currentness of legal knowledge are paramount.

While this study provides valuable insights, opportunities for further research abound. First, the document representation and text embedding possibilities discussed here are not exhaustive. For example, the dataset's documents did not adhere to a consistent structure. Therefore, we could not experiment with additional document representation approaches, such as paragraphs or sections.

Investigating with recurrent neural networks (RNN) and long short-term memory (LSTM) networks was held for future research. The effectiveness of these methods for evaluating text similarity is uncertain as they have not yet been substantially investigated, especially with small datasets like ours. Specialized models such as Law2Vec, a Word2Vec model pre-trained on a substantial legal corpus, hold promise as a potential candidate for future research. However, it could not be incorporated into our current investigation because it focuses on English-written text. Also, further examination of Transformers such as BERT in domain-specific tasks, with potential fine-tuning and contextual pre-training, may reveal untapped potential. The intricacies of legal language and context may require tailored pre and post-processing steps to harness the full power of these models.

Finally, in this study's approach, potential similar cases represent a list sorted according to similarity. Additionally, the list is restricted to a predetermined number of samples. Future research may assess the efficacy of portraying all cases in two-dimensional or three-dimensional space to bypass this limitation. In order to achieve this objective, it is necessary to investigate the influence of dimensionality reduction techniques on text embeddings.

Funding sources

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under the project - UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

CRediT authorship contribution statement

Hugo Mentzingen: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nuno António:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Fernando Bacao:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Marcio Cunha:** Writing – review & editing, Conceptualization.

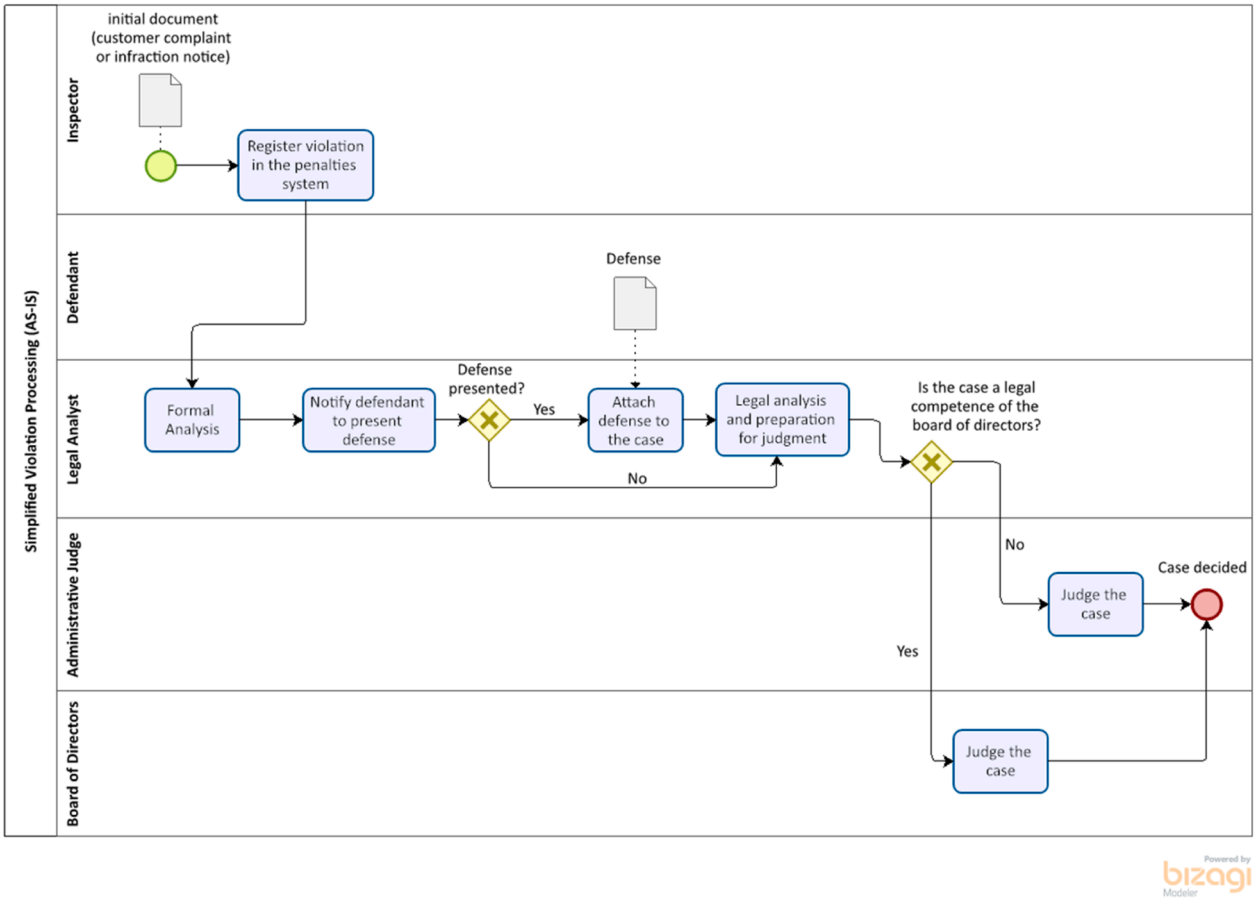
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge Brazil's Superintendency of Private Insurance for supporting and providing data for this work.

Appendix 1. Infraction examination business process



Appendix 2. Detailed results

The following table presents the results[2] in detail, according to the experimental setup described in Section 3.4. The three best results for each metric combination (mAR, Recall@5, mAP, and Precision@5) and assessment set (Experts one, two, and three, and the Experts' mean) are highlighted in bold.

[2] The Full, Full without stemming, Concepts, Concepts & Relations and Sentences corpora are represented by the F, F (NS), C, C&R, and S labels. For continuous bag-of-words and skip-gram methods, we used the CBOW and SG tags. Unigrams, Bigrams and Trigrams are represented respectively by Uni, Bi and Tri.

Table 5
Complete results set for the second study.

							Experts Mean		Expert 1		Expert 2		Expert 3	
Model	**Training Corpus**	**Evaluation Corpus**	**Topics**	**Dimensions**	**Algorithm**	**N–Grams**	**mAR**	**Recall@5**	**mAR**	**Recall@5**	**mAR**	**Recall@5**	**mAR**	**Recall@5**
W2V	F	C & R	–	100	CBOW	–	**0.705**	0.389	0.746	0.456	0.867	0.764	0.750	0.523
W2V	C	C	–	100	CBOW	–	**0.705**	0.357	0.743	0.431	0.856	0.710	**0.772**	0.526
W2V	C	C	–	100	SG	–	**0.701**	**0.393**	0.747	**0.486**	0.862	0.721	0.753	0.511
W2V	F	C & R	–	300	CBOW	–	0.701	0.370	0.740	0.449	0.859	0.674	0.753	0.509
W2V	F	F	–	300	SG	–	0.699	0.359	0.751	0.435	0.871	0.731	0.735	0.501
W2V	F	C	–	100	CBOW	–	0.699	0.371	0.734	0.430	0.869	0.730	0.756	0.498
W2V	F	C	–	300	CBOW	–	0.699	0.378	0.729	0.452	0.865	0.716	0.760	0.509
W2V	C & R	C & R	–	300	CBOW	–	0.698	0.358	0.733	0.436	0.863	0.725	0.747	0.536
W2V	F	F	–	100	CBOW	–	0.697	0.330	**0.761**	0.420	0.876	0.684	0.725	0.481
W2V	C	C	–	300	SG	–	0.696	0.391	0.741	**0.480**	0.858	**0.770**	0.761	0.533
W2V	F	F	–	300	CBOW	–	0.696	0.357	**0.755**	0.438	**0.878**	0.721	0.724	0.509
W2V	C	C	–	300	CBOW	–	0.695	0.355	0.737	0.422	0.847	0.684	0.745	0.509
TF–IDF	F	F	–	–	–	Uni & Bi	0.695	0.385	**0.759**	0.457	**0.884**	0.752	0.733	0.493
W2V	C & R	C & R	–	100	CBOW	–	0.694	0.366	0.734	0.450	0.860	0.726	0.750	0.527
W2V	F	C & R	–	300	SG	–	0.691	0.373	0.728	0.437	0.866	0.765	0.748	0.516
W2V	F	F	–	100	SG	–	0.691	0.355	0.737	0.435	0.856	0.742	0.733	0.505

(*continued on next page*)

Table 5 (*continued*)

W2V	F	C	–	300	SG	–	0.685	0.355	0.718	0.423	0.859	0.748	0.752	0.520
TF–IDF	C & R	C & R	–	–	–	Uni & Bi	0.684	0.376	0.744	0.443	0.867	0.710	0.724	0.492
TF–IDF	C	C	–	–	–	Uni & Bi	0.681	0.391	0.741	0.467	0.875	0.758	0.731	0.512
W2V	F	C & R	–	100	SG	–	0.681	0.378	0.713	0.414	0.836	0.741	0.749	0.530
W2V	F	C	–	100	SG	–	0.681	0.355	0.710	0.399	0.838	0.714	0.750	0.519
W2V	C & R	C & R	–	300	SG	–	0.681	0.381	0.730	**0.468**	0.873	**0.768**	0.741	0.528
LDA	S	S	20	–	–	–	0.680	0.245	0.683	0.242	0.761	0.433	0.725	0.353
LDA	S	S	80	–	–	–	0.680	0.310	0.703	0.366	0.819	0.621	0.742	0.422
D2V	F	F	–	200	–	–	0.680	0.326	0.705	0.379	0.855	0.676	0.742	0.491
W2V	C & R	C & R	–	100	SG	–	0.679	0.370	0.726	0.437	0.858	0.762	0.742	0.525
LDA	S	S	40	–	–	–	0.678	0.330	0.686	0.356	0.799	0.537	0.762	0.443
TF–IDF	F	C & R	–	–	–	Uni & Bi	0.676	0.375	0.735	0.444	0.869	0.727	0.718	0.487
D2V	C & R	C & R	–	200	–	–	0.675	0.344	0.703	0.395	0.852	0.667	0.750	0.461
D2V	C & R	C & R	–	100	–	–	0.674	0.323	0.703	0.369	0.854	0.667	0.744	0.457
TF–IDF	F	C	–	–	–	Uni & Bi	0.673	0.360	0.731	0.436	0.877	0.728	0.725	0.473
D2V	F	F	–	100	–	–	0.671	0.356	0.698	0.424	0.855	0.666	0.732	0.491
D2V	C & R	C & R	–	300	–	–	0.670	0.347	0.700	0.390	0.852	0.681	0.742	0.479
D2V	F	F	–	300	–	–	0.669	0.340	0.693	0.393	0.861	0.690	0.733	0.485
T2V	Pre–trained	F (NS)	–	–	–	–	0.669	0.352	0.731	0.434	0.845	0.666	0.725	0.465
W2V	F	S	–	300	CBOW	–	0.667	0.320	0.695	0.355	0.856	0.629	0.725	0.465
W2V	S	S	–	100	CBOW	–	0.667	0.323	0.695	0.364	0.851	0.650	0.729	0.457
W2V	F	S	–	100	CBOW	–	0.664	0.320	0.694	0.379	0.849	0.649	0.721	0.453

(*continued on next page*)

Table 5 (*continued*)

W2V	S	S	–	300	CBOW	–	0.662	0.306	0.694	0.366	0.853	0.632	0.726	0.424
W2V	S	S	–	300	SG	–	0.662	0.334	0.701	0.408	0.846	0.683	0.720	0.475
D2V	F	C & R	–	200	–	–	0.661	0.387	0.688	0.446	0.840	0.711	0.752	0.516
D2V	C	C	–	300	–	–	0.660	0.348	0.694	0.400	0.857	0.700	0.745	0.477
D2V	F	C & R	–	300	–	–	0.660	0.378	0.679	0.420	0.851	0.706	0.757	0.507
W2V	F	S	–	100	SG	–	0.656	0.320	0.697	0.382	0.840	0.665	0.714	0.443
W2V	S	S	–	100	SG	–	0.656	0.337	0.697	0.413	0.848	0.672	0.717	0.468
TF–IDF	F	F	–	–	–	Bi & Tri	0.655	**0.395**	0.710	0.462	0.874	**0.769**	0.706	0.498
D2V	C	C	–	200	–	–	0.653	0.334	0.685	0.407	0.846	0.672	0.736	0.470
W2V	F	S	–	300	SG	–	0.653	0.334	0.697	0.398	0.851	0.695	0.710	0.477
TF–IDF	F	S	–	–	–	Uni & Bi	0.650	0.334	0.712	0.411	**0.880**	0.683	0.702	0.452
D2V	F	C	–	200	–	–	0.645	0.367	0.686	0.434	0.850	0.691	0.738	0.489
BERT	Pre–trained	F	–	–	–	–	0,645	0,290	0,667	0,347	0,685	0,346	**0,763**	**0,541**
D2V	F	C & R	–	100	–	–	0.644	0.381	0.668	0.439	0.856	0.732	0.739	0.508
D2V	C	C	–	100	–	–	0.644	0.324	0.675	0.373	0.839	0.636	0.728	0.460
TF–IDF	F	C & R	–	–	–	Bi & Tri	0.644	0.334	0.689	0.385	0.844	0.675	0.711	0.445
TF–IDF	F	C	–	–	–	Bi & Tri	0.643	0.333	0.687	0.379	0.850	0.653	0.713	0.458
LDA	S	S	10	–	–	–	0.642	0.267	0.653	0.284	0.762	0.450	0.690	0.353
W2V	Pre–trained	F (NS)	–	300	SG	–	0.641	0.314	0.684	0.381	0.837	0.678	0.721	0.473
TF–IDF	C	C	–	–	–	Bi & Tri	0.638	**0.395**	0.698	0.462	0.856	0.748	0.691	0.507
TF–IDF	C & R	C & R	–	–	–	Bi & Tri	0.637	0.376	0.698	0.430	0.850	0.742	0.683	0.498
D2V	F	C	–	300	–	–	0.635	0.368	0.669	0.439	0.838	0.686	0.731	0.478
D2V	F	C	–	100	–	–	0.635	0.377	0.670	0.444	0.837	0.697	0.721	0.483

(*continued on next page*)

Table 5 (*continued*)

TF–IDF	F	S	–	–	–	Bi & Tri	0.632	0.349	0.679	0.422	0.865	0.715	0.705	0.473
TF–IDF	S	S	–	–	–	Uni & Bi	0.631	0.316	0.684	0.367	0.856	0.670	0.689	0.423
W2V	Pre–trained	F (NS)	–	300	CBOW	–	0.629	0.325	0.674	0.399	0.843	0.694	0.720	0.489
W2V	Pre–trained	F (NS)	–	100	SG	–	0.629	0.316	0.673	0.380	0.821	0.668	0.708	0.464
W2V	Pre–trained	F (NS)	–	100	CBOW	–	0.622	0.307	0.668	0.371	0.806	0.630	0.710	0.467
TF–IDF	S	S	–	–	–	Bi & Tri	0.622	0.316	0.670	0.386	0.845	0.667	0.692	0.433
D2V	S	S	–	300	–	–	0.614	0.194	0.638	0.223	0.754	0.431	0.695	0.310
D2V	S	S	–	100	–	–	0.610	0.198	0.637	0.224	0.758	0.445	0.691	0.325
D2V	S	S	–	200	–	–	0.610	0.196	0.637	0.224	0.760	0.423	0.696	0.327
BM25	F	C	–	–	–	–	0.601	0.263	0.622	0.277	0.704	0.340	0.642	0.304
BM25	F	F	–	–	–	–	0.599	0.262	0.617	0.273	0.701	0.340	0.640	0.304
BM25	C	C	–	–	–	–	0.598	0.266	0.619	0.281	0.705	0.340	0.640	0.303
BM25	F	C & R	–	–	–	–	0.596	0.263	0.618	0.277	0.697	0.340	0.636	0.304
BM25	C & R	C & R	–	–	–	–	0.595	0.266	0.617	0.281	0.693	0.340	0.637	0.304
BERT	Pre–trained	C	–	–	–	–	0.589	0,285	0,682	0,431	0,626	0,328	**0,768**	**0,562**
BERT	Pre–trained	C & R	–	–	–	–	0.589	0,269	0,682	0,409	0,616	0,306	0,754	**0,570**
D2V	F	S	–	200	–	–	0.574	0.164	0.582	0.188	0.705	0.353	0.683	0.296
D2V	F	S	–	100	–	–	0.572	0.166	0.582	0.195	0.706	0.357	0.674	0.294
D2V	F	S	–	300	–	–	0.572	0.168	0.581	0.193	0.710	0.357	0.678	0.299
BM25	S	S	–	–	–	–	0.566	0.185	0.602	0.209	0.696	0.391	0.595	0.276
LDA	C	C	10	–	–	–	0.512	0.287	0.557	0.332	0.673	0.496	0.556	0.352
LDA	C & R	C & R	10	–	–	–	0.511	0.293	0.587	0.362	0.686	0.474	0.520	0.320

(*continued on next page*)

Table 5 (*continued*)

LDA	F	C	10	–	–	–	0.506	0.292	0.566	0.334	0.701	0.540	0.531	0.359
LDA	F	C & R	10	–	–	–	0.501	0.273	0.574	0.331	0.797	0.614	0.591	0.393
LDA	F	C	20	–	–	–	0.435	0.249	0.489	0.287	0.589	0.437	0.508	0.370
LDA	F	F	10	–	–	–	0.427	0.221	0.488	0.277	0.647	0.463	0.517	0.330
LDA	F	F	20	–	–	–	0.418	0.253	0.482	0.307	0.706	0.585	0.551	0.411
LDA	C	C	20	–	–	–	0.409	0.303	0.472	0.353	0.684	0.625	0.493	0.439
LDA	C & R	C & R	20	–	–	–	0.407	0.209	0.421	0.239	0.678	0.487	0.525	0.331
LDA	C & R	C & R	40	–	–	–	0.376	0.238	0.420	0.275	0.589	0.465	0.441	0.342
LDA	F	C & R	40	–	–	–	0.374	0.273	0.428	0.316	0.622	0.509	0.478	0.382
LDA	C	C	40	–	–	–	0.361	0.308	0.422	0.375	0.681	0.629	0.478	0.418
LDA	F	C	40	–	–	–	0.359	0.223	0.400	0.258	0.626	0.536	0.509	0.368
LDA	C & R	C & R	80	–	–	–	0.356	0.302	0.413	0.367	0.681	0.640	0.477	0.429
LDA	F	F	80	–	–	–	0.355	0.294	0.418	0.359	0.589	0.564	0.418	0.392
LDA	F	C & R	20	–	–	–	0.352	0.237	0.414	0.298	0.611	0.562	0.435	0.368
LDA	F	F	40	–	–	–	0.339	0.265	0.388	0.313	0.586	0.506	0.422	0.338
LDA	F	C & R	80	–	–	–	0.311	0.221	0.369	0.273	0.595	0.511	0.419	0.355
LDA	F	C	80	–	–	–	0.308	0.240	0.366	0.280	0.513	0.443	0.389	0.325
LDA	C	C	80	–	–	–	0.280	0.255	0.336	0.308	0.654	0.643	0.415	0.409
LDA	F	S	80	–	–	–	0.076	0.063	0.098	0.084	0.200	0.200	0.160	0.160
LDA	F	S	20	–	–	–	0.075	0.063	0.096	0.084	0.200	0.200	0.160	0.160
LDA	F	S	10	–	–	–	0.075	0.063	0.096	0.084	0.200	0.200	0.160	0.160
LDA	F	S	40	–	–	–	0.074	0.063	0.096	0.084	0.200	0.200	0.160	0.160
BM25	F	S	–	–	–	–	0.070	0.063	0.091	0.084	0.200	0.200	0.160	0.160

References

Aggarwal, C. C. (2016). Evaluating recommender systems. In C. C. Aggarwal (Ed.), *Recommender systems: The textbook* (pp. 225–254). Springer International Publishing. https://doi.org/10.1007/978-3-319-29659-3_7.

Amaral-Garcia, S. (2021). Administrative courts. *Encyclopedia of Law and Economics*, 1–8. https://doi.org/10.1007/978-1-4614-7883-6_578-2

Angelov, D. (2020). *Top2Vec: Distributed representations of topics*. 1–25. http://arxiv.org/abs/2008.09470.

Arora, J., Patankar, T., Shah, A., & Joshi, S. (2020). Artificial intelligence as legal research assistant. In *, 2826. CEUR Workshop Proceedings* (pp. 60–65).

Batalli, M., & Pepaj, I. (2022). *Citizens' right to seek judicial review of administrative acts and its impact on governance reforms*. https://doi.org/10.22495/cgobrv6i2p8.

Berente, N., Seidel, S., & Safadi, H. (2019). Data-driven computationally intensive theory development. *Information Systems Research, 30*(1), 50–64. https://doi.org/10.1287/ISRE.2018.0774

Bhattacharya, P., Ghosh, K., Ghosh, S., Pal, A., Mehta, P., Bhattacharya, A., & Majumder, P. (2019). FIRE 2019 AILA track: Artificial intelligence for legal assistance. In *, 2517. ACM International Conference Proceeding Series* (pp. 4–6). https://doi.org/10.1145/3368567.3368587

Bhattacharya, P., Mehta, P., Ghosh, K., Ghosh, S., Pal, A., Bhattacharya, A., & Majumder, P. (2020a). FIRE 2020 AILA track: Artificial intelligence for legal assistance. In *ACM International Conference Proceeding Series* (pp. 1–3). https://doi.org/10.1145/3441501.3441510

Bhattacharya, P., Mehta, P., Ghosh, K., Ghosh, S., Pal, A., Bhattacharya, A., & Majumder, P. (2020b). Overview of the FIRE 2020 AILA track: Artificial intelligence for legal assistance. In *, 2826. CEUR Workshop Proceedings* (pp. 1–11).

Biel, Ł., & Kockaert, H.J. (2023). *Handbook of terminology* (F. Steurs & H. J. Kockaert, Eds.; Vol. 3). John Benjamins Publishing Company. https://doi.org/10.1075/hot.3.

Chalkidis, I. (2018). *Law2Vec: Legal word embeddings*. https://archive.org/details/Law2Vec.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 1724–1734). https://doi.org/10.3115/v1/d14-1179

Cochran, W. G. (1977). *Sampling techniques* (3rd Edition). John Wiley & Sons, Ltd.

Cornell University Law School. (2022). *administrative law*. Legal Information Institute. https://www.law.cornell.edu/wex/administrative_law.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding. In *, 1. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 4171–4186). https://github.com/tensorflow/tensor2tensor.

Devlin, J., Chang, M.-W. W., Lee, K., & Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In *, 1. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 4171–4186). https://arxiv.org/abs/1810.04805v2.

Di Nunzio, G. M. G. M. (2020). A study on lemma vs stem for legal information retrieval using R tidyverse. IMS UniPD @ AILA 2020 Task 1. In *, 2826. CEUR Workshop Proceedings* (pp. 54–59).

Domingues, M. (2022). *dominguesm/legal-bert-base-cased-ptbr*. Hugging Face. https://huggingface.co/dominguesm/legal-bert-base-cased-ptbr.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, Article 101994. https://doi.org/10.1016/J.IJINFOMGT.2019.08.002

Fagan, F., & Levmore, S. (2019). The impact of artificial intelligence on rules, standards, and judicial discretion. *Southern California Law Review, 93*(1), 1–36. https://doi.org/10.2139/SSRN.3362563

Fon, V., & Parisi, F. (2006). Judicial precedents in civil law systems: A dynamic analysis. *International Review of Law and Economics, 26*(4), 519–535. https://doi.org/10.1016/j.irle.2007.01.005

Frankenreiter, J., & Nyarko, J. (2022). Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom Ed.)*. https://doi.org/10.2139/ssrn.4027030. *Forthcoming*.

Gao, J., Ning, H., Sun, H., Liu, R., Han, Z., Kong, L., & Qi, H. (2019). FIRE2019@AILA: Legal retrieval based on information retrieval model. In *, 2517. CEUR Workshop Proceedings* (pp. 64–69).

Goebel, R., Kano, Y., Kim, M.-Y., Rabelo, J., Satoh, K., & Yoshioka, M. (2023). Summary of the competition on legal information, extraction/entailment (COLIEE) 2023. In *19th International Conference on Artificial Intelligence and Law, ICAIL 2023 - Proceedings of the Conference* (pp. 472–480). https://doi.org/10.1145/3594536.3595176

Gomez, A. R. (2021). Demand side justice. *Georgetown Journal on Poverty Law and Policy, XXVIII*(3), 411–436.

Greene, W. H. (2017). *Econometric analysis* (8th Ed.,, *1*. Pearson https://books.google.com/books/about/Econometric_Analysis.html?id=-WFPYgEACAAJ.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., & Aluisio, S. (2017). *Portuguese word embeddings: Evaluating on word analogies and natural language tasks*. http://arxiv.org/abs/1708.06025.

Henkel, M., Perjons, E., & Sneiders, E. (2017). Examining the potential of language technologies in public organizations by means of a business and IT architecture model. *International Journal of Information Management, 37*(1), 1507–1516. https://doi.org/10.1016/j.ijinfomgt.2016.05.008

Hu, W., Zhao, S., Zhao, Q., Sun, H., Hu, X., Guo, R., Li, Y., Cui, Y., & Ma, L. (2022). BERT_LF: A similar case retrieval method based on legal facts. *Wireless Communications and Mobile Computing, 2022*. https://doi.org/10.1155/2022/2511147

Kar, A. K., Angelopoulos, S., & Rao, H. R. (2023). Guest Editorial: Big data-driven theory building: Philosophies, guiding principles, and common traps. *International Journal of Information Management, 71*(April), Article 102661. https://doi.org/10.1016/j.ijinfomgt.2023.102661

Kim, M. Y., Rabelo, J., Goebel, R., Yoshioka, M., Kano, Y., & Satoh, K. (2023). COLIEE 2022 summary: Methods for legal document retrieval and entailment. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13859 LNAI* (pp. 51–67). https://doi.org/10.1007/978-3-031-29168-5_4

Kulkarni, Y. H. Y. H., Patil, R., & Shridharan, S. (2017). Detection of catchphrases and precedence in legal documents. In *, 2036. CEUR Workshop Proceedings* (pp. 86–89).

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*(1), Article 100008. https://doi.org/10.1016/j.jjimei.2021.100008

Kumar, S., Reddy, P. K., Reddy, V. B., & Singh, A. (2011). Similarity analysis of legal judgments. In *Compute 2011 - 4th Annual ACM Bangalore Conference* (pp. 3–6). https://doi.org/10.1145/1980422.1980439

Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights, 1*(2), Article 100017. https://doi.org/10.1016/J.JJIMEI.2021.100017

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *, 4. 31st International Conference on Machine Learning, ICML 2014* (pp. 2931–2939).

Leburu-Dingalo, T., Motlogelwa, N. P., Thuma, E., & Modungo, M. (2020). UB at fire 2020 precedent and statute retrieval. In *, 2826. CEUR Workshop Proceedings* (pp. 12–17). https://www.ub.bw.

Li, H., Su, W., Wang, C., Wu, Y., Ai, Q., & Liu, Y. (2023). *THUIR@COLIEE 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval*. https://arxiv.org/abs/2305.06812v1.

Liu, L., Liu, L., & Han, Z. (2020). Query revaluation method for legal information retrieval. In *, 2826. CEUR Workshop Proceedings* (pp. 18–21). https://trec.nist.gov/pubs/trec16/appendices/measures.pdf.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development, 1*(4), 309–317. https://doi.org/10.1147/rd.14.0309

Lv, Y., & Zhai, C. (2011). Lower-bounding term frequency normalization. In *International Conference on Information and Knowledge Management, Proceedings* (pp. 7–16). https://doi.org/10.1145/2063576.2063584

Ma, Y., Shao, Y., Liu, B., Liu, Y., Zhang, M., & Ma, S. (2021). Retrieving legal cases from a large-scale candidate corpus. In *Proceedings of the Eighth International Competition on Legal Information Extraction/Entailment, COLIEE2021*.

Mandal, A., Chaki, R., Saha, S., Ghosh, K., Pal, A., & Ghosh, S. (2017). Measuring similarity among legal court case documents. In *ACM International Conference Proceeding Series* (pp. 1–9). https://doi.org/10.1145/3140107.3140119

Mandal, A., Ghosh, K., Ghosh, S., & Mandal, S. (2021). Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law, 29*(3), 417–451. https://doi.org/10.1007/s10506-020-09280-2

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Marshall, S. W. (2005). Prevalence and incidence. *Encyclopedia of Social Measurement*, 141–147. https://doi.org/10.1016/B0-12-369398-5/00144-4

Martin, P. W. (2008). Reconfiguring law reports and the concept of precedent for a digital age. *Villanova Law Review, 53*(1), 1–46. https://digitalcommons.law.villanova.edu/vlr/vol53/iss1/1.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software, 2*. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software, 3*(29), 861. https://doi.org/10.21105/joss.00861

Mcintyre, G. (2020). The impact of a lack of consistency and coherence: How key decisions of the International criminal court have undermined the court's legitimacy. *Questions of International Law, 67*, 25–57. www.icc-cpi.int/RelatedRecords/CR2018_02989.PDF.

Mentzingen, H., António, N., & Bacao, F. (2023). Automation of legal precedents retrieval: findings from a literature review. *International Journal of Intelligent Systems, 2023*, 1–22. https://doi.org/10.1155/2023/6660983

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A Meeting of SIGDAT, a Special Interest Group of the ACL Held in Conjunction with ACL 2004* (pp. 404–411).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Nason, S. (2018). *Administrative justice can make countries fairer and more equal – if it is implemented properly*. The Conversation. https://theconversation.com/administrative-justice-can-make-countries-fairer-and-more-equal-if-it-is-implemented-properly-108238.

Perlingeiro, R. (2014). Brazil's administrative justice system in a comparative context. *Revista de Investigações Constitucionais, 1*(3), 33–58. https://doi.org/10.5380/RINC.V1I3.40517

Popova, O., Maroz, R., & Gámez, M.A.Q. (2021). The undeniable benefits of court automation. *Let's talk development*. https://blogs.worldbank.org/developmenttalk/undeniable-benefits-court-automation.

Rabelo, J., Goebel, Randy, Kim, M.-Y., Kano, Y., Yoshioka, M., & Satoh, K (2022). Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *The Review of Socionetwork Strategies, 16*(1), 111–133. https://doi.org/10.1007/S12626-022-00105-Z. *2022 16:1*.

Rabelo, J., Kim, M.-Y., & Goebel, R. (2023). Semantic-based classification of relevant case law. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13859 LNAI* (pp. 84–95). https://doi.org/10.1007/978-3-031-29168-5_6

Rabelo, J., Kim, M. Y., Goebel, R., Yoshioka, M., Kano, Y., & Satoh, K. (2021). COLIEE 2020: Methods for legal document retrieval and entailment. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12758 LNAI* (pp. 196–210). https://doi.org/10.1007/978-3-030-79942-7_13

Ranera, L. T. B., Solano, G. A., & Oco, N. (2019). Retrieval of semantically similar philippine supreme court case decisions using Doc2Vec. In *2019 International Symposium on Multimedia and Communication Technology (ISMAC)* (pp. 1–6). https://doi.org/10.1109/ISMAC.2019.8836165

Rhode, D. L. (2004). *Access to justice*. Oxford University Press.

Richardson, L. (2007). *BeautifulSoup*. https://www.crummy.com/software/BeautifulSoup/.

Rigoni, A. (2014). Common-law judicial reasoning and analogy. *Legal Theory, 20*(2), 133–156. https://doi.org/10.1017/S1352325214000044

Robertson, S., Zaragoza, H., Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

Roitblat, H. L., Kershaw, A., & Oot, P. (2010). Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology, 61*(1), 70–80. https://doi.org/10.1002/asi.21233

Schröder, G., Thiele, M., & Lehner, W. (2011). *Setting goals and choosing metrics for recommender system evaluations*. 811.

Shahade, A. K., Walse, K. H., Thakare, V. M., & Atique, M. (2023). Multi-lingual opinion mining for social media discourses: An approach using deep learning based hybrid fine-tuned smith algorithm with adam optimizer. *International Journal of Information Management Data Insights, 3*(2), Article 100182. https://doi.org/10.1016/J.JJIMEI.2023.100182

Shinyama, Y., Guglielmetti, P., & Marsman, P. (2019). *pdfminer.six*. https://github.com/pdfminer/pdfminer.six.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21. https://doi.org/10.1108/00220410410560573

Struijk, M., Ou, C. X. J., Davison, R. M., & Angelopoulos, S. (2022). Putting the IS back into IS research. *Information Systems Journal, 32*(3), 469–472. https://doi.org/10.1111/ISJ.12368

Susskind, R. (2020). The future of courts. *The Practice, 6*(5). https://thepractice.law.harvard.edu/article/the-future-of-courts/.

Thenmozhi, D., Kannan, K., & Aravindan, C. (2017). A text similarity approach for precedence retrieval from legal documents. In *FIRE (Working Notes)* (pp. 90–91). http://ceur-ws.org/Vol-2036/T3-9.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5999–6009. https://doi.org/10.48550/arxiv.1706.03762

Velicogna, M. (2007). Justice systems and ICT What can be learned from Europe? *Utrecht Law Review, 3*(1), 129. https://doi.org/10.18352/ULR.41

Vogel, F., Hamann, H., & Gauer, I. (2017). *Computer-assisted legal linguistics: Corpus analysis as a new tool for legal studies*. https://doi.org/10.1111/lsi.12305.

Westermann, H., Savelka, J., & Benyekhlef, K. (2021). Paragraph similarity scoring and fine-tuned bert for legal information retrieval and entailment. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12758 LNAI* (pp. 269–285). https://doi.org/10.1007/978-3-030-79942-7_18

Wilcox, R. (2015). Inferences about the skipped correlation coefficient: Dealing with heteroscedasticity and non-normality. *Journal of Modern Applied Statistical Methods, 14*(2), 2–8, 10.56801/10.56801/V14.I.769.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2019). *Multilingual universal sentence encoder for semantic retrieval*. arXiv. https://doi.org/10.48550/ARXIV.1907.04307.

Zarindast, A., Sharma, A., & Wood, J. (2021). Application of text mining in smart lighting literature - an analysis of existing literature and a research agenda. *International Journal of Information Management Data Insights, 1*(2), Article 100032. https://doi.org/10.1016/J.JJIMEI.2021.100032

Zhang, N. N., Pu, Y. F. Y.-F., Yang, S. Q. S.-Q., Zhou, J.-L. J. L., & Gao, J.-K. J. K. J.-K. (2017). An ontological chinese legal consultation system. *IEEE Access : Practical Innovations, Open Solutions, 5*, 18250–18261. https://doi.org/10.1109/ACCESS.2017.2745208

Zhao, Z., Ning, H., Liu, L., Huang, C., Kong, L., & Han, Y. (2019). FIRE2019 @ AILA : Legal information retrieval using improved BM25. In *FIRE (Working Notes), December 2019* (pp. 12–15).