Artificial Intelligence and Law
https://doi.org/10.1007/s10506-025-09446-w

ORIGINAL RESEARCH

LegalAnalytics: bridging visual explanations and workload streamline in Brazilian Supreme Court appeals

Lucas Resck[1] · Felipe Moreno-Vera[1] · Tobias Veiga[1] · Gerardo Paucar[1] · Ezequiel Fajreldines[3] · Guilherme Klafke[3] · Luis G. Nonato[2] · Jorge Poco[1]

Accepted: 19 February 2025

Abstract
The Brazilian Supreme Court serves as the highest judicial authority in Brazil and is responsible for adjudicating constitutional matters presented as extraordinary appeals. These appeals undergo a rigorous screening process guided by established legal principles known as Topics of General Repercussion. Seeking to streamline this procedure, we developed LegalAnalytics to explore the research question: *Can machine learning and explainable AI techniques enhance the classification of appeals in legal workflows?* LegalAnalytics harnesses advanced natural language processing algorithms and classification models to categorize each appeal according to the most pertinent topics accurately. In addition, it incorporates LIME (Local Interpretable Model-agnostic Explanations) to highlight the key sections of an appeal and compare them with relevant precedents. This approach ensures a transparent justification for every classification. The system is thoughtfully designed with a user-friendly interface tailored for public servants, judges, and lawyers. Extensive testing with dozens of legal experts confirmed the effectiveness of LegalAnalytics, with consistently positive feedback underscoring its significant practical impact.

Keywords Visual analytics · Legal document classification · Natural language processing · Explainable AI

1 Introduction

The Brazilian Supreme Court (STF) holds the highest position within the hierarchy of the Brazilian judicial system. Among its various responsibilities, the STF judges constitutional matters that originate from lower courts. Most of the processes reaching the Supreme Court come in the form of an Extraordinary Appeal (RE—Portuguese abbreviation for "Recurso Extraordinário"), with thousands of applications filed each year, creating sizeable workload for the court. Legislators introduced the prerequisite of general repercussion to reduce this workload, meaning that potential

Extended author information available on the last page of the article

Published online: 11 July 2025

constitutional violations should have a substantial impact to be assessed by STF. After accepting and judging a case, the STF enacts a precedent summarized in a *topic*, which is meant to be applied by lower courts. Future appeals face a screening to identify, among other things, if their content fits into an existing topic. If the public servants in charge of the screening process deem a topic applicable, the appeal goes through a simplified procedure, not reaching the Justices' offices, thus reducing the number of cases in the STF.

Screening Extraordinary Appeals is just one example of a procedure adopted by the STF to manage its workload. In fact, automation of procedures has been a central focus throughout the Brazilian judiciary, leading to the development of several data science and machine learning initiatives within the judicial system (Araujo et al., 2020; Salomão et al., 2023). The primary objective of these initiatives is not to replace human decision-making but rather to improve efficiency and reliability in legal procedures. However, ensuring that computational tools, particularly those based on machine learning, operate in an unbiased and fair manner is a nontrivial task, which adds challenge for professionals utilizing such tools.

The present work aims to tackle the hurdles discussed above by introducing LegalAnalytics, a system specifically developed to aid in the screening of Extraordinary Appeals. LegalAnalytics builds on natural language processing and classification models to assess the applicability of judicial topics to REs. Unlike other systems under consideration in Brazilian courts (see Sect. 2), the proposed tool supports a decision and provides explanations for it, streamlining the screening process while enhancing confidence in the outcomes. By prioritizing transparency, our approach aligns with a range of Brazilian and international recommendations regarding the use of artificial intelligence, which is crucial in the context of the judicial system. Furthermore, LegalAnalytics has been meticulously designed to ensure its interface is user-friendly and accessible to a diverse audience, including public servants, judges, and lawyers. In this context, appealing lawyers can use LegalAnalytics to identify relevant GR topics, helping manage client expectations and prepare stronger arguments for distinguishing cases or justifying a different outcome, especially given the limited opportunities to contest STF's topic application. For appealed parties, early knowledge of the GR topic aids in preparing more effective counter-arguments or counter-appeals, improving the overall quality of legal reasoning. To evaluate the effectiveness of the proposed system, we conducted a comprehensive evaluation involving dozens of expert users. Their feedback effectively demonstrated the usefulness and benefits of using LegalAnalytics in practice (demo and additional material are available at http://visualdslab.com/papers/LegalAnalytics/).

Therefore, the main contributions of this work can be summarized as follows:

- A novel methodology for classifying Extraordinary Appeals based on the most probable applicable Topics of General Repercussion;
- The integration of an explanation mechanism to justify classification decisions, which is crucial for bringing confidence in the outcomes, addressing concerns about applying AI in the legal domain;
- A meticulously designed user-friendly interface conceived for a diverse audience;

- LegalAnalytics, a web-based visual analytics system that simplifies the analysis of REs, related topics, and similar legal processes;
- A powerful and precise prediction system that can assist lawyers in developing more effective reasoning to differentiate the case at hand from established precedents or to highlight the applicability of consolidated judicial principles;
- A comprehensive and thorough evaluation involving dozens of experts to confirm the usefulness and benefits of LegalAnalytics.

2 Related work

In recent years, there has been a growing interest in the application of AI to the legal field. The literature on applied artificial intelligence and machine learning in law is vast, and we point the reader to the surveys by Katz et al. (2023), Zhong et al. (2020), and Atkinson et al. (2020) for more comprehensive discussions. To better contextualize our contribution, we focus the discussion on systems built upon AI resources for Law analytics, emphasizing applications in the Brazilian legal system.

2.1 Artificial intelligence in law

Systems. Several systems have been proposed to address legal problems with machine learning and AI. The work most similar to ours is LegalVis (Resck et al., 2023), a visual analytics system developed to assist judicial experts in analyzing precedent citations within legal documents. LegalVis employs machine learning and explainability methods to infer and explain precedent citations in documents in a carefully designed visual interface. However, instead of Topics of General Repercussion, LegalVis approaches another type of Brazilian precedent called “binding precedent.”[1] A topic is a concrete decision with significant relevance in Brazilian society, while a binding precedent is a general, abstract summary that uniformizes the jurisprudence about a juridical theme. Furthermore, LegalVis is an academic prototype, with a sophisticated though complex visual interface. The proposed LegalAnalytics system, in contrast, has been designed with a focus on end users, fulfilling their quotidian domains (e.g., uploading their own RE file) with a simple and intuitive visual interface. CLAUDETTE (Lippi et al., 2019) is a web server that detects and categorizes unfair clauses in online terms of service using a diverse set of classification methods such as support vector machines, convolutional neural networks, and long short-term memory neural networks. The CLAUDETTE allows users to input terms of service through copy-and-paste, detecting unfair clauses (and their categorization) on a sentence level. In contrast, LegalAnalytics suggests topics based on entire documents and provides explanations for its decisions at the

[1] “Súmula Vinculante,” in its original name in Portuguese. We provide a brief explanation about it in Sect. 3.3.

paragraph or sentence level, eliminating the need for users to copy and paste text manually.

Tools and Search Engines. AI has also been utilized to assist users in exploring and identifying similar legal cases. For example, Bluetick (n.d.) provides suggestions and highlights similarities between documents during searches. In Brazil, notable examples of such tools include Buscador Dizer o Direito (n.d.), which offers annotated legal documents, and OABJuris (n.d.), which leverages AI to deliver relevant suggestions. More sophisticated platforms used in Brazil include Finch Platform (n.d.) and Legal One (Thomson Reuters, n.d.). In the US, there are Lexis (n.d.), Westlaw (Thomson Reuters, n.d.), and Casetext (n.d.). The latter, in particular, employs Transformer-based models (Vaswani et al., 2017) to find similar cases. We draw inspiration from those tools to design LegalAnalytics, incorporating a functionality to search for similar documents using textual and vector comparisons in latent space (Sect. 5.3).

Explainable AI in Law. Some works rely on logistic regression weights and attention scores of Transformer-based models (Zhao et al., 2023) for explaining the outcomes of machine learning models. Several of those works focus on emphasizing relevant words (Resck et al., 2023; Caled et al., 2019), although there are alternatives that extend attention scores to encompass entire sentences (Zhao et al., 2023). Support vector machine kernel weights (Aletras et al., 2016) and random forest rules (Arriba-Pérez et al., 2022; González-González et al., 2023) have also been employed for explanation purposes. More sophisticated mechanisms include intermediate, interpretable labels predicted by a secondary model that feeds the original, primary model, from which explanations are extracted (Lyu et al., 2022; Wu et al., 2022; Bhambhoria et al., 2022; Zhong et al., 2020). Our approach, in contrast, utilizes a separate explanation technique to extract the most important parts of the text (Sect. 5.2). Similarly to Resck et al. (2023) and Bhambhoria et al. (2021) we use LIME (Ribeiro et al., 2016) to derive the explanations. Resck et al. (2023), in particular, relies on LIME to explain the most important parts of a legal text in the context of precedent citations. However, their machine learning task differs from ours by dividing the text into sentences rather than paragraphs. Alternative explanation methods include Grad-CAM (Selvaraju et al., 2017), Integrated Gradients (Sundararajan et al., 2017), and input gradients (Benedetto et al., 2023; Semo et al., 2022) for enabling explanations.

2.2 Brazilian legal AI

The Brazilian Judiciary possesses characteristics that make it particularly well-suited for integrating and expanding the use of Artificial Intelligence (AI) to streamline and enhance judicial processes. Chief among these is the substantial volume of legal cases it handles. According to the National Council of Justice (CNJ), up to the end of 2022, Brazil had a staggering 81.4 million ongoing cases, with 21.7% of them currently suspended.

Another important factor is the prevailing trend toward digitalization, which has been boosted by initiatives such as Justice 4.0 and the 100% Digital Court program.

Since 2010, a total of 215 million cases have been processed electronically, with certain branches of the judiciary already achieving full digitalization (National Council of Justice, 2023).

More recently, the Brazilian Judiciary has experienced a significant increase in the adoption of AI systems. Following a mandate from the National Council of Justice, all judicial entities are required to report the progress of their AI projects and submit the developed models to a centralized repository called Sinapses (Brasil, 2020). Sinapses serves as a nationwide platform for control, governance, and information sharing.

As of May 2022, the Panel of AI Projects in the Judiciary reported 111 ongoing projects, with 63 already implemented and 42 registered in the Sinapses database (National Council of Justice, 2022). The primary motivation for most projects is to enhance productivity (84.6%), with the majority being developed by the institution's internal teams (54.1%). Python is the predominant programming language used (86.5%), and text analysis techniques are employed in most projects (88.3%). In total, 53 courts reported involvement in AI project development, with special recognition given to the Court of Justice of the State of Rondônia. This court has been a pioneer in implementing key AI models that were later adopted and disseminated by the National Council of Justice to other judicial entities.

The regulations of the National Council of Justice and discussions among academics and practitioners demonstrate a concern for quality, transparency, explainability, and respect for fundamental rights in the development of these systems. For example, Salomão et al. (2023) conducted a panoramic study of the level of governance of the systems registered on the Sinapses platform in light of the existing rules. In the following, we compare the proposed LegalAnalytics tool with other systems, considering their key characteristics.

VICTOR is an AI system launched in 2018, fully implemented by the Brazilian Supreme Court in 2020, and integrated into the Court's internal procedural management platform (STF Digital) (Salomão et al., 2023). It represented a pioneering AI solution within the Brazilian Judiciary, being the first widely publicized and extensively studied product for case screening (Araujo et al., 2020; Brazilian Supreme Court, 2021). Its primary objective is to identify topics of general repercussion applicable to extraordinary appeals arriving at the Supreme Court, a task that inspired the development of LegalAnalytics. Given the lack of standardization of appeals and the prevalence of scanned physical documents, the tool also incorporates optical character recognition (OCR) functionality to segregate relevant documents for classification and, ultimately, suggest the repercussion topic.

The LegalAnalytics system shares similarities with VICTOR; however, the differences are clearly notable. While the latter was trained using supervised learning on a dataset comprising 22,000 petitions from the period of 2014 to 2017, focusing on the 27 most prevalent topics of general repercussion, the former utilized both supervised (classification) and unsupervised (textual embedding similarity) learning on the textual dataset referenced in Sect. 4.3. This dataset consists, after careful preprocessing, of 10,710 documents spanning from 1959 to 2022, targeting the 30 most prevalent topics. Both systems encountered limitations in topic diversity due to the scarcity of court decisions available for training across the less prevalent topics.

Although the STF has publicized VICTOR's statistics, limited information about its interface and user interaction is available. In a video presentation,[2] a brief glimpse of the tool is provided; however, it does not allow verification of any explainability mechanisms. Additionally, STF's press releases do not mention a meaningful way for users to assess the consistency of the results generated by VICTOR. Given this context, we conclude that VICTOR lacks any significant explanation mechanism.

In contrast, LegalAnalytics was designed to offer a range of potential topics based on the highest probability criterion while highlighting the document's most relevant sections for classification under each topic. Furthermore, LegalAnalytics includes a feature for clustering similar cases based on their textual content. It is worth noting that this clustering capability was also introduced to the Supreme Court in 2023 through another tool, the VitorIA system (Brazilian Supreme Court, 2023). However, to the best of our knowledge, VitorIA does not predict case outcomes and is limited to the clustering task. Moreover, LegalAnalytics was developed strongly emphasizing user interface design, ensuring ease of use and a clear understanding of its functionalities. Dozens of experts in the field have evaluated the solution's effectiveness, providing overwhelmingly positive feedback.

RAFA 2030 (Artificial Networks Focused on the 2030 Agenda) is another tool the Brazilian Supreme Court developed and implemented in 2022 (Salomão et al., 2023). Its source code is open and available on the project's dedicated website (Brazilian Supreme Court, 2022). The tool aims to perform a multilabel classification of cases reaching the Court, assigning one of the 17 Sustainable Development Goals (SDGs) of the United Nations to each case based on the text of the initial petitions and case judgments. Clustering cases according to the SDGs enables judges, for example, to construct thematic agendas or align rulings with socioeconomic development strategies. The presentation of multiple labels is a feature also found in LegalAnalytics, which operates similarly to the RAFA 2030 system, albeit with different objectives.

ATHOS is an AI system developed by the Brazilian Superior Court of Justice and implemented in 2019 (Salomão et al., 2023). Its purpose is to categorize and group similar cases by analyzing the text of petitions and decisions. The tool conducts clustering using a model that represents the text as vectors in a 300-dimensional space, trained on a dataset of over 300,000 documents. The vector representation enables comparisons based on distance. ATHOS resembles LegalAnalytics in the way the findings are presented, enabling result filtering and the search for similar cases. However, LegalAnalytics incorporates a side-by-side comparison window for each text, allowing users to see the relevant parts of the text the model considers. This feature aims to provide users with justification for the model's identification of similar text segments.

BEM-TE-VI is an AI system developed by the Brazilian Superior Labor Court, utilizing data visualization and classification to screen cases. The model was trained on a dataset comprising 5 million cases processed by the Court and their corresponding decisions between 2018 and 2020 (Salomão et al., 2023). Its primary

[2] https://www.youtube.com/watch?v=_gjqAYq_-zY

objective is to offer decision predictions, recommend relevant advisors, and analyze admissibility requirements. Similar to the LegalAnalytics tool, the BEM-TE-VI system provides a probability of an outcome, presenting the result as a percentage accuracy. However, akin to ATHOS, it lacks a feature for comparing case texts with indications of relevant passages.

ALEI (Intelligent Legal Analysis) is a system implemented by the Federal Regional Court of the 1st Region in 2022, designed to categorize cases and propose draft decisions based on precedents from higher courts (Salomão et al., 2023). Developed using supervised learning techniques, it relies on annotations from samples of appeals to identify "appeal objects," which represent the topics of the appeals as perceived by the Court's staff. The tool, developed in collaboration with the University of Brasília, draws inspiration from several functionalities of the VICTOR system. The feature of decision drafting may be incorporated into LegalAnalytics in the future, further streamlining its usefulness.

3 Problem description

This section introduces the problem of applying Topics of General Repercussion to Extraordinary Appeals that reach the Supreme Court. We also present the requirements for an automated system to assist in this problem.

3.1 Extraordinary appeal

Its primary purpose is to enable the Supreme Court to review decisions made by lower courts. Under the 1988 Constitution, this procedure is outlined in Article 102, item III. It is used to challenge decisions that contradict the Constitution, declare a federal law or treaty unconstitutional, contest government actions that violate constitutional principles, or affirm the validity of local laws based on differing interpretations of federal legislation.

Extraordinary appeals are presented in lower courts and receive a first assessment. In general, the judgment of these appeals follows a similar flow in the various Brazilian courts, which can be divided in two steps:

1st phase in lower courts. After the constitutional appeal is presented, a specialized body of the State Court carries out an admissibility judgment, checking whether requirements are met.

2nd phase in the STF. If the lower court determines that there is neither a binding precedent nor justifiable grounds for denying admissibility, the constitutional appeal is elevated to the Supreme Court. Alternatively, lower courts may deny the constitutional appeal on other grounds, such as weak precedents. In such cases, parties can present an interlocutory appeal (called *Agravo em Recurso Extraordinário*, or ARE for short), forcing the original appeal to be judged by the higher court.[3]

[3] Both *Recursos Extraordinários* and *Agravos em Recurso Extraordinário* have virtually the same content, especially in the context of this search effort—to automate the admissibility exam.

Once in STF, the appeal follows a screening procedure with a new assessment of admissibility, now carried out by departments linked to the Presidency. If the appeal meets the requirements, it goes to the Justices' offices for further processing.

Historically, extraordinary appeals have posed a bottleneck in the Court's workload. Since 2006, the Court has received almost 1 million constitutional appeals (Brazilian Supreme Court, n.d.). A significant reduction took place from 2008, when 10.5 thousand appeals were received by the Court, as opposed to the 60 thousand received in 2006. The number increased again in 2012, varying between 60 thousand and 80 thousand yearly appeals. More specifically, in 2022, the STF received 49,533 Extraordinary Appeals and interlocutory appeals; in 2023, the number was 54,974 (Brazilian Supreme Court, 2024).

3.2 Topics of general repercussion

In response to the "crisis of the Brazilian Supreme Court," the Extraordinary Appeal underwent significant changes. Constitutional Amendment 45/2004 introduced a new requirement for these appeals: the General Repercussion (GR) of the constitutional issues discussed in the case. The purpose of this change is twofold. First, GR prescribes that constitutional matters should be relevant. Second, cases decided under GR should form a strong precedent to prevent STF from discussing the same subject twice.

In the 15 years since this change, enforcement and adjudication procedures have been improved, starting with procedural changes enacted by the Regimental Amendment No. 21/07. The General Repercussion was quickly associated with groups of similar appeals and the judgment of paradigmatic cases. Each Topic of General Repercussion[4] has a paradigmatic case selected by the lower courts. The decision established under GR should be applied promptly by the lower courts and can be applied individually by justices. The GR also made it possible to suspend procedures identical to those of Extraordinary Appeals accepted by the STF. This measure allowed lower courts to await the final judgment and stop sending similar cases to the Supreme Court.

Along with the GR, significant technological advances occurred in the STF, such as creating the Virtual Plenary. Initially, it was used for justices to express their opinion on the existence of General Repercussions. New features were incorporated, such as the reaffirmation of consolidated jurisprudence, making the system more transparent and allowing discussions to be monitored in real time.[5]

Gradually, STF expanded Virtual Plenary, allowing justices to assess the merits of cases. Statistics released by the Court since 2020 reflect this expansion: the STF created a website dedicated to demonstrating the benefits of the Virtual Plenary (Brazilian Supreme Court, n.d.), showing that in 2021 and 2022 more than 98% of cases, i.e., more than 27 thousand decisions, were judged in a virtual environment.

[4] In Portuguese, it is called "Tema de Repercussão Geral," with a literal translation "Theme of General Repercussion." However, STF has been using the "Topic" translation in its English publications in the last years (Brazilian Supreme Court, 2022a, 2022b).

[5] Ongoing judgments on general repercussions can be followed on the STF's own website, available at https://portal.stf.jus.br/jurisprudenciaRepercussao/tesesJulgamento.asp.

Despite implementing the electronic lawsuit and the Virtual Plenary, the volume of constitutional appeals still soared. To curb this, the STF decided to strengthen the General Repercussion mechanism. First, the number of topics soared: in January 2024, the STF recorded 1,294 diferent topics (Brazilian Supreme Court, n.d.). Fig. 1 illustrates the number of topics admitted by the STF annually. As of March 13, 2024, four new topics have been admitted this year. Additionally, nearly 50% of the topics were admitted between 2008 and 2013. The great diversity of topics made it complex to assess whether each representative appeal fits into a previous topic or gives rise to a new one.

In 2016, the STF made changes to its internal organization. Resolution No. 586/2016 created a screening sector that should, among other things, check if a constitutional appeal is discussing matters already decided under a topic of GR. If this is the case, the constitutional appeal faces an abbreviated conclusion, and it is not forwarded to justices' offices. This organizational change was highly successful, and recent reports show that up to 60 percent of constitutional appeals end in this early evaluation (Brazilian Supreme Court, n.d.).

The Victor System (Araujo et al., 2020), launched in 2018, is an AI tool developed to assist in screening lawsuits in this early phase. Created as a joint work between the startup Legal Labs and the University of Brasília, Victor analyzes appeals sent to the Court, classifying them into existing topics based on textual analysis. Initially, it categorized 27 topics (Brazilian Supreme Court, 2021), reducing the average screening time from 44 minutes to 5 seconds (Prescott and Mariano, 2019). Although composing the series of technological innovations brought by the Court to improve efficiency, Victor represents a new chapter as it is the first assistive technology. As we will discuss in Sect. 2, despite its effectiveness, Victor has limitations, especially in presenting justifications for its decisions and a user-friendly interface.

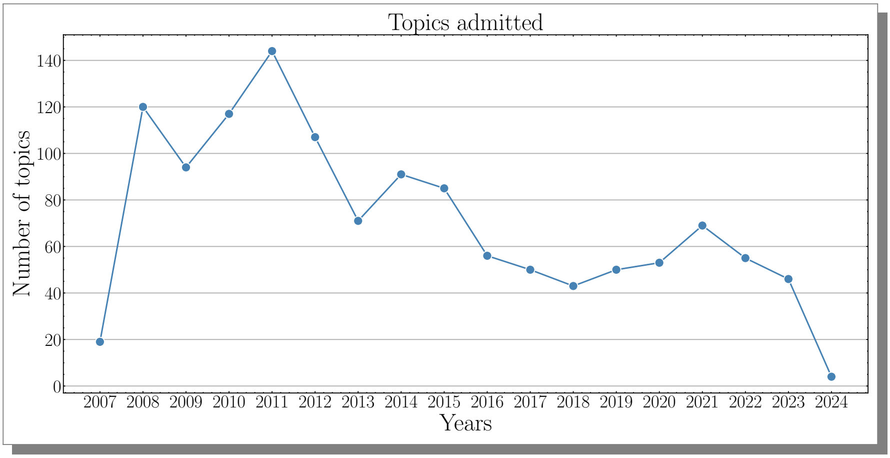


Fig. 1 Number of topics by year of admission. As of March 13, 2024, we note that four new topics have been admitted this year. Source: Adapted from "Plataforma Corte Aberta do STF" (Brazilian Supreme Court, n.d.)

3.3 Application of topics by STF

A **topic**, in the context of REs, refers to the grouping of similar appeals according to the legal issue at stake. These topics, expressed in short texts, highlight the legal controversy involved in a process. Topic 1, for example, has the following formulation: "Base for calculating PIS and COFINS on imports." Its description, in turn, can be accessed on the STF portal.[6] Topic 1 groups all the appeals that discuss, to some extent, the basis for calculating taxes on imports, questioning the respective law. As it was judged by the STF, there is already a solution to the corresponding legal problem. This solution is called **thesis**,[7] which must be applied by the lower courts and can be applied by STF justices in individual appeals.

For the sake of clarity, the procedures discussed above can be summarized as follows:

General Repercussion (GR). It is a requirement for the admissibility (acceptance) of Extraordinary Appeals, which must present an important legal issue that transcends the individual process.

Topic of GR. It is a short text that summarizes the legal issue submitted for judgment with GR, allowing to group similar appeals under it.

Representative appeal of the controversy. It is the appeal that represents a set of similar appeals. Once judged, it allows the application of the understanding to the set of appeals.

Thesis. It is the understanding that the STF has on the legal issue that must be applied to all cases that fall under the same topic, whether before or after the trial. This understanding may include the application of the law, its unconstitutionality, or its interpretation, among other things.

Súmula Vinculante. Súmulas vinculantes are the strongest precedent in the Brazillian system. They bind every jurisdiction and governmental authority, while general repercussions bind only the judiciary. Decisions made against a *súmula vinculante* find a fast track to be challenged in STF, aggravating the court overload. Because of this, STF has been very shy in enacting newer *súmulas vinculantes*, practically stopping doing so after the advent of general repercussion.

According to the Open Court platform (Brazilian Supreme Court, n.d.), as of March 13, 2024, 854 out of 1,294 topics currently recognized by the STF had GRs confirmed, and only two are currently being analyzed by the Court (Fig. 2-a). Then, 711 out of the 854 topics with recognized GR have already been judged, and 143 cases were or are subject to the appeal process, which represents a source of controversy (Fig. 2-b). Of the 711 already judged, 555 are on the merits, i.e., when there is a discussion on the legal issue (upholding or not granting), and the remaining 156 need a reconfirmation. In total, 1149 paradigm processes have already become final;

[6] Available at https://portal.stf.jus.br/jurisprudenciaRepercussao/verAndamentoProcesso.asp?incidente=2549049&numeroProcesso=559937 &classeProcesso=RE&numeroTema=1.

[7] For Topic 1, the thesis is: "The part of art. 7th, I, of Law 10,865/2004, which adds to the calculation basis of the so-called PIS/COFINS-Importation the value of ICMS levied on customs clearance and the value of the contributions themselves."

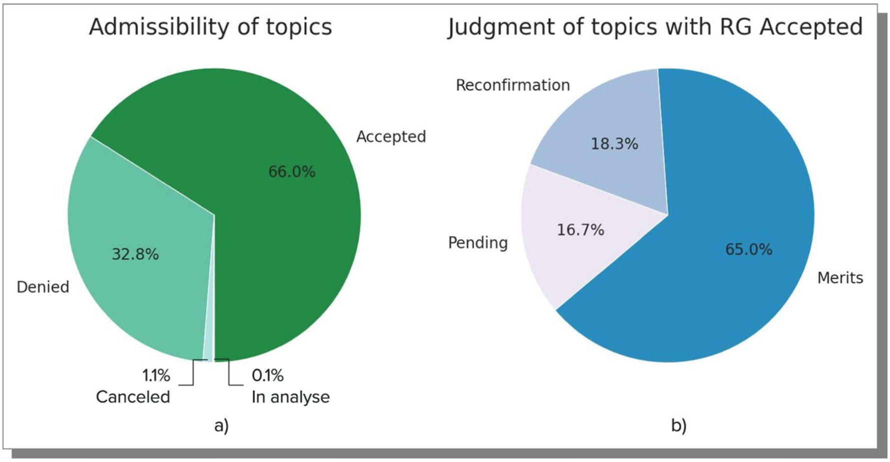


Fig. 2 **a** Shows the current proportions of accepted, denied, canceled, and under-analysis topics in STF, **b** Depicts the proportions of topics with recognized General Repercussion and the pending and under reconfirmation ones. Source: Adapted and translated from "Plataforma Corte Aberta do STF" (Brazilian Supreme Court, n.d.)

that is, they have definitively ended, with the average time for a decision being 3 years and 9 months.

3.4 STF's pipeline of extraordinary appeals

An Extraordinary Appeal undergoes a four-stage pipeline when arriving in STF, namely:

1st stage - Analysis of the objective requirements of the appeals. STF public servants analyze whether the appeal meets objective requirements, such as respect for the deadline, payment of fees, quality and intelligibility of the procedural document, etc. If an appeal does not pass this stage, it is denied, and the case is returned to the lower court.

2nd stage - Analysis of the existence of recognized General Repercussion. A specific department of the Court analyzes whether the appeal discusses a legal issue that corresponds to any GR Topics. Constitutional appeals form complex narratives, sometimes portraying more than one subject. This means that more than one topic can be applied. The combination of applicable topics is also variable.

We should note that the Victor system (Araujo et al., 2020) was designed to help at this stage, screening and analyzing the text in search of similarities with previous topics. If an appeal does not pass this stage because it fits into an existing issue, the STF returns the case to the lower court to await judgment (if the case has yet to be judged) or apply the understanding (if it already exists). If both scenarios, the procedure ends quickly, at least from the perspective of STF.

No topics being recognized means that there is a new controversy at hand. In this scenario, STF describes the controversy as a new topic, and the appeal goes on for further processing.

3rd stage - Analysis of the existence of consolidated jurisprudence or precedent applicable to the case. In the third stage, the public servants of the STF analyze whether the appeal discusses a legal issue that has already been resolved by the Court through a precedent[8] (for example, a binding precedent) or dominant jurisprudence. If an appeal fails to progress at this stage due to its prior resolution, it is deemed already adjudicated and returned to the court of origin.

The main difference between the second and the third stage is appealability. There is no appeal for rulings that establish that a topic is applicable: parties must wait for the lower courts to issue a new ruling in such cases. This means that the workload is redirected to lower courts, ending the tasks for STF. Rulings made in the third stage, although final, are appealable. Moreover, the STF must rule on the appeal.

This explains why STF is pushing for more general repercussion topics, as it allows for a leaner ruling that implicates less workload for the Court.

4th stage - Normal processing of the appeal by the offices. The appeal goes to the justices' offices to undergo a new admissibility examination and follow the standard processing, which consists of assigning a new topic, judgment in the Virtual Plenary regarding the GR, existence of a constitutional issue, and reaffirmation of consolidated jurisprudence, and later the judgment on the merits.

The number of appeals returned from the second stage is significant. To give an example, in 2021, 15,182 appeals were returned to the lower courts, 48% (7,285 processes) due to the existence of general repercussions in just 10 topics.[9] The 2020 Activity Report (Brazilian Supreme Court, 2021) provides even greater detail on the screening process and its results, showing that ten hypotheses were responsible for most denied appeals.

3.5 System requirements

Section 3.4 presents the four stages that an incoming Extraordinary Appeal faces when arriving at the STF. LegalAnalytics aims to assist in the second stage, in which the existence of a Topic of General Repercussion is verified.

Developing software that incorporates artificial intelligence and data visualization to assist specialists in a given area (in this case, Law) requires a detailed definition of what this software needs to account for. A set of requirements has been defined from a series of interviews with legal experts and weekly meetings with lawyers collaborating on this work, mainly accounting for stage 2. The proposed

[8] "Precedent," in this case, is the translation of "Súmula," a legal tool used in Brazilian courts to formalize court understandings, as usual precedents are not binding in civil law countries.

[9] In descending order: Topics 660, 339, 1,119, 800, 1,114, 288, 424, 913, 793, and 766. (Brazilian Supreme Court, 2022).

LegalAnalytics system has been designed and built based on these requirements, which we've included below.

Suggestion of Relevant Topics. The systems should be able to suggest the results of an admissibility exam receiving only the lawsuit documents. This means it should not require further input from users or other materials such as doctrine or jurisprudence.

In implementing this, we struggled due to the complexity of lawsuits. Especially at the latter stages, lawsuits can be made of dozens of documents. To solve this, we end up restricting our tool to General Repercussion prerequisite, as public servants in the STF, when analyzing it, consider mainly the constitutional appeal. Moreover, a ruling based on a General Repercussion topic is enough to hinder the constitutional appeal, even if it would face problems with other prerequisites, such as weak precedents (Santos, 2024). The system can rely on Extraordinary Appeal petitions documents when suggesting topics. The system must also communicate its confidence in the suggestion, emphasizing the topics with greater certainty of application.

Identification of Similar Appeals. The system must pinpoint appeals similar to the RE. Besides efficiency, one of the goals of General Repercussion is to assure equal treatment to constitutional appeals. As such, presenting similar appeals that share the conclusion suggested by the algorithm reinforces reliability. In particular, it should be possible for the user to identify how the higher court applies topics to similar cases. When topics are divergent, similar appeals should act as a form of controversy, presenting different points of view regarding the application of the topics that have been suggested. Finally, it should be possible to filter similar appeals by other characteristics such as reporting justice, date of publication, and applied topics. This allows users to explore the history of topics applications in similar cases.

Explainability of System Suggestions. Explaining the decisions of automated algorithms (in our case, the suggestion of pertinent topics and the identification of similar appeals) is of paramount importance in sensitive contexts such as legal procedures. Therefore, the system must be able to explain the reasons that led to such decisions clearly. The system must explain why a topic was suggested based on the content of the process text and why a similar process was identified based on the text of both documents. Explanations of the system should not be understood *a priori* as a reinforcement of the algorithms' decision but as an explanation of its operation to communicate confidence to the user that the system works well "for the right reasons."[10]

4 Dataset

To train the classifiers (see Sect. 5.1) and populate the system (see Sect. 6), it is essential to compile a dataset of Extraordinary Appeals previously adjudicated by the STF, along with information on the relevant topics applied to each case.

[10] In Appendix A, we verified that, yes, the explanations also reinforce the suggestion of pertinent topics, that is, the explanations are "correct."

We assembled a comprehensive dataset by collecting documents and data from 1,512,599 processes, which were then processed for use. The methodologies for data collection and processing are detailed in this section.

4.1 Data collection

To construct the database, we initiated the process by collecting all case records registered with the Presidency of the Court, utilizing the STF distribution minutes consultation tool.[11] We then accessed the electronic records of these cases, enabling the capture of structured information from the documents contained within these cases. Specifically, we collected documents whose titles have references to RE or ARE,[12] indicating cases that qualify for judgment under the General Repercussion regime. Ultimately, our collection consisted of structured information on the cases and text files of appeals in PDF format. Considering the vast volume of cases at the STF, including every document was unnecessary. Instead, we selected documents pertaining to approximately 52% of the cases adjudicated by the STF from 1959 to 2022. This effort resulted in a compilation of 1,512,599 cases with structured information and 2,803,510 PDF documents. Additionally, we gathered the STF decisions and judicial orders, which will be analyze to identify the topics applied in these cases. In summary, we collected appeal petitions, process movement, and archives of decisions and judicial orders.

4.2 Data processing

We divide the data pre-processing into four steps: process filtering, text extraction, text cleaning, and topic identification.

Processes Filtering. After collecting 1,512,599 cases and 2,803,510 PDF documents, we focused on “complete” cases. A single RE petition, a Topic of General Repercussion, and a predictable procedural sequence characterize these. Specifically, this means the appeal reaches the STF, undergoes the screening process, and has a topic subsequently applied. We filtered appeal-type cases from this dataset, pairing the appeal texts with their corresponding Supreme Court topics. This resulted in 110,812 cases, including 105,126 of the RE and ARE types.

Text Extraction. Downloaded PDFs often contain a mix of text, images, watermarks, and sensitive data, such as personal information. However, files from the STF platform lack standardization, with document formats varying by state. Some documents exceed 360 pages, and many predating 2015 suffer from poor scan quality or remain undigitized. To better understand these challenges, we manually reviewed a subset of 50 PDF documents to analyze their content. This review revealed that many documents included attached materials such as

[11] Available at https://portal.stf.jus.br/atadistribuicao/.

[12] ARE are Appeals in Extraordinary Appeals, as described in Sect. 3.1. From now on, we refer to both documents as “appeal” or “RE” only.

scanned pages, invoices, and extraneous information. Common issues included poor scan quality, digital watermarks, and footnote signatures.

To address these challenges, we used PyMuPDF to distinguish between high- and low-quality documents. We implemented a language verification step to assess whether the extracted text consisted of Portuguese words or was dominated by symbols. Using a threshold of 75%, we evaluated the proportion of Portuguese words and symbols in the extracted text per page to determine the document's quality average. Approximately 45% of the documents were classified as having poor scan quality. For high-quality documents, we used PyMuPDF to extract text directly. For poor-quality documents, we applied OCR techniques to extract the text and reevaluated its quality.

Text Cleaning. After text extraction, we identified document pages that did not meet the language threshold, even when using PyMuPDF or OCR techniques. Through manual exploration, we observed the presence of electronic or digital signatures and watermark seals within the documents. To clean the text, we manually identified various elements, including watermarks, digital signatures, home addresses (CEPs), personal identification numbers (CPFs), links, and attached images.

We then implemented regular expressions to locate and remove these elements. Comprehensive details on the regular expressions can be found in Appendix B. This process significantly improved the quality and readability of the extracted text.

Topic Identification. Information on STF-applied topics can be found in structured process records or in the texts of decisions and judicial orders. Accurately identifying topics is challenging due to variations in how decisions are written. By analyzing structured records and applying regular expressions to search decision texts, we reduced the original dataset of 105,126 RE and ARE processes to 34,028, where a topic could be identified. For a detailed discussion of the regular expressions used in this process, please refer to Appendix C.

4.3 Dataset preparation

After curating an appropriate dataset of processes and document texts, our next step was to create the labeled dataset, which consists of input texts paired with their corresponding topic labels. This requires a dataset containing only one appeal text per case or process.

Using this criterion, we reduced the dataset from 34,028 documents to 10,710. We also manually reviewed the selected documents and found that many contained irrelevant scanned content, such as images and invoices. As a result, we removed 477 documents that were longer than 50 pages. Furthermore, we analyzed the identified topics in the processes, revealing that the most frequently applied topics were 660, 800, 339, and 810. Figure 3 illustrates the distribution of the top 30 topics within this subset, highlighting a significant imbalance among the labels.

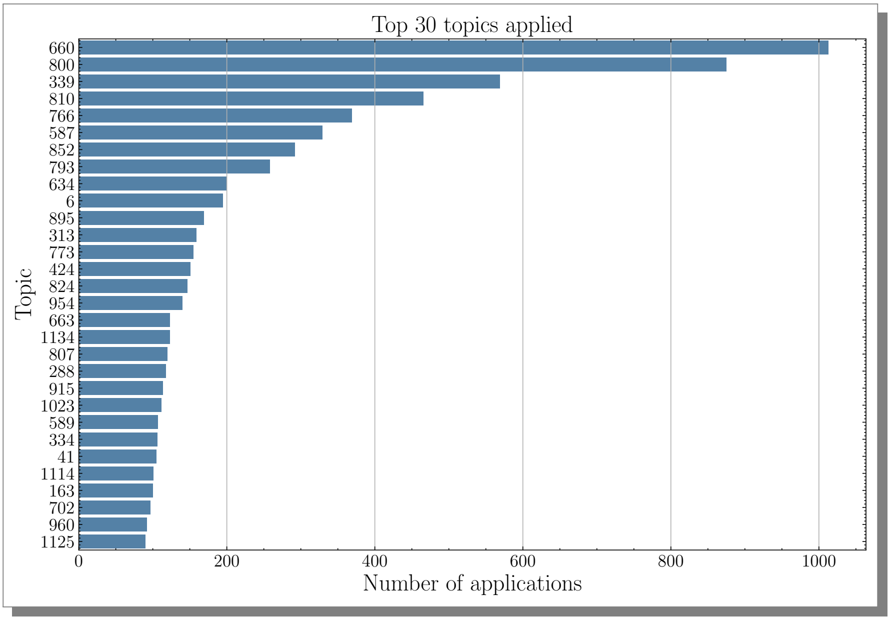


Fig. 3 Number of applications per topic (top-30 topics) in the subset of 10,233 processes. More than one topic can be applied in the same process

5 Machine learning methodology

As outlined in Sect. 3.5, a key functionality of the LegalAnalytics system is to process appeals and recommend Topics of General Repercussion using machine learning techniques. In addition, the system provides explanations for the algorithm's decisions. The following subsections provide detailed descriptions of these processes. Implementing machine learning models, their explainability components, and the analysis of similar appeals are all conducted using Python.

5.1 Topic classification

As detailed in Sect. 4, we refined a subset of processes so that each Extraordinary Appeal includes the *text of the appeal petition* as input and the *topic applied by the STF* as the target. We aim to classify these appeal texts—and, by extension, the appeal processes—into specific topics.

Given the uneven distribution of topic applications within the training subset (as shown in Fig. 3), we adopted a focused approach. We meticulously selected the 30 most frequently applied topics to train our machine learning models. This selection was based on the frequency of occurrence of each topic, ensuring that our analysis focused on the most representative topics. These topics span diverse subjects, including labor, social security, and tax issues, ensuring comprehensive coverage across key legal areas.

5.1.1 Multilabel classification

In our approach, we adopted multilabel classification (Tsoumakas and Katakis, 2007). This method differs from traditional multiclass classification, where a model is typically trained to classify an appealing text into a single topic. Instead, multilabel classification allows each topic to be classified independently, enabling multiple topics to apply simultaneously to a single appeal text or none at all. To implement this, we utilized binary classifiers for each topic. Specifically, each topic and its corresponding model were trained using the subset of data described in Sect. 4 indicating whether a specific topic applies to each process.

It is important to note that all selected topics (see Fig. 3) are relatively rare, with some represented in as few as 1% of our dataset of 10,233 processes. This scarcity poses a significant challenge for model training. However, despite these hurdles, the results discussed in Sect. 5.1.4 demonstrate that the models achieved satisfactory performance.

5.1.2 Models

The models used in our prototype comprise two main components: **vectorization** and **classification**. In the vectorization phase, texts are transformed into vectors within a high-dimensional vector space. Following this, in the classification phase, these vectors are evaluated to determine whether a specific topic applies or not. Before proceeding with vectorization using the TF-IDF technique (Term Frequency-Inverse Document Frequency) (Leskovec et al., 2020), it is essential to perform standard text preprocessing. This includes converting text to lowercase, removing punctuation and other non-standard characters (such as large numbers and extra spaces), eliminating stop words, and applying lemmatization. The decision for these steps followed a traditional text pre-processing pipeline (Resck et al., 2023) and a manual data analysis.

Vectorization employs the TF-IDF[13] method, which results in vectors with very high dimensionality (tens of thousands of coordinates). Previous work (Resck et al., 2023) has already shown that the TF-IDF method is effective for classifying Brazilian Supreme Court documents, particularly when compared to neural network-based embeddings and large language models (LLMs). Given the high performance of the TF-IDF method (Sect. 5.1.4), the data imbalance (Sect. 4), the computational cost of training more complex models such as LLMs, and the extensive experiments conducted by Resck et al. (2023) comparing these methods in the same context, we opted to use the TF-IDF method for vectorization. After vectorization, the vectors are normalized so that each has a variance of one across all documents.[14] These normalized vectors are then used as input for classification via generalized linear

[13] The number n of n-grams was chosen between 1 and 2 using cross-validation.

[14] The application (or not) of a normalization of the vectors was included in the optimization of the hyperparameters through cross-validation. Normalization did not include subtracting the mean as it hampers the vectors' sparsity.

models, specifically logistic regression (Kleinbaum and Klein, 2010), implemented using Python's scikit-learn library. Data was split into training, validation, and test sets using a 70/10/20% ratio. The split was random and stratified for each class. Validation was used for probability calibration and selection of the probability threshold used in the evaluation of discrete metrics (e.g., F1-score). With the training split, we performed 5-fold stratified cross-validation (Refaeilzadeh et al., 2009) to search for the models' optimal hyperparameters and ensure the preservation of class distributions across all folds. We used class weights to address the imbalanced dataset, ensuring that the model can learn from the minority class examples. Whether considering class weights when training the model was included in the cross-validation process. Finally, the cross-validation optimized the model's hyperparameters to maximize the Area Under the Precision-Recall Curve (AUPRC) metric, detailed in Sect. 5.1.4, which is more sensitive to imbalanced datasets than simply optimizing accuracy. The full specifications of the model, including all hyperparameters and the cross-validation process, are outlined in Appendix D.

Additionally, to enhance the accuracy of the logistic regressions' probability outputs, we employed a probability calibration technique on the independent validation subsets of the data (not used in training), specifically the method proposed by Platt (1999).

Why linear models? The decision to utilize linear models resulted from a comprehensive evaluation of other models, such as random forests and language models. Focusing on a particular topic (800), we trained a varied number of traditional machine learning models (e.g., SVM, boosting, random forest, and perceptron). Additionally, we fine-tuned a language model, BERT, pretrained in Portuguese. To overcome the limitation of BERT's input size, we also used a sliding window approach to classify the text. Among all models, the logistic regression presented one of the most competitive performances (0.95 accuracy, 0.99 AUC and 0.84 average precision) while being one of the fastest and simplest models. For complete results, refer to Appendix G. LegalAnalytics' visual interface enforces fast and small machine learning models, as the system is a web browser tool that may be deployed on a local server with no GPUs. In particular, the explainer tool used (LIME) requires thousands of model inferences in real-time for each sample to generate an explanation, which restricts the use of a light model such as logistic regression. Our findings are consistent with previous work in literature that tested similar models in similar datasets of Brazilian precedents and showed that linear models are effective for classifying Brazilian Supreme Court documents, particularly when compared to neural network-based embeddings and large language models (LLMs) (Resck et al., 2023).

5.1.3 Thresholds

Following the calibration phase, the probabilities generated by our models are used to assess the relevance of a topic to a particular process within the LegalAnalytics interface. Specifically, these probabilities help to categorize the topics as **very relevant** (high probability), **relevant** (medium probability), or **not very relevant** (low probability). To establish precise thresholds for these categories, we conducted a detailed analysis focusing on maintaining precision levels of 90% for "very relevant" and 50% for "relevant" across all 30 models. This approach ensures that when a topic is identified as "very relevant", there is a 90% probability that this assessment is accurate (similarly, a 50% likelihood for "relevant"). Based on this analysis, we set the probability thresholds for each classifier. For instance, Topic 766 has 65% for "very relevant" and 20% for "relevant". This methodical setting of thresholds aims to optimize the accuracy and utility of the system in practical applications.

The selected thresholds are intended solely for visualization purposes and do not influence model performance or interpretability. They serve as illustrative benchmarks to demonstrate the models' output and performance clarity for users. The overarching goal is to ensure the model achieves a level of performance that is meaningful and interpretable for end-users.

5.1.4 Results

We used each of the 30 selected topics to train a model and then evaluated these models on test subsets. Subsequently, we calibrated the models. Figure 4 shows the final performance of each model in the test data, sorted by the average precision metric. The key metrics presented in the graphs include the following:

- **Average precision (AUPRC)**: This metric represents the area under the precision-recall curve by varying the model's decision threshold. It quantifies the balance between the model's accuracy in predicting positive instances (precision) and its ability to identify all actual positive cases (recall). The closer this metric is to 100%, the more effective the model. For context, a random classifier typically scores between 1% and 10% on this metric due to data imbalance.
- **Area under the ROC curve (AUC)**: This metric is the area under the receiver operating characteristic curve, which plots the true positive rate against the false positive rate at various threshold settings. It measures the model's ability to differentiate between the application (positive) and non-application (negative) of a topic, with a random classifier achieving an AUC of 50%.

The data shown in Fig. 4 reveal that most of the models achieved high AUPRC scores and all exhibited high AUC values. Although a few models recorded lower AUPRC scores, they were significantly higher than those of a random classifier and maintained very high AUC values. This indicates that all models effectively distinguish between positive and negative cases and most accurately identify positive instances. The results are particularly encouraging for the simpler models that used

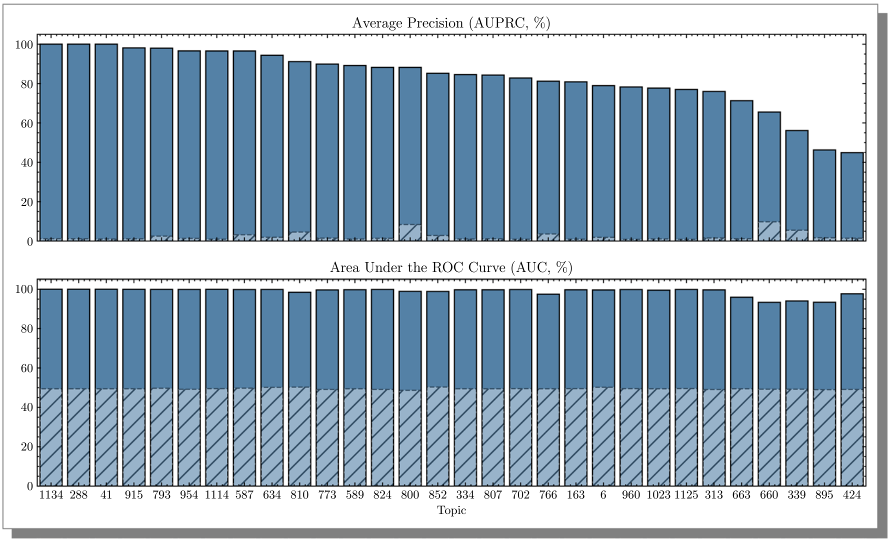


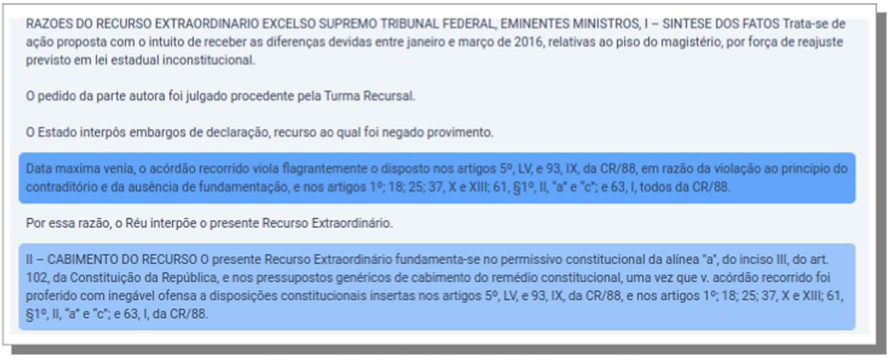
Fig. 4 Models performance for the 30 selected topics under the AUPRC and AUC metrics, sorted by AUPRC. The blue bars represent the performance value, and the hashed bars indicate the baseline performance (dummy classifier with 'stratified' strategy; refer to scikit-learn's `DummyClassifier`'s strategies: https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html).

bag-of-words vectorization and classification through generalized linear models (Sect. 5.1.2). These models performed exceptionally well, achieving AUPRC scores above 80% and, in many instances, nearing 100%. This high level of performance is notable, especially considering the scarcity of data—some models were trained with as few as 60 positive examples. Other studies have documented similar success stories with bag-of-words models outperforming more complex algorithms (Resck et al., 2023; Domingues, 2021; Avinash and Sivasankar, 2019; Zhu et al., 2016; Wang et al., 2017). Additionally, these simpler models benefit from fast execution times, facilitating easier implementation and explanation of results (see Sect. 5.2). For comprehensive details on the test results of all models, please see Appendix F.

5.2 Explainability

A mere suggestion from a machine learning model about the applicable topics for a process is insufficient. As stipulated in the system requirements (Sect. 3.5), the algorithm decision must include an explanation for legal professionals to utilize this technique effectively. For instance, from a text containing dozens or hundreds of paragraphs, it is crucial to pinpoint the most significant paragraphs that influence the model's suggestion to apply a specific topic (as illustrated in Fig. 5).

Several approaches are possible to explain the decisions of text classification models. In this study, we employ the *Local Interpretable Model-agnostic Explanations (LIME)* (Ribeiro et al., 2016). LIME operates by randomly omitting paragraphs from the input text and observing changes in the model's output



RAZOES DO RECURSO EXTRAORDINARIO EXCELSO SUPREMO TRIBUNAL FEDERAL, EMINENTES MINISTROS, I – SINTESE DOS FATOS Trata-se de ação proposta com o intuito de receber as diferenças devidas entre janeiro e março de 2016, relativas ao piso do magistério, por força de reajuste previsto em lei estadual inconstitucional.

O pedido da parte autora foi julgado procedente pela Turma Recursal.

O Estado interpôs embargos de declaração, recurso ao qual foi negado provimento.

Data maxima venia, o acórdão recorrido viola flagrantemente o disposto nos artigos 5º, LV, e 93, IX, da CR/88, em razão da violação ao princípio do contraditório e da ausência de fundamentação, e nos artigos 1º; 18; 25; 37, X e XIII; 61, §1º, II, "a" e "c"; e 63, I, todos da CR/88.

Por essa razão, o Réu interpõe o presente Recurso Extraordinário.

II – CABIMENTO DO RECURSO O presente Recurso Extraordinário fundamenta-se no permissivo constitucional da alínea "a", do inciso III, do art. 102, da Constituição da República, e nos pressupostos genéricos de cabimento do remédio constitucional, uma vez que v. acórdão recorrido foi proferido com inegável ofensa a disposições constitucionais insertas nos artigos 5º, LV, e 93, IX, da CR/88, e nos artigos 1º; 18; 25; 37, X e XIII; 61, §1º, II, "a" e "c"; e 63, I, da CR/88.

Fig. 5 Example of presenting the explanation using a color scale. The darker the blue, the more important the paragraph is for the model's decision to suggest the application of a certain Topic of General Repercussion. In this example, the topic in question is the 339. The document snippet is in Brazilian Portuguese but its content is not relevant for the understanding of the explanation

probabilities. It then constructs a simpler, interpretable linear regression model to predict the model's output based on the presence of paragraphs. From this, we derive linear regression coefficients, which serve as importance scores for each paragraph in relation to the model's final prediction. These importance scores are visually represented using a color scale on the document background, highlighting the paragraphs that significantly impact the model's decision (see Fig. 5).

LIME offers several advantages over traditional interpretation methods, such as analyzing the coefficients of logistic regression classifiers (Sect. 5.1.2). One of its key strengths is its ability to provide adjustable granularity in explanations. Users can choose to focus on words, paragraphs, or sentences, simply by selecting the level of perturbation detail. In addition, LIME provides individualized explanations for each output, facilitating tailored explanations for each case. Being model-agnostic, LIME does not depend on a specific model architecture, which simplifies system implementation and facilitates the application of this methodology to various types of models. Compared to similar explanation methods like SHAP (Lundberg and Lee, 2017), we found LIME easier to implement despite their theoretical equivalence in certain contexts (Lundberg and Lee, 2017). The choice of LIME was also guided by previous work (Resck et al., 2023). In our system, LIME highlights 14 paragraphs for explanation, focusing on those with "positive" importance, meaning that they positively influence the model's decision. On average, about 10 paragraphs are highlighted. For a detailed evaluation of these explanations, please refer to Appendix A.

We acknowledge that simply highlighting features with a high impact on a classification result represents a relatively low bar for defining explanations for end users. However, this practice remains widely accepted in prior work as a reasonable means of offering interpretability in text classification tasks (Chan et al., 2022; Resck et al., 2024). Particularly in the legal domain, LIME's paragraph-level explanation mimics how public servants analyze and annotate legal documents, making it a suitable choice for our system.

5.3 Similar appeals

After proposing a topic application and explaining this suggestion, the system must also identify similar appeals to the one being analyzed. This feature is crucial because it either supports the algorithmic decision or highlights discrepancies, enhancing the decision-making process.

To find similar appeals, we start by converting the text of the appeal under analysis and those in the database into vector representations using TF-IDF vectorization (Leskovec et al., 2020) with 30,000 features, including n-grams between 1 and 2. This is followed by dimensionality reduction through Truncated Singular Value Decomposition (SVD) (Halko et al., 2011) to 200 components and subsequent scaling and normalization. These vectorization steps are implemented using the scikit-learn library (Pedregosa et al., 2011).

The method we use to measure the similarity between appeals is straightforward and effective. We calculate the cosine similarity metric (Cer et al., 2018) between their vector representations; appeals with vectors that form a smaller angle are considered to be more similar. The system compares the vector of the appeal under analysis against all vectors in the database and identifies the ten most similar appeals, providing a clear and concise result.

We provide detailed comparisons between the documents to elucidate why certain appeals are deemed similar. This involves generating vector representations for individual paragraphs from both the analyzed appeal and each similar appeal. We compute the cosine similarity for each pair of paragraphs—one from the analyzed appeal and one from a similar appeal—and organize the results into a similarity matrix. Using a Linear Sum Assignment optimizer, we determine the most closely matched pairs of paragraphs, ensuring no paragraph is repeated within the same document. We apply heuristics to discard paragraph pairs with similarity scores below 0.85 and paragraphs shorter than 80 characters, focusing only on the most relevant comparisons.

6 LegalAnalytics

The previous sections detailed the motivation and technical description of our methodology. This section presents a system prototype that employs the described methodology, highlighting a carefully designed user interface. The interface consists of three main screens: a search screen and two screens for analyzing a selected process, encompassing the system functionalities discussed earlier. We detail each screen and its components below (for implementation details, refer to Appendix E). All system screen figures are in English for a broader audience; for the original screens in Brazilian Portuguese, refer to Appendix B. Each screen includes a feedback button, allowing users to report any issues or suggestions for improvement.

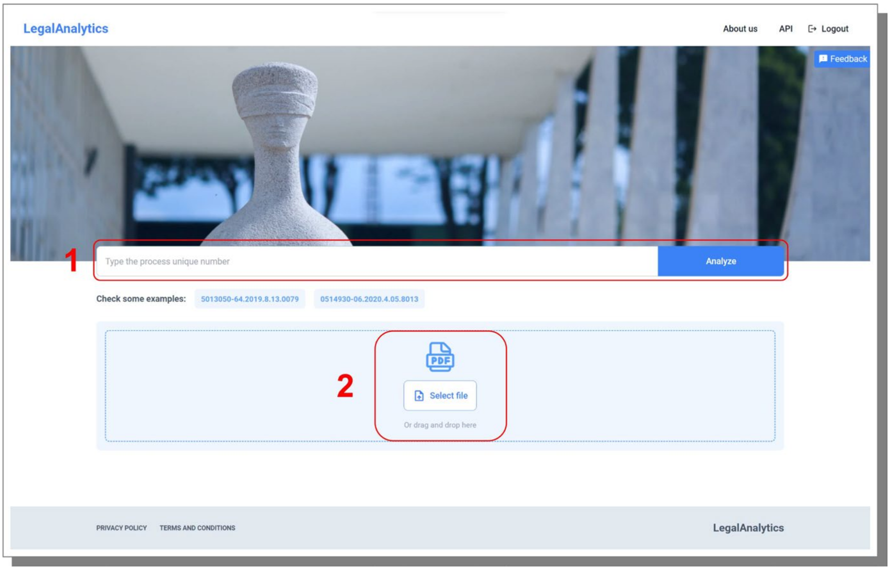


Fig. 6 Appeal Selection screen and its components

6.1 Appeal selection

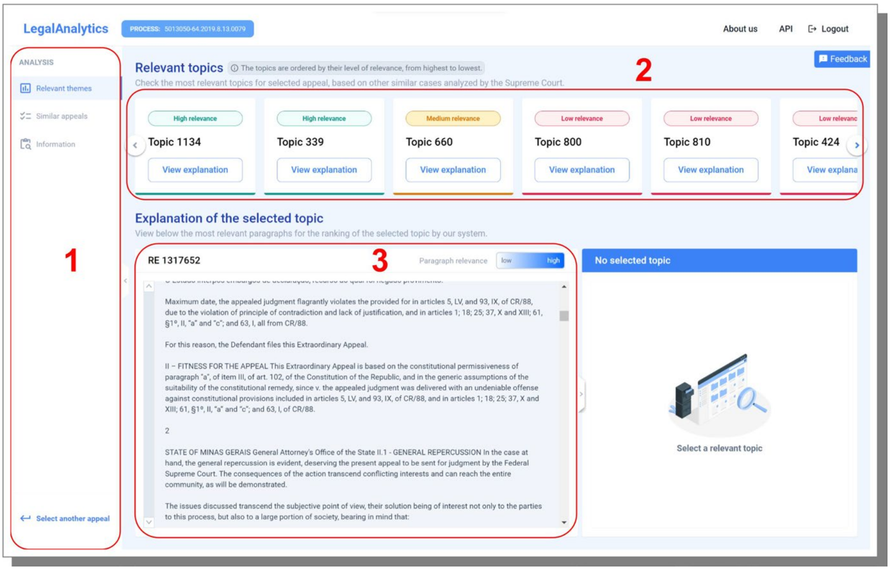
On the Appeal Selection screen, the judicial expert (from now on referred to as the user) selects the appeal they wish to analyze (see Fig. 6).

The main components of this screen are as follows:

1. **Search bar**: The user enters the unique process number to search for and analyze the desired process appeal. If the search is successful, the user is redirected to the *Relevant Topics* screen (Sect. 6.2).
2. **PDF upload**: The user clicks the upload button, which opens the file selection window in their browser. The selected file must be an RE petition in PDF format. If the file is processed successfully, the user is redirected to the *Relevant Topics* screen.

6.2 Relevant topics

Upon the user's active selection of the desired appeal from the previous screen, they are directed to the Relevant Topics screen. This screen presents the appeal's classification result into the 30 Topics of General Repercussion (Sect. 5.1) and includes an explanation component.



Relevant topics

The topics are ordered by their level of relevance, from highest to lowest.

Check the most relevant topics for selected appeal, based on other similar cases analyzed by the Supreme Court.

Explanation of the selected topic

View below the most relevant paragraphs for the ranking of the selected topic by our system.

RE 1317652

Maximum date, the appealed judgment flagrantly violates the provided for in articles 5, LV, and 93, IX, of CR/88, due to the violation of principle of contradiction and lack of justification, and in articles 1; 18; 25; 37, X and XIII; 61, §1°, II, "a" and "c"; and 63, I, all from CR/88.

For this reason, the Defendant files this Extraordinary Appeal.

II – FITNESS FOR THE APPEAL This Extraordinary Appeal is based on the constitutional permissiveness of paragraph "a", of item III, of art. 102, of the Constitution of the Republic, and in the generic assumptions of the suitability of the constitutional remedy, since v. the appealed judgment was delivered with an undeniable offense against constitutional provisions included in articles 5, LV, and 93, IX, of CR/88, and in articles 1; 18; 25; 37, X and XIII; 61, §1°, II, "a" and "c"; and 63, I, of CR/88.

2

STATE OF MINAS GERAIS General Attorney's Office of the State II.1 - GENERAL REPERCUSSION In the case at hand, the general repercussion is evident, deserving the present appeal to be sent for judgment by the Federal Supreme Court. The consequences of the action transcend conflicting interests and can reach the entire community, as will be demonstrated.

The issues discussed transcend the subjective point of view, their solution being of interest not only to the parties to this process, but also to a large portion of society, bearing in mind that:

Fig. 7 Example of a Relevant Topics screen, in which the user chose the appeal with unique process number 5013050-64.2019.8.13.0079 and appeal number 1317652

6.2.1 Result of relevant topics

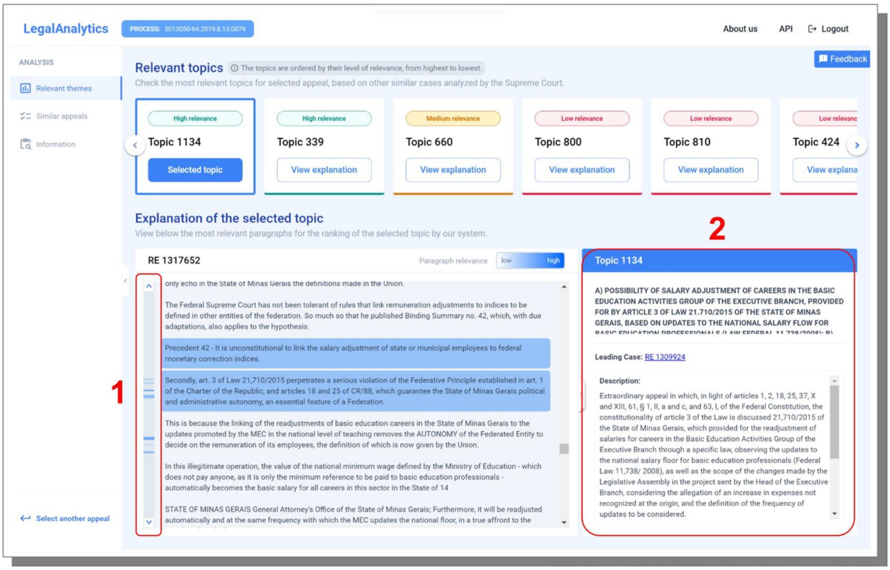
The relevant and suggested topics are displayed and ordered according to their degree of relevance (Fig. 7).

The main components of this screen are:

1. **Side Navigation Bar**: This bar allows users to navigate between relevant topic suggestions, similar appeals, and search for other processes (returning to the initial Appeal Selection screen).
2. **Carousel of Topics**: This feature presents a list of inferred topics for the appeal, ordered from most relevant (highest returned probability) to least relevant. The user has the power to select a topic of interest, enabling them to delve deeper into the model explanation.
3. **Process Content**: Contains the appeal's content. The header includes the appeal number ID (e.g., in Fig. 7, the number ID is 1317652). The file name is displayed in the header if a document is uploaded in PDF format.

6.2.2 Explanation of the selected topic

When a topic is selected on the Relevant Topics screen, several components update to display the model explanation. Relevant excerpts from the appeal are highlighted with colors (see Fig. 8). After selecting a topic, the main components of the screen are:



Relevant topics

The topics are ordered by their level of relevance, from highest to lowest.

Check the most relevant topics for selected appeal, based on other similar cases analyzed by the Supreme Court.

Explanation of the selected topic

View below the most relevant paragraphs for the ranking of the selected topic by our system.

RE 1317652

only echo in the State of Minas Gerais the definitions made in the Union.

The Federal Supreme Court has not been tolerant of rules that link remuneration adjustments to indices to be defined in other entities of the federation. So much so that he published Binding Summary no. 42, which, with due adaptations, also applies to the hypothesis:

Precedent 42 - It is unconstitutional to link the salary adjustment of state or municipal employees to federal monetary correction indices.

Secondly, art. 3 of Law 21,710/2015 perpetrates a serious violation of the Federative Principle established in art. 1 of the Charter of the Republic, and articles 18 and 25 of CR/88, which guarantee the State of Minas Gerais political and administrative autonomy, an essential feature of a Federation.

This is because the linking of the readjustments of basic education careers in the State of Minas Gerais to the updates promoted by the MEC in the national level of teaching removes the AUTONOMY of the Federated Entity to decide on the remuneration of its employees, the definition of which is now given by the Union.

In this illegitimate operation, the value of the national minimum wage defined by the Ministry of Education - which does not pay anyone, as it is only the minimum reference to be paid to basic education professionals - automatically becomes the basic salary for all careers in this sector in the State of 14

STATE OF MINAS GERAIS General Attorney's Office of the State of Minas Gerais; Furthermore, it will be readjusted automatically and at the same frequency with which the MEC updates the national floor, in a true affront to the

Topic 1134

A) POSSIBILITY OF SALARY ADJUSTMENT OF CAREERS IN THE BASIC EDUCATION ACTIVITIES GROUP OF THE EXECUTIVE BRANCH, PROVIDED FOR BY ARTICLE 3 OF LAW 21.710/2015 OF THE STATE OF MINAS GERAIS, BASED ON UPDATES TO THE NATIONAL SALARY FLOW FOR BASIC EDUCATION PROFESSIONALS (LAW FEDERAL 11.738/2008); B)

Leading Case: RE 1309924

Description:

Extraordinary appeal in which, in light of articles 1, 2, 18, 25, 37, X and XIII, 61, § 1, II, a and c, and 63, I, of the Federal Constitution, the constitutionality of article 3 of the Law is discussed 21,710/2015 of the State of Minas Gerais, which provided for the readjustment of salaries for careers in the Basic Education Activities Group of the Executive Branch through a specific law, observing the updates to the national salary floor for basic education professionals (Federal Law 11,738/ 2008), as well as the scope of the changes made by the Legislative Assembly in the project sent by the Head of the Executive Branch, considering the allegation of an increase in expenses not recognized at the origin, and the definition of the frequency of updates to be considered.

Fig. 8 Screen of Relevant Topics but with the topic furthest to the left (i.e. most pertinent) selected for a detailed analysis

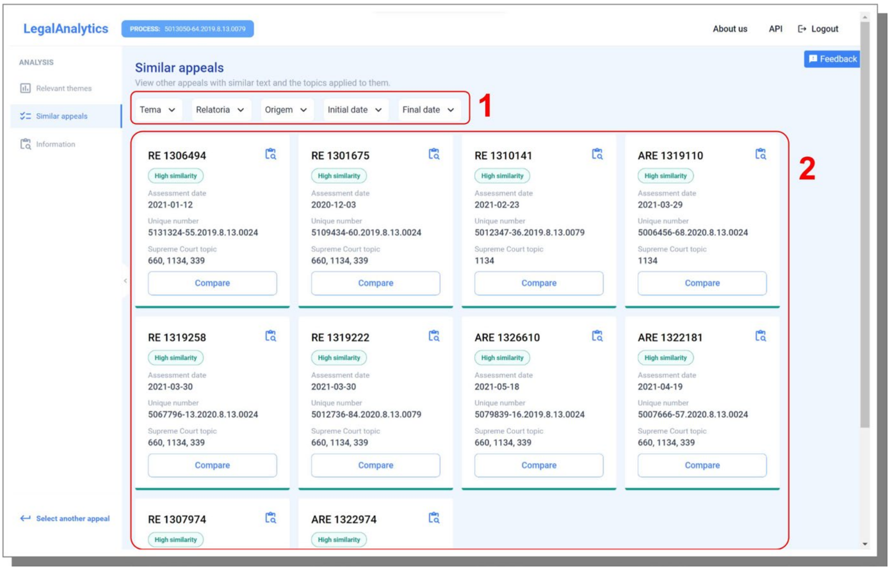
1. **Quick Paragraph Navigation Bar**: This bar identifies the most relevant paragraphs in the document. Stronger colors indicate higher relevance. The corresponding paragraphs are aligned for easy viewing when the user clicks on a highlighted region.
2. **Information About the Selected Topic**: This section details the selected topic, including the title, a link to the Leading Case on the STF website, the description, and the thesis (Sect. 3.3).

6.3 Similar appeals

The side navigation bar on the Relevant Topics screen guides users to the Similar Appeals screen. This screen presents the results of identifying similar appeals, provides filters for refining searches, and includes a visual component for comparing similar documents.

6.3.1 Results from similar appeals

On this screen, we display the functionality of similar documents by presenting a list of other processes whose RE text is similar to the RE text selected for analysis (see Fig. 9). The main components of this screen are:



Similar appeals

View other appeals with similar text and the topics applied to them.

RE 1306494 — High similarity — Assessment date 2021-01-12 — Unique number 5131324-55.2019.8.13.0024 — Supreme Court topic 660, 1134, 339

RE 1301675 — High similarity — Assessment date 2020-12-03 — Unique number 5109434-60.2019.8.13.0024 — Supreme Court topic 660, 1134, 339

RE 1310141 — High similarity — Assessment date 2021-02-23 — Unique number 5012347-36.2019.8.13.0079 — Supreme Court topic 1134

ARE 1319110 — High similarity — Assessment date 2021-03-29 — Unique number 5006456-68.2020.8.13.0024 — Supreme Court topic 1134

RE 1319258 — High similarity — Assessment date 2021-03-30 — Unique number 5067796-13.2020.8.13.0024 — Supreme Court topic 660, 1134, 339

RE 1319222 — High similarity — Assessment date 2021-03-30 — Unique number 5012736-84.2020.8.13.0079 — Supreme Court topic 660, 1134, 339

ARE 1326610 — High similarity — Assessment date 2021-05-18 — Unique number 5079839-16.2019.8.13.0024 — Supreme Court topic 660, 1134, 339

ARE 1322181 — High similarity — Assessment date 2021-04-19 — Unique number 5007666-57.2020.8.13.0024 — Supreme Court topic 660, 1134, 339

RE 1307974 — High similarity

ARE 1322974 — High similarity

Fig. 9 Similar Appeals screen without any search filter defined

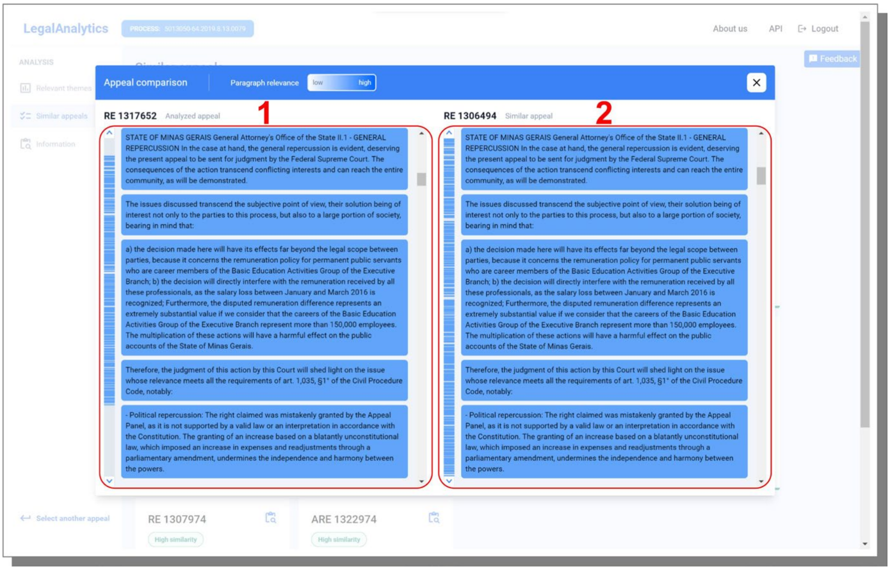
1. **Search Filters**: These filters narrow the search for recommended appeals. Filters include topic number, rapporteur judge, origin (federative unit), start date, and end date.
2. **Similar Appeals**: A list of appeals that resemble the analyzed appeal and match the user-defined filters.

    (a) Each similar appeal is displayed on a card containing basic information such as the name of the RE document, date of assessment, unique process number, and STF-applied topics.
    (b) The user can click "compare" on an appeal card to open a modal that compares the analyzed text with the text on the card (similar appeal) and explains the similarity.

6.3.2 Comparison with similar appeals

On the Similar Appeals screen, when a similar appeal is selected for comparison, the system displays a modal highlighting the similarity explanation's functionality. It indicates which parts of the compared appeals contain similar paragraphs (see Fig. 10). After selecting a similar appeal, the main components of the screen are:

1. **Appeal Being Analyzed**: This component helps users quickly identify similar paragraphs in both documents. It has two subcomponents:



RE 1317652 Analyzed appeal

STATE OF MINAS GERAIS General Attorney's Office of the State II.1 - GENERAL REPERCUSSION In the case at hand, the general repercussion is evident, deserving the present appeal to be sent for judgment by the Federal Supreme Court. The consequences of the action transcend conflicting interests and can reach the entire community, as will be demonstrated.

The issues discussed transcend the subjective point of view, their solution being of interest not only to the parties to this process, but also to a large portion of society, bearing in mind that:

a) the decision made here will have its effects far beyond the legal scope between parties, because it concerns the remuneration policy for permanent public servants who are career members of the Basic Education Activities Group of the Executive Branch; b) the decision will directly interfere with the remuneration received by all these professionals, as the salary loss between January and March 2016 is recognized; Furthermore, the disputed remuneration difference represents an extremely substantial value if we consider that the careers of the Basic Education Activities Group of the Executive Branch represent more than 150,000 employees. The multiplication of these actions will have a harmful effect on the public accounts of the State of Minas Gerais.

Therefore, the judgment of this action by this Court will shed light on the issue whose relevance meets all the requirements of art. 1,035, §1° of the Civil Procedure Code, notably:

- Political repercussion: The right claimed was mistakenly granted by the Appeal Panel, as it is not supported by a valid law or an interpretation in accordance with the Constitution. The granting of an increase based on a blatantly unconstitutional law, which imposed an increase in expenses and readjustments through a parliamentary amendment, undermines the independence and harmony between the powers.

RE 1306494 Similar appeal

STATE OF MINAS GERAIS General Attorney's Office of the State II.1 - GENERAL REPERCUSSION In the case at hand, the general repercussion is evident, deserving the present appeal to be sent for judgment by the Federal Supreme Court. The consequences of the action transcend conflicting interests and can reach the entire community, as will be demonstrated.

The issues discussed transcend the subjective point of view, their solution being of interest not only to the parties to this process, but also to a large portion of society, bearing in mind that:

a) the decision made here will have its effects far beyond the legal scope between parties, because it concerns the remuneration policy for permanent public servants who are career members of the Basic Education Activities Group of the Executive Branch; b) the decision will directly interfere with the remuneration received by all these professionals, as the salary loss between January and March 2016 is recognized; Furthermore, the disputed remuneration difference represents an extremely substantial value if we consider that the careers of the Basic Education Activities Group of the Executive Branch represent more than 150,000 employees. The multiplication of these actions will have a harmful effect on the public accounts of the State of Minas Gerais.

Therefore, the judgment of this action by this Court will shed light on the issue whose relevance meets all the requirements of art. 1,035, §1° of the Civil Procedure Code, notably:

- Political repercussion: The right claimed was mistakenly granted by the Appeal Panel, as it is not supported by a valid law or an interpretation in accordance with the Constitution. The granting of an increase based on a blatantly unconstitutional law, which imposed an increase in expenses and readjustments through a parliamentary amendment, undermines the independence and harmony between the powers.

RE 1307974

ARE 1322974

Fig. 10 Similar Appeals screen when the user chooses to compare the analyzed appeal (left side) with a similar appeal (right side). The quick navigation bar to the left of each text indicate there are several similar paragraphs between both appeals

(a) **Document Content**: Highlights the most similar paragraphs. When the user clicks on a relevant paragraph, it aligns at the top of the page alongside the corresponding similar paragraph in the similar appeal.

(b) **Relevant Paragraph Quick Navigation Bar**: Allows the user to quickly identify very similar paragraphs in this document compared to the other document. Clicking on a colored region will slide the text to the corresponding paragraph's location.

2. **Similar Appeal**: Functions similarly to component 1 but focuses on the reference of the similar appeal.

7 System evaluation with experts

An evaluation was conducted with expert users to verify the system's effectiveness in terms of requirements (Sect. 3.5), interface (Sect. 6), and features (Sect. 6). The following sections describe the evaluated users, the evaluation protocol, and an analysis of the results.

7.1 Users

We recruited 16 expert users to evaluate the system and methodology through connections established by specialist lawyers who collaborated on the project.

The participants for the evaluation were selected based on their experience in the legal field and their availability to participate. To minimize bias, participants were second-degree contacts of the lawyers collaborating on the project. Specifically, the lawyers (co-authors) requested assistance from colleagues, who then extended invitations to their peers, including legal professionals and students. To ensure impartiality: participants were anonymous, and there were no incentives provided for a positive evaluation of the system; the evaluators did not participate in the system's development.

Most of the users were between 25 and 34 years old, with three users aged between 35 and 44 years old and three users aged between 45 and 60 years old. Except for one user, all reported having experience researching, analyzing, and/or writing REs and RG topics or at least being familiar with these concepts. Among the 16 participants, 6 of them declared having between 1 and 5 years of experience, while 3 had more than 5 years. Fig. 11 presents the distribution of education and occupation of the users who participated in the evaluation.

7.2 Evaluation protocol

The user evaluation occurred in three stages. First, users were invited to watch an explanatory video outlining the task and features of the tool and to complete a short demographic survey. The users then performed two tasks (T1 and T2) to explore and familiarize themselves with the system. Finally, users completed four quantitative and qualitative evaluation questionnaires (E1, E2, E3 and E4). The details of each task and questionnaire are provided below.

7.2.1 Task T1

Guided exploration of the system's functionalities using an RE request already included in the database and extensively analyzed. The steps for this task are the following.

1. Access the RE using the unique process number 5013050-64.2019.8.13.0079, available as an example on the system's home screen (Appeal Selection screen, Fig. 6);
2. Identify the most pertinent topic suggested by the system;
3. Identify a paragraph of the text of the document suggested by the tool as important for the application of that topic and evaluate their agreement with the paragraph suggestion;

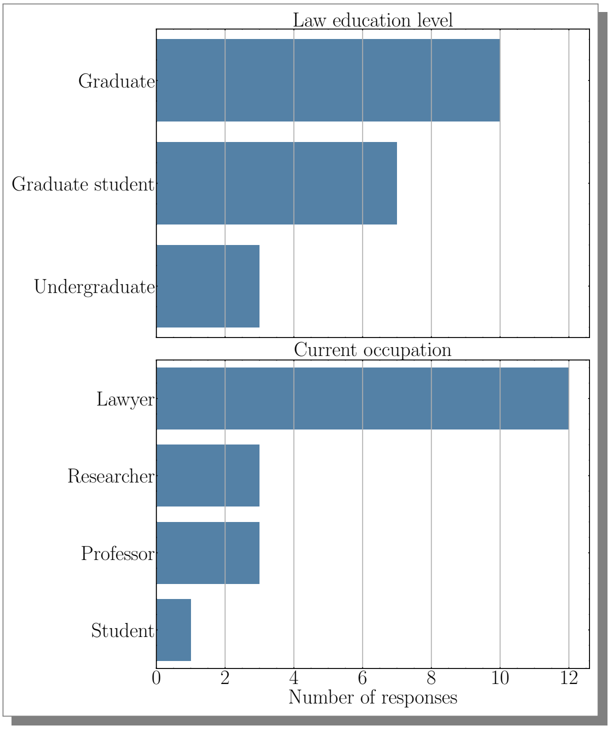


Fig. 11 Education and occupation of users participating in the system evaluation

4. Find an appeal categorized as "very similar," filed in the state of Minas Gerais and submitted in 2020, where the topic above was applied, and provide its unique process number;
5. Identify a paragraph from the original process suggested by the system as very similar between the original appeal and the similar appeal;
6. Provide feedback.

7.2.2 Task T2

Free exploration of the system's functionalities by uploading an RE petition file of the user's choice. If the user does not have an RE file, a list of pre-selected petitions was provided for free selection. The steps for this task are:

1. Upload the RE petition file;
2. Identify a topic, preferably relevant or very relevant, for analysis;
3. Identify a paragraph in the document suggested by the tool as important for the application of that topic and evaluate their agreement with the paragraph suggestion;
4. Check for similar processes.

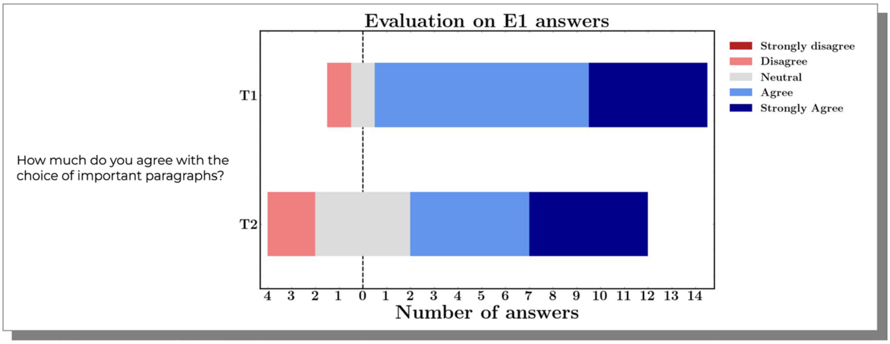


Fig. 12 Answers from Evaluation E1 on the relevance of the paragraphs suggested as justification to a topic

The results of the two tasks are presented in Fig. 12.

7.2.3 Evaluation E1

Quantitative questions (Likert scale) to determine whether the user agrees with the paragraphs identified as important for the topics chosen in tasks T1 and T2.

7.2.4 Evaluation E2

Comparison with other similar systems through mostly qualitative questions. The user would only answer this questionnaire if familiar with another similar system. The questions were:

1. *"What other tool(s) or methodology(ies) that allow similar analyzes to those carried out do you know?"*
2. *"Regarding your preference between LegalAnalytics and the other tool(s) you know:"* (Likert scale).
3. *"What are the advantages that LegalAnalytics has over these other tool(s)?"*
4. *"What are the disadvantages that LegalAnalytics has in relation to this other tool(s)?"*

7.2.5 Evaluation E3

Assessment of the ease of the system's functionalities through quantitative questions (Likert scale). The features investigated include:

1. Access to the process text from the process number;
2. Upload PDF of an appeal;
3. Identification of topics relevant to a process;
4. Identification of relevant paragraphs of the process in relation to a topic;

5. Identification of similar appeals;
6. Filtering similar appeals of interest;
7. Identification of why a process is similar;
8. Sending feedback.

Users also had the opportunity to leave free comments.

7.2.6 Evaluation E4

General quantitative questions (Likert scale) regarding the user's perception of the system. The questions assess:

1. Usefulness of LegalAnalytics;
2. Ease of learning to use LegalAnalytics;
3. Potential of LegalAnalytics to reduce the time required to analyze appeals;
4. Ease of use of LegalAnalytics;
5. Intuitiveness of the LegalAnalytics interface.

Finally, the user could leave a final comment.

For the analysis of quantitative questions using the Likert scale, we assume that the scale values have numerical equivalence ranging from 1 to 5, with higher values associated with more positive evaluations and the value 3 representing a neutral response.

7.3 Results

Evaluation E1 assessed the quality of paragraph suggestions provided to users during Tasks T1 (guided system exploration) and T2 (free exploration). Figure 12 presents the distribution of responses for the system (Sect. 7.2). The average response score was significantly greater than 3 (neutral score) in both tasks: for each question, in a *t*-test with a null hypothesis of the mean score being equal to 3 and an alternative hypothesis of the mean being greater than 3, the p-value was less than 0.05, effectivelly rejecting the null hypothesis. These results indicate that users generally agree with the paragraphs deemed relevant by the system.

In **Evaluation E2**, only one user reported familiarity with another similar system and no preference between the tools. Due to the low number of responses, no definitive conclusions can be drawn about comparing the systems. However, this may suggest limited competition from similar systems within the Brazilian legal context (Sect. 2).

Evaluation E3 assessed the system's features and yielded an average score of at least 4.8 across all questions, indicating that users generally found the tool "easy" or "very easy" to use. The mean score for each question was significantly greater than 3 (p-value < 0.0002 for a *t*-test for each question). Figure 13 displays the distribution of responses for each question.

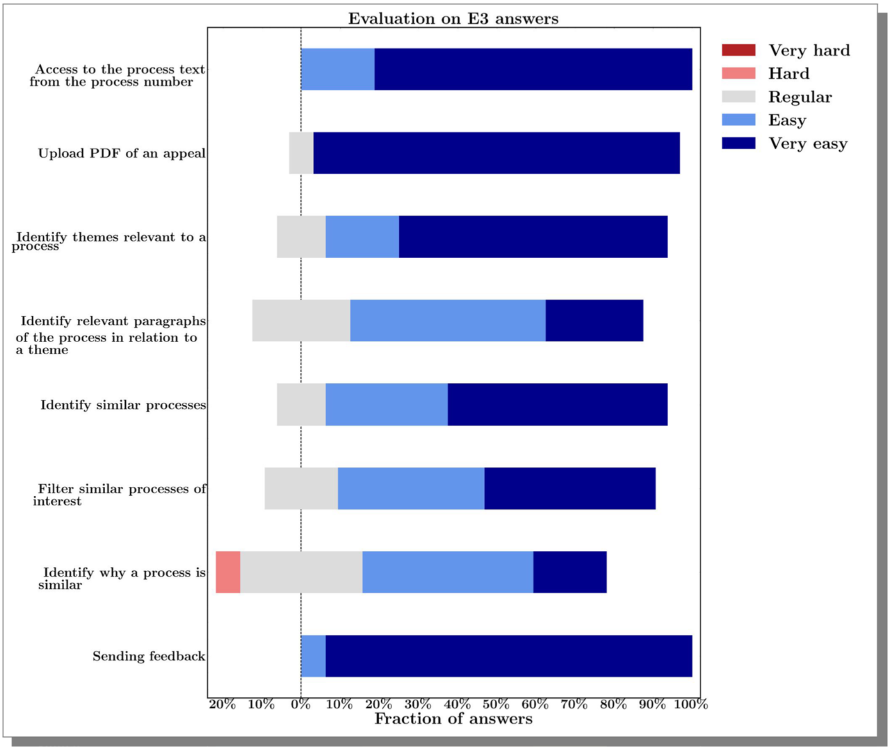


Fig. 13 Answers from the Evaluation E3 on the ease of use of the system's features

In **Evaluation E4**, the average score for all questions was at least 4.05, indicating that users generally perceived the tool as (i) useful, (ii) easy to learn, (iii) having high potential, (iv) easy to use, and (v) intuitive (Fig. 14). The mean score for each question was significantly greater than 3 (p-value $< 1e^{-8}$ for a *t*-test for each question). Figure 14 presents the distribution of responses for each question.

Finally, users were allowed to provide additional comments on the system. Some of the highlighted comments include the following.

- *"The tool is very good!"*
- *"Overall, I found the system very easy."*
- *"I loved the solution, when can I sign up? Congratulations!"*
- *"Very good UX."*
- *"Separating the Analysis into Relevant Topics and Similar Resources makes access much easier and more organized."*

There were also some suggestions:

- *"Please place the color field going from green to red, as this nuance of blue tones does not make it easier to see."*

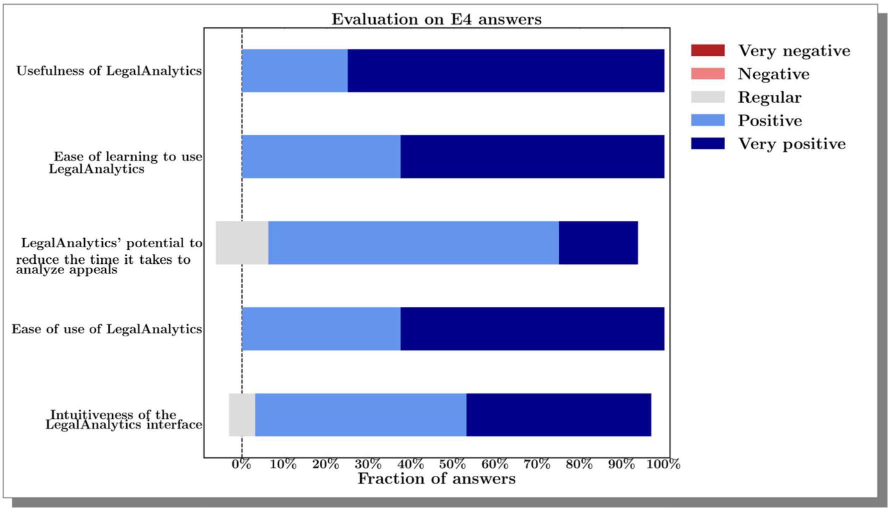


Fig. 14 Response to the Evaluation E4 on the general perception of the system by users

- *"The only observation I made in the feedback on the system itself, regarding date selection, is that I think it can be improved."*
- *"For the date filter you have to click many times. Perhaps the option to type the date or another way to select days, months and years independently could also help in the search."*
- *"The platform is very interesting. I would like to know if there is research into adding concentrated control of constitutionality to the system (ADI, ADC, ADPF)."* Abbreviations are types of law mechanisms in Brazil.

The main comments focused on interface details and certain features, such as the suggestion of paragraphs, which not always aligned with user expectations. It is important to note that these comments were individual and not unanimous. In general, the quantitative evaluations (Figs. 13 and 14) demonstrate the high quality of the system's features. The suggestions will be considered in future work.

8 Conclusion

The high volume of Extraordinary Appeals submitted to the Brazilian Supreme Court each year has necessitated the adoption of computational tools, particularly machine learning models, to assist in screening cases and verifying their validity. However, many of these tools lack transparency, raising concerns about their fairness and potential biases.

The proposed LegalAnalytics system has been designed to address these transparency issues by suggesting Topics of General Repercussion for Extraordinary Appeals while clearly explaining the model's reasoning. LegalAnalytics highlights the most relevant sections of the analyzed document that influenced the model's decision, providing users with a deeper understanding of the process. Results show that LegalAnalytics achieves both high accuracy and high-quality explanations.

Despite its effectiveness, LegalAnalytics has some limitations. For instance, the system currently focuses on the 30 most frequently applied topics, which presents challenges when extending coverage to less common topics due to the scarcity of labeled data for these cases. Additionally, the system relies on a text-to-text comparison approach that does not fully capture semantic meaning. As a result, it struggles when an appeal uses language significantly different from that of previous cases. While this issue is rare—given the standardized nature of constitutional appeals—it can occasionally lead to misclassifications.

Future developments will focus on expanding the range of topics the system can process. We are also working on enhancing its explainability by integrating advanced explanation techniques with generative models (e.g., ChatGPT-like systems). This hybrid approach aims to provide richer, more nuanced explanations, further improving LegalAnalytics' transparency and usability.

Appendix A evaluation of explanations

The goal of providing explanations (Sects. 3.5 and 5.2) is not convincing the user that the model is correct, but rather helping the user understand the model's decision-making process. Given that, it is not necessarily true that explanations extracted with LIME (Sect. 5.2) must be "good," i.e., it is not necessarilythe case that they must actually help convince the user that the model made the right decision for the right reasons—it may happen that the model made the correct decision for the wrong reasons and this is reflected in the explanations. In such a case, the problem would not be with the explanations, but they would lose their usefulness in the system.

To evaluate the correctness of explanations (usually called *plausibility;* DeYoung et al., 2020), an evaluation of the explanations was carried out with the lawyers involved in this project: 25 RE documents were selected for each of three chosen topics (339, 793, and 824), these topics being applied by the STF and also suggested by the model, to be explained and presented to lawyers, who were responsible for evaluating the quality of the explanations. A small number of documents with more than one of these topics were also selected (4 documents).

The lawyers pointed out the number of paragraphs selected by LIME that really corroborate the decision to apply the RG topic in question (metric known as *precision*) and also a qualitative assessment of the quality of the explanation (excellent, good, regular, bad, or very bad). The results are compiled in Table 1 below.

We noticed that the results of the evaluation by the lawyers were positive, with a high accuracy rate (*precision*) and very good qualitative evaluations. The lawyers also made some pertinent comments about the explanations, the main ones being:

- Some explanations could contain fewer paragraphs;
- LIME is not able to capture all paragraphs relevant to the topic in all documents;
- The formatting of paragraphs impairs the explanation in some cases;

Table 1 Results of evaluation of explanations by lawyers

Metric		Evaluator 1	Evaluator 2
Average hit rate		86.26%	84.34%
Quality	Great	50	45
	Good	31	27
	Regular	2	7
	Bad	0	4
	Terrible	0	0

- LIME often selects pre-questioning paragraphs (not exactly the topic), trial summary, and headings.
- The documents of topic 824 are very similar.

It is important to note that these comments were not generalized, but specific to some topics and documents. Regarding the documents on topic 824 in particular, the fact that they were very similar showed that, despite the explanations being different, they were still considered good. Finally, it is also important to point out that the evaluation of the explanations used the models and explanations of a developing version of the system prototype, not the final one. However, this version under development was evaluated and tested and also achieved good performance (26 out of 30 topics had models with *balanced accuracy* higher than 80%, and 29 higher than 60%).

Appendix B regular expressions for remove watermarks, signatures, and personal information

Full details of the regular expression used for each case:

1. For attached images, we identified two types of meta-data:

   (a) < image : \w{1, }, * width : \d{1, }, height : \d{1, }, bpc : \d{1, }>.
   (b) < image : \w{1, }(. * ?), * width : \d{1, }, height : \d{1, }, bpc : \d{1, } > .

2. For personal information, we use the expressions:

   (a) CEP : CEP[: \s] * \d{2, }[.] * \d{3, } − \d{3}.
   (b) CPF : CPF[: \s] * \d{3}.\d{3}.\d{3} − \d{2}.

3. For watermarks, we identified 3 types:

   (a) Impresso por:\d{3}.\d{3}.\d{3}−\d{2}.{0, 25}\nEm:\d{2}/\d{2}/\d{4}−\d{2}: \d{2}: \d{2}\n.

   (b) . * Impresso por :* . *\n. *\n. * Em :\d{2}/\d{2}/\d{4} – \d{2} : \d{2} : \d{2}por : \d{3}.\d{3}.\d{3} – \d{2}.
   (c) Para conferir o original, acesse o site, informe o processo \d{1, } – \d{1, }.\d{1, }.\d{1, }.\d{1, }.\d{1, } e o código \w{1, }\d{1, }.

4. For digital signatures, we identify common expression to search and remove:

   (a) Documento recebido eletronicamente da origem.
   Documento assinado digitalmente [,]* conforme MP n°
   (b) \d{1, }.\d{1, }–\d{1, }/\d{1, }[\s * de \d{1, }/\d{1, }/\d{1, }] * [,].
   * [\s*Lein° \d{1, }.\d{1, }/\d{1, }] * .
   Este documento é cópia do original as sin ado digitalmente
   (c) por ([^ +.]). Protocolado em \d{2, }/\d{2, }/\d{2, } às \d{2, } : \d{2, } : \d{2, }, sob o número \w{1, }\d{1, }

5. For links, we perform a recursive search between the link and the previous and posterior lines. We apply this because links can be separated into multiple lines. We use the expressions.*< link >* . *, previous.. * \n, and posterior \n. *. In addition, we filter digital links to signatures with the expression: \n. * .jus.br. * \n.

Appendix C regular expressions for label extraction

From the 10,233 samples in our labeled dataset, we could extract the labels from the process records for 8,206 samples. The remaining 2,027 samples had the labels extracted from the PDF documents using regular expressions, as mentioned in Sect. 4.2.

For the label extraction of these samples with no available label metadata, we used the monocratic decisions and rulings ("despachos"), which are the documents used to communicate the decisions of the justices to return the appeal to the lower courts based on the general repercussion topic. To extract the labels from the monocratic decisions and rulings, we used the following regular expressions (with ignore case flag) to handle both the singular ("tema") and plural ("temas") forms of the word "tema" (topic):

- (? <![a – z0 – 9])tema[^\d]{1, 8}(\d{0, 1}\.{0, 1}\d{1, 3})
  (? <![*a* – *z*0 – 9])temas[^\d]{1, 8}\d{0, 1}\.{0, 1}\d{1, 3}
- (? : [, en.º °\n]{1, 6}\d{0, 1}\.{0, 1}\d{1, 3})+

The regular expressions were carefully designed based on exploring a comprehensive set of document texts. Even though the regular expressions may seem complex, they are derived from a simple pattern of "tema 800" and updated to handle exceptional cases. We developed a number of unit tests to ensure the regular expressions correctly identified the topics in the most complex cases. For example, consider the following text: "Temas n° 8.000, 9, 100, 1.800 e nº°76, sistemas 33 e 44, claramente temas n° 1, 1004, 1.009, n° 3 e 788” The regular expression correctly identifies “8.000”, “9”, “100”, “1.800”, “76”, “1”, “1004”, “1.009”, “3”, “788” and correctly does not identify “33” and “44”. We did not perform a quantitative evaluation of the regular expressions, but we are confident in their accuracy based on the extensive exploration and unit testing performed.

Appendix D complete model specification

Below is the complete specification of the classification models described in Sect. 5.1.2, in scikit-learn format. We sort the model details following the number of applications exhibited in Fig. 3. Each topic has its own model, and the specification of the models may vary between them according to the hyperparameters selected in the cross-validation process. In particular, the `n_gram_range` varies between `(1, 1)` and `(1, 2)`, the `scaler` may or may not be used, `C` varies in `numpy.logspace(-4, 4, 10)`, and `class_weight` varies between `None` and ’`balanced`’.

- Topic 660

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=2.782559402207126,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
    cv='prefit')
```

- Topic 800

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=166.81005372000558,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
    cv='prefit')
```

- Topic 339

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=2.782559402207126,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
    cv=’prefit’)
```

- Topic 810

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
     cv=’prefit’)
```

- Topic 766

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2)),
            (’logistic_regression’, LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
   cv=’prefit’)
```

- Topic 587

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('scaler', StandardScaler(with_mean=False)),
            ('logistic_regression', LogisticRegression(C=0.000774263682681127,
                        class_weight='balanced',
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 852

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2)),
            ('logistic_regression', LogisticRegression(C=2.782559402207126,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
    cv='prefit')
```

- Topic 793

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2)),
            ('logistic_regression', LogisticRegression(C=21.54434690031882,
                        class_weight='balanced',
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 634

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=166.81005372000558,
                        class_weight=’balanced’,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
   cv=’prefit’)
```

- Topic 6

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2)),
            (’logistic_regression’, LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
    cv=’prefit’)
```

- Topic 895

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
   cv=’prefit’)
```

- Topic 313

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=0.3593813663804626,
                        class_weight='balanced',
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 773

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=166.81005372000558,
                        class_weight='balanced',
                        dual=True,
                        max_iter=1000.0,
                          random_state=42,
                          solver='liblinear'))]),
     cv='prefit')
```

- Topic 424

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2)),
            ('logistic_regression', LogisticRegression(C=166.81005372000558,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 824

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=166.81005372000558,
                        class_weight=’balanced’,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
   cv=’prefit’)
```

- Topic 954

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                       min_df=2,
                       ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=0.000774263682681127,
                       class_weight=’balanced’,
                       dual=True,
                         max_iter=1000.0,
                         random_state=42,
                         solver=’liblinear’))]),
     cv=’prefit’)
```

- Topic 663

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=1291.5496650148827,
                        class_weight=’balanced’,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
    cv=’prefit’)
```

- Topic 1134

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                       min_df=2)),
            (’logistic_regression’, LogisticRegression(C=10000.0,
                       class_weight=’balanced’,
                       dual=True,
                       max_iter=1000.0,
                       random_state=42,
                       solver=’liblinear’))]),
   cv=’prefit’)
```

- Topic 807

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2)),
            (’logistic_regression’, LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
    cv=’prefit’)
```

- Topic 288

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                       min_df=2)),
            (’logistic_regression’, LogisticRegression(C=0.0001,
                       class_weight=’balanced’,
                       dual=True,
                       max_iter=1000.0,
                       random_state=42,
                       solver=’liblinear’))]),
   cv=’prefit’)
```

- Topic 915

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=0.005994842503189409,
                        class_weight='balanced',
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 1023

```
CalibratedClassifierCV(
  base_estimator=Pipeline(
    steps=[
      ('tf_idf', TfidfVectorizer(lowercase=False,
                    min_df=2,
                    ngram_range=(1, 2))),
      ('logistic_regression', LogisticRegression(C=2.782559402207126,
                    class_weight='balanced',
                    dual=True,
                    max_iter=1000.0,
                    random_state=42,
                    solver='liblinear'))]),
    cv='prefit')
```

- Topic 589

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 334

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=166.81005372000558,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
    cv=’prefit’)
```

- Topic 41

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            (’logistic_regression’, LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
    cv=’prefit’)
```

- Topic 1114

```
    CalibratedClassifierCV(
        base_estimator=Pipeline(
            steps=[
                (’tf_idf’, TfidfVectorizer(lowercase=False,
                        min_df=2)),
                (’logistic_regression’, LogisticRegression(C=10000.0,
                        class_weight=’balanced’,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver=’liblinear’))]),
       cv=’prefit’)
```

- Topic 163

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2)),
            ('logistic_regression', LogisticRegression(C=21.54434690031882,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
    cv='prefit')
```

- Topic 702

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2,
                        ngram_range=(1, 2))),
            ('logistic_regression', LogisticRegression(C=10000.0,
                        class_weight='balanced',
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
    cv='prefit')
```

- Topic 960

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2)),
            ('logistic_regression', LogisticRegression(C=166.81005372000558,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
   cv='prefit')
```

- Topic 1125

```
CalibratedClassifierCV(
    base_estimator=Pipeline(
        steps=[
            ('tf_idf', TfidfVectorizer(lowercase=False,
                        min_df=2)),
            ('logistic_regression', LogisticRegression(C=166.81005372000558,
                        dual=True,
                        max_iter=1000.0,
                        random_state=42,
                        solver='liblinear'))]),
            cv='prefit')
```

Appendix E implementation details

To develop the system, we used the Model-View-Controller (MVC) development pattern in order to have better control over the data flow and how to display it. The system architecture is represented in Fig. 15. The technologies used are made up of three main components. The components are database, backend, and frontend. The database is based on ElasticSearch.

The database component manages and solves all requests made by the back-end, which receives and sends information to the front-end. The ElasticSearch database is an unstructured database (it is not organized into tables), through which it is possible to store complex information such as texts, vectors, etc.

The back-end was developed using Python as a programming language and Flask as a web framework for quickly creating a REST service. We have the advantage of creating an API that can be consulted without needing the tool. Furthermore, Python-Flask allows us to have quick and easy compatibility with

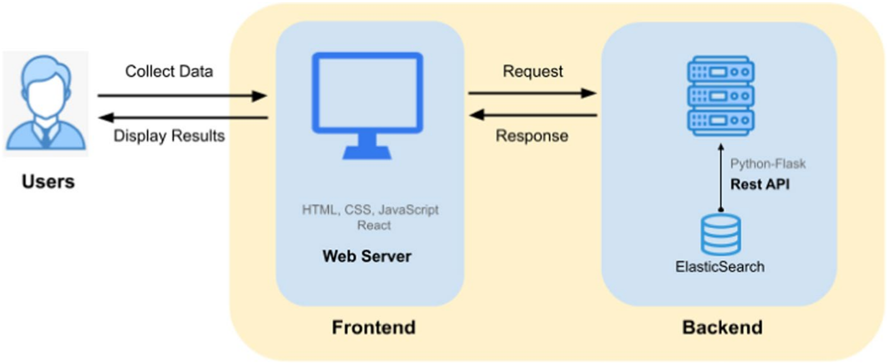


Fig. 15 LegalAnalytics system architecture diagram

the classification models used to infer related topics. It also allows us to quickly implement services for already detailed databases, quickly controlling the flow of information requested by the tool or external queries. The advantage of using this framework is the flexibility when adding, updating, or removing features or components, for example, protocol security using flask-CORS, number of requests per second with flask-limit, and organization when creating the Rest API routing.

The front-end presented in Sect. 6, defined as User Interface, was developed using React, JavaScript/TypeScript, and Node.JS. All components used are installed on a main server protected with a username and password for greater security. Through the front-end, the user has screens and functionalities as described in Sect. 6.

Finally, we employ docker containers to deploy all the systems separately. Each component mentioned above is implemented as a container, all connected to the same virtual network to ensure better communication between each service.

Appendix F classification performance

The following table presents the complete performance results of the machine learning models described in Sect. 5.1.2. In addition to the AUPRC and AUC for testing, the table presents the metrics for training and for a random *"dummy"* classifier with '`stratified`' strategy[15]. In addition, F_1-score and G-mean (geometric mean) metrics are also included with thresholds adjusted in an evaluation subset. It is important to emphasize again that the models evaluated are prior to the calibration process (Sect. 5.1.2), therefore the associated thresholds are prior as well.

[15] Refer to scikit-learn's `DummyClassifier`'s strategies: https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html.

Topic	AUPRC			F$_1$-score				AUC			G-mean				Test support
	train	test	dummy	threshold	train	test	dummy	train	test	dummy	threshold	train	test	dummy	
topic_1134	100.00	100.00	1.22	99.98	100.00	100.00	0.00	100.00	100.00	49.46	99.98	100.00	100.00	0.00	25.0
topic_288	99.94	100.00	1.17	50.31	97.62	100.00	0.00	100.00	100.00	49.46	50.31	99.97	100.00	0.00	24.0
topic_41	99.91	100.00	1.03	97.14	70.80	92.31	0.00	100.00	100.00	49.46	0.50	99.24	99.16	0.00	21.0
topic_915	98.52	98.09	1.12	57.44	98.14	95.45	0.00	99.98	99.96	49.46	57.44	99.36	95.55	0.00	23.0
topic_793	98.87	97.98	2.53	97.44	98.06	84.44	1.94	99.98	99.94	49.71	4.21	99.45	99.60	13.69	52.0
topic_954	96.87	96.59	1.37	49.03	92.06	94.55	0.00	99.96	99.88	49.21	49.03	94.19	96.34	0.00	28.0
topic_1114	100.00	96.57	0.98	47.32	100.00	90.00	0.00	100.00	99.97	49.51	47.32	100.00	94.82	0.00	20.0
topic_587	100.00	96.55	3.21	99.89	98.23	95.38	2.90	100.00	99.84	49.75	99.89	98.25	96.87	17.10	66.0
topic_634	99.97	94.36	1.96	85.53	98.59	91.14	2.38	100.00	99.89	50.18	2.19	99.66	97.13	15.64	40.0
topic_810	99.77	91.13	4.57	7.91	91.83	82.93	5.24	99.99	98.48	50.31	5.37	99.42	94.89	22.63	93.0
topic_773	99.08	89.89	1.51	98.76	96.30	78.57	0.00	99.99	99.66	49.13	98.76	98.10	84.18	0.00	31.0
topic_589	96.72	89.17	1.03	80.30	84.85	80.00	0.00	99.96	99.76	49.46	4.75	99.70	92.31	0.00	21.0
topic_824	99.69	88.24	1.42	79.90	94.06	84.06	0.00	100.00	99.87	49.13	79.90	99.91	99.73	0.00	29.0
topic_800	99.99	88.24	8.40	20.10	98.31	83.25	6.21	100.00	98.92	48.66	8.25	99.73	95.95	23.92	175.0
topic_852	92.85	85.21	2.85	14.34	86.65	76.19	3.45	99.80	98.87	50.32	1.40	97.87	96.44	18.31	58.0
topic_334	99.04	84.57	1.03	4.32	79.14	65.31	0.00	99.99	99.72	49.46	4.32	99.72	87.03	0.00	21.0
topic_807	97.91	84.30	1.17	13.25	90.32	81.48	0.00	99.98	99.73	49.46	1.35	99.54	97.31	0.00	24.0
topic_702	100.00	82.80	0.93	97.73	100.00	81.08	0.00	100.00	99.83	49.53	97.73	100.00	88.79	0.00	19.0
topic_766	98.38	81.20	3.59	30.59	93.67	72.99	2.63	99.91	97.47	49.43	1.88	96.44	92.72	16.12	74.0
topic_163	93.15	80.84	0.98	14.35	77.97	71.43	0.00	99.93	99.74	49.51	5.12	99.36	94.38	0.00	20.0
topic_6	98.77	78.96	1.91	9.31	80.00	77.55	2.41	99.98	99.64	50.21	9.31	99.51	98.19	15.84	39.0
topic_960	98.50	78.26	0.88	51.88	93.23	75.68	0.00	99.99	99.85	49.53	2.43	99.77	99.73	0.00	18.0
topic_1023	89.52	77.70	1.07	89.35	86.03	70.00	0.00	99.92	99.51	49.46	3.88	97.61	97.27	0.00	22.0
topic_1125	99.95	76.99	0.88	99.69	76.47	64.52	0.00	100.00	99.88	49.56	99.69	78.68	74.48	0.00	18.0

Topic	AUPRC			F_1-score				AUC			G-mean				Test support
	train	test	dummy	threshold	train	test	dummy	train	test	dummy	threshold	train	test	dummy	
topic_313	80.44	75.96	1.56	99.51	15.00	17.14	0.00	99.73	99.67	49.08	39.32	99.30	99.45	0.00	32.0
topic_663	99.59	71.29	1.22	99.92	55.46	36.36	0.00	100.00	95.95	49.46	1.63	99.37	95.08	0.00	25.0
topic_660	89.55	65.54	9.80	27.45	84.87	59.49	8.46	98.70	93.32	49.25	9.87	93.29	86.30	27.47	203.0
topic_339	82.99	56.11	5.51	93.01	0.00	0.00	4.29	99.13	94.01	49.24	6.34	95.32	87.66	20.32	114.0
topic_895	98.48	46.27	1.66	3.77	75.40	37.65	0.00	99.97	93.37	48.98	0.67	92.76	86.75	0.00	34.0
topic_424	99.75	44.89	1.47	99.37	10.71	0.00	0.00	100.00	97.67	49.13	0.18	97.01	94.34	0.00	30.0

Appendix G results for other models

In this section, we present detailed results for different model architectures that we tested during the development of the LegalAnalytics system. Here, we focus on the popular Topic 800, which is the second most frequent topic in the dataset (Fig. 3), had a good performance in preliminary classification tests (Fig. 4) and was deemed relevant in a manual exploration of the data. The data used was the machine learning dataset described in Sect. 4 in its version of size 10,710 with test corresponding to 20% of the data. We tested the performance of the following models:

- Boosting, distance-based, ensemble methods, linear models, Naive Bayes, SVM, and tree-based, all of them with TF-IDF vectorization (unigram and dimensionality 90659). When available, we performed cross-validation with the average precision metric. The choice for these models followed their easily available implementation (Pedregosa et al., 2011). We randomly shuffled the data and stratified train-test by the target variable. We present the results in Table 2.

Table 2 Comparison of various machine learning models with TF-IDF vectorization, sorted by columns in their order of appearance

Model	Average precision	Acc.	Balanced accuracy	F1 score	ROC AUC
XGBoost	0.89	0.97	0.88	0.82	0.99
LightGBM	0.88	0.97	0.9	0.83	0.99
Calibrated Classifier CV	0.86	0.96	0.87	0.81	0.99
Logistic Regression	0.84	0.95	0.79	0.71	0.99
Random Forest	0.84	0.94	0.71	0.58	0.98
AdaBoost	0.82	0.96	0.86	0.77	0.98
Extra Trees	0.82	0.93	0.69	0.54	0.98
Bagging Classifier	0.79	0.95	0.82	0.74	0.96
K-Nearest Neighbors	0.77	0.95	0.85	0.75	0.96
Bernoulli Naive Bayes	0.61	0.93	0.81	0.64	0.96
Decision Tree	0.57	0.95	0.85	0.74	0.85
Extra Tree	0.35	0.92	0.75	0.55	0.75
Dummy Classifier	0.1	0.9	0.5	0	0.5
Stochastic Gradient Descent	–	0.97	0.89	0.83	–
Linear SVM	–	0.97	0.88	0.82	–
Passive Aggressive	–	0.96	0.88	0.8	–
SVM	–	0.96	0.85	0.79	–
Ridge Classifier	–	0.96	0.85	0.78	–
Ridge Classifier CV	–	0.96	0.85	0.78	–
Perceptron	–	0.96	0.84	0.75	–
Nearest centroid	–	0.9	0.93	0.66	–

Table 3 Results for the BERT model

Model	Accuracy	Balanced accuracy	F1 score
BERT	0.93	0.93	0.64
BERT (sliding window)	0.94	0.94	0.68

- Transformer encoder (BERT) pretrained in Portuguese (`neuralmind/bert-base-portuguese-cased`[16]), 3 epochs, batch size of 128, learning rate of $5e^{-5}$, and class weights. To overcome BERT's limitation of input size of 512 tokens, we also tested a sliding window with a stride of 64 for train and test with a test prediction aggregation by maximum probability. The test set corresponds to 20% of the data with the most recent documents. We adjusted the threshold in a small validation set to maximize balanced accuracy. We present the results in Table 3.

Appendix H user interface in portuguese

The following figures present the original user interface (Sect. 6) in Brazilian Portuguese. (See Figs. 16, 17, 18, 19 and 20).

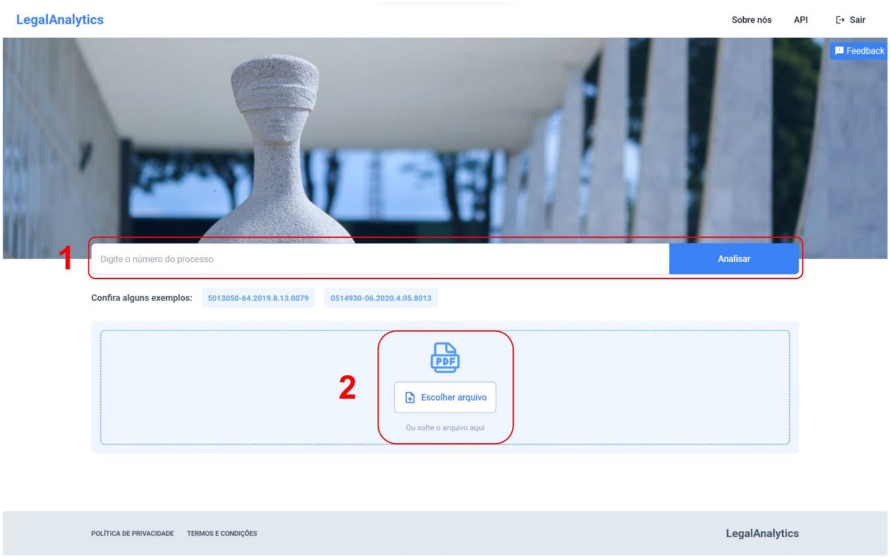
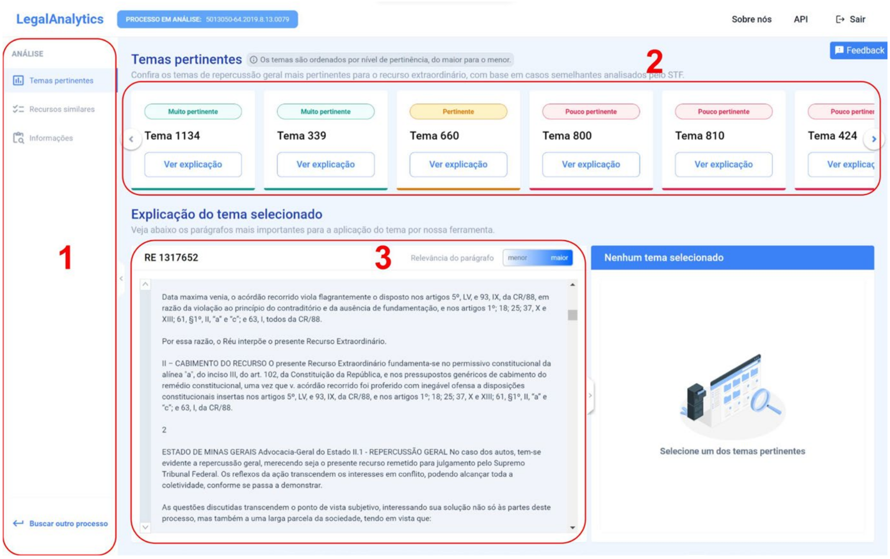


Fig. 16 Original Appeal Selection Screen (Fig. 6) in Portuguese

[16] Available at https://huggingface.co/neuralmind/bert-base-portuguese-cased.



Temas pertinentes

Os temas são ordenados por nível de pertinência, do maior para o menor.

Confira os temas de repercussão geral mais pertinentes para o recurso extraordinário, com base em casos semelhantes analisados pelo STF.

Explicação do tema selecionado

Veja abaixo os parágrafos mais importantes para a aplicação do tema por nossa ferramenta.

RE 1317652

Data maxima venia, o acórdão recorrido viola flagrantemente o disposto nos artigos 5º, LV, e 93, IX, da CR/88, em razão da violação ao princípio do contraditório e da ausência de fundamentação, e nos artigos 1º; 18; 25; 37, X e XIII; 61, §1º, II, "a" e "c"; e 63, I, todos da CR/88.

Por essa razão, o Réu interpõe o presente Recurso Extraordinário.

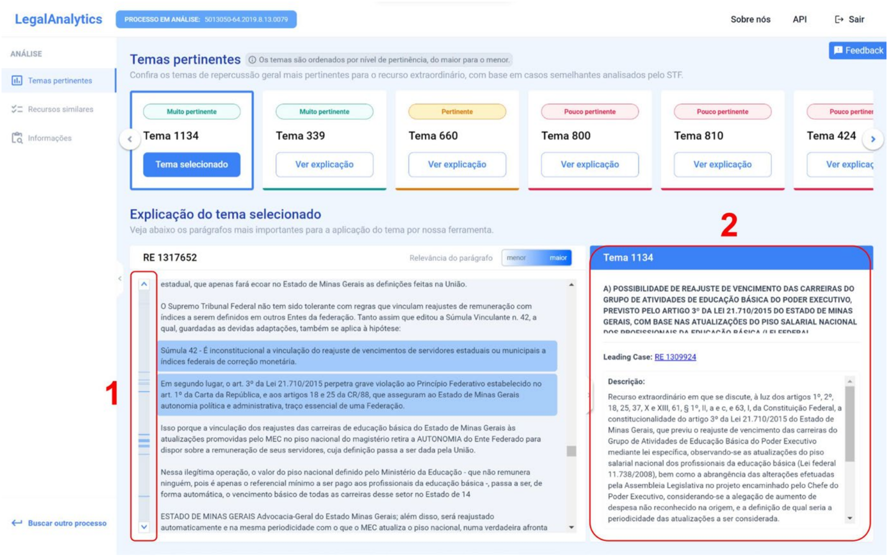
II – CABIMENTO DO RECURSO O presente Recurso Extraordinário fundamenta-se no permissivo constitucional da alínea "a", do inciso III, do art. 102, da Constituição da República, e nos pressupostos genéricos de cabimento do remédio constitucional, uma vez que v. acórdão recorrido foi proferido com inegável ofensa a disposições constitucionais insertas nos artigos 5º, LV, e 93, IX, da CR/88, e nos artigos 1º; 18; 25; 37, X e XIII; 61, §1º, II, "a" e "c"; e 63, I, da CR/88.

2

ESTADO DE MINAS GERAIS Advocacia-Geral do Estado II.1 - REPERCUSSÃO GERAL No caso dos autos, tem-se evidente a repercussão geral, merecendo seja o presente recurso remetido para julgamento pelo Supremo Tribunal Federal. Os reflexos da ação transcendem os interesses em conflito, podendo alcançar toda a coletividade, conforme se passa a demonstrar.

As questões discutidas transcendem o ponto de vista subjetivo, interessando sua solução não só às partes deste processo, mas também a uma larga parcela da sociedade, tendo em vista que:

Fig. 17 Original Relevant Topics screen (Fig. 7) in Portuguese



Temas pertinentes

Os temas são ordenados por nível de pertinência, do maior para o menor.

Confira os temas de repercussão geral mais pertinentes para o recurso extraordinário, com base em casos semelhantes analisados pelo STF.

Explicação do tema selecionado

Veja abaixo os parágrafos mais importantes para a aplicação do tema por nossa ferramenta.

RE 1317652

estadual, que apenas fará ecoar no Estado de Minas Gerais as definições feitas na União.

O Supremo Tribunal Federal não tem sido tolerante com regras que vinculam reajustes de remuneração com índices a serem definidos em outros Entes da federação. Tanto assim que editou a Súmula Vinculante n. 42, a qual, guardadas as devidas adaptações, também se aplica à hipótese:

Súmula 42 - É inconstitucional a vinculação do reajuste de vencimentos de servidores estaduais ou municipais a índices federais de correção monetária.

Em segundo lugar, o art. 3º da Lei 21.710/2015 perpetra grave violação ao Princípio Federativo estabelecido no art. 1º da Carta da República, e aos artigos 18 e 25 da CR/88, que asseguram ao Estado de Minas Gerais autonomia política e administrativa, traço essencial de uma Federação.

Isso porque a vinculação dos reajustes das carreiras de educação básica do Estado de Minas Gerais às atualizações promovidas pelo MEC no piso nacional do magistério retira a AUTONOMIA do Ente Federado para dispor sobre a remuneração de seus servidores, cuja definição passa a ser dada pela União.

Nessa ilegítima operação, o valor do piso nacional definido pelo Ministério da Educação - que não remunera ninguém, pois é apenas o referencial mínimo a ser pago aos profissionais da educação básica -, passa a ser, de forma automática, o vencimento básico de todas as carreiras desse setor no Estado de 14

ESTADO DE MINAS GERAIS Advocacia-Geral do Estado Minas Gerais; além disso, será reajustado automaticamente e na mesma periodicidade com o que o MEC atualiza o piso nacional, numa verdadeira afronta

Tema 1134

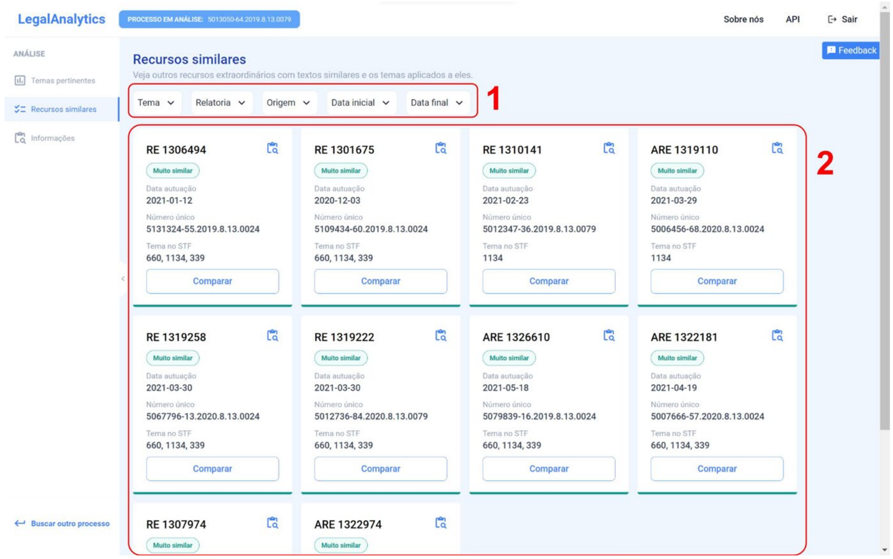
A) POSSIBILIDADE DE REAJUSTE DE VENCIMENTO DAS CARREIRAS DO GRUPO DE ATIVIDADES DE EDUCAÇÃO BÁSICA DO PODER EXECUTIVO, PREVISTO PELO ARTIGO 3º DA LEI 21.710/2015 DO ESTADO DE MINAS GERAIS, COM BASE NAS ATUALIZAÇÕES DO PISO SALARIAL NACIONAL DOS PROFISSIONAIS DA EDUCAÇÃO BÁSICA (LEI FEDERAL

Leading Case: RE 1309924

Descrição:

Recurso extraordinário em que se discute, à luz dos artigos 1º, 2º, 18, 25, 37, X e XIII, 61, § 1º, II, a e c, e 63, I, da Constituição Federal, a constitucionalidade do artigo 3º da Lei 21.710/2015 do Estado de Minas Gerais, que previu o reajuste de vencimento das carreiras do Grupo de Atividades de Educação Básica do Poder Executivo mediante lei específica, observando-se as atualizações do piso salarial nacional dos profissionais da educação básica (Lei federal 11.738/2008), bem como a abrangência das alterações efetuadas pela Assembleia Legislativa no projeto encaminhado pelo Chefe do Poder Executivo, considerando-se a alegação de aumento de despesa não reconhecido na origem, e a definição de qual seria a periodicidade das atualizações a ser considerada.

Fig. 18 Original Relevant Topics screen (Fig. 8) in Portuguese after selecting a topic



Recursos similares

Veja outros recursos extraordinários com textos similares e os temas aplicados a eles.

Tema | Relatoria | Origem | Data inicial | Data final

RE 1306494
Muito similar
Data autuação
2021-01-12
Número único
5131324-55.2019.8.13.0024
Tema no STF
660, 1134, 339
Comparar

RE 1301675
Muito similar
Data autuação
2020-12-03
Número único
5109434-60.2019.8.13.0024
Tema no STF
660, 1134, 339
Comparar

RE 1310141
Muito similar
Data autuação
2021-02-23
Número único
5012347-36.2019.8.13.0079
Tema no STF
1134
Comparar

ARE 1319110
Muito similar
Data autuação
2021-03-29
Número único
5006456-68.2020.8.13.0024
Tema no STF
1134
Comparar

RE 1319258
Muito similar
Data autuação
2021-03-30
Número único
5067796-13.2020.8.13.0024
Tema no STF
660, 1134, 339
Comparar

RE 1319222
Muito similar
Data autuação
2021-03-30
Número único
5012736-84.2020.8.13.0079
Tema no STF
660, 1134, 339
Comparar

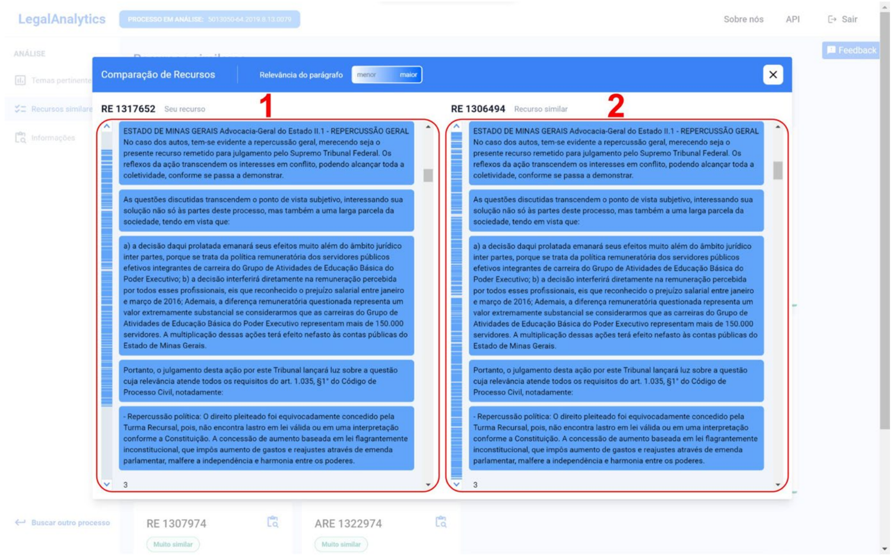
ARE 1326610
Muito similar
Data autuação
2021-05-18
Número único
5079839-16.2019.8.13.0024
Tema no STF
660, 1134, 339
Comparar

ARE 1322181
Muito similar
Data autuação
2021-04-19
Número único
5007666-57.2020.8.13.0024
Tema no STF
660, 1134, 339
Comparar

RE 1307974
Muito similar

ARE 1322974
Muito similar

Fig. 19 Original Similar Appeals screen (Fig. 9) in Portuguese



RE 1317652 Seu recurso

ESTADO DE MINAS GERAIS Advocacia-Geral do Estado II.1 - REPERCUSSÃO GERAL No caso dos autos, tem-se evidente a repercussão geral, merecendo seja o presente recurso remetido para julgamento pelo Supremo Tribunal Federal. Os reflexos da ação transcendem os interesses em conflito, podendo alcançar toda a coletividade, conforme se passa a demonstrar.

As questões discutidas transcendem o ponto de vista subjetivo, interessando sua solução não só às partes deste processo, mas também a uma larga parcela da sociedade, tendo em vista que:

a) a decisão daqui prolatada emanará seus efeitos muito além do âmbito jurídico inter partes, porque se trata da política remuneratória dos servidores públicos efetivos integrantes de carreira do Grupo de Atividades de Educação Básica do Poder Executivo; b) a decisão interferirá diretamente na remuneração percebida por todos esses profissionais, eis que reconhecido o prejuízo salarial entre janeiro e março de 2016; Ademais, a diferença remuneratória questionada representa um valor extremamente substancial se considerarmos que as carreiras do Grupo de Atividades de Educação Básica do Poder Executivo representam mais de 150.000 servidores. A multiplicação dessas ações terá efeito nefasto às contas públicas do Estado de Minas Gerais.

Portanto, o julgamento desta ação por este Tribunal lançará luz sobre a questão cuja relevância atende todos os requisitos do art. 1.035, §1° do Código de Processo Civil, notadamente:

- Repercussão política: O direito pleiteado foi equivocadamente concedido pela Turma Recursal, pois, não encontra lastro em lei válida ou em uma interpretação conforme a Constituição. A concessão de aumento baseada em lei flagrantemente inconstitucional, que impôs aumento de gastos e reajustes através de emenda parlamentar, malfere a independência e harmonia entre os poderes.

3

RE 1306494 Recurso similar

ESTADO DE MINAS GERAIS Advocacia-Geral do Estado II.1 - REPERCUSSÃO GERAL No caso dos autos, tem-se evidente a repercussão geral, merecendo seja o presente recurso remetido para julgamento pelo Supremo Tribunal Federal. Os reflexos da ação transcendem os interesses em conflito, podendo alcançar toda a coletividade, conforme se passa a demonstrar.

As questões discutidas transcendem o ponto de vista subjetivo, interessando sua solução não só às partes deste processo, mas também a uma larga parcela da sociedade, tendo em vista que:

a) a decisão daqui prolatada emanará seus efeitos muito além do âmbito jurídico inter partes, porque se trata da política remuneratória dos servidores públicos efetivos integrantes de carreira do Grupo de Atividades de Educação Básica do Poder Executivo; b) a decisão interferirá diretamente na remuneração percebida por todos esses profissionais, eis que reconhecido o prejuízo salarial entre janeiro e março de 2016; Ademais, a diferença remuneratória questionada representa um valor extremamente substancial se considerarmos que as carreiras do Grupo de Atividades de Educação Básica do Poder Executivo representam mais de 150.000 servidores. A multiplicação dessas ações terá efeito nefasto às contas públicas do Estado de Minas Gerais.

Portanto, o julgamento desta ação por este Tribunal lançará luz sobre a questão cuja relevância atende todos os requisitos do art. 1.035, §1° do Código de Processo Civil, notadamente:

- Repercussão política: O direito pleiteado foi equivocadamente concedido pela Turma Recursal, pois, não encontra lastro em lei válida ou em uma interpretação conforme a Constituição. A concessão de aumento baseada em lei flagrantemente inconstitucional, que impôs aumento de gastos e reajustes através de emenda parlamentar, malfere a independência e harmonia entre os poderes.

3

Fig. 20 Original Similar Appeals screen (Fig. 10) in Portuguese after selecting a process for comparison

Acknowledgements This work was supported by the National Council for Scientific and Technological Development (CNPq) under Grants #311144/2022-5 and #307184/2021-8, the Carlos Chagas Filho Foundation for Research Support of Rio de Janeiro State (FAPERJ) under Grant #E-26/201.424/2021, the São Paulo Research Foundation (FAPESP) under Grants #2021/07012-0 and #2022/09091-8, the Fundação Getulio Vargas under the Mario Henrique Simonsen scholarship (Ezequiel Fajreldines) and other grants, and the Coordination for the Improvement of Higher Education Personnel (CAPES). The views, opinions, hypotheses, and conclusions or recommendations expressed in this material are those of responsibility of the author(s) and do not necessarily reflect the views of the funding agencies and institutions they are affiliated with.

Data availability We regret that we cannot release the data used for this research due to legal restrictions. This data contains sensitive information about judicial processes in Brazil and is protected under Brazilian law (Law No. 13793/2019).

References

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European court of human rights: a natural language processing perspective. Peer J computer science. 2:e93

Araujo PHL, Campos TE, Braz FA, Silva NC (2020) VICTOR: a Dataset for Brazilian Legal Documents Classification. In: proceedings of the 12th language resources and evaluation conference, pp. 1449–1458. European language resources association, Marseille, France. https://aclanthology.org/2020.lrec-1.181

Arriba-Pérez F, García-Méndez S, González-Castaño FJ, González-González J (2022) Explainable machine learning multi-label classification of Spanish legal judgements. J King Saud Univ- Comput Info Sci 34(10, Part B), 10180–10192 https://doi.org/10.1016/j.jksuci.2022.10.015

Atkinson K, Bench-Capon T, Bollegala D (2020) Explanation in AI and law: past, present and future. Artif Intell 289:103387

Avinash M, Sivasankar E (2019) A Study of Feature Extraction Techniques for Sentiment Analysis. In: Abraham A, Dutta P, Mandal JK, Bhattacharya A, Dutta S (eds) Emerging Technologies in Data Mining and Information Security. Springer Singapore, Singapore, pp 475–486. https://doi.org/10.1007/978-981-13-1501-5_41

Benedetto I, Koudounas A, Vaiani L, Pastor E, Baralis E, Cagliero L, Tarasconi F (2023) PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction. In: Ojha, A.K., Doğruöz, A.S., Da San Martino, G., Tayyar Madabushi, H., Kumar, R., Sartori, E. (eds.) Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pp. 1401–1411. Association for Computational Linguistics, Toronto, Canada.https://doi.org/10.18653/v1/2023.semeval-1.194. https://aclanthology.org/2023.semeval-1.194

Bhambhoria R, Dahan S, Zhu X (2021) Investigating the state-of-the-art performance and explainability of legal judgment prediction. In: Proceedings of the Canadian conference on artificial intelligence. Canadian Artificial Intelligence Association (CAIAC), Online. https://doi.org/10.21428/594757db.a66d81b6. https://caiac.pubpub.org/pub/hj1na0xf/release/1

Bhambhoria R, Liu H, Dahan S, Zhu X (2022) Interpretable low-resource legal decision making. Proceedings of the AAAI conference on artificial intelligence 36:11819–11827

Bluetick: Bluetick Vinden zonder zoeken [Bluetick Find without searching (free translation)]. https://www.bluetick.nl/

Brasil: Resolução no 332 de 21/08/2020. dispõe sobre a ética, a transparência e a governança na produção e no uso de inteligência artificial no poder judiciário e dá outras providências. DJe/CNJ, no 274 (2020). National Council of Justice

Brazilian Supreme Court (2021) Activity report of the Brazilian Supreme Court, 2020. Technical Report Year 2020, Supremo Tribunal Federal, Brasília, Brazil. Original title in Portuguese: "Relatório de atividades do Supremo Tribunal Federal, 2020.". https://bibliotecadigital.stf.jus.br/xmlui/handle/123456789/2779

Brazilian Supreme Court (2021) Victor Project advances in research and development to identify themes of general repercussion. Original title in Portuguese: "Projeto Victor avança em pesquisa e desenvolvimento para identificação dos temas de repercussão geral". https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=471331&ori=1

Brazilian Supreme Court (2022) Activity report of the Brazilian Supreme Court, 2021. Technical Report Year 2021, Supremo Tribunal Federal, Brasília, Brazil. Original title in Portuguese: "Relatório de atividades do Supremo Tribunal Federal, 2021.". http://bibliotecadigital.stf.jus.br/xmlui/handle/123456789/3775

Brazilian Supreme Court (2024) Activity report of the Brazilian Supreme Court, 2023. Technical Report Year 2023, Supremo Tribunal Federal, Brasília, Brazil (2024). Original title in Portuguese: "Relatório de atividades do Supremo Tribunal Federal". http://bibliotecadigital.stf.jus.br/xmlui/handle/123456789/5941

Brazilian Supreme Court (2023) Justice Rosa Weber launches robot VictórIA for judicial processes clustering and classification. Original title in Portuguese: "Ministra Rosa Weber lança robô VitórIA para agrupamento e classificação de processos". https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=507426&tip=UN

Brazilian Supreme Court (2022) Documentation RAFA. Original title in Portuguese: "Documentação RAFA 2030". https://agenda2030rafa.github.io/rafa_documentacao/

Brazilian Supreme Court (July 2022) STF Bulletin. Technical Report Edition 3, Brazilian Supreme Court, Brasília, Brazil. https://www.stf.jus.br/repositorio/cms/portalStfInternacional/portalStfAgenda_en_us/anexo/STF_Bulletin_edition_3_2022_1682022.pdf

Brazilian Supreme Court (October 2022) STF Bulletin. Technical Report Edition 4, Brazilian Supreme Court, Brasília, Brazil. https://www.stf.jus.br/arquivo/cms/jurisprudenciaInternacional/anexo/STF_Bulletin_4.pdf

Brazilian Supreme Court: Digital Library. Original title in Portuguese: "Biblioteca Digital". https://bibliotecadigital.stf.jus.br/xmlui/handle/123456789/5941

Brazilian Supreme Court: Open Court. Original title in Portuguese: "Corte Aberta". https://transparencia.stf.jus.br/extensions/corte_aberta/corte_aberta.html

Brazilian Supreme Court: Virtual Plenary. Original title in Portuguese: "Plenário Virtual". https://portal.stf.jus.br/hotsites/plenariovirtual/

Buscador Dizer o Direito: Buscador Dizer o Direito: Encontre jurisprudência comentada do STF e do STJ e muito mais [Search engine Saying the Law: Find commented jurisprudence of the STF and STJ and much more (free translation)]. https://www.buscadordizerodireito.com.br/

Caled D, Won M, Martins B, Silva MJ (2019) A Hierarchical Label Network for Multi-label EuroVoc Classification of Legislative Contents. In: Doucet A, Isaac A, Golub K, Aalberg T, Jatowt A (eds.) Digital Libraries for Open Knowledge. Lecture Notes in Computer Science, vol. 11799, pp. 238–252. Springer, Cham. https://doi.org/10.1007/978-3-030-30760-8_21. https://link.springer.com/chapter/10.1007/978-3-030-30760-8_21

Casetext: Compose: Craft exceptional briefs, without the busy work. https://compose.law/

Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung Y-H, Strope B, Kurzweil R (2018) Universal Sentence Encoder. arXiv:1803.11175 [cs]

Chan A, Sanjabi M, Mathias L, Tan L, Nie S, Peng X, Ren X, Firooz H (2022) UNIREX: A Unified Learning Framework for Language Model Rationale Extraction. In: Proceedings of BigScience Episode #5 - Workshop on Challenges & Perspectives in Creating Large Language Models, pp. 51–67. Association for Computational Linguistics, virtual+Dublin.https://doi.org/10.18653/v1/2022.bigscience-1.5. https://aclanthology.org/2022.bigscience-1.5 Accessed 2023-08-25

DeYoung J, Jain S, Rajani NF, Lehman E, Xiong C, Socher R, Wallace BC (2020) ERASER: A Benchmark to Evaluate Rationalized NLP Models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458. Association for Computational

Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.408. https://aclanthology.org/2020.acl-main.408

Domingues LER (December 2021) Inferring and explaining potential citations to binding precedents in Brazilian Supreme Court Decisions. BSc thesis, Fundação Getulio Vargas, Rio de Janeiro, Brazil. http://bibliotecadigital.fgv.br:80/dspace/handle/10438/31845

Finch Platform: Finch - Simplificando o mundo jurídico [Finch - Simplifying the legal world (free translation)]. https://finchsolucoes.com.br/

González-González J, Arriba-Pérez F, García-Méndez S, Busto-Castiñeira A, González-Castaño FJ (2023) Automatic explanation of the classification of Spanish legal judgments in jurisdiction-dependent law categories with tree estimators. J King Saud Univ- Comput Info Sci 35(7):101634. https://doi.org/10.1016/j.jksuci.2023.101634

Halko N, Martinsson PG, Tropp JA (2011) Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev 53(2):217–288. https://doi.org/10.1137/090771806

Katz DM, Hartung D, Gerlach L, Jana A, Bommarito II MJ (2023) Natural Language Processing in the Legal Domain. arXiv:2302.12039 [cs]. https://doi.org/10.48550/arXiv.2302.12039

Kleinbaum DG, Klein M (2010) Logistic regression. Springer, New York

Leskovec J, Rajaraman A, Ullman JD (2020) Mining of Massive Datasets, 3rd edn

Lexis: Lexis - Online legal research. https://www.lexisnexis.com/en-us/products/lexis.page

Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz H-W, Sartor G, Torroni P (2019) CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artif Intell Law 27(2):117–139. https://doi.org/10.1007/s10506-019-09243-2

Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017), vol. 30. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html Accessed 2021-11-15

Lyu Y, Wang Z, Ren Z, Ren P, Chen Z, Liu X, Li Y, Li H, Song H (2022) Improving legal judgment prediction through reinforced criminal element extraction. Info Proc Manag 59(1):102780. https://doi.org/10.1016/j.ipm.2021.102780

National Council of Justice: CNJ Bulletin. Technical report, National Council of Justice, Brasília, Brazil (2023). https://www.cnj.jus.br/wp-content/uploads/2023/08/justica-em-numeros-2023.pdf

National Council of Justice: Panel of AI Projects in the Judiciary Branch - 2022. Original title in Portuguese: "Painel de Projetos de IA no Poder Judiciário - 2022". https://www.cnj.jus.br/sistemas/plataforma-sinapses/paineis-e-publicacoes/

OABJuris: OABJuris Jurisprudência de uma forma mais ágil e eficaz [OABJuris Jurisprudence in a more agile and effective way (free translation)]. https://jurisprudencia.oab.org.br/

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Platt J et al (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 10(3):61–74

Prescott R, Mariano R (2019) Victor, the STF's AI, reduced task time from 44 minutes to five seconds. Original title in Portuguese: "Victor, a IA do STF, reduziu tempo de tarefa de 44 minutos para cinco segundos". https://www.convergenciadigital.com.br/Inovacao/Victor%2C-a-IA-do-STF%2C-reduziu-tempo-de-tarefa-de-44-minutos-para-cinco-segundos-52015.html

Refaeilzadeh P, Tang L, Liu H (2009) Cross-Validation. In: LIU, L., ÖZSU, M.T. (eds.) Encyclopedia of Database Systems, pp. 532–538. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

Resck LE, Ponciano JR, Nonato LG, Poco J (2023) LegalVis: Exploring and Inferring Precedent Citations in Legal Documents. IEEE Transactions on Visualization and Computer Graphics 29(6):3105–3120. https://doi.org/10.1109/TVCG.2022.3152450. Presented at IEEE VIS: Visualization & Visual Analytics 2022. Date of Publication: 18 February 2022

Resck L, Raimundo MM, Poco J (2024) Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024, pp. 4190–4216. Association for Computational Linguistics, Mexico City, Mexico. https://doi.org/10.18653/v1/2024.findings-naacl.262. Also presented as a poster at the LatinX in NLP at NAACL 2024 workshop. https://aclanthology.org/2024.findings-naacl.262 Accessed 2024-07-01

Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. KDD '16, pp. 1135–1144. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/2939672.2939778

Salomão LF, Tauk CS, Leme E, Loss J, Nunes D, Bragança F, Coelho JL, Braga R (2023) Artificial Intelligence: Technology Applied to Conflict Management Within the Scope of the Brazilian Judiciary, 3rd edn. Fundação Getulio Vargas, Rio de Janeiro, Brazil. Original text in Portuguese: "Inteligência Artificial: Tecnologia Aplicada à Gestão dos Conflitos no Âmbito do Poder Judiciário Brasileiro". https://ciapj.fgv.br/sites/ciapj.fgv.br/files/relatorio_ia_3a_edicao_0.pdf

Santos EFd (2024) Os riscos e possibilidade da inteligencia artificial no poder judiciario: estudo de caso do projeto victor

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. IEEE, Venice, Italy. https://doi.org/10.1109/ICCV.2017.74. ISSN: 2380-7504. https://ieeexplore.ieee.org/document/8237336

Semo G, Bernsohn D, Hagag B, Hayat G, Niklaus J (2022) ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US. In: Aletras, N., Chalkidis, I., Barrett, L., Goanţă, C., Preoţiuc-Pietro, D. (eds.) Proceedings of the Natural Legal Language Processing Workshop 2022, pp. 31–46. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid). https://doi.org/10.18653/v1/2022.nllp-1.3. https://aclanthology.org/2022.nllp-1.3

Sundararajan M, Taly A, Yan Q (2017) Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR, Sydney, Australia. ISSN: 2640-3498. https://proceedings.mlr.press/v70/sundararajan17a.html

Thomson Reuters: Legal One, a solução jurídica que se adequa à sua realidade [Legal One, the legal solution that suits your reality (free translation)]. https://www.thomsonreuters.com.br/pt/juridico/legal-one.html

Thomson Reuters: Westlaw - Legal research tools & platforms. https://legal.thomsonreuters.com/en/products/westlaw

Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Warehousing Mining (IJDWM) 3(3):1–13

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is All You Need. In: Proceedings of the 31st international conference on neural information processing systems, pp. 6000–6010. Curran Associates Inc., Long Beach, California, USA. https://doi.org/10.5555/3295222.3295349

Wang Y, Zhou Z, Jin S, Liu D, Lu M (2017) Comparisons and Selections of Features and Classifiers for Short Text Classification. In: IOP Conference Series: Materials Science and Engineering, vol. 261, p. 012018. IOP Publishing, Hawaii, USA. https://doi.org/10.1088/1757-899X/261/1/012018

Wu Y, Liu Y, Lu W, Zhang Y, Feng J, Sun C, Wu F, Kuang K (2022) Towards Interactivity and Interpretability: A Rationale-based Legal Judgment Prediction Framework. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 conference on empirical methods in natural language processing, pp. 4787–4799. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. https://doi.org/10.18653/v1/2022.emnlp-main.316. https://aclanthology.org/2022.emnlp-main.316

Zhao Q, Gao T, Guo N (2023) Legal judgment prediction via legal knowledge fusion and prompt learning. Rochester, New York

Zhong H, Wang Y, Tu C, Zhang T, Liu Z, Sun M (2020) Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. Proceedings of the AAAI conference on artificial intelligence 34:1250–1257. https://doi.org/10.1609/aaai.v34i01.5479

Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 5218–5230. Association for computational linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.466. https://aclanthology.org/2020.acl-main.466

Zhu W, Zhang W, Li G-Z, He C, Zhang L (2016) A study of damp-heat syndrome classification using Word2vec and TF-IDF. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM), pp. 1415–1420. IEEE, Shenzhen, China. https://doi.org/10.1109/BIBM.2016.7822730. https://ieeexplore.ieee.org/document/7822730

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

Authors and Affiliations

Lucas Resck[1] · **Felipe Moreno-Vera**[1] · **Tobias Veiga**[1] · **Gerardo Paucar**[1] · **Ezequiel Fajreldines**[3] · **Guilherme Klafke**[3] · **Luis G. Nonato**[2] · **Jorge Poco**[1]

✉ Lucas Resck
lucas.domingues@fgv.edu.br

Felipe Moreno-Vera
felipe.moreno@fgv.br

Tobias Veiga
tobsv21@gmail.com

Gerardo Paucar
carlos.malqui@fgv.br

Ezequiel Fajreldines
ezequiel.santos@fgv.br

Guilherme Klafke
guilherme.klafke@fgv.br

Luis G. Nonato
gnonato@icmc.usp.br

Jorge Poco
jorge.poco@fgv.br

[1] School of Applied Mathematics, Fundação Getulio Vargas, 190 Praia de Botafogo,
Rio de Janeiro, RJ 22250-900, Brazil

[2] Institute of Mathematical and Computer Sciences, University of São Paulo, 400 Avenida
Trabalhador São-carlense, São Carlos, SP 13566-590, Brazil

[3] Law School, Fundação Getulio Vargas, 233 Rua Rocha, São Paulo, SP 01313-020, Brazil