University of Brasília

Faculty of Technology
Department of Electrical Engineering

A new machine learning based framework to classify and analyze industry-specific regulations

Letícia Moreira Valle

A thesis submitted to the University of Brasília in partial fulfillment of the requirement for the degree of Doctor of Electrical Engineering.

Supervisor
Prof. Dr. Ugo Silva Dias

Brasília
2023

University of Brasília

Faculty of Technology
Department of Electrical Engineering

A new machine learning based framework to classify and analyze industry-specific regulations

By
Letícia Moreira Valle

A thesis submitted to the University of Brasília in partial fulfillment of the requirement for the degree of Doctor of Electrical Engineering.

Examination Board

Prof. Dr. Ugo Silva Dias,
EnE/University of Brasília
Supervisor

Prof. Dr. Georges Daniel Amvame Nze,
EnE/University of Brasília
PPGEE member

Prof. Dr. Fernando de Barros Filgueiras,
Federal University of Goiás
External member

Prof. Dr. Diego Lisboa Cardoso,
Federal University of Pará
External member

Publication: PPGEE 196/23

Brasília, May 2023

Acknowledgments

First and foremost, I would like to express my deepest gratitude to the University of Brasília for giving me the theoretical knowledge, endurance, and endless curiosity I needed to undertake this research project. In particular, I would like to thank my advisor, Prof. Ugo Silva Dias, for his patience and guidance over the last few years, always helping me to solve problems and overcome difficulties in the best possible way. In addition, my special thanks go to Prof. Daniel Guerreiro and Silva, who was always available to discuss ideas and techniques during this entire research project.

Also, this endeavor would not have been possible without the support of the National School of Public Administration (ENAP), which financially supported this research and believed in the potential of its results. I would like to extend my sincere thanks to Mr. Leonardo Monteiro Monastério, former General Coordinator of Data Science, for introducing me to this challenge and putting me in touch with the people needed to make it happen, and Pedro Masson, current General Coordinator of Data Science, who accompanied me in the development of the research, facilitating talks with the regulatory agencies, and in communicating the results. Furthermore, I would be remiss in not mentioning the research team of undergraduate and graduate students who at some point worked on the development of this study; your support was essential for the construction of this work.

Finally, I also want to thank my family, especially my parents Francisco and Valéria Valle, my brother Fernando César, and my husband Victor Ferreira, for their enormous patience and the invaluable love and support they have given me throughout my life and especially in difficult times.

Abstract

Government transparency and openness are crucial elements in advancing the modernization of the state. The combination of transparency and digital information has given rise to the concept of Open Government, which increases citizen understanding and monitoring of government actions, improving the quality of public services and the government decision making process. To enhance legislative transparency and comprehension of the Brazilian regulatory process and its characteristics, this study introduces RegBR, the first national framework designed to centralize, classify, and analyze regulations from the Brazilian government.

RegBR employs automated ETL routines, data mining and machine learning techniques to construct a centralized database of Brazilian federal legislation. The framework evaluates various natural language processing (NLP) models in a text classification task on a novel Portuguese legal corpus and conducts regulatory analysis based on metrics that pertain to linguistic complexity, restrictiveness, popularity, and industry-specific citation relevance. This study proposes and presents metrics policymakers can use to assess their own work, thereby increasing the openness and transparency of the public process while also facilitating new research in the area of Brazilian regulatory impact.

According to Google Analytics data, the popularity metric and the regulatory flow pages rank as the fourth and fifth most visited web pages in the Infogov website, respectively, indicating significant interest in the information made available by RegBR. Specifically, these two pages alone received over 800 unique views within the first two months of 2023.

Keywords: Government Transparency, Natural Language Processing (NLP), Text Classification, Machine Learning, Regulation Metrics

Resumo

A transparência e abertura do governo são elementos cruciais no avanço da modernização do estado. A combinação entre transparência e informação digital deu origem ao conceito de Governo Aberto, que aumenta o entendimento do cidadão sobre processos governamentais e como consequência o monitoramento das ações do governo, o que, por sua vez, melhora a qualidade dos serviços públicos e do processo decisório governamental. Com o objetivo de melhorar a transparência legislativa e a compreensão do processo regulatório brasileiro e suas características, este estudo apresenta o RegBR, a primeira estrutura nacional projetada para centralizar, classificar e analisar as regulamentações do governo brasileiro.

O RegBR emprega rotinas de ETL automatizadas e técnicas de mineração de dados e aprendizado de máquina para construir um banco de dados centralizado da legislação federal brasileira. A estrutura avalia vários modelos de processamento de linguagem natural (NLP) em uma tarefa de classificação de texto em um novo corpus jurídico português e realiza análises regulatórias com base em métricas que dizem respeito à complexidade linguística, restritividade, popularidade e relevância de citação específica dos setores da economia. Este estudo propõe e apresenta métricas que podem ser usadas pelos formuladores de políticas para avaliar seu próprio trabalho, aumentando assim a abertura e a transparência do processo público, além de facilitar novas pesquisas na área de impacto regulatório brasileiro.

De acordo com os dados do Google Analytics, as paginas contendo a métrica de popularidade e o fluxo regulatório são a quarta e a quinta páginas mais visitadas no site Infogov, respectivamente, indicando interesse significativo nas informações disponibilizadas pela RegBR. Especificamente, essas duas páginas sozinhas receberam mais de 800 visualizações únicas nos dois primeiros meses de 2023.

Keywords : Transparência Governamental, Processamento Natural de Linguagem (PNL), Classificação de Texto, Aprendizado de Máquina, Métricas de Regulação

Contents

List of Figures

List of Tables

1 Introduction

In recent years, information and communication technologies (ICTs) have played a key role in promoting openness and transparency in government activities [1]. These aspects are key elements to increase trust in government, informed decision-making, and democratic participation [2, 3].

The ICT concept is actually an umbrella term that includes any sort of technological tools and resources used to store, process, transmit, share or exchange information, e.g. software applications and operational systems, web based information and applications such as websites and blogs, broadcast technologies including radio, television or podcasts and others infrastructure and components that enable modern computing [4]. Besides being essential in modern daily-basis life, these tools are a cost-effective and convenient way to promote openness and transparency by allowing the citizen to track activities of common interest and by allowing monitoring, controlling and discussing behaviors and tendencies [5]. In fact, in the last decades, the scope of e-government studies is expanding to consider not only government basic operations and service delivery but also to enable citizen participation and engagement using technology tools [6].

In the context of government regulations, ICTs provide new tools for governments to manage regulatory information, to advance public access to regulations, and to improve the transparency of the regulatory process. This is particularly important as the impacts and consequences of government regulations have been studied for decades, and they are considered a crucial policy tool for tackling market inefficiencies [7], to foster economic growth and develop a more prosperous society. However, the complexity of the market and political processes are some of the challenges that are encountered in discussing and evaluating regulations [8]. For instance, regulations can increase the barrier to entry a specific market and, therefore, make it less efficient [9].

With the general goal of creating a public tool that helps to improve the transparency of Brazilian regulatory data, this thesis research presents the RegBr framework, a tool that enables citizens to track regulations of common interest and the policy maker to measure the quantity and quality of the regulation that he/she produces and that serve as a database for future studies on Brazilian regulation.

Within this context, RegBR also leverages the BOLD concept, or Big and Open Linked Data, when creating a network of linked regulatory data that can be easily queried and visualized by citizens, policymakers, and researchers.

BOLD is a new paradigm in data management that emphasizes the interoperability and integration of data from diverse sources, based on the concepts of Open Data, Linked Data and Big Data and having the potential to transform government and its interactions with the public [10]. In the context of regulatory data, BOLD can be particularly useful in promoting transparency and accountability by enabling citizens and policymakers to access and analyze regulatory data from diverse sources. By adopting the BOLD concept, RegBR can provide citizens with a powerful tool to monitor and access centralized Brazilian regulation, being able to interact with the tool via the government website Infogov, https://infogov.enap.gov.br/, a federal government data portal, or even download the database in order to create research using the RegBR's metadata.

The work presented in this thesis aims to provide a comprehensive overview of the RegBR tool and demonstrate its potential to improve the transparency and accountability of the regulatory process in Brazil, as well as to serve as a database for future studies on Brazilian regulation.

1.1 Research subject

The main purpose of this study is the creation of a new machine learning based framework to classify and analyze industry-specific regulations. In addition to the general concern related to increasing Brazilian legislative transparency, this study also aims to make contributions in machine learning, especially in the legal domain. Despite recent advancements in Natural Language Processing (NLP) applications, their usage in the legal domain is still relatively under-explored [11]. Some of the challenges in this area include the scarcity of relevant labeled documents, the financial and time related cost of classifying these documents (often depending on a domain-expert such as a lawyer or law student) and the documents' length, typically longer than the standard length used for training NLP models, such as tweets, customer reviews, and other smaller documents. Despite these constraints, the application of machine learning techniques in the law domain is recently gaining ground.

For instance, [12] conducted a comparative study on the performance of various machine learning algorithms in classifying judgments of the Singapore Supreme Court written in English. Similarly, [13] presented results of machine learning algorithms in the task of predicting the field of law to which a case belongs.

Another common NLP application in the law domain is the prediction of court ruling decisions. For example, [14] used extremely randomized trees to predict the US Supreme Court's rulings and, more recently, [15] tackled the task of predicting patent litigation and time to litigation. Finally, [16] proposed a model to predict the verdicts of the European Court of Human Rights (ECRH).

Regarding the application on the regulatory field, an integrated approach that covers the management of regulations, efficient access, and retrieval of regulatory information is often lacking [17]. The creation of an information infrastructure that allows government agents and the general public to compare and contrast different regulatory documents will improve the understanding of regulations and increase government transparency.

Some recent studies and projects are advancing this area of governance. One example of such work is the American initiative RegData [18], created by Mercatus Center at George Mason University, with the goal of quantifying federal regulations by industry and by the regulatory agency for all federal regulations of the United States. The metrics expanded in RegData include a measurement of the applicability of each regulation to each one of the industries that comprise the US economy using information from the regulatory text. Their work acts as both a framework and a database to analyze regulations in the US. This methodology was also expanded to other countries, such as Australia [19] and Canada [20]. This genre of work is especially interesting as once researchers achieve to objectively measure regulations, policymakers can use this information to gather insights about their impacts and even to evaluate their own work, in terms of whether the legislative characteristics are in accordance with what was initially planned by the government.

Some other organizations are also developing frameworks to help countries evaluate the design and implementation of their regulatory policy. One of the best examples is the Regulatory Policy Committee from the Organization for Economic Cooperation and Development (OECD). Its objective is to assist countries in building and strengthening capacity for regulatory quality and regulatory reform [21].

Among several indicators compiled by OECD that capture the level of anti- competitive regulation in the economy, the most relevant is the Product Market Regulation (PMR). The economy wide PMR indicator covers state control, barriers to entrepreneurship, and barriers to trade and investment [22]. However, in the context of developing economies, the regulation impacts are not as well-known and studied as in developed countries. Specifically, there is no framework to study and analyze regulations tendencies in Brazil in the relevant literature. The country figured on the bottom three countries in the economy wide PMR indicator for 2018 [23].

1.2 Motivation

In the past decade, governments around the world have been focusing efforts on trying to predict the impacts of a new regulation before it is actually published, but they have been paying remarkably little attention to analyzing regulations after adoption or to evaluating the impacts of the procedures and practices that govern the regulatory process itself [21].

As Brazil ranks at the 46th place out of 48 countries evaluated by OECD in terms of regulatory performance, the country urgently needs scientific contributions that study regulatory policies and help to metrify the legislative process, to allow policymakers to better understand whether regulations efficiently achieve their intended goals or to prioritize regulations that may need reforms.

In addition to promoting studies that support improvements in the regulatory performance, the government should use ICTs to promote openness and transparency about their legislative operations and decisions. As stated before, despite recent advances in information technology, an integrated approach that covers the management of regulations, efficient access, and retrieval of regulatory information is often lacking.

As previously mentioned, some recent studies and projects are advancing this area of governance and that is particularly important as several relationships between industry regulation and economic interests can be drawn from analyzing data. This thesis proposes a framework applicable to Brazil called RegBR, produced in partnership with the National School of Public Administration (Escola Nacional de Administração Pública, ENAP), which aims to produce relevant information on the national regulatory situation. Instead of responding to a citizen's demand to access some information, RegBR already deliver information to the citizens in a simple and visually friendly way, centralizing information of different sourcers, compiling results and reducing access costs.

In addition to the direct use by the population, RegBR can have several applications to the federal government. First, the framework and its data can subsidize new regulatory studies. For instance, the Brazilian Public Service Journal (*Revista do Serviço Público*) opened a call for papers using RegBR as its data source.

Second, RegBR can assist regulatory agencies decision makers in measuring their own work, i.e., the tool allows the heads of regulatory agencies to measure what their organization produces in terms of volume and characteristics of regulations. It allows managers to have concrete parameters to quantify and monitor regulations, such as: restrictiveness, measure of interest, influence of the regulated sector in the economy, and linguistic complexity of the regulations. In this context, the decision maker who wants, for example, to make some specific sector of the economy less regulated, can use RegBR to evaluate how the regulations produced by his organization behave over time.

Third, RegBR can also be used as a monitoring tool. The framework allows the Brazilian Federal Government to monitor its regulatory production, measuring lengthwise the number of acts that have been produced. A salient consideration pertains to the use of RegBR in measuring the quantity of normative instruments and their legal status. This instrument holds the potential for employment by the Federal Government analyzing the impact of the Decree 10.139 and other analogous endeavors with comparable objectives.

Also, RegBR can be used as a comparative apparatus allowing the Brazilian Federal Government to compare its normative production by industry, by regulator and by metrics with distinct countries that already have similar metrics, like United States, Canada and Australia through RegData initiative for example.

Then, this study presented in this thesis can be used as a predictor of regulatory governance design in Brazil by facilitating analysis of regulatory trends and practices, as well as identifying areas and economic sectors where regulation improvements may be needed. Lastly, the proposed framework can facilitate the achievement of the proposal of the Brazilian Law 12.527 from 2011, the Access to information law, by ensuring access to data and fomenting active transparency to public information.

1.3 Research objectives

Following the elucidation of this study research subjects and motivations, the broad aim of this investigation is to develop the first national framework to centralize, classify and analyze regulations from the Brazilian government. This goal is further supported by the following specific objectives.

- Improve the understanding of regulatory trends and helps policymakers to better identify and prioritize regulations that may need reforms.
- Be a a public data tool that aims to increase government transparency and facilitate the access to public information since federal regulations used in this work were gathered and centralized in a public database for ease of access. Therefore, legislative transparency increases together with the popular awareness about the legislation.
- Contribute to the literature of Portuguese text classification field at large by benchmarking the different text classification techniques employed, such as Bag-of-Words (BoW), Word Embedding and Transfer Learning against a test subset of the federal normative legislation corpus acquired, therefore providing a blueprint to future developments in the field.
- Create a framework that presents a clearer picture of the Brazilian regulatory scenario from the methods and metrics proposed by this work. Propose a work process that adequately integrates said framework with the processes of extracting, transforming, analyzing and making the information available for download and consultation in a public government site (https://infogov.enap.gov.br/regbr)

1.4 Methodology

For the purposes outlined above, this work aggregates and processes data from many different decentralized sources and applies many different ETL (Extract, Transform, and Load) techniques. It builds bots and data pipelines responsible for scrapping and cleaning data from the official government websites of the leading regulatory agencies in Brazil to consolidate a novel database of federal legislation that could be used in the regulatory metrics analysis.

After the data extracting and cleaning process, the proposed framework applies NLP techniques to classify federal legislation regarding their CNAE's areas[1]. In that context, this work contributes to the literature of Portuguese NLP at large by benchmark the main different text classification techniques employed, such as Bag-of-Words (BoW) [24], Word Embedding [25], and Transfer Learning [26] against a test subset of the federal normative legislation corpus acquired, therefore providing a blueprint to future developments in the field. We consider this an essential contribution to the NLP field in Portuguese, which, unlike its English counterpart, does not have many contributions in the area.

In the context of normative acts text classification, we can formally define $d_i \in$ D as a document from a set of normative act texts $D = \{d_0, d_1, d_2, ..., d_n\}$ and $c_i \in$ C as a label from a set of labels $C = \{c_0, c_1, c_2, ..., c_{18}\}$, which represents the eighteen different classes based on the Brazilian Institute for Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística*, IBGE) economic sector classification. Hence, we define Text Document Classification (TDC) as the task of assigning d_i to c_i in order to structure dataset efficiently and accurately [27].

After classifying the federal legislation into the affected economic sectors, this work applies different proposed metrics that measure different aspects of regulation, which can be monitored by different economic sectors, such as linguistic complexity,

[1] The CNAE (in portuguese *Classificação Nacional de Atividades Econômicas*) is used to divide economic activities into different sectors.

restrictiveness, citation influence and measure of interest over time. For the creation of metrics that allows regulators and policymakers to better identify and prioritize regulations that may need reform, we propose a set of metrics $M = \{m_0, m_1, m_2, m_3\}$, described as:

- Restrictiveness (m_0): indicates the regulatory restriction counts and how the sectors of the economy have become regulated by more or less restrictive laws over the years. This metric is adapted from [18].
- Industry citation relevance (m_1): Calculates the relevance of regulations to economic sectors and industries, based on the frequency of citations of these sectors' keywords in the general context of normative acts. This metric is also adapted from [18], including modifications presented in Section 6.3 of this work.
- Measure of interest (m_2): Indicates how popular a law is for the population, based on the active search for that law on Google, and the frequency of citations of the law in the Official Gazette of the Federal Government. This metric is a novel contribution from RegBR.
- Linguistic complexity (m_3): Uses the median sentence length, the frequency of conditional terms, and Shannon's entropy to measure the linguistic complexity of a document. These metrics are adapted from [20].

Next sections present in more detail the research design and the data modeling techniques used in this work.

1.5 Presentations and Publications

The RegBr framework was conceptualized and developed as the main output of this doctoral research project in a partnership between the University of Brasilia (UnB) and the National School of Public Administration (ENAP). After its first version release, in July

2021, some lectures and presentations were held by me at forums focused on Machine Learning and Government innovation. Among them, the following are highlighted:

- *RegBR: Mais de Cinco Décadas de Atos Normativos Federais* presented as a webinar for the Institute of Applied Economic Research (IPEA) on 9th september 2021.
- *RegBr: Mais de Cinco Décadas de Normativos Federais* presented at the 7th International Seminar on Data Analysis event organized by the Union Court of Auditors (TCU) on 22nd October 2021.

Its second version was released in May 2022 after a few rounds of conversations with Regulatory Agencies to understand which improvements would bring more value to the framework. At that time, two articles were under review and were eventually accepted for publication.

- *RegBR: A novel Brazilian government framework to classify and analyze industry-specific regulations* - The main article of this research, containing the groundwork of the study, the technical presentation of text classification and metrics creation and its first results. The article was accepted for publication at PLOS ONE (A1) on September 19th 2022 and published on September 28th 2022 at https://doi.org/10.1371/journal.pone.0275282.
- *RegBR: An Application Overview* - A paper focused on RegBR applications for the Brazilian Federal Government. The article was accepted for publication at the Public Service Magazine (RSP) (A4) on July 7th 2022 and does not have a definite date for publication, what should happen by the first semester of 2023.

Soon, we expect that new updates on the framework will include new features that could feed more discussions and publications about RegBr.

1.6 Organization

Next Chapter brings a more detailed exposition of the national and international literature about transparency and regulatory impact studies, as well as an literature review on Natural Language Processing, from the classic preprocessing techniques developed throughout the last 3 or 4 decades, to the present state-of-the art with transformers and other complex Natural Language Understanding models. The third Chapter will discuss the concept of text classification and the results and discussion of the normative text classification applied in the normative acts and Chapter 4 will introduce the regulatory metrics conception, results, and discussion. Chapter 5 discuss six of many applications of the RegBr on the federal government. Finally, Chapter 6 concludes this work with the major conclusions remarks and next steps.

2 Background and Related Work

This section presents a brief review of related literature on the use of information and communication technologies to promote openness and transparency tools for society and of the impact of regulations as policy tools. Also, this section also presents a summary of NLP, a subfield of linguistics, computer science, and artificial intelligence areas, and of Text classification techniques, widely used in this doctoral research.

2.1 Transparency, Openness and Open Government

2.1.1 ICTs as transparency tools

The usage of ICTs to promote openness and transparency has been increasing in recent years, as it is often considered cost-effective and convenient [1]. E-government initiatives aim to improve the efficiency, effectiveness, and quality of government services and activities by employing ICTs. They have modified not only the organization of activities and processes within organizations but also the public perception on governmental institutions and policy-making decisions [28, 29]. Some benefits of ICTs adoption include promoting good governance, strengthening reform-oriented initiatives and enhancing relationships between government and citizens, with successful cases from governments across the Americas, Asia, and Europe [30].

The scope of e-government studies is expanding to consider not only government basic operations and service delivery, but also to enable citizen participation and engagement using technology tools [6]. In other words, the interactions between governments and non-state stakeholders is gaining more importance recently. For example, some governments are adopting social media to disseminate complementary information and provide participation channels to citizens, with the goal to increase perceptions of government transparency [31]. This is a consequence of making up-to-date government information easily accessible and providing more interactions of the government agencies with the public. These are some efforts that are part of the open government concept, which is based on making public information access freely and more efficient.

Open Government can also be defined as using compatible standards and architectures to improve people's access to data [32]. Finally, being an open government is not only related to making information available but also opening a communication channel with the citizens, informing and receiving feedback, and, more importantly, acting based on feedback. This process puts governments into a more active and collaborative role and such process is critical to reach the full open government state [6]. For instance, designing more intuitive websites with a focus on creating reliable and publicly accessible infrastructure that "exposes" the underlying data is essential for the success of open government initiative [33].

Indeed, intuitive web applications are essential in bridging the gap between the government and the public [34, 35]. The use of these data instruments can influence government policy-making and provide more interaction with the public, as they can be interactive and used to release information for both governmental decision-makers and the public [36, 37]. Additionally, dashboards can also be used to verify data integrity and quality, an essential aspect in decision-making contexts [38]. These tools help to increase transparency, governance and trust in the government [39]. Moreover, they allow citizens to participate in the decision-making process for new public policies.

2.1.2 Big and Open Linked Data for transparency and informed decision

Big and Open Linked Data (BOLD) refers to the massive amounts of data that are being generated by various sources and made available for public use. The term *linked* refers to the fact that the data is interconnected, with each piece of information being linked to other pieces of data, creating a vast network of information [40] .

BOLD operates on the principle of transparency, which means that the data is made available to the public in a way that is accessible, understandable, and usable [10]. Providing information alone is not sufficient and mechanisms are necessary for ensuring that the information can be easily accessible, processed and interpreted [41]. The accessibility can be achieved using standardized formats, such as RDF and JSON, which allow the data to be easily exchanged and processed by different systems.

The goal of BOLD is to provide a wealth of information that can be used by individuals, organizations, and governments to make informed decisions. This data can be used to track trends, monitor progress, and identify areas where improvements are needed.

In the context of transparency, BOLD plays an important role in ensuring that the public has access to information that is relevant to their lives. For example, government agencies may use BOLD to publish information about their activities, such as budgets, expenditures, and contracts. This allows citizens to see how their tax dollars are being spent and to hold their government accountable for its actions.

Similarly, businesses may use BOLD to share information about their products, services, and operations. This can help consumers make more informed purchasing decisions and hold companies accountable for their actions.

Overall, BOLD is a powerful tool for promoting transparency and empowering individuals and organizations to make informed decisions. As more and more data become available, it is essential that we continue to develop new ways to analyze and interpret this information, so that we can unlock its full potential for the benefit of society.

2.1.3 Review of regulatory impact studies

Government regulations are considered a crucial policy tool for addressing market inefficiencies [7], as they can affect economic agents in different ways. They are intended to correct market failures and, therefore, to increase economic efficiency and growth. In practical terms, the government can also intervene and regulate cases where people should have access to certain services and goods regardless of the ability to pay, such as health care and education services.

Most of the impacts are the consequence of constraints or expansions of their legal rules, and this intervention can play a critical role in successful development efforts on the economy [42, 43]. However, the complexity of the market and other political processes often result in regulations that are not honorably created [44, 45]. In other words, regulations driven by particular interest of specific groups lobbying for legislative changes that result in personal gain, e.g., *rent-seeking* behavior [46]. Even when this is not the case, regulations may result in unintentional consequences [8] or may do so at an unsatisfactory cost in terms of economic distortion.

Regulations can act as a potential aid or risk to every industry in the economy. By acting as force that has power to stimulate or restrain, to take or give resources, the government can help or hurt a vast number of industries at its discretion. Thus, it is crucial to evaluate the causal effect of regulations on the economy sectors. However, the studies that examine these impacts are often focused on one specific regulation or sector. Some examples include examining the impact of liquidity regulation on the banking sector [47], the differences in regulation for collaborative economy peer-to-peer accommodation in different European cities [48], and regulatory risk and the resilience of new sustainable business models in the energy sector [49] in Germany. Compared to thousands of existing regulations that govern a country's economy, these studies are comparatively limited in scope.

One alternative to study and analyze the impact of regulations is to create indicators based on the time evolution of the number of legal documents that specify regulations.

Some works such as [50, 51] use data from the Code of Federal Regulations (CFR), an annual publication that contains all regulations issued at the federal level, to create quantitative measurements of federal or state regulations created or in effect in the United States, each year.

These metrics are expanded in RegData [18] and include a measurement of the applicability of each regulation to each one of the industries that comprise the US economy using information from the regulatory text. As stated before, their work acts as both a framework and a database to analyze regulations in the US and this methodology was also expanded to other countries, such as Australia [19] and Canada [20].

As stated in Research subject session, some organizations, like OECD, are developing frameworks to help countries evaluate the design and implementation of their regulatory policy. Also, there have been noteworthy developments to address regulatory issues in some economy sectors.

One example is the logistic infrastructure sector [52]. Despite these developments, it is still imperative to adopt a more strategic perspective on logistics infrastructures. The authors provide several recommendations for enhancing regulation in logistics infrastructure in Brazil.

Another sector in which regulations are constantly being debated in Brazil is the sanitation sector [53]. The main challenge is regarding the government level at which conceding authority should reside and how private operators can fulfill social objectives [54]. Some argue that these issues are not the crucial barriers to the sector's development when one looks at the operators' productivity performance. Instead, they suggest that operators dissipate their productivity potential and apply higher tariffs in the absence of efficiency incentives. That is to say that the debate over the regulatory framework should be redirected to focus on what instruments should be put in place to create incentives to efficiency and increase sharing of the resulting gains with users.

On a different note, the role of government regulation in ecological restoration is discussed in [55]. More specifically, the work focus on the state of São Paulo's goal to increase the effectiveness of tropical forest restoration projects. The authors argue that some points about ecological restoration are still unclear. In other words, there is not a firm consensus yet. Some of these points include whether to set goals for preservation in legal instruments or if legislation on this topic should be delayed until adequate scientific knowledge is available.

It is important to note that although Brazil has regulatory impact studies for specific sectors of the economy, in context of developing economies, including Brazil, the regulation impacts are not as well-known and studied as in developed countries. Specifically, there is not yet a general framework to study and analyze regulations in Brazil.

2.2 Natural Language Processing

Natural language processing is a subarea of Artificial Intelligence (AI) that exists for more than 50 years, having roots in the linguistics field [56]. NLP techniques aim to understand and respond to text or voice data and in the same way humans' beings do.

With the accelerated growth in the use of this area of knowledge in the last decade, many different applications have emerged. Here are a few prominent examples:

- *Smart assistants:* Amazon's Alexa or Apple's Siri are great examples of smart assistants that recognizes speech patterns thanks to voice recognition, inferring meaning and providing a useful response to that interpretation. Modern speech recognition techniques also use context information [57] to make advanced conversations, often applying humor and sarcasms to answer questions. In this way, interactions are expected to grow more and more with the advancement of NPL and speech recognition techniques [57].
- *Search results:* Google's and Bing's search engine are a classic example of search engines applications, they try to understand text inputs and make search more natural and relevant. NPL can interpret search queries written by costumers and

display relevant results based on similar search behaviors or user intent. With the technology advance, it is also expected to have more entity-based search results models replacing classical phrase-based indexing and ranking.

- *Language translation:* Trying to translate a text into another language using only a dictionary without an online language translation can be a real challenge. Different languages frequently don't allow for straight translation and have different orders for sentence structure. In this way, different methods are being created, like rule-based, statistical and example-based machine translation [58] to make text translation more convenient and grammatically correct.

- *Sentiment Analysis:* To analyze costumers social media interactions, such as comments and reviews, brand name mentions and products ranking can help brands to understand how a marketing campaign is doing or even monitor costumers' issues before they became public in order provide a better costumer experience. Also, sentiment analyses can help brands to understand what their pros and cons in costumer view are, helping the brand to improve its product quality.

- *Predictive text:* Some features like autocorrect, autocomplete and predictive texts are super popular and present in today's smartphones, being a huge helper in finishing words and suggesting new ones to make users gain time in texting messages or even write a more understandable one. Like in others NPL applications, predictive texts will learn from the user behavior and customize the suggestions to user's personal language, being very personal to each user experience.

- *Text Classification:* One of the significant tasks in natural language processing with extensive applications [59]. Generally, they classify text inputs into predetermined categories using several different possible approaches like rule based, machine learning or both. Spam detection and movies categorization are common examples of such applications. Next session will delve deeper into this type of application.

2.2.1 The text classification problem

Automatically categorize a text into predetermined categories is a challenge that has gained more and more applications in industry and academia [60]. The text classification techniques typically apply machine learning algorithms to classify documents into the predetermined categories. In fact, Machine Learning and NLP techniques work together to detect and automatically classify patterns from different types of documents, but before applying these advanced techniques in the input text, several preprocessing steps must be performed to improve the text quality.

Text preprocessing techniques

As stated, before applying sophisticate techniques to extract value from texts, it is necessary to pre-process them and clean useless information like characters, numbers and format symbols. Some of the most common preprocessing techniques are described as follow.

- *Tokenization:* Is a fundamental task in any NPL pipeline that consists of separating a piece of text into smaller units called tokens, that can be words, characters or subwords. In general, the tokenizer breaks unstructured data into chunks of information that can be considered discrete elements. The occurrences of these elements can be used as a vector that represents the document.
- *Stopword removal:* Stop words removal is a very important preprocessing step, consisting in removing the words that don't add to the overall meaning of the text. In fact, stop words are available in abundance in any language and by removing these words, typically articles and pronouns, we remove the low-level information in the text to give more focus to the important information.
- *Stemming:* It is common in a text to have different forms of a word, such *organize, organizes,* and *organizing* or derivationally related words with similar meanings, such as *democracy, democratic,* and *democratization*. In many cases, it is useful

to reduce inflectional forms and derivationally related forms of a word to a common base form [61]. With this preprocessing technique, a model can learn that these words are somehow similar and are used in a similar context. Stemming then chops off the ends of the words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes, not always coming up with a stem (word root) that exists grammatically.

- *Lemmatization:* Similar in goal to the stemming preprocessing technique, Lematization uses vocabulary and morphological analysis of words to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma [61]. In this case, the lemma is a grammatically correct word.

2.2.2 Text classification main approaches

The problem design choices have a strong influence on the performance of the learning system trying to solve it [62]. For text classifications problems, the method used to transform words into a numeric representation suitable for the classifier is crucial to determine its efficiency. This article presents three main approaches to represent text data: BoW, Word embedding, and Transfer Learning, each presenting different characteristics. We evaluated these three approaches by increasing complexity and recentness.

Historically, one of the first methodologies used to deal with text data was the BoW approach [24], which consists in transforming each word in a feature and its value is based on the number of times it occurs. The term frequency–inverse document frequency (TF-IDF) [63] is one of the most used BoW method. The main idea of TF-IDF is to increase the value of a feature based on the frequency it appears in a document and based on the inverse document frequency of the same word across all documents. This approach achieved fairly robust results in several tasks and there is a good theoretical basis for its effectiveness [64].

Despite its efficacy, TF-IDF has a serious drawback: it does not capture semantic or syntactic information of words, i.e., there is no relation between the meaning of a specific word and the value it assumes. To deal with this limitation, several word embeddings methods were proposed. Word embeddings capture both semantic and syntactic information of words. In this approach, each word is represented by a multi-dimensional vector, where each entry represents information about that word meaning and context. Some of the most used word embeddings methods are *word2vec* [25], *GloVe* [65] and *fastText* [66].

These word representations are interesting because they explicitly encode many linguistic regularities and patterns, some of them can be represented as linear translations. For instance, the result of a vector calculation such as vec("King") - vec("Man") + vec("Woman") will result in a vector close to vec("Queen"). These approaches achieved impressive results, surpassing BoW in several tasks.

However, one of the limitations of conventional word embeddings is that they are often pre-trained on text corpus from co-occurrence statistics. In other words, they are applied in a context free manner. The word "bank" in "new bank account" and "power bank" would be represented by the same vector encoding. The solution to this problem is based on training contextual representations on text corpus. This was first achieved by training a deep bidirectional language model [67] on a large text corpus. In this case, each word is assigned to an embedding that is a function of the entire input sentence and not only on the specific word.

This approach is expanded by the Bidirectional Encoder Representations from Transformers (BERT)[26], where the authors designed BERT to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. Therefore, this pre-trained model can be easily fine-tuned for a wide range of tasks. The obtained results outperformed previous state-of-the-art models in several tasks and this approach was used in multiple subsequent works [68, 69, 70].

In this study, these three approaches of increasing complexity and recentness were evaluated. The classifiers used for each approach are briefly explained below and more details about the implementation and parameters can be found in Session 3.

3 Normative acts text classification

This session focuses on exploring the theoretical framework used for text classification utilizing machine learning algorithms to categorize text data into pre-defined classes. Specifically, this framework is applied to the analysis of normative acts to identify the sector of the economy that would be impacted by such regulations. The use of various classifiers such as Support Vector Machines (SVMs), word embedding based techniques with deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) and even transformers-based models is explored, and the performance of each algorithms is benchmarked using standard evaluation metrics such as accuracy and F1-score.

In the next chapter, after the identification of the best performing model, a set of regulation metrics is constructed by analyzing the normative acts that have been classified into each sector.

3.1 Machine learning techniques in legal text classification context

As stated earlier in the text, most of the work done in regulatory impact is concerned with one specific sector. A broader analysis depends on having information about which sector a given regulation affects. If this information is not available, one can use text classification to divide regulations into different sectors.

Text classification methods have been successfully applied to several NLP tasks and applications ranging from prediction of academic performance [71], software design pattern selection [72], to learning the taxonomy from a set of text documents [73]. Despite several critical applications of NLP, the classification methods in the legal domain are still relatively under-explored.

One of the reasons this area remains relatively unexplored is the scarcity of labeled documents in the law domain if we use as a comparison the amount of data necessary to train deep learning models. This limitation can be even more aggravated in smaller jurisdictions like Singapore [12], where the number of cases is quite limited.

Another restriction is a consequence of the application in the legal domain. As opposed to labeling images, legal documents are harder to classify and, most of the time, the classification will depend on domain-expert (such as a lawyer or a legal professional) to classify them. That results in a labeling process that may become very expensive in terms of time and financial resources.

The final aspect pertains to the typically extended length of legal texts, which surpasses the standard size employed in common NLP tasks, such as tweets, customer reviews, and other comparatively concise documents. Nevertheless, despite this challenge, the utilization of machine learning techniques within the legal domain has recently experienced an increase in popularity.

As an illustration, a comparative analysis was carried out by [12] to evaluate the efficacy of different machine learning algorithms in categorizing verdicts rendered by the Singapore Supreme Court in the English language. They employed linear models based on BoW, models based on word embedding, and more complex language models such as BERT [26] to classify the data into 31 different legal areas relevant to Singapore's standard law system. They found that classical machine learning models outperform the more recent deep learning-based classifiers on specific metrics, suggesting that the impressive emerging results from these models in several NLP tasks may not carry well into the legal domain without any additional research work.

Argumentation can significantly impact law practices [74] to investigate to which extent an algorithm can identify argumentative propositions in legal text, their argumentative function, and structure. The data used was composed of legal texts extracted from the European Court of Human Rights (ECHR), and their goal was to classify argumentative vs. non-argumentative sentences.

Based on the association between a legal text and its domain label in a database of legal texts, [75] presents a classification approach to identify the relevant domain to which a specific legal text belongs. The features were created by first using TF-IDF to transform the text into numeric features, and the number of features was then reduced by using the information gain as a pruning threshold. A SVM classified the data with a polynomial kernel, and the authors evaluated the results for identifying topics covered by a piece of legislation and for the classification of individual articles.

Some of the works presented in the literature aim to predict court ruling decisions. For instance, [14] employed extremely randomized trees to anticipate the rulings of the US Supreme Court, while [15] recently addressed the task of forecasting patent litigation and the time to litigation. Lastly, [16] proposed a model to forecast the verdicts of the European Court of Human Rights (ECRH).

Another example includes [13], where the authors present results of machine learning algorithms in the task of predicting the decisions of the French Supreme Court and the law area to which a case belongs.

In [76], the authors use three different approaches for judgment prediction. The first classifiers used TF-IDF to extract word features and a SVM [77] as the classification model. They also use FastText [78], a simple and efficient approach for text classification based on Ngrams. The models achieved many results on the accuracy of charges prediction and relevant law articles prediction, but poor results in recall and precision scores.

More similarly to this thesis study, [79] creates a multilabel corpus of legal provisions in contracts. That is achieved by first crawling and scraping Security and Exchange

Commission (SEC) filings and later using machine learning to classify the extracted corpus.

3.2 Research design & data modeling

The RegBr database was designed to store all the necessary information about the economy sectors classification and the creation of general metrics. The use of non-relational databases, such as MongoDB or ElasticSearch, was considered, but for reasons of tool familiarity and systems standardization, PostgreSQL [80] was chosen as the database management system, without prejudice. To construct a more mature data environment, the database has two levels, one intermediate data layer containing the raw data and a second layer containing the transformed data after aggregation and transformation routines. The collection of all federal regulatory acts in a single, centralized, and automated database is one of the main RegBR contributions as it reduces the barrier to data release [81] and increase government transparency, based on BOLD concepts. The compiled database will be available for other researchers who want to use the Federal Regulation data compiled to develop research or other related activities.

3.2.1 Implementation of ETL routines

An essential phase of this research consists of creating ETL routines for extracting, transforming, and loading Brazilian federal normative acts into a database. One of the main difficulties in implementing this activity is that Brazilian regulatory norms are not centralized in a single source. Before starting to collect and extracting federal laws in a decentralized manner, a study was carried out to verify the existence of a centralized database that could already exist in this context. A project called LEXML [82], from the Brazilian Federal Senate, was considered, but because it only provides metadata information, it could not be incorporated on RegBR, once it needs the full text information of the normative acts. In addition, the output results do not help in the development of this work, since LEXML does not structure the results by types of normative acts.

By the time of the thesis release, the returned results using LEXML present all the legislation related to the searched term. Some metadata information can be used in the project, such as the publication date, when the law starts to take effect and where the law came from, but this information is not a priority. In that way, although the LEXML tool does not help with obtaining the text of the law itself, it may be a future possibility to obtain specific metadata information.

In this context, the presented study has implemented an ETL (extract, transform, load) pipeline to handle different data sources with the aid of scraping robots and Apache Airflow [83], which manages the data collection routines. The pipeline utilizes various tools, including Python programming language [84], the *Beautiful Soup* [85] and *Selenium* [86] libraries, and Apache Airflow, an open-source workflow management platform that schedules and executes tasks based on predefined dependencies and triggers. Figure 3.1 illustrates the high level abstraction structure developed to carry out the ETL process.

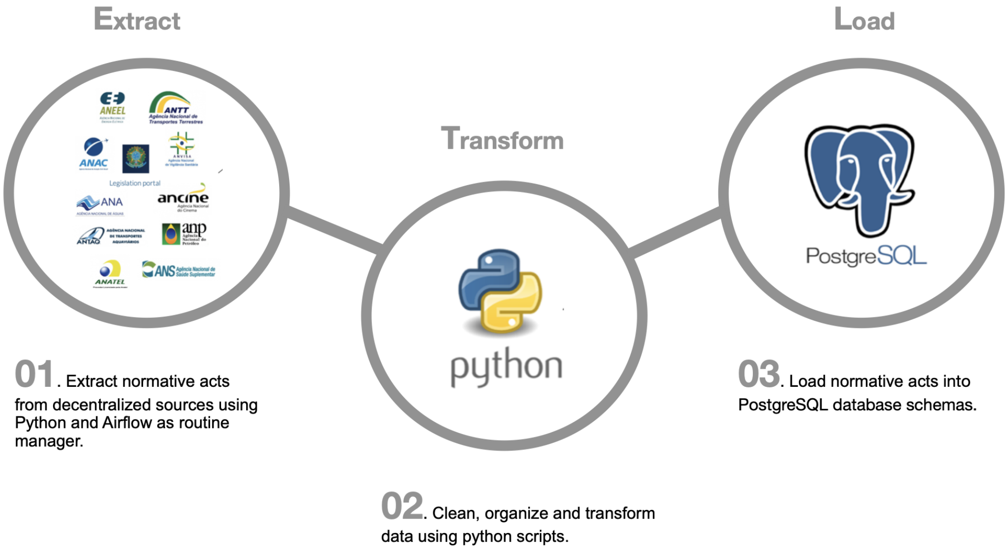

Extract
Transform
Load
ANEEL
ANTT
ANAC
Legislation portal
ANA
ancine
ANTAQ
anp
ANATEL
ANS
python
PostgreSQL
01. Extract normative acts from decentralized sources using Python and Airflow as routine manager.
02. Clean, organize and transform data using python scripts.
03. Load normative acts into PostgreSQL database schemas.

Figure 3.1: RegBR ETL organization.

Apache Airflow utilizes directed acyclic graphs (DAGs) to orchestrate workflow management. Tasks and dependencies are defined in Python, and Airflow automates the scheduling and execution process. DAGs can be triggered on a defined schedule, such as hourly or daily, or based on external events. In this study, Airflow managed the data extraction of 12 different sources, writing the extracted material into an intermediate raw data schema. Following the centralization of normative acts into a raw data schema, data cleaning and organization is performed through a series of data transformations. All the text preprocessing is done using the python package nltk. This process involves removing the stop-words, transforming each word in tokens, and stemming the text. The transformed data is then stored in a distinct PostgreSQL database schema that is specifically designed for readily usable data. This schema contains data that can be shared with other researchers and serves as the primary data source for subsequent analyses. Figure 3.2 illustrates the different Airflow DAGs created in this study.

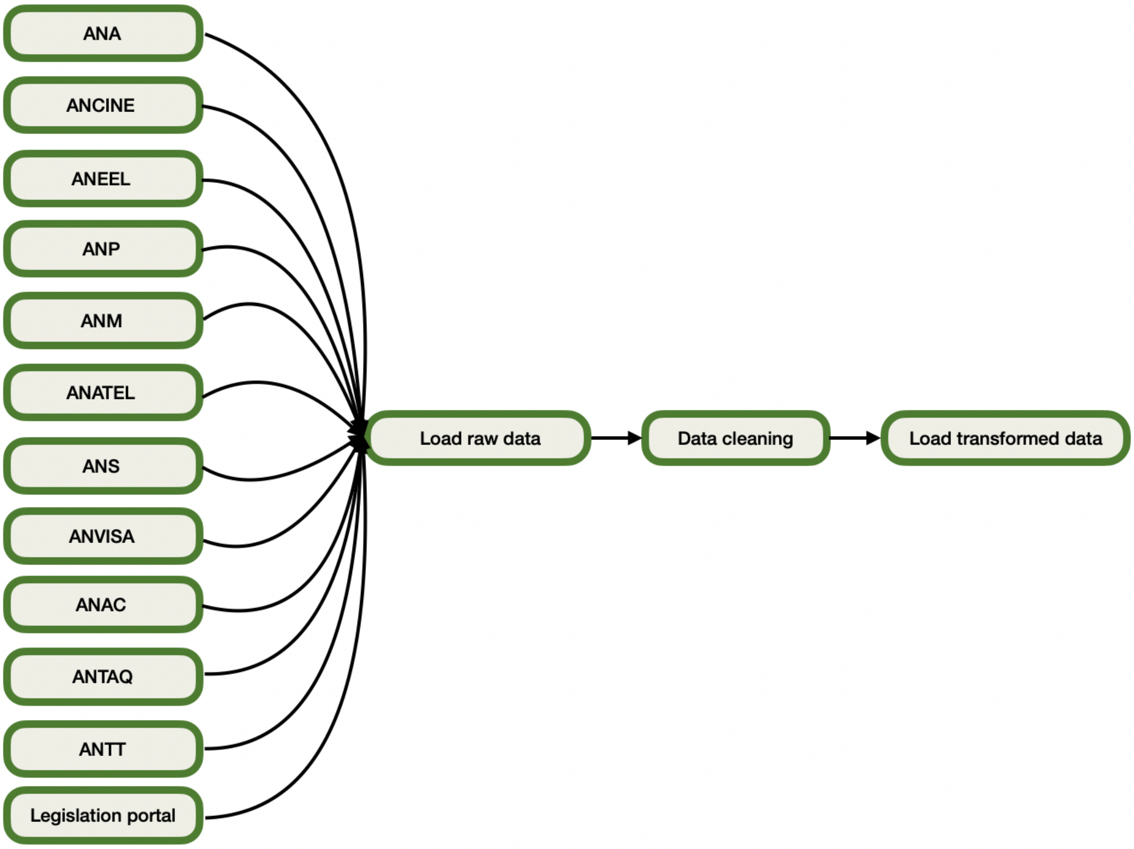


Figure 3.2: Airflow DAGs

As stated, the normative acts data used in this study was collected through web scraping techniques utilizing robots created with tools such as Selenium and Beautiful Soup, which were orchestrated by Apache Airflow. The data consisted of both text and metadata from normative acts, which were extracted from the websites of regulatory agencies and from the legislation portal. However, a significant challenge faced during the data collection phase was the necessity of developing 12 different robots, one for each regulatory agency and the legislation portal, due to the distinct website architectures of each site. The research team had to contact the regulatory agencies responsible for the websites to determine the relevant normative acts and available information. Furthermore, each agency provided varying levels of information, making it challenging to centralize the data into a unified database.

Another obstacle encountered during data collection was the existence of certain normative acts in image format, necessitating the use of OCR technologies to extract text from the images, increasing the complexity of the code.

Due to the amount of regulations that were being extracted from the websites of regulatory agencies and the legislation portal, some of these sources required the resolution of a captcha after certain interactions. This posed a significant challenge in the extraction of data. As a workaround, *time sleep* functions with varying intervals were implemented at different stages of the extraction process to mimic human interaction and thereby facilitate the successful extraction of data.

It is important to note that the utilization of web scraping techniques in the ETL routines carries a risk of code failure in the event of future modifications in the HTML of the web pages. Nonetheless, it was the sole feasible method for data extraction given the absence of an API or data centralization.

To enhance the ETL routines, an infrastructure architecture was established to ensure greater reliability and consistency. This architecture was implemented on the Amazon Web Services (AWS) platform. Two servers compose the main infrastructure: a server responsible for the application and scheduling of scripts (ETL) and a server storing database information. Figure 3.3 shows that the key infrastructure components are

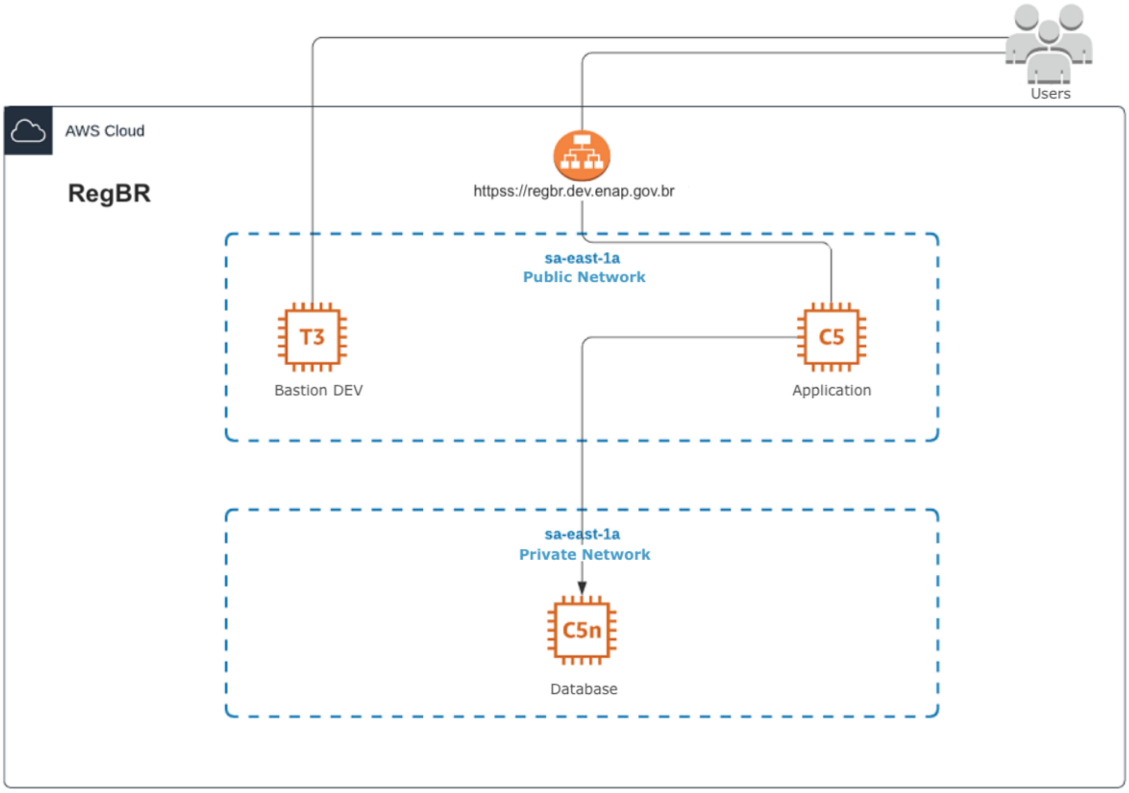


Figure 3.3: AWS infrastructure architecture.

- Bastion Server: Responsible for concentrating SSH access to Development environments. Amazon EC2 T3 Instances - T3.nano with 2Vcpu and 6GB memory.
- Application: Amazon EC2 C5 instances - c5.2clarge composed by 8 cpus, 16GB of RAM and 90 GB of disk (20 GB for the operating system and 70 GB for the application). Apache Airflow and Python 3 were installed along with the required dependencies.
- Database: Amazon EC2 C5 instances - c5.xlarge composed by 4 cpus, 8GB of RAM and 40GB of disk (20 GB for the operating system and 20 GB for Database). An instance of the PostgreSQL 12 database was installed.

The Direct access to the environment is allowed only through the Bastion Server.

3.2.2 Dataset annotation

An essential procedure for classifying normative acts into the different economic sectors that they regulate using text classification techniques, is the creation of a labeled dataset to be used in the models training phase.

The dataset annotation was carried out by a legal professional with expert domain knowledge on normative acts structure and content. Given the limited resources available for the text classification task, we have opted to use one person for the labeling task. This decision was based on several justifications, including cost-saving, time-saving, expertise and consistency, as using one person to perform the labeling we can ensure consistency in the labeling process, what can help to minimize errors and ensure high-quality data. Furthermore, different annotators may have better knowledge about different parts of the input space and therefore be inconsistently accurate across the task domain [87]. Additionally, the person performing the labeling has unique knowledge and insights that make them the best fit for the task. We believe that this approach will provide reliable results while making the most efficient use of our available resources.

The National Classification of Economic Activities (*Classificação Nacional de Atividades Econômicas*, CNAE) [88] is the official categorization, adopted by IBGE, to produce statistics by type of economic activity and by Public Administration, identifying economic activities in registrations of legal entities.

The CNAE is structured in twenty-one main categories, called sections, which in turn have four additional hierarchical levels: division, group, class, and subclass. The fifth level, designated subclass, is defined for use by the Public Administration. Table 3.1 shows an example of the CNAE structure for the ’Human health and social services’ section.

For research reasons, only the first hierarchy level is used in this work, resulting in 21 different classes of economy sectors. Moreover, for the purposes of this research, some classes were not relevant as they present low frequency and/or are very similar to other sectors. Thus, to simplify the classification process, we decided to perform the

Table 3.1: CNAE's Structure Example - Human health and social services

Hierarchy		
	Q	Human health and social services
Section:	86	Human health care activities
⌊ Division:	86.1	Hospital care activities
⌊ Group:	86.10-1	Hospital care activities
⌊ Class	86.10-	Hospital care activities, except emergency room and emergency care units
⌊ Subclass:	1/01	

following aggregations:

- Classes 7 (trade and repair of vehicles), 9 (accommodation and meals) and 12 (real estate activities) of the CNAE were merged in 7: Commerce, Accommodation and Food, and Real Estate Services.
- Classes 19 (Other service activities), 20 (Domestic services) and 21 (International organizations) were merged in 17: Other services.

In addition to these aggregations, we added an extra class (the 18th) to indicate legislation that does not deal with regulatory activities. In this study, we use class and economy sector interchangeably. Classes definition illustrated in Table 3.2, that presents a simpler view of the final 18 classes considered for classification. The labels in the annotation process were defined according to the main economic sector affected by each normative act analyzed.

The norm types defined as *Ordinances* and *Resolutions*, which correspond to about 60% of the total normative set, do not needed to be annotated and classified since the regulatory agency that regulates the normative act is known, as well as its economic area of activity. From the remaining normative acts which needed to be classified, 20% was annotated by the consultant to create the training and test dataset.

Table 3.2: Final classes

Class	Definition
1	Agriculture, livestock, and forest
2	Extractive industry
3	Transformation industry
4	Electricity and gas
5	Water, sewage, and waste
6	Construction
7	Commerce, accommodation & food and real estate services
8	Transportation, storage and mail
9	Information and communication
10	Financial, insurance and related services
11	Professional, scientific, and technical activities
12	Administrative activities and complementary services
13	Public administration, defense, and social security
14	Education
15	Human health and social service
16	Arts, culture, sports and recreation
17	Other services
18	Non-regulatory

3.2.3 Normative acts data structure

The RegBr v2.0, launched in April 2022, corpus comprises about 52,000 normative acts of the Brazilian federal legislation written in Portuguese, divided in ten types, since 1891. Table 3.3 presents the studied types of normative acts in Portuguese and in English.

Table 3.3: Normative acts information

Type	Normative act in Portuguese	Normative act in English
1	*Emenda Constitucional*	Constitutional Amendment
2	*Lei Ordinária*	Laws
3	*Decreto-lei*	Decree Law
4	*Medidas provisória*	Provisional measure
5	*Lei complementar*	Supplementary Law
6	*Decreto*	Decree
7	*Resolução*	Resolution
8	*Portaria*	Ordinance
9	Instrução Normativa	Normative Instruction
10	Súmula	Precedent

The first six types come from the Brazilian Legislation Portal [89] in HTML format, while the remaining four normative act types come from decentralized sources of the electronic portals of the 11 Brazilian regulatory agencies in various formats, including PDF, HTML, and images. It is important to note that the median length of a normative act is about 469 tokens, which is significantly longer than the typical customer review or news article commonly found in datasets for benchmarking machine learning models on text classification.

The dataset contains about 8 thousand labeled acts divided into 18 classes. It includes all legislation available on the internet from the cited sources containing normative acts since the end of the 19th century. For information, we divided labeled data into 75% training and 25% testing.

3.2.4 Main approaches

To construct a more resilient benchmark for this specific text classification scenario, various approaches were implemented. The ensuing subsections delineate the primary characteristics employed to formulate these models.

Statistical models

In what we define as statistical models, the documents words were transformed in features using TF-IDF, while the classifiers were implemented using Scikit-learn [90]. Some of the most used classifiers in machine learning applications such as the Logistic Regression, SVM with linear kernel and Gradient Boosting Classifier [91] were applied. Additionally, we also evaluated the Ridge classifier, which essentially treats the classification problem as a regression problem by predicting continuous target values and then assigning labels based on those values.

Considering the potential of fully connected neural networks, a model utilizing a fully connected (FC) neural network using the TF-IDF features was also incorporated.

Furthermore, an SVM classifier with features extracted through Latent Semantic Analysis (LSA) [92] was included as it is considered a classical approach in legal text classification, according to [12]. Since the dataset was imbalanced, i.e., the different classes are not approximately equally represented, the Synthetic Minority Over-sampling Technique (SMOTE) [93] was employed to enhance performance by generating samples from minority classes. Hyperparameters for each classifier were determined using grid search, which was also employed to identify the optimal parameters for the TF-IDF vectorizer, including the maximum number of features, cut-off frequencies, and n-gram range.

It is also important to mention that an alternative vectorizer, Count-Vectorizer, was examined but consistently underperformed in comparison to TF-IDF.

Word embedding models

In addition to statistical models, this study also evaluated word embeddings approaches. Statistical models rely on sparse vectors, which are high-dimensional and have many zero values, making them computationally inefficient. Word embeddings, on the other hand, produce dense vectors that encode the semantic meaning of words in a more compact and meaningful manner, making them easier and faster to process.

Also, statistical models typically treat each word as independent of all other words in the sentence or document, whereas word embeddings capture the contextual relationships between words, which is crucial for natural language understanding. By embedding words in a higher-dimensional space, word embeddings allow for the recognition of word similarity and relatedness based on their semantic meanings, whereas statistical models are often limited to word frequency counts and co-occurrence patterns.

In that sense, word vectors pre-trained on large corpora have been shown to capture syntactic and semantic word properties. This capability was leveraged by using *word2vec* [25] and *GloVe* [65], both with 300 dimensions, pre-trained on a Portuguese corpus [94].

These embeddings were used by two different neural network classifiers; a Convolutional Neural Network (CNN) [95] and a Long Short Term Memory (LSTM) network [96]. These architecture were chosen as they are often used in NLP applications [97] and they are also some of the building blocks of more complex deep learning models.

Transfer learning models

Recently, impressive results were achieved by language models that were pre-trained on large unlabeled corpora and then fine-tuned for specific tasks. Generally, transfer learning models are an asset when dealing with smaller datasets, since its models are pre-trained on large datasets and can be fine-tuned on smaller, task-specific datasets, resulting in better performance on tasks with limited data availability, such as in the legal domain.

Also, transfer learning models have the ability to capture context. Models such as BERT and GPT are trained on large amounts of text data using self-supervised learning techniques that allow them to capture the contextual relationships between words and sentences. This enables them to perform well on tasks that require understanding the meaning of text in context, different from using word embedding who is limited to capturing the semantic relationships between words and may not be as effective at capturing the nuances of language.

Therefore, two variants of BERT [26] known as $BERT_{base}$ (12-layers; 110M parameters) and $BERT_{large}$ (24-layers; 340M parameters) were evaluated in this study.

However, one of the limitations of BERT is the self-attention transformer architecture [98] which only accepts up to 512 tokens. Since some legal texts are longer than this, a ULM-FiT model, which accepts longer inputs due to its stacked-LSTM architecture [99], was also employed.

3.3 Results and discussion

Over the last few decades, especially with recent breakthroughs in NLP and text mining, text classification applications have been widely studied and implemented [100]. One area that has grown and gained relevance in recent years is related to the classification of legal texts [12], which usually have complex and technical language, imposing manual classification tasks for a select group of jurists with specific domain knowledge. Still, legal texts are composed of large amounts of words and content, making it infeasible to perform the classification task manually and efficiently.

With regards to the structure of a text classification problem, it can be broken down into the following four phases: Feature extraction, dimensionality reduction, approach and model selection, and evaluation [100].

- Feature Extraction: Since texts are unstructured data, they must be cleaned and converted into structured feature space before applying classification algorithms. Feature extraction also reduces the number of features in a dataset by creating new features from the existing ones. Standard techniques of feature extraction include One Hot Encoding [24], Term Frequency-Inverse Document Frequency (TF-IDF) [63, 101], and text embedding techniques such as Word2Vec [25].

- Dimensionality reduction: Optional phase defined as the data transformation from a high-dimensional space into a low-dimensional space. It is used in order to reduce the time and space complexity in some applications, especially when using data sets containing many unique words.

- Approach and model selection: The most critical step in the classification pipeline. At this point, it is important to have a solid understanding of the main approaches and models used in the literature for the application of interest.

- Evaluation methods: Used to assess the model's performance, it is the final part of the classification pipeline. There are several available metrics to evaluate models, such as accuracy, F-measure, recall, precision, and ROC curve. The most popular metric in the context of imbalanced text classification is F-measure since accuracy can undervalue how well classifiers are doing on minority classes. In contrast, F-measure balances precision and recall of classifiers on each class [102]. For these reasons, RegBR uses F1 score and accuracy as its main evaluation metrics.

The results of the text classification benchmark performed in this study are presented in two parts: first using the entire set of normative acts starting with the first normative act available on the digital platforms used as the source, in 1891, and soon after, using only normative acts from 1964 onward in order to delimit a clear and more linguistically homogeneous temporal scope, since the vocabulary used at the beginning of 20th century is considerably different from the current one when we refer to legal texts.

Another reason for picking 1964 is that it represents a milestone in Brazilian political and economic history as a military dictatorship was established in 1964, a period from which many normative acts remain valid until now, even after the promulgation of the 1988 Constitution. This year was also chosen among several tested years for presenting the best trade-off between including the most normative acts in the training phase and also having a consistent linguistic style between the text analyzed.

The models and hyper-parameters used in this legal text classification benchmark study are presented in Table 3.4.

It is important to mention why such different models were chosen to compose the benchmark. Firstly, these models are widely used and well-established in machine learning applications, including text classification. They have demonstrated good performance across a range of tasks and datasets, making them suitable baseline models for comparison. Secondly, each of these models has a different underlying assumption and structure. The diversity in model assumptions and structures provides a more comprehensive evaluation of the text classification problem. Thirdly, these models are relatively easy to implement, and their results are interpretable, making them suitable for benchmarking purposes. Popular libraries such as Scikit-learn provide built-in function

Table 3.4: Hyperparameters used to generate the results

Models	Hyperparameters
Tf-idf vectorizer	Maximum number of features = 10,000 / n-gram range = (1,1) / no constraints in the frequency of words
Logistic Regression (LR)	Regularization coefficient = 2
Ridge Classifier (RC)	Regularization coefficient = 1
Oversampling	Over-sampling using SMOTE and cleaning using Tomek links with default parameters
SVM	Linear Kernel, Regularization coefficient = 0.5
XGBoost	XGBoost with 100 estimators
SVM + LSA (500)	LSA with 500 topics / SVM with linear kernel with regularization coefficient = 1
Word2vec	300 dimensions
GloVe	300 dimensions
LSTM	Embedding layer followed by a bi-directional with 64 hidden units followed by a fully connected layer with 256 units with dropout = 0.1 and a final layer with 64 hidden units.
CNN	Embedding layer followed by 4 convolutional layers with 36 filters of varying kernel size (1,2,3,5) followed by a fully connected layer with 144 hidden units with dropout = 0.1
BERT	BERT pre-trained (12 layers with 110 million parameters for the base model, and 24 layers with 335 million parameters for the large model) using Portuguese corpus [103], followed by a final fully connected layer with dropout = 0.05
ULM-FiT	One AWD-LSTM layer [104] followed by 4 QRNN layers [105] with dropout and a final fully connected layer
NN + TF-IDF	Fully connected network with 25 hidden units in the first layer followed by batch normalization, dropout = 0.25 and a final layer with 18 units

for these models, making their implementation straightforward. Lastly, these models offer various hyperparameters that can be tuned to optimize their performance on the specific text classification task. This adaptability also makes them suitable for benchmarking. In that regard, tables 3.5 and 3.6 evaluate these models using all data, and data after 1964, respectively.

Across the models implemented, word embedding models consistently underperformed the statistical and transfer learning models on accuracy. Regarding the F1-score, word embedding models performed worse, on average, than the statistical models. This result can be observed because the pre-trained word embedding models used in the study were trained on a large generic body of Brazilian Portuguese, and European Portuguese, of different sources and genres, from USP Word Embeddings Repository [94], and not in a specific and targeted set of legal and bureaucratic texts of the federal government.

Analyzing transfer learning approaches, BERT models performed slightly worse than the best statistical models, and MultiFit had the worst performance overall, in both

Table 3.5: Classification Results with all data

Models	**Accuracy**	**Average F1-score**
Logistic Regression (LR)	62.64 ± 1.03%	0.571 ± 0.093
Ridge Classifier (RC)	63.77 ± 0.94%	0.59 ± 0.006
LR$_{SMOTE}$	63.57 ± 0.83%	0.597 ± 0.009
SVM	63.96 ± 1.19%	0.592 ± 0.013
XGBoost	60.94 ± 0.79%	0.553 ± 0.013
Ensemble$_{RC,LR_{SMOTE}}$	63.59 ± 0.82%	**0.598** ± **0.09**
Ensemble$_{RC,SVM}$	**64.06** ± **1.18%**	0.592 ± 0.10
SVM$_{LSA}$	59.50 ± 4.7%	0.538 ± 0.045
LSTM$_{word2vec}$	57.08 ± 1.02%	0.5043 ± 0.03
CNN$_{word2vec}$	59.99 ± 0.47%	0.541 ± 0.072
LSTM$_{GloVe}$	57.48 ± 0.38%	0.5151 ± 0.088
CNN$_{GloVe}$	59.48 ± 0.52%	0.543 ± 0.064
BERT$_{base}$	61.84 ± 0.85 %	0.551 ± 0.024
BERT$_{large}$	48.70 ± 1.19 %	0.382 ± 0.067
ULM-FiT	55.29 ± 1.03%	0.526 ± 0.055
FC Neural Network$_{TF-IDF}$	58.58 ± 0.9%	0.541 ± 0.011

Table 3.6: Classification Results with data post-1964

Models	**Accuracy**	**Average F1-score**
Logistic Regression (LR)	65.93 ± 1.25%	0.575 ± 0.015
Ridge Classifier (RC)	67.97 ± 0.95%	0.612 ± 0.015
LR$_{SMOTE}$	66.15 ± 0.011%	0.609 ± 0.013
SVM	67.72 ± 1.31%	**0.619** ± **0.013**
XGBoost	63.91 ± 1.11%	0.568 ± 0.015
Ensemble$_{RC,LR_{SMOTE}}$	64.96 ± 1.83%	0.591 ± 0.022
Ensemble$_{RC,SVM}$	**67.97** ± **1.09%**	0.616 ± 0.015
SVM$_{LSA}$	61.51 ± 3.94%	0.531 ± 0.031
LSTM$_{word2vec}$	58.37 ± 1.02%	0.521 ± 0.012
CNN$_{word2vec}$	61.66 ± 0.92%	0.565 ± 0.079
LSTM$_{GloVe}$	59.78 ± 1.36%	0.533 ± 0.014
CNN$_{GloVe}$	61.21 ± 1.57%	0.565 ± 0.016
BERT$_{base}$	62.21 ± 0.94 %	0.514 ± 0.061
BERT$_{large}$	52.72 ± 0.89 %	0.428 ± 0.056
ULM-FiT	58.14 ± 0.92%	0.538 ± 0.033
FC Neural Network$_{TF-IDF}$	63.66 ± 0.94%	0.569 ± 0.020

accuracy and F1-score metrics. The causes of inferior performance with respect to statistical models could be attributed to the limited domain-specific knowledge and the small number of pre-trained language models available in Portuguese. In this work, we used neuralmind [103] BERT language model and FastAI ULMFit based language model, both trained in a wide range of texts but not necessarily linked to legal texts.

Transfer learning models are trained on large, general-purpose datasets, and this specific used pre trained model may not have sufficient domain-specific knowledge to handle the nuances and complexities of the legal domain. Legal language often contains jargon, technical terms, and complex sentence structures that may be unfamiliar to the pre-trained model, resulting in lower performance.

Although considering only post 1964 material to train and test the model makes us lose about 40% of the normative acts, an improvement in the performance metrics is noticeable, with both accuracy and F1 score metrics improving. This improvement may be partly explained by the fact that the language observed in the texts is more homogeneous.

Additional analyzes were performed to handle the fact that the dataset was unbalanced. Two different methodologies were used in order to obtain similar numbers of examples for the different classes. First, an undersampling method was applied by randomly eliminating examples from the most numerous classes. The performance obtained was not satisfactory, possibly due to the decrease in the training dataset. Then, the dataset was balanced using SMOTE [93], which oversampled the examples of the least represented classes. In this case, the results were slightly worse than the classification using the unbalanced data, indicating limitations of the technique for the problem at hand. There are several reasons why SMOTE can have not so good results in unbalanced datasets, including overfitting since SMOTE can generates synthetic data points that are too similar to the original minority class resulting in lack of diversity and reduced generalization performance on new data and also including the generation of noisy data points that are not representative of the true distribution of the minority class.

Overall, while transfer learning models have shown superior performance on many natural language processing tasks, including text classification, their performance may be limited in the legal domain due to the unique characteristics of legal language and the limited availability of labeled data. As noted by [106], BERT models may not be well-suited for domain-specific legal text tasks. While this does not preclude their use, it does increase the complexity involved.

As stated, there are several potential explanations for BERT's inability to outperform models such as SVM and Ridge classifier in our text classification task. An additional possible reason is BERT's possible difficulty in learning accurate representations for underrepresented classes in datasets with significant class imbalance, like our dataset. Also, although BERT has demonstrated remarkable effectiveness in general language representation tasks, it is trained on unlabeled, generic sources like Wikipedia and open-source articles, resulting in a lack of domain-specific knowledge, as the dataset used in this study.

Moreover, a limitation of the BERT model is its inability to handle texts exceeding 512 tokens, which, in our case, constitutes more than half of the corpus. This constraint may contribute to the comparatively weaker performance observed relative to other models. Alternative architectures within the BERT family, which can accommodate longer texts, may potentially yield improved performance. Such investigations could be pursued in future research.

In cases like the one in this study, a carefully designed statistical model with appropriate feature engineering and data preprocessing may outperform transfer learning models. The ensemble model of Ridge classifier and SVM emerged as the best performing approaches on both accuracy and macro averaged F1 scores. It is also important to notice that the statistical models are significantly faster for training and testing when compared to the implementations using deep neural networks.

The results obtained with the Ensemble$_{RC,SVM}$, with around 68% of accuracy and 62% of F1-score, for a larger number of classes, is an exciting result and is similar to what was shown in other legal text classification benchmarks in other languages [12].

Due to its good performance and low computational cost, we employ the Ensemble$_{RC,SVM}$ as the model used to classify regulations into different economic sectors. Moreover, this model also allows us to verify the most relevant terms in classifying each sector, which improves the model’s transparency and interpretability.

4 Regulatory metrics conception

This section introduces metrics that allow regulators and policymakers to better identify and prioritize regulations that may need reforms. In this sense, this work provides a variety of quantitative data and indicators, including

- Regulatory stock analysis over time, making the federal regulatory flow transparent
- Restrictiveness metric, indicating the regulatory restriction counts;
- Industry citation relevance metric, that calculates the frequency of industry-relevant terms in the context of federal regulations among the economic sectors considered;
- Popularity metric, indicating how popular a law is for the general population and for the Federal government;
- Linguistic complexity metric, measuring the linguistic complexity of a normative act.

All the metrics and indicators presented in this section are calculated based on legislative documents obtained via the automated ETL routines presented in section *3.2.1 Implementation of ETL routines* and classified into sectors of economy using the best performing model presented in section *3.3 Results and discussion*. The entire ETL system and text classification task is updated every 3 months, being the study results always current and available to the public and decision makers to use.

4.1 Regulatory stock analysis over time

The concept of *Regulatory stock Management* is not new in the Brazilian government context. Nonetheless, it gained more relevance since the publication of Decree 10.139, of 28 November 2019, which imposes the review and consolidation of all normative acts with a hierarchy lower than the decree by the end of 2021. In addition, as determined by the decree, each federal administration body and entity must divide all its normative acts by thematic relevance and review them by steps [107].

To adapt to this new context, some regulatory agencies are establishing working groups for quantitative mapping of regulatory stock. In this context, RegBR brings a general analysis of the Brazilian regulatory stock filtered by sector of the economy or by regulatory agency, to assist regulatory authorities in managing the country's regulatory stock and better adequate the regulatory process to international quality parameters.

Figure 4.1 illustrates the quantitative analysis of the normative acts *Resolutions*, *Ordinances*, *Normative instructions* and *'Precedents'* by agency, while Figure 4.2 presents the data filtered by type of normative actor. It is interesting to note that the creation of Brazilian regulatory agencies took place around the 2000s, with the creation of the first regulations on the same period. The graphs below can be found at https://infogov.enap.gov.br/regbr/fluxo-regulatorio.

The information presented in the aforementioned figures can be filtered by economic sector and normative act situation and is available for public consultation, increasing government transparency and allowing easy access to information for citizens and policy makers interested in monitoring their work metrics.

4.2 Regulation restrictiveness metric

Recently, regulatory reforms have gained increasing attention in the political and economic context [108] and, consequently, researchers have been trying to introduce

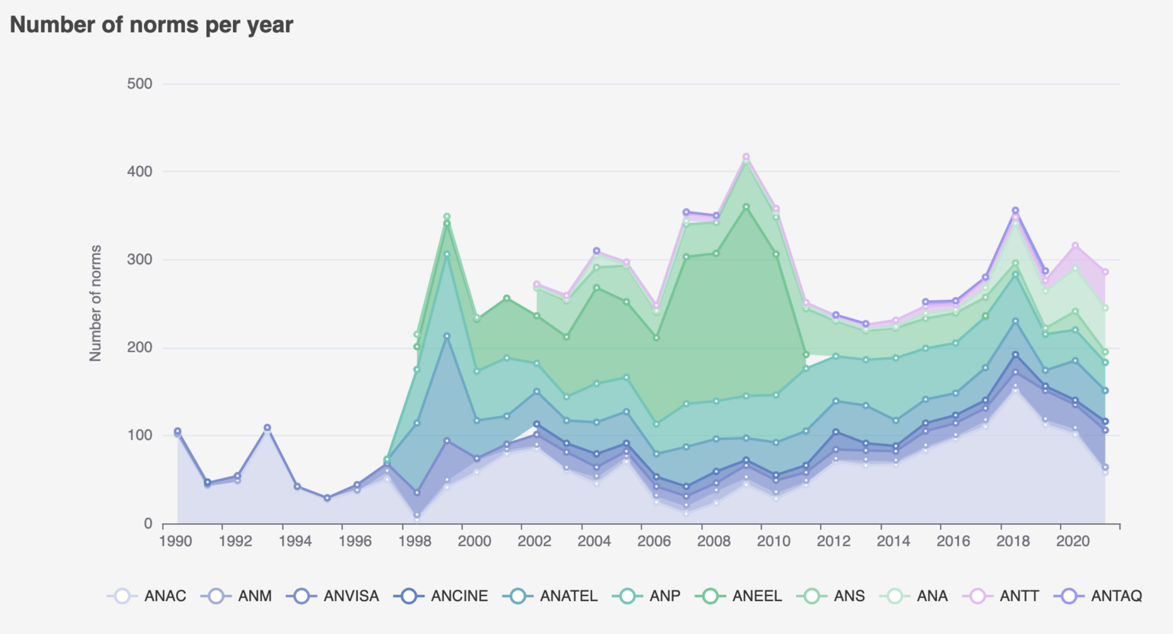


Figure 4.1: Normative acts of regulatory agencies over time.

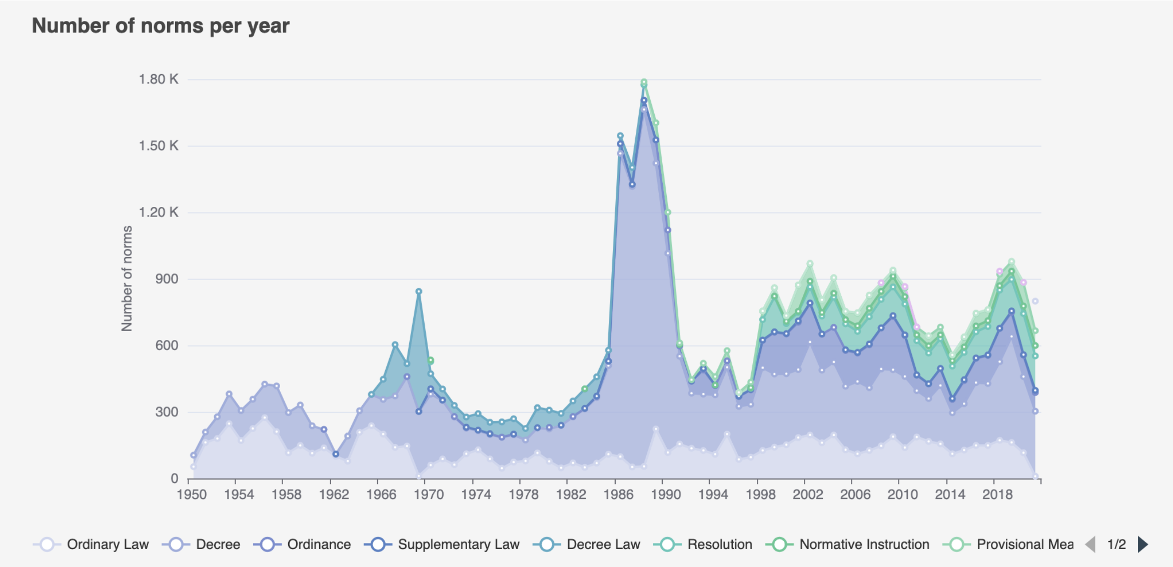


Figure 4.2: Normative acts of different normative types over time.

simple and direct forms to quantify regulations and perform ex-post evaluation [109].

One of the first methods used for this purpose was based on counting the law codes pages [18]. This method, due to its simplicity, does not always correctly represent the complexity or the importance of the laws since long texts are not necessarily stricter than short and concise texts. In addition, the fact that some texts use more tables, graphs, diagrams, and annexes, which disproportionately increases the number of pages, and these changes in formatting styles over time can negatively influence the use of this metric as the main method for quantifying regulatory power of a law.

In order to overcome the problem of different text formatting, [110] proposed using file size data for quantification purposes, however, the presence of large graphics and tables can still bias this measure.

As a methodology capable of overcoming these problems, RegData US [18] proposed the use of word count to quantify the restrictiveness of a piece of regulation.

Regulatory restrictions are defined as words and phrases in a regulatory text context that indicate specific obligations or prohibitions [111]. As normative texts are intended to restrict or expand legal scopes, these texts often use certain verbs and adjectives such as ‘shall’ and ‘must’. The restriction metric is then measured by the total number of occurrences of restrictive words in a set of laws within the body of the normative act.

In the last few years, in the Brazilian context, we have seen a greater number of enacted normative acts. For this reason, RegBR proposes a slight modification in the original metric of law restrictiveness of RegData. In addition to counting the restrictive words in a set of normative acts, this number is divided by the number of normative acts in the set, by economic sector in each year, thus obtaining a metric for the average number of restrictive words by normative acts over the years for each studied economic sector. In this way, we can understand whether the average number of restrictive words by normative acts has increased, decreased or remained constant over time.

Thus, the regulation restrictiveness metric for RegData BR is defined as

$$\text{restrictiveness}(\text{year},\ \text{economic sector}) = \left(\frac{\sum \text{restrictive word count}}{\sum \text{law}}\right), \tag{4.1}$$

where the word counts are defined by a list[1] of restrictive words in Portuguese that intend to restrict or expand legal scopes. This list was proposed by the ENAP’s researchers and validated by law professionals with vast experience in the legislative field.

[1] In portuguese, the restrictive words are: *vetado, vedado, defeso, proibido, negado, determina, obriga, ordena, impõe, limita, delimita, demarca, restringe, confina, reduz, define, deve, deverá, precisa* and *necessita.*

After generating the time-series with the average number of restrictive words per law and per year for each economic sector, a stationarity statistical test is performed on the restrictiveness metric over time to assess the existence of trends in the data [112, 113]. This would indicate possible increases or decreases in the average strictness of the laws over time. For this purpose, the Augmented Dickey Fuller (ADF) test [114] was applied as well as the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [115] together [116]. In general, if the results of both tests suggest that the series is stationary, we can consider its stationarity with high confidence [116]. In simple words, we may infer whether the mean function of the series is constant or not. Table 4.1 [117] shows the possible results with p-value of 0.05 for each test based on its null hypothesis.

Table 4.1: Results interpretations.

	ADF	KPSS
p <0.05	Stationary	Non-stationary
p >0.05	Failed to reject the stationary hypothesis	Failed to reject the non-stationary hypothesis

From the results of Table 4.1, Table 4.2 [117] summarizes the possible interpretations when the two tests are combined. The series is called trend stationary when it becomes stationary after removing the trend. Similarly, difference stationary implies that the series requires differencing the series to make it stationary.

Table 4.2: Outcomes interpretations of combined ADF and KPSS tests.

		KPSS	
		p >0.05	p <0.05
ADF	p <0.05	Stationary	Inconclusive
	p >0.05	Inconclusive	Non-stationary

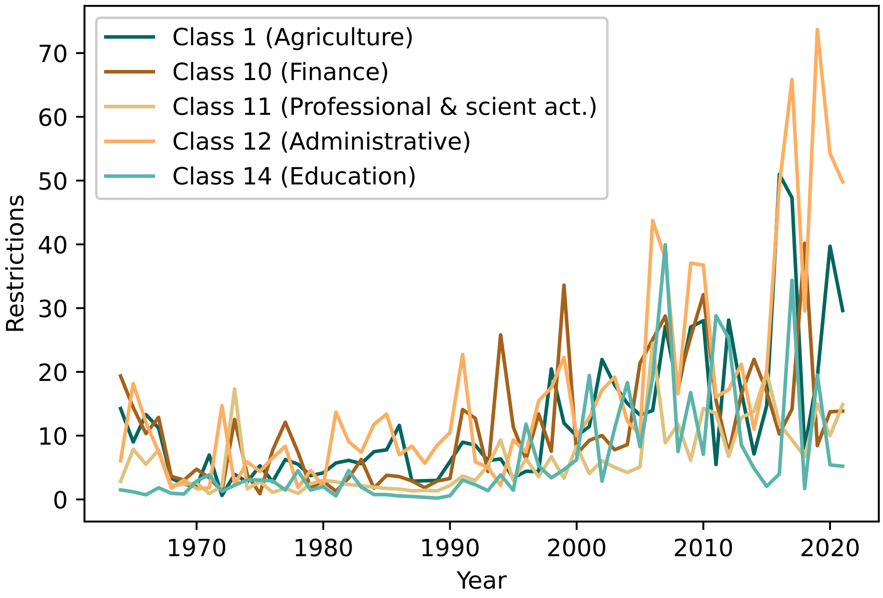
The experiment was conducted first generating the time series for the metric of word restrictiveness per law by year and by economic sector. Next, we applied the ADF and KPSS tests to each time series corresponding to each economic sector. If the ADF test failed to reject the stationary hypothesis ($p > 0.05$) and the result for the KPSS test was non-stationary (p < 0.05), we considered the time series non stationary, as indicated in Table 4.2. Finally, Table 4.3 presents the results of the stationary tests to all classes.

Table 4.3: Classification Results with all data

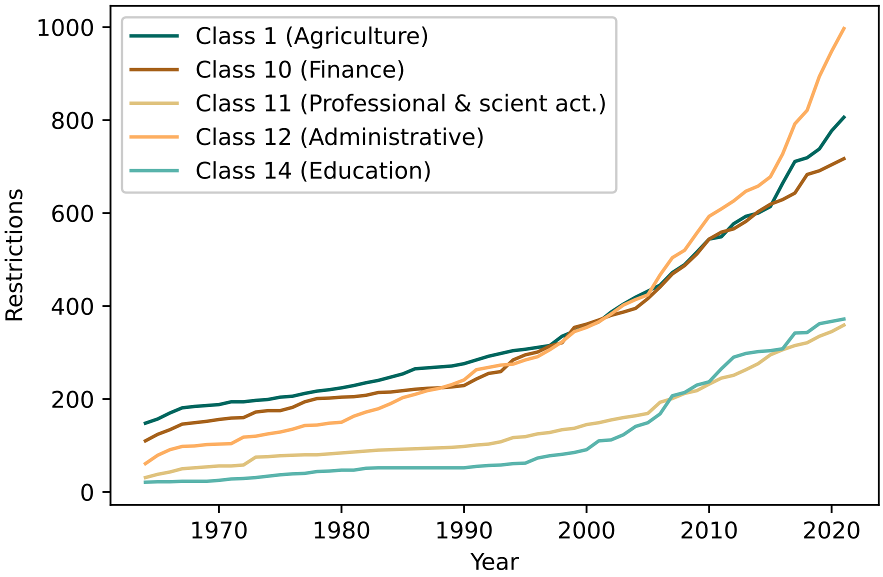
Classes	ADF	KPSS	Result
Class 1	0.993189	0.010000	Non-stationary
Class 2	0.063195	0.100000	Non-stationary
Class 3	0.039528	0.011266	Inconclusive
Class 4	0.243227	0.044970	Non-stationary
Class 5	0.567624	0.022370	Non-stationary
Class 6	0.000037	0.072917	Stationary
Class 7	0.025908	0.059229	Stationary
Class 8	0.344502	0.024106	Non-stationary
Class 9	0.101205	0.010000	Non-stationary
Class 10	0.816323	0.010000	Non-stationary
Class 11	0.692446	0.01000	Non-stationary
Class 12	0.999052	0.010000	Non-stationary
Class 13	0.025713	0.010000	Inconclusive
Class 14	0.978855	0.010000	Non-stationary
Class 15	0.382618	0.010000	Non-stationary
Class 16	0.316495	0.010000	Non-stationary
Class 17	0.004189	0.100000	Stationary
Class 18	0.260208	0.012934	Non-stationary

For these non-stationary series, a polynomial fitting was performed to model and remove the trend afterwards by subtracting the values of the fitting curve from the original time series. Finally, we applied again both tests to each de-trended time series. All de-trended time series tested positive to stationarity, confirming the initial time series status of non-stationarity with a trend curve.

As a result of the experiment, 12 out of 17 economic sectors tested to be non-stationary, presenting a positive trend of increasing the number of restrictive words over the years: Agriculture (1), Extractive Industry (2), Electricity and gas (4), Water and sewage (5), Transportation (8), Information and communication (9), Finance (10), Professional Scientific activities (11), Administrative activities (12), Education (14), Human health (15) and Arts, culture, sports, and recreation (16). Figure 4.3 shows the restrictions word count per laws over time, since 1964, for five of the economic sectors discussed in a cumulatively and not cumulatively way.



(a) Brazilian restrictive word count per law, 1964 - 2021.



(b) Brazilian restrictive word count, cumulatively, 1964 - 2021.

Figure 4.3: Brazilian restrictive word count

4.3 Industry citation relevance

The industry citation relevance metric measures the influence of the CNAE's economic sectors, also called industries, based on their citation frequency in the general corpus of normative acts. If words directly related to a particular economic sector are used frequently throughout the entire corpus, that sector is understood to show more relevance than an economic sector that is not frequently cited, which may indicate which sectors have been prioritized in the context of regulatory legislation.

To calculate the industry citation relevance metric by year, for each industry we sum the total occurrences of the specific strings terms that represent the industry in that year and divide by the total number of words in the normative acts corpus from that specific year as

$$\text{relevance}(\text{year}, \text{industry}) = \frac{\sum \text{industry specific strings on corpus}}{\sum \text{corpus words}}. \tag{4.2}$$

Then, a normalization is applied to the metric values by dividing them by the maximum value in that year to obtain a range between 0 and 1, where 1 represents the most relevant sector in that year.

The industry specific strings were derived from the most relevant words[2] for the Ensemble$_{RC,SVM}$ text classification model, the top performing method. The top 10 words that represent each industry are unique, i.e., they do not affect the metric for other industries because there is no overlap between different sectors.

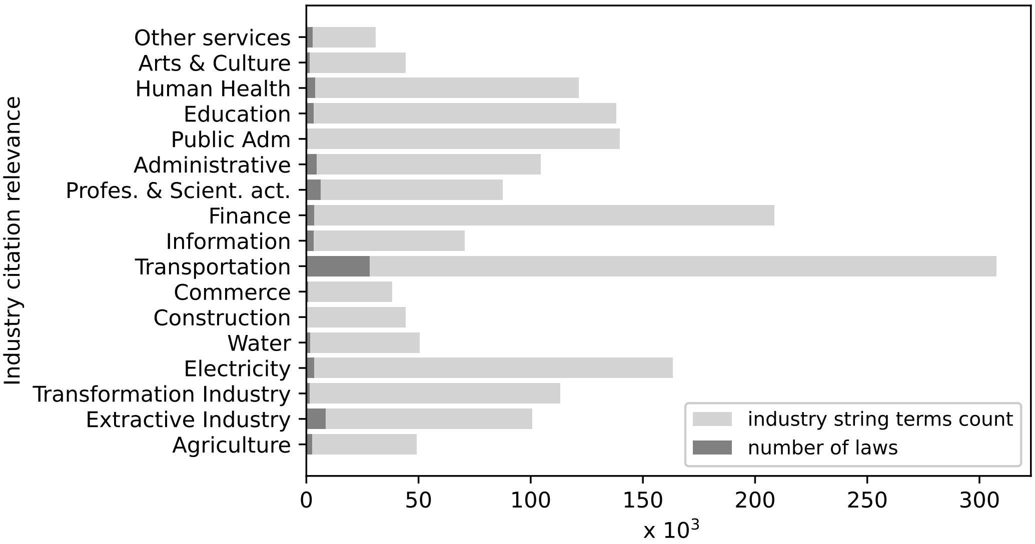


Figure 4.4: industry citation relevance string terms count.

[2] Formally words whose corresponding weights had most significant values.

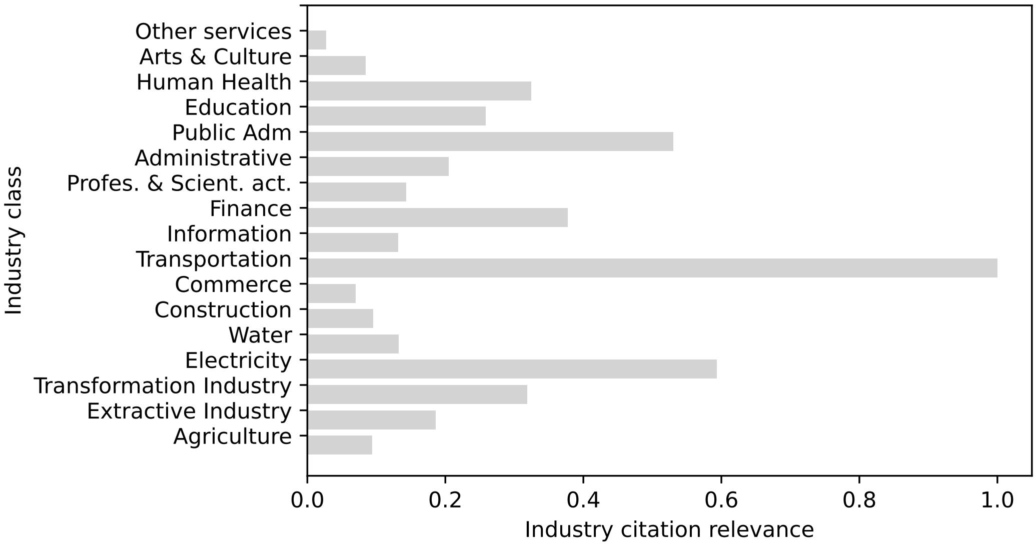


Figure 4.5: Industry citation relevance metric for the year of 2020.

Next, Figure 4.5 displays the relevance metric of each CNAE industry. The bars represent the number of occurrences of the industry-specific search strings divided by the corpus's word count for 2020. In that year, the most relevant industries in the context of regulatory acts are 'Transport' (code 08) and 'Electricity' (code 04) and, curiously, the 'Health' sector was only in 5th place, even with the incidence of the COVID-19 pandemic during the year concerned.

In addition, we can check the relevance metric values for each industry over the years. Figure 4.6 shows the relevance metric for the economic sectors since 1964.

The industry citation relevance presented in Figure 4.6 correlates with some historical events described as follows. For instance, it is interesting to observe that between 2001 and 2004, due to the frequent energy blackouts and need for energy rationing, a regulatory reform was initiated in the Brazilian electrical sector [118], including the creation of the National Electric Energy Agency (ANEEL), which led to the increase in the relevance of this sector in the federal normative context in the aforementioned period, as shown in Figure 4.6 on class 4.

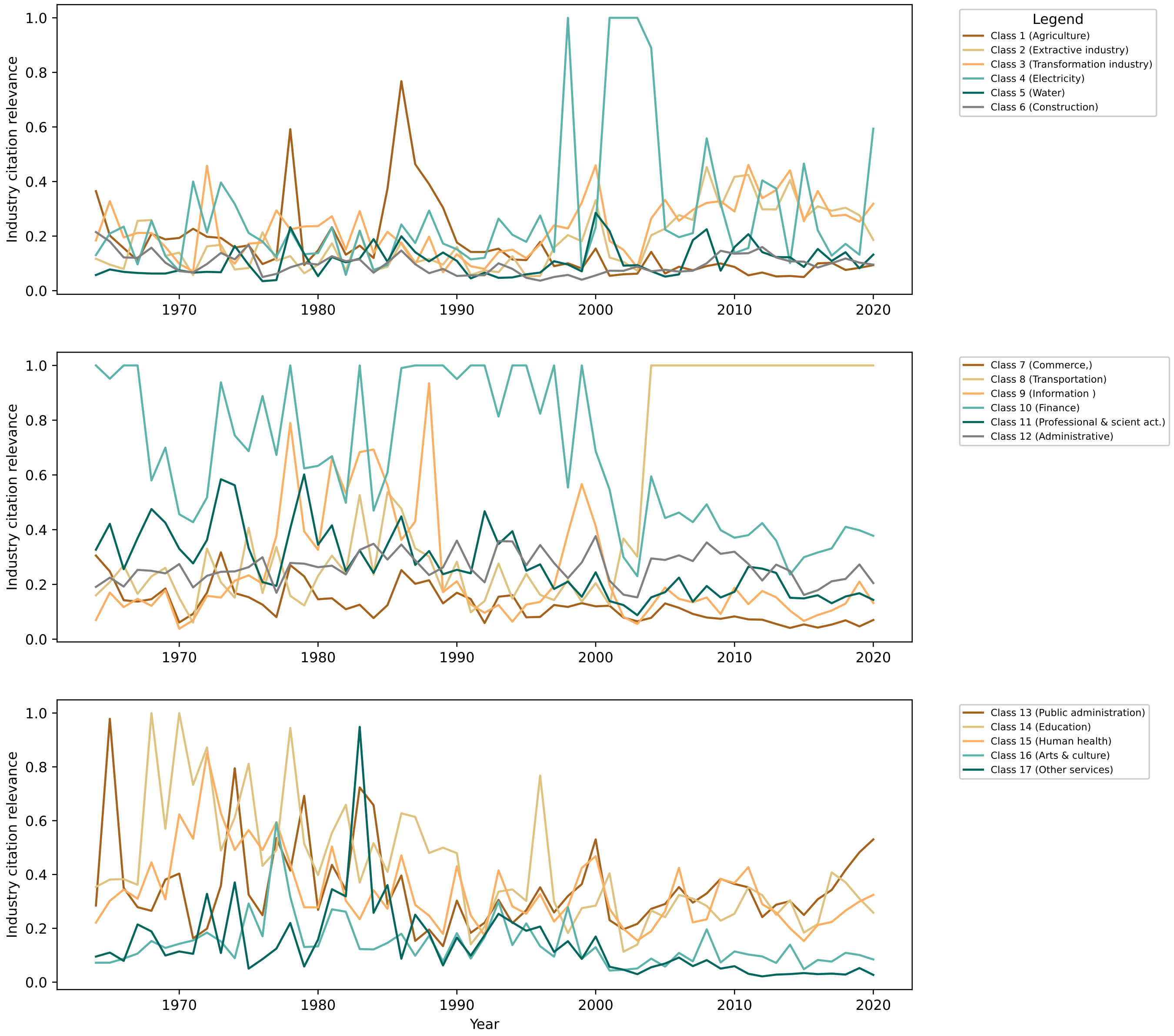


Figure 4.6: Industry citation relevance frequencies.

On a similar note, the increase in transport industry citation relevance since 2000 was due to the creation of the National Land Transport Agency (ANTT) and the National Waterway Transport Agency (ANTAQ) both in 2001, and the creation of the National Civil Aviation Agency (ANAC) in 2005, strongly regulating the transport sector in Brazil. Figure 4.6 illustrates this behavior.

In the opposite direction, in the 80s and 90s, the Brazilian economy was marked by crises and hyperinflation. In 1986, the then-president José Sarney launched the

Cruzado Plan, the country's largest economic stabilization plan at that time. Several economic measures such as currency changes and freezing of wages, prices, and exchange rates were taken during the same year. With the return of hyper-inflation months later, several other plans were implemented until economic stabilization in the late 1990s as a result of the Real Plan [119], the thirteenth economic plan for stabilizing the Brazilian economy since the early 1980s, implemented by the Itamar Franco administration in 1994. Figure 4.6 shows that during these two decades, the finance sector was very relevant in the federal normative context, decreasing its relative relevance from the 2000s.

This type of analysis can help citizens to assess government's priorities, increasing transparency. For example, after a pandemic, one would expect a higher of industry citation relevance on the human health sector. By looking at how one sector compares related to others, the population can make the government accountable for its prioritization.

4.4 Normative act popularity

In order to allow the government to gain insights about the regulatory topics of greatest interest to its population and to its internal administration, a novel metric is also proposed, which is the normative act popularity. This new metric indicates how popular is a normative act concerning a specific group. This metric aims to indicate which are the most popular normative acts, and consequently, the ones that generate the most interest from the general population and the federal government.

Population Normative act popularity

The population normative act popularity is calculated based on the population active search for specific normative acts on Google, a traditional search engine for the general population. In this context, we used information from the Google Trends engine alongside rules to search for normative acts and get their search frequencies.

Google first launched Google Trend in 2006 to analyze the popularity of top search topics in the Google Search platform across various regions and languages starting from 2004 [120]. This tool has been used in different applications over the years, such as prediction of the stock market behavior [121] and tourism patterns [122], it was also used to understand the behavior of epidemiological diseases [123] and to calculate search popularity of professional cycling [120].

Google Trends measures search popularity in relative terms based on a randomly drawn sample, normalizing search data to make comparisons between different terms easier. For search popularity calculation, each data point is divided by the total searches of the geography and time range it represents to compare relative popularity [124]. It is important to note that Google Trends search popularity ranges from 0 to 100, and it only shows data for popular terms, so search terms with low volume are set to 0 [124].

For RegBR, the search strings used to calculate Google Trends search popularity were formed by the following elements: the act type followed by its number, a slash, and the publication year. This is the most usual way of researching a normative act on search tools. Examples of the search strings are *Law 8.112/1990* and *Constitutional Amendment 20/1998*. It is important to indicate that the normative act popularity was calculated to the first six normative act types, excluding ’Ordinances’, ’Resolutions’, ’Normative Instructions’ and ’Precedents’ since its search strings are not standardized with the other normative acts types.

As search parameters, it was used ’BR’ as geo-attribute and the last ten years as timeframe. It is worth mentioning that for normative acts existing for less than ten years, only the actual existence of the normative act was considered in calculating their average popularity.

Table 4.4: Average popularity of normative acts on Google Trends

Normative act name and subject	Average interest
Law 13.982/2020 - COVID-19 emergency aid	89.9
Law 13.979/2020 - COVID-19 emergency measures	49.8
Law 13.467/2017 - Labor Reform	36.9
Law 13.146/2015 - Statute of People with Disabilities	34.0
S.L.[a] 123/2006 - Statute of Micro and Small Business	31.8
Law 11.343/2006 - Public Policies on Drugs	27.8
Law 10.826/2003 - Disarmament Statute	25.7
Law 8.112/1990 - Legal regime for federal civil servants	23.0
Law 8.666/1993 - Public bids and contracts	22.8
Law 11.101/2005 - Company Rehab and Bankruptcy	22.1
Law 12.799/2013 - Registration fee exemption for exams	20.5
Law 14.010/2020 - COVID-19 pandemic law	20.4
S.L. 101/2000 - Fiscal responsibility law	19.1
Law 12.016/2009 - Writ of Mandamus law	18.0
C.A.[b] 87/2015 - Interstate taxes and operations	16.9

[a] Supplementary Law.
[b] Constitutional Amendment.

As a result, Table 4.4 show the average popularity for the 15 more popular normative acts. Not surprisingly, the most popular normative acts for the general public are concerned with important social aspects such as the COVID-19 pandemic crises, the labor reform and the statutes of people with disabilities, micro and small business, and disarmament proposal.

Government Normative act popularity

On the other hand, the popularity metric in the context of the Federal Government is based on the frequency of normative acts citations in the Official Gazette of the Federal Government (DOU).

To implement this search, a data extractor of all contents of sections 2 and 3 from DOU was implemented since 2001, when its digital form became available. Acts located in section 2 deal with publications relating to public servants, such as appointments and designations of commissioned positions, while section 3 is meant to publish notices, contracts, amendments, cancellations, agreements, concessions, among others. Since section 1 is intended to publish normative acts itself, such as laws, decrees, resolutions, normative instructions, ordinances, and other normative acts of general interest, this session is not used to quantify the citations of normative acts in itself.

After extraction, the text content is structured, the citation frequency for each normative act is calculated and then normalized to obtain a popularity metric value between 0 and 100.

As a result, Figure 4.5 shows the average popularity for the 15 most popular normative acts on DOU. The most popular normative acts for the Federal Government are those related to public administration, such as rules that regulate bidding and contracts at the federal level, dispose of internal rules for the Union's Court of Auditors, and establish the legal regime of the public civil servants of the Union.

Table 4.5: Average popularity of normative acts on DOU.

Normative act name and subject	Average interest
Law 8.666/1993 - Public Adm bids and contracts	100.0
Law 8.443/1992 - Federal Audit Court	94.6
Law 8.112/1990 - Regime for servants of the Union	80.5
Law 10.520/2002 - Law of the Auction	49.5
Law 13.303/2016 - State Law	46.1
Law 10.887/2004 - Retirement of the public servants	41.0
C.A. 41/2003 - Social security of the public servants	33.3
C.A. 47/2005 - Retirement of the public servants	24.4
Law 11.416/2006 - Judicial servants careers	19.2
Law 13.135/2015 - Pension in case of death	17.5
S.L. 123/2006 - Statute of Micro and Small Business	13.9
Law 12.772/2012 - Federal Magisterial careers plan	10.2
Decree 7.892/2013 - Price Registration System	8.1
Decree 5.450/2005 - Federal auction purchases	7.7
Law 13.979/2020 - COVID-19 emergency measures	7.2

It is important to mention that the values of the two popularity metrics are not comparable since the data from Google Trends does not represent the frequency of citations of a normative act but rather the result of a Google calculation method that considers factors such as Geolocation. In addition, both metrics are not calculated over the same time interval, making them incomparable among themselves.

4.5 Linguistic complexity

The last implemented metric is related to the complexity of each regulation, and it is relevant for several reasons as more complex regulations may force regulated entities to employ more personal and spend more time understanding them. Moreover, it can also make it less accessible to the public as the language gets too specific.

Inspired by [20], three different metrics were employed to compare regulations' complexity. The first metric is the median sentence length. The median is used to avoid the effects of outliers that can appear in parsing sentences in law documents, such as tables or other bodies of text. The main assumption is that longer sentences tend to be more challenging to understand and, consequently, increase the document's complexity.

The second metric employed is Shannon's entropy, a measure of the average information of a single message from a given source [125]. It can be interpreted as measuring the frequency that new ideas (or words) are introduced in documents. Consequently, simpler and more focused documents have a lower entropy score than more complex documents. The entropy can be defined as

$$H(X_j) = -\sum_{i=1}^{N} p(x_{i,j}) log_2(p(x_{i,j})), \tag{4.3}$$

in which X_j denotes the j-th document, $p(x_{i,j})$ indicates the probability/frequency of the word $x_{i,j}$ occurring in document j and N is the total number of words in document j.

The final metric is the frequency of conditional words in the text, which counts the number of branching words (in English, those are words such as "if", "but", and "provided") found in any given part in the text. Since we are working with Portuguese text, we adapted the conditional terms to words that denote a similar branching idea [3].

The goal of evaluating these three metrics is to understand if regulations are getting more complex by analyzing if the ideas are extended and wordy. Recall that the following analysis is based on the classification results obtained in Section 3.3. Therefore, the accuracy of the results is directly related to the quality of the classification procedure.

It is worth mentioning that the three-complexity metrics considered are not directly related. In other words, they express the complexity of a document by measuring different parameters. Therefore, one metric can increase over time while the other two may present different behaviors. So, to assess if the regulations from a sector became complex, we have to analyze the three metrics together.

For example, if the number of conditional words for a given sector increases, the entropy can decrease if the total number of words is the same as the idea is being presented using a smaller word variation.

The main goal of the complexity analysis is to understand how it evolved over time. It can be analyzed on a macro-level by grouping all sectors and calculating the median for each metric over time. The median is used to avoid the influence of possible outliers in the observed period. Moreover, to have comparable results, each metric was standardized (zero mean and unit variance) and shifted by subtracting the first value of the series, so each metric starts at zero. The results are shown in Figure 4.7.

It is evident that the overall complexity increased in the observed period. The conditional count metric was reduced starting in the middle of the 1990s, but its value is still higher than the beginning of the period. On the other hand, both entropy and median sentence length increased substantially starting in 2000.

However, it is essential to note that different sectors of the economy can present different regulatory dynamics. Thus, the overall complexity might not present the whole picture of a sector regulation over time. Moreover, since each sector would have three curves representing different metrics, the complexity trend could end up being

[3] We used the following words as conditional words:*"se", "caso", "quando", "dado que", "desde que", "a menos que", "a não ser que", "embora", "ainda que" "mesmo que", "posto que"* and *"em que"*.

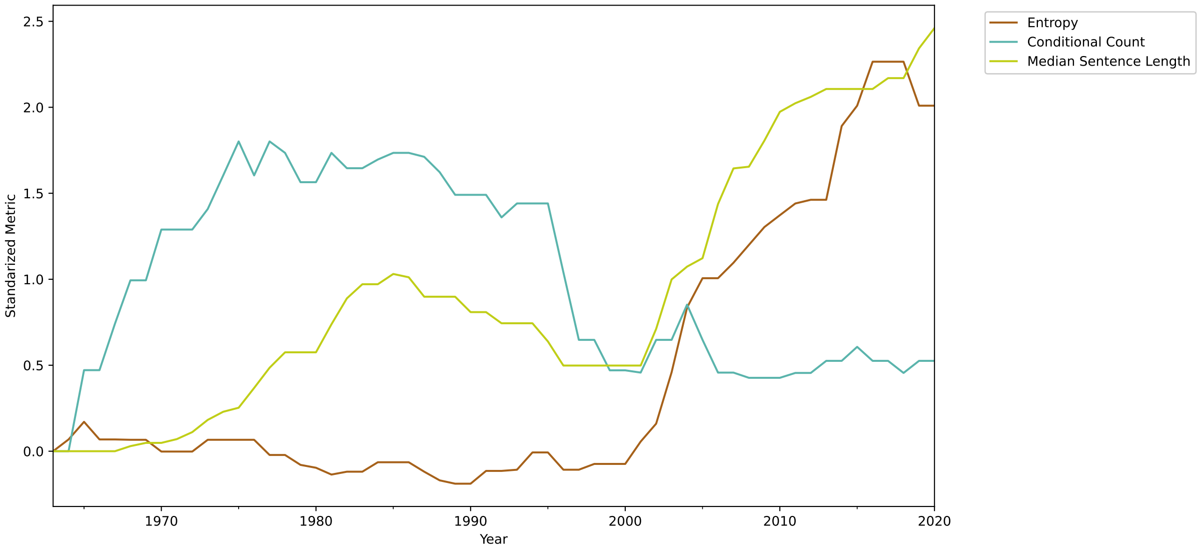


Figure 4.7: Median metrics from all sectors grouped together, a moving average of 14 years was used to smooth out the curves.

challenging to present. To represent each sector as a single time-series, a kernel principal component analysis [126] (KPCA) was employed to reduce the three metrics to one dimension.

Since the metrics are not necessarily correlated, the same is true with respect to its one-dimensional projection. Therefore, only sectors that present significant levels of correlation between the KPCA projection and their complexity metrics are considered, which implies that an increase in the projected complexity is related to an increase in the complexity metrics. The results are presented in Figure 4.8.

Most of the observed sectors had an increase in regulation complexity throughout the period from 1964 to 2020. There are two distinct moments where regulations from most sectors become more complex. First around 1970 and then in the year 2000. Since the curves were smoothed using a moving average, the variations are not reflected instantly. So these two periods can be attributed to the new form of government established in 1964 and the re-democratization that started in 1985.

As stated before, the results presented in this section should not be considered as facts. They are based on classifications from a model that can (and most likely will) make mistakes, and on noisy metrics that try to represent the complexity of a text.

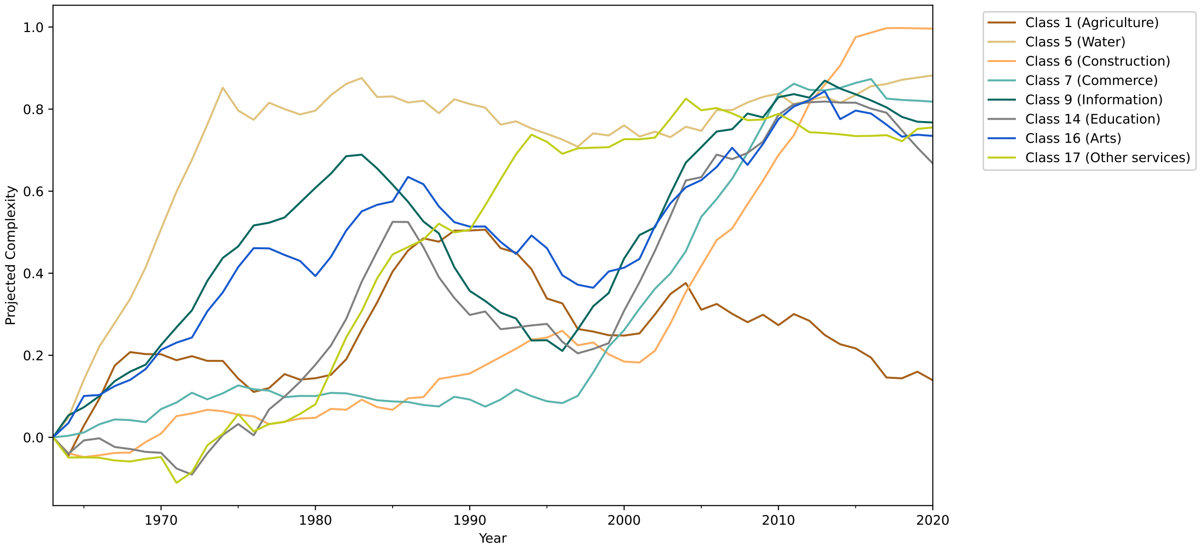


Figure 4.8: Complexity projection, a moving average of 14 years was used to smooth out the curves.

Instead, they can be used as a decision support tool for policymakers; in other words, they can help gain insights on how regulations evolved over time and how they can be improved.

5 RegBR applications

In this section, six of many applications of RegBr framework on the federal government are discussed: RegBr as a regulatory base for studies; RegBr as a feedback mechanism to regulatory agencies; RegBr as a monitoring tool; RegBr as a comparative apparatus; RegBr as a predictor of regulatory governance design in Brazil and RegBr as a transparency instrument. All of these applications are related to the usage of RegBr by citizens, organizations, and by the Federal Government itself.

Even though RegBr is a framework with all the features described in previous sections, the applications stated in this section bring another look to its usage and focus on how this research object can be useful in a practical way.

5.1 Regulatory base to studies

Since RegBr is a framework that quantifies regulations produced by the Brazilian government across the years, this tool became an important dataset that subsidizes regulatory studies. For instance, the *Revista do Serviço Público (Public Service Journal)* recently opened a call of papers to prospect studies that uses RegBr as its data source.

Additionally, the Brazilian Executive branch, particularly regulatory agencies, has been increasingly adopting the practice of regulatory impact analysis (RIA) prior to deliberation on new regulations, and regulatory results assessments (RRA) following regulation implementation. ANEEL, the National Agency of Electric Energy, is an example of a regulatory agency successfully implementing these practices to evaluate regulations in the energy economic sector [127].

Scholars seeking to better comprehend the impact of regulations in Brazil can utilize RegBr as a valuable data source for both RIA and RRA. These developments demonstrate a positive shift towards evidence-based policymaking and reducing negative externalities. RegBr's data can be utilized to assess the effectiveness of these practices and identify areas for improvement.

5.2 Feedback mechanism to regulatory agencies

The RegBr framework comprises four metrics that evaluate the quality of regulations generated by Brazilian regulatory agencies based on their popularity, restrictiveness, influence on the economic sector, and linguistic complexity. The linguistic complexity metric, in particular, is of great importance to regulatory agencies, as it serves as a gauge of transparency and effectiveness. It assesses the comprehensibility of normatives to citizens and regulated entities, and regulations with high linguistic complexity may pose a challenge to their interlocutors' understanding, ultimately impacting their effectiveness.

In 2021, the National Agency of Waters (Agência Nacional de Águas - ANA) utilized the RegBr metrics to assess the linguistic complexity of their normatives. Following a meeting with RegBr researchers, ANA staff confirmed their concern that the linguistic complexity of their regulations might be impeding public understanding.

This instance illustrates how regulatory agencies are utilizing RegBr as a feedback mechanism to evaluate and improve their regulatory processes.

5.3 Monitoring tool

RegBR can also be used as a monitoring tool. The framework allows the Brazilian Federal Government to monitor its regulatory production, measuring lengthwise the number of acts that have been produced.

Regarding the task of monitoring the number of normative that are being produced and have been produced, the federal government signed the Decree nº 10.139 from 2019 November (Decreto nº 10.139 de Novembro de 2019), that establishes that every organization in the Federal Government must revoke normatives that no longer had applicability. This initiative aimed to reduce the number of normative in the legal system.

Since RegBR quantifies the number of normative and their legal status, this tool can be used by the Federal Government to analyze the impact of the Decree 10.139 and other initiatives that have similar goals.

Moreover, RegBR can assist decision makers in measuring their own work, providing a framework that allows the heads of regulatory agencies to measure what their organization produces in terms of volume and characteristics of regulations. The use of standard metrics can be interesting to have a clearer view of historical trends in the context of a specific regulatory agency or to compare different regulatory agencies in terms of their produced normative characteristics.

5.4 Comparative apparatus

RegBR is an initiative inspired by RegData, from Mercatus Center, which aims to quantify the regulation produced by the Brazilian Federal Government. Like RegData, RegBR analyzes data by regulator and industry and measures regulatory restrictiveness, a key metric for assessing the impact of regulations on economic growth and development.

Updates in RegData are launched at “quantgov.org" site, which also brings information about other English-speaking countries such as Canada and Australia that also implemented a similar framework. With RegBR, the Brazilian government can compare its regulatory environment with these countries and identify areas where it can improve its regulatory policies and practices.

By using RegBR, the Brazilian government can gain valuable insights into how its regulatory environment affects different industries and identify potential areas for improvement. Additionally, comparing its regulatory production and restrictiveness with other countries can help Brazil benchmark itself and learn from best practices implemented by other nations.

5.5 Predictor of regulatory governance design in Brazil

As the regulatory governance concept involves a range of activities including the development of regulations, the monitoring and enforcement of regulatory compliance, and the evaluation of the effectiveness of regulatory policies and practices, the RegBR framework has the potential to be a valuable predictor of regulatory governance design in Brazil.

By providing policymakers and researchers with access to centralized regulatory data, the tool can facilitate analysis of regulatory trends and practices, as well as identify areas and economic sectors where regulation improvements may be needed.

One of the key advantages of RegBR is that it can provide policymakers with compiled information on the quantity and quality of regulatory activity in Brazil. By monitoring changes in regulatory activity over time, policymakers can identify areas where regulatory reform may be needed to improve the quality and effectiveness of the regulatory system. This can be particularly valuable in identifying emerging regulatory issues and addressing them before they become major problems.

In addition, by providing citizens with access to regulatory data and facilitating public discussion of regulatory issues, RegBr can help to build trust in the regulatory process and promote greater public participation in regulatory decision-making.

5.6 Transparency instrument

Last but not least, RegBR achieves one proposal of Law nº 12.527, from 2011 November (Lei nº 12.527 de Novembro de 2011), known in Brazil as *Access to Information Law*, which states that:

> "art. 3 The procedures provided for in this Law are intended to ensure the fundamental right of access to information and must be carried out in accordance with the basic principles of public administration and with the following rules:
>
> . . .
>
> IV - fostering the development of a culture of transparency in public administration;"

Based on the principles of *fostering the culture of transparency*, RegBr gathers and centralizes federal regulations on a public database, with intuitive navigation and easy data visualization.

It is important to note that all the normative analyzed by RegBR had already been published in the Official Gazette of the Federal Government (DOU) and gathered as text on the government's official website. Nevertheless, RegBR goes a step further as regards accessibility. The tool transforms massive amounts of data that are being generated by various sources into a data base and subsequently displays the results from the data analysis in a simple format, making the new information public and interactive on the website Infogov. In this context, RegBr leverages the already discussed

BOLD concept to create a network of linked regulatory data that can be easily queried and visualized by citizens, policymakers, and researchers.

Probably the most important application of this framework is to ensure easy access to data and to foment active transparency, which means providing information to the public without being requested by a specific citizen.

6 Conclusion

This doctoral thesis presents a novel active transparency framework, RegBr, designed to facilitate regulatory analysis and monitoring of legislative metrics over time. The framework also includes a benchmark for text classification of Brazilian federal normative legislation into economic sectors since 1964, using data collected from decentralized sources and classified using state-of-the-art natural language processing models.

RegBr implements various metrics, such as text linguistic complexity, law popularity, law restrictiveness, and citation relevance of each industry regulated by the law corpus, to evaluate the Brazilian regulatory stock. These metrics are tracked over time and provide essential information for identifying and prioritizing regulations that require reforms. Moreover, policymakers can use these metrics to measure their own work. The metrics provided can also be used for future comparative analyses of government changes and their relationship with federal regulations produced by the legislative branch.

An additional significant contribution of this work is the centralized database compiled, containing different regulatory acts from 1891 to the present day. This database will be made available to researchers and professionals for further investigation, facilitating transparency and reducing future costs for obtaining data and disseminating information.

An essential objective of this research is to advance the democratization of information. To achieve this, it is crucial to effectively measure access to the produced content. To this end, RegBR is accessible as a tool through the Infogov website. According to Google Analytics data, a web analytics service that provides statistics and basic analytical tools for search engine optimization (SEO) and marketing purposes, the *popularity metric*, and the *regulatory flow* pages rank as the fourth and fifth most visited web pages in the Infogov website, respectively. This indicates significant interest in the information made available by RegBR. Specifically, these two pages alone received over 800 unique views within the first two months of 2023.

The author aims to increase openness and transparency of the public process and to support new studies in the area of Brazilian regulatory impact. Besides the national impact, this framework has the potential to be replicated in other Portuguese-speaking countries. Additionally, the proposed metrics can be adapted to many different languages and legal corpora, enabling a worldwide comparison of regulatory evolution.

6.1 Limitations and Suggestions for Future Research

Every research has inherent limitations that need to be acknowledged, and this study is no exception. Despite concerted efforts to minimize them, certain limitations must be acknowledged.

During this doctoral research program, two iterations of the RegBR framework were developed, with the second iteration boasting a broader scope encompassing 10 different types of regulations and 11 regulatory agencies. In future research, it is proposed that continued collaboration with regulatory agencies will strengthen the standardization of regulations in their websites, facilitating the centralization of new regulations within the RegBR database. In addition, broadening the study's scope by including new types of regulatory acts and regulatory agencies is also a will.

Also, the study recognized a potential limitation in the data collection phase, where the use of web scraping techniques in the ETL routines could result in code failures if there are modifications in the HTML of the web pages in the future. Nonetheless, due to the absence of an API or data centralization, web scraping was considered the only feasible method for data extraction. However, to ensure the reliability and robustness of the system, the RegBR's development will continue after this dissertation research, and all required code maintenance will be conducted.

Another limitation refers to the fact that normative texts are by nature extensive, and several studies have highlighted the constraints of Transfer Learning algorithms, such as BERT, in effectively classifying lengthy texts. Specifically, due to the quadratic increase in memory and processing time, the commonly used limit for applications using BERT is around only 512 tokens.

Although diverse models and NLP methodologies were employed in this study, it is by no means exhaustive, and there may be other models such as GPT-4 or ensemble approaches that can enhance the findings and strengthen the text classification benchmark.

Finally, the development of new regulatory metrics is also a future goal, as they can be instrumental in aiding decision-makers and enhancing the regulatory impact analysis. Future metrics can also provide a more comprehensive understanding of the costs and benefits of regulatory policies, including their economic, social, and environmental impact

Bibliography

[1] Bertot, John C, Paul T Jaeger e Justin M Grimes: *Using icts to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies*. Government information quarterly, 27(3):264–271, 2010.

[2] Shuler, John A, Paul T Jaeger e John Carlo Bertot: *Implications of harmonizing e-government principles and the federal depository library program (fdlp)*. Government Information Quarterly, 27(1):9–16, 2010.

[3] Cuillier, David e Suzanne J Piotrowski: *Internet information-seeking and its relation to support for access to government records*. Government Information Quarterly, 26(3):441–449, 2009.

[4] Heeks, Richard: *Information and communication technologies, poverty and development*. Development Informatics Working Paper no. 5.

[5] Bertot, John C., Paul T. Jaeger e Justin M. Grimes: *Using icts to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies*. Government Information Quarterly, 27(3):264–271, 2010, ISSN 0740-624X. https://www.sciencedirect.com/science/article/pii/S0740624X10000201.

[6] Abu-Shanab, Emad A: *Reengineering the open government concept: An empirical support for a proposed model*. Government Information Quarterly, 32(4):453–463, 2015.

[7] Pigou, Arthur Cecil: *The economics of welfare*. Palgrave Macmillan, 2013.

[8] Peltzman, Sam: *The effects of automobile safety regulation*. Journal of political Economy, 83(4):677–725, 1975.

[9] Dal Bó, Ernesto: *Regulatory capture: A review*. Oxford review of economic policy, 22(2):203–225, 2006.

[10] Janssen, Marijn e Jeroen van den Hoven: *Big and open linked data (bold) in government: A challenge to transparency and privacy?* Government Information Quarterly, 32(4):363–368, 2015, ISSN 0740-624X. https://www.sciencedirect.com/science/article/pii/S0740624X15001069.

[11] Sulea, Octavia Maria, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu e Josef van Genabith: *Exploring the use of text classification in the legal domain*, 2017. https://arxiv.org/abs/1710.09306.

[12] SOH, Jerrold Tsin Howe, How Khang LIM e Ian Ernst CHAI: *Legal topic classification: A comparative study of text classifiers on singapore supreme court judgments.(2019)*. Em *Proceedings of the Natural Legal Language Processing Workshop*, páginas 67–77, 2019.

[13] Şulea, Octavia Maria, Marcos Zampieri, Mihaela Vela e Josef van Genabith: *Predicting the law area and decisions of french supreme court cases*. Em *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, páginas 716–722, 2017.

[14] Katz, Daniel Martin, Michael J. Bommarito, II e Josh Blackman: *A general approach for predicting the behavior of the supreme court of the united states*. PLOS ONE, 12(4):1–18, abril 2017. https://doi.org/10.1371/journal.pone.0174698.

[15] Wongchaisuwat, Papis, Diego Klabjan e John O McGinnis: *Predicting litigation likelihood and time to litigation for patents*. Em *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, páginas 257–260, 2017.

[16] Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro e Vasileios Lampos: *Predicting judicial decisions of the european court of human rights: A natural language processing perspective*. PeerJ Computer Science, 2:e93, 2016.

[17] Law, Kincho H, Gloria Lau, Shawn Kerrigan e Julia A Ekstrom: *Regnet: Regulatory information management, compliance and analysis*. Government Information Quarterly, 31:S37–S48, 2014.

[18] Al-Ubaydli, Omar e Patrick A McLaughlin: *Regdata: A numerical database on industry-specific regulations for all united states industries and federal regulations, 1997–2012*. Regulation & Governance, 11(1):109–123, 2017.

[19] McLaughlin, Patrick A, Oliver Sherouse e Jason Potts: *Regdata: Australia*. Mercatus Research Paper, 2019.

[20] McLaughlin, Patrick A, Stephen Strosko e Laura Jones: *Regdata canada: A snapshot of regulatory restrictions in canada's provinces*. Mercatus Center, George Mason University, Regulatory Snapshot, 2019.

[21] Parker, David e Colin Kirkpatrick: *Measuring regulatory performance*. The economic impact of, 2012.

[22] Égert, Balázs e Isabelle Wanner: *Regulations in services sectors and their impact on downstream industries: The oecd 2013 regimpact indicator*. Relatório Técnico, OECD, 2016.

[23] Vitale, Cristiana, Rosamaria Bitetti, Isabelle Wanner, Eszter Danitz e Carlotta Moiso: *The 2018 edition of the oecd pmr indicators and database: Methodological improvements and policy insights*. Relatório Técnico, OECD, 2020.

[24] Salton, Gerard e Christopher Buckley: *Term-weighting approaches in automatic text retrieval*. Information processing & management, 24(5):513–523, 1988.

[25] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado e Jeff Dean: *Distributed representations of words and phrases and their compositionality*. Em *Advances in neural information processing systems*, páginas 3111–3119, 2013.

[26] Devlin, Jacob, Ming Wei Chang, Kenton Lee e Kristina Toutanova: *BERT: Pre-training of deep bidirectional transformers for language understanding*. Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 4171–4186, Minneapolis, Minnesota, junho 2019. Association for Computational, Linguistics.

[27] Akhter, M. P., Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood e M. T. Sadiq: *Document-level text classification using single-layer multisize filters convolutional neural network*. IEEE Access, 8:42689–42707, 2020.

[28] Orlikowski, Wanda J e Stephen R Barley: *Technology and institutions: What can research on information technology and research on organizations learn from each other?* MIS quarterly, páginas 145–165, 2001.

[29] Mitchell, Victoria L e Robert W Zmud: *The effects of coupling it and work process strategies in redesign projects*. Organization Science, 10(4):424–438, 1999.

[30] Shim, Dong Chul e Tae Ho Eom: *E-government and anti-corruption: Empirical analysis of international data*. Intl Journal of Public Administration, 31(3):298–316, 2008.

[31] Song, Changsoo e Jooho Lee: *Citizens' use of social media in government, perceived transparency, and trust in government*. Public Performance & Management Review, 39(2):430–453, 2016.

[32] Fishenden, Jerry e Mark Thompson: *Digital government, open architecture, and innovation: why public sector it will never be the same again*. Journal of public administration research and theory, 23(4):977–1004, 2013.

[33] Robinson, David, Harlan Yu, William P Zeller e Edward W Felten: *Government data and the invisible hand*. Yale JL & Tech., 11:159, 2008.

[34] Matheus, Ricardo, Marijn Janssen e Devender Maheshwari: *Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities*. Government Information Quarterly, 37(3):101284, 2020.

[35] Velcu-Laitinen, Oana e Ogan M Yigitbasioglu: *The use of dashboards in performance management: Evidence from sales managers.* International Journal of Digital Accounting Research, 12, 2012.

[36] Maheshwari, Devender e Marijn Janssen: *Dashboards for supporting organizational development: principles for the design and development of public sector performance dashboards*. Em *proceedings of the 8th international conference on theory and practice of electronic governance*, páginas 178–185, 2014.

[37] Brown, Alan, Jerry Fishenden, Mark Thompson e Will Venters: *Appraising the impact and role of platform models and government as a platform (gaap) in uk government public service reform: Towards a platform assessment framework (paf)*. Government Information Quarterly, 34(2):167–182, 2017.

[38] Chengalur-Smith, InduShobha N, Donald P Ballou e Harold L Pazer: *The impact of data quality information on decision making: an exploratory analysis*. IEEE Transactions on Knowledge and Data Engineering, 11(6):853–864, 1999.

[39] Allio, Michael K: *Strategic dashboards: designing and deploying them to improve implementation*. Strategy & Leadership, 2012.

[40] Christian Bizer, Tom Heath e Tim Berners-Le: *Linked Data: The Story so Far*. Massachusetts Institute of Technology, USA, 2011.

[41] Matheus, Ricardo e Marijn Janssen: *Transparency dimensions of big and open linked data*. Em Janssen, Marijn, Matti Mäntymäki, Jan Hidders, Bram Klievink, Winfried Lamersdorf, Bastiaan van Loenen e Anneke Zuiderwijk (editores): *Open and Big Data Management and Innovation*, páginas 236–246, Cham, 2015. Springer International Publishing.

[42] Stiglitz, Joseph E: *Markets, market failures, and development*. The American Economic Review, 79(2):197–203, 1989.

[43] Coates, Dennis e Jac C Heckelman: *Interest groups and investment: a further test of the olson hypothesis*. Public Choice, 117(3):333–340, 2003.

[44] Stigler, George J: *The theory of economic regulation*. The Bell journal of economics and management science, páginas 3–21, 1971.

[45] McChesney, Fred S: *Rent extraction and rent creation in the economic theory of regulation*. The Journal of Legal Studies, 16(1):101–118, 1987.

[46] Rowley, Charles, Robert D Tollison e Gordon Tullock: *The political economy of rent-seeking*, volume 1. Springer Science & Business Media, 2013.

[47] Banerjee, Ryan N e Hitoshi Mio: *The impact of liquidity regulation on banks*. Journal of Financial intermediation, 35:30–44, 2018.

[48] Dredge, Dianne, Szilvia Gyimóthy, Andreas Birkbak, Torben Elgaard Jensen e Anders Madsen: *The impact of regulatory approaches targeting collaborative economy in the tourism accommodation sector: Barcelona, berlin, amsterdam and paris*. Impulse Paper, 1(9), 2016.

[49] Leisen, Robin, Bjarne Steffen e Christoph Weber: *Regulatory risk and the resilience of new sustainable business models in the energy sector*. Journal of Cleaner Production, 219:865–878, 2019.

[50] Crews, Clyde Wayne: *Ten thousand commandments: An annual snapshot of the federal regulatory state*. Cato Institute, 2002.

[51] Coffey, Bentley, Patrick A McLaughlin e Robert D Tollison: *Regulators and redskins*. Public Choice, 153(1):191–204, 2012.

[52] Lodge, Martin, Christian Van Stolk, Julia Batistella-Machado, Daniel Schweppenstedde e Martin Stepanek: *Regulation of logistics infrastructure in Brazil*. RAND, 2017.

[53] Sousa, Ana Cristina A. de e Nilson do Rosário Costa: *Política de saneamento básico no brasil: discussão de uma trajetória*. História, Ciências, Saúde-Manguinhos, 23(Hist. cienc. saude-Manguinhos, 2016 23(3)):615–634, Jul 2016, ISSN 0104-5970. https://doi.org/10.1590/S0104-59702016000300002.

[54] Motta, Ronaldo Seroa da e Ajax Moreira: *Efficiency and regulation in the sanitation sector in brazil*. Utilities Policy, 14(3):185–195, 2006.

[55] Aronson, James, Pedro HS Brancalion, Giselda Durigan, Ricardo R Rodrigues, Vera L Engel, Marcelo Tabarelli, José MD Torezan, Sergius Gandolfi, Antônio CG de Melo, Paulo Y Kageyama *et al.*: *What role should government regulation play in ecological restoration? ongoing debate in são paulo state, brazil*. Restoration Ecology, 19(6):690–695, 2011.

[56] Nadkarni, Prakash M, Lucila Ohno-Machado e Wendy W Chapman: *Natural language processing: an introduction*. Journal of the American Medical Informatics Association, 18(5):544–551, setembro 2011, ISSN 1067-5027. https://doi.org/10.1136/amiajnl-2011-000464.

[57] Baber, C.: *Developing Interactive Speech Technology*, página 13–18. Taylor Francis, Inc., USA, 1993, ISBN 074840127X.

[58] Middi, Venkata Sai Rishita, Middi Raju e Tanvir Ahmed Harris: *Machine translation using natural language processing*. MATEC Web of Conferences, 277:02004, janeiro 2019.

[59] Bhavani, A. e B. Santhosh Kumar: *A review of state art of text classification algorithms*. Em *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, páginas 1484–1490, 2021.

[60] Ittoo, Ashwin, Le Minh Nguyen e Antal van den Bosch: *Text analytics in industry: Challenges, desiderata and trends*. Computers in Industry, 78:96–107, 2016, ISSN 0166-3615. https://www.sciencedirect.com/science/article/pii/S0166361515300646, Natural Language Processing and Text Analytics in Industry.

[61] Christopher D. Manning, Prabhakar RaghavanHinrich Schütze: *Introduction to Information Retrieval*. 2008.

[62] Arboretti, Rosa, Riccardo Ceccato, Luca Pegoraro e Luigi Salmaso: *Design choice and machine learning model performances*. Quality and Reliability Engineering International, 38(7):3357–3378, may 2022.

[63] Jones, Karen Sparck: *A statistical interpretation of term specificity and its application in retrieval*. Journal of documentation, 1972.

[64] Robertson, Stephen: *Understanding inverse document frequency: on theoretical arguments for idf*. Journal of documentation, 2004.

[65] Pennington, Jeffrey, Richard Socher e Christopher D Manning: *Glove: Global vectors for word representation*. Em *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, páginas 1532–1543, 2014.

[66] Joulin, Armand, Edouard Grave, Piotr Bojanowski e Tomas Mikolov: *Bag of tricks for efficient text classification*. Em *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, páginas 427–431. Association for Computational Linguistics, April 2017.

[67] Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee e Luke Zettlemoyer: *Deep contextualized word representations*. Em *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 2227–2237, 2018.

[68] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer e Veselin Stoyanov: *Roberta: A robustly optimized BERT pretraining approach*. CoRR, abs/1907.11692, 2019.

[69] Sanh, Victor, Lysandre Debut, Julien Chaumond e Thomas Wolf: *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter*. Em *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, página 0, 2019.

[70] Wang, Wei, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng e Luo Si: *Structbert: Incorporating language structures into pre-training for deep language understanding*. Em *International Conference on Learning Representations*, página 0, 2020. https://openreview.net/forum?id=BJgQ4lSFPH.

[71] Giannakas, F, C Troussas, I Voyiatzis e C Sgouropoulou: *A deep learning classification framework for early prediction of team-based academic performance*. Applied Soft Computing, 106:107355, 2021.

[72] Hussain, Shahid, Jacky Keung, Muhammad Khalid Sohail, Arif Ali Khan e Manzoor Ilahi: *Automated framework for classification and selection of software design patterns*. Applied Soft Computing, 75:1–20, 2019.

[73] Paukkeri, Mari Sanna, Alberto Pérez García-Plaza, Víctor Fresno, Raquel Martínez Unanue e Timo Honkela: *Learning a taxonomy from a set of text documents*. Applied Soft Computing, 12(3):1138–1148, 2012.

[74] Palau, Raquel Mochales e Marie Francine Moens: *Argumentation mining: the detection, classification and structure of arguments in text*. Em *Proceedings of the 12th international conference on artificial intelligence and law*, páginas 98–107, 2009.

[75] Boella, Guido, Luigi Di Caro e Llio Humphreys: *Using classification to support legal knowledge engineers in the eunomos legal document management system*. Em *Fifth international workshop on Juris-informatics (JURISIN)*, páginas 245–283. Citeseer, 2011.

[76] Yuan, Lufeng, Jun Wang, Shifeng Fan, Yingying Bian, Binming Yang, Yueyue Wang e Xiaobin Wang: *Automatic legal judgment prediction via large amounts of criminal cases*. Em *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, páginas 2087–2091, 2019.

[77] Suykens, Johan AK e Joos Vandewalle: *Least squares support vector machine classifiers*. Neural processing letters, 9(3):293–300, 1999.

[78] Joulin, Armand, Edouard Grave e Piotr Bojanowski Tomas Mikolov: *Bag of tricks for efficient text classification*. EACL 2017, página 427, 2017.

[79] Tuggener, Don, Pius von Däniken, Thomas Peetz e Mark Cieliebak: *Ledgar: a large-scale multi-label corpus for text classification of legal provisions in contracts*. Em *12th Language Resources and Evaluation Conference (LREC) 2020*, páginas 1228–1234. European Language Resources Association, 2020.

[80] Rowe, L. A. e M. R. Stonebraker: *The POSTGRES Data Model*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989, ISBN 0558600000.

[81] Conradie, Peter e Sunil Choenni: *On the barriers for local government releasing open data*. Government Information Quarterly, 31:S10–S17, 2014.

[82] Government, Brazilian Federal: *Lexml project*, 2020. https://www.lexml.gov.br/, acesso em 2021-04-16.

[83] Foundation, The Apache Software: *Apache airflow*. https://airflow.apache.org/, acesso em 2021-04-16.

[84] Van Rossum, Guido e Fred L. Drake: *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009, ISBN 1441412697.

[85] Richardson, Leonard: *Beautiful soup documentation*. April, 2007.

[86] Gojare, Satish, Rahul Joshi e Dhanashree Gaigaware: *Analysis and design of selenium webdriver automation testing framework*. Procedia Computer Science, 50:341–346, 2015, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050915005396, Big Data, Cloud and Computing Challenges.

[87] Yan Yan, Rómer Rosales, Glenn Fung Ramanathan Subramanian e Jennifer Dy: *Learning from multiple annotators with varying expertise*. Machine Learning.

[88] IGBE: *Introdução à classificação nacional de atividades econômicas - cnae versão 2.0*. Relatório Técnico, IBGE, 2007.

[89] Government, Brazilian Federal: *Brazilian legislation portal*, 2021. http://www4.planalto.gov.br/legislacao/, acesso em 2021-03-06.

[90] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg *et al.*: *Scikit-learn: Machine learning in python*. the Journal of machine Learning research, 12:2825–2830, 2011.

[91] Friedman, Jerome H: *Greedy function approximation: a gradient boosting machine*. Annals of statistics, páginas 1189–1232, 2001.

[92] Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer e Richard Harshman: *Indexing by latent semantic analysis*. Journal of the American society for information science, 41(6):391–407, 1990.

[93] Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall e W Philip Kegelmeyer: *Smote: synthetic minority over-sampling technique*. Journal of artificial intelligence research, 16:321–357, 2002.

[94] Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva e Sandra Aluísio: *Portuguese word embeddings: Evaluating on word analogies and natural language tasks*. Em *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, páginas 122–131, 2017.

[95] Krizhevsky, Alex, Ilya Sutskever e Geoffrey E Hinton: *Imagenet classification with deep convolutional neural networks*. Advances in neural information processing systems, 25:1097–1105, 2012.

[96] Hochreiter, Sepp e Jürgen Schmidhuber: *Long short-term memory*. Neural computation, 9(8):1735–1780, 1997.

[97] Goldberg, Yoav: *Neural network methods for natural language processing*. Synthesis lectures on human language technologies, 10(1):1–309, 2017.

[98] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin: *Attention is all you need*. Em *Advances in neural information processing systems*, páginas 5998–6008, 2017.

[99] Howard, Jeremy e Sebastian Ruder: *Universal language model fine-tuning for text classification*. Em *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 328–339, 2018.

[100] Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes e Donald Brown: *Text classification algorithms: A survey*. Information, 10(4), 2019, ISSN 2078-2489. https://www.mdpi.com/2078-2489/10/4/150.

[101] Sabbah, Thabit, Ali Selamat, Md Hafiz Selamat, Fawaz S Al-Anzi, Enrique Herrera Viedma, Ondrej Krejcar e Hamido Fujita: *Modified frequency-based term weighting schemes for text classification*. Applied Soft Computing, 58:193–206, 2017.

[102] Forman, George e Martin Scholz: *Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement abstract*. SIGKDD Explorations, 12:49–57, janeiro 2010.

[103] Souza, Fábio, Rodrigo Nogueira e Roberto Lotufo: *BERTimbau: pretrained BERT models for Brazilian Portuguese*. Em *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, página 0, 2020.

[104] Merity, Stephen, Nitish Shirish Keskar e Richard Socher: *Regularizing and optimizing LSTM language models*. Em *International Conference on Learning Representations*, página 0, 2018. https://openreview.net/forum?id=SyyGPP0TZ.

[105] Bradbury, James, Stephen Merity, Caiming Xiong e Richard Socher: *Quasi-Recurrent Neural Networks*. International Conference on Learning Representations (ICLR 2017), 2017.

[106] Westermann, Hannes, Jaromir Savelka e Karim Benyekhlef: *Paragraph similarity scoring and fine-tuned bert for legal information retrieval and entailment*. Em Okazaki, Naoaki, Katsutoshi Yada, Ken Satoh e Koji Mineshima (editores): *New Frontiers in Artificial Intelligence*, páginas 269–285, Cham, 2021. Springer International Publishing.

[107] Legislation Portal, Planalto: *Decree no. 10.139, of november 28, 2019*, 2019. http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/decreto/D10139.htm, acesso em 2021-02-05.

[108] Castro, Marcus de: *New legal approaches to policy reform in Brazil*. University of Brasília Law Journal, vol. 1, june 2014, 1, janeiro 2014.

[109] OECD: *Regulatory performance: Ex-post evaluation of regulatory policies.* Proceedings from OECD expert meeting., 2003.

[110] Mulligan, Casey e Andrei Shleifer: *The extent of the market and the supply of regulation*. Quarterly Journal of Economics, 120(4):1445–1473, 2005.

[111] McLaughlin, Patrick: *Regdata canada: A data-driven approach to regulatory reform*. Relatório Técnico, George Mason University, Mercatus Center, 2019.

[112] Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane e Graeme Hirst: *Measuring emotion in parliamentary debates with automated textual analysis*. PLOS ONE, 11:1–18, dezembro 2016. https://doi.org/10.1371/journal.pone.0168843.

[113] Oreiro, José Luis, Luciano Luiz Manarin e Paulo Gala: *Deindustrialization, economic complexity and exchange rate overvaluation: the case of Brazil (1998-2017)*. PSL Quarterly Review, 73(295):313–341, 2020.

[114] Dickey, David A. e Wayne A. Fuller: *Distribution of the estimators for autoregressive time series with a unit root*. Journal of the American Statistical Association, 74(366a):427–431, 1979. https://doi.org/10.1080/01621459.1979.10482531.

[115] Kwiatkowski, Denis, Peter C.B. Phillips, Peter Schmidt e Yongcheol Shin: *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?* Journal of Econometrics, 54(1):159–178, 1992, ISSN 0304-4076. https://www.sciencedirect.com/science/article/pii/030440769290104Y.

[116] Kmenta, Jan: *E. kocenda and a. cerný, elements of time series econometrics: An applied approach, karolinum press, charles university, prague (2007) isbn 978-80-246-1370-3 (228 pp)*. Economic Systems, 33(2):185–187, 2009. https://EconPapers.repec.org/RePEc:eee:ecosys:v:33:y:2009:i:2:p:185-187.

[117] Seabold, Skipper e Josef Perktold: *Statsmodels: Econometric and statistical modeling with python*. Proceedings of the 9th Python in Science Conference, 2010, janeiro 2010.

[118] Jardini, J.A., Dorel Ramos, J. Martini, L. Reis e C. Tahan: *Brazilian energy crisis*. Power Engineering Review, IEEE, 22:21 – 24, maio 2002.

[119] Averbug, André: *The brazilian economy in 1994–1999: from the real plan to inflation targets*. World Economy, 25(7):925–944, 2002.

[120] Genoe, Alexander, Ronald Rousseau e Sandra Rousseau: *Applying google trends' search popularity indicator to professional cycling*. Journal of Sports Economics, 22:152700252098832, janeiro 2021.

[121] Preis, Tobias, Helen Moat e H. Stanley: *Quantifying trading behavior in financial markets using google trends*. Scientific Reports, 3:1684, abril 2013.

[122] Siliverstovs, Boriss e Daniel S. Wochner: *Google trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from swiss tourism regions*. Journal of Economic Behavior & Organization, 145:1–23, 2018, ISSN 0167-2681. https://www.sciencedirect.com/science/article/pii/S0167268117302937.

[123] Seifter, Ari, Alison Schwarzwalder, Kate Geis e John Aucott: *The utility of "google trends" for epidemiological research: Lyme disease as an example*. Geospatial Health, 4(2):135–137, May 2010. https://geospatialhealth.net/index.php/gh/article/view/195.

[124] trends.google.com: *Google trends*, 2012. http://trends.google.com/trends.

[125] Shannon, Claude Elwood: *A mathematical theory of communication*. ACM SIGMOBILE mobile computing and communications review, 5(1):3–55, 2001.

[126] Schölkopf, Bernhard, Alexander Smola e Klaus Robert Müller: *Kernel principal component analysis*. Em *International conference on artificial neural networks*, páginas 583–588. Springer, 1997.

[127] André Ramon Silva Martins, Carmen Silvia Sanches e Thelma Maria Melo Pinheiro: *Iniciativas para a institucionalização do uso de evidências no processo regulatório na aneel - um estudo de caso de agência reguladora.* Políticas Públicas e o uso de evidência no Brasil: conceitos, métodos, contextos e práticas - IPEA, 2022.