










Natural Language Processing Applied in the Context of Economic Defense: A Case Study in a Brazilian Federal Public Administration Agency

Vanessa Coelho Ribeiro¹^a, Jeanne Louize Emygdio¹^b, Guilherme Pereira Paiva¹^c,
Bruno Justino Garcia Praciano¹^d, Valério Aymoré Martins¹^e, Edna Dias Canedo^{1,2}^f,
Fábio Lúcio Lopes Mendonça¹^g, Rafael Timóteo de Sousa Júnior¹^h
and Ricardo Staciarini Puttini¹ⁱ

¹National Science and Technology Institute on Cyber Security, Electrical Engineering Department, University of Brasília (UnB), P.O. Box 4466, Brasília DF, Brazil

²Department of Computer Science, University of Brasília (UnB), P.O. Box 4466, Brasília DF, Brazil

Keywords: Natural Language Processing, Artificial Intelligence, Public Administration Agency, Jurisprudence, Antitrust.

Abstract: Natural Language Processing (NLP) and Machine Learning (ML) resources can be used in Jurisprudence to deal more accurately with the large volume of documents and data in this context to provide speed to the execution of processes and greater accuracy to judicial decisions. This article aims to present applied research with a qualitative approach and exploratory objective, technically characterized as a case study. The research was conducted in a Brazilian federal public administration agency to verify the existence of antitrust practices in the pharmaceutical field and the monitoring of such practices by the institution. To this end, a methodological path was established based on three stages: building the corpus, running the NLP pipeline and consultation of the results in the Jurisprudence Search System (BJ System). In compliance with the objective of the case study, it was possible to identify the performance of the agency around the domain elicited, as well as indications of the existence of antitrust practices, since the 276 documents retrieved from the BJ system relate directly to routine processes executed by the agency, either in the sense of investigation, trial or analysis of the business practices.


1 INTRODUCTION


Natural Language Processing (NLP) ((ISO) and (IEC), 2022); ((ISO) and (IEC), 2021) comprises a branch of studies that originate from the articulation of theories, methods, and technologies fundamentally derived from Computer Science, Artificial Intelligence, and Linguistics to establish effective communication between humans and machines employing natural language.

In the last few decades, various initiatives have been implemented or are under development in Jurisprudence Search, using techniques from text mining, machine learning, natural language processing, and neural networks (Loutsaris and Charalabidis, 2020).


In the legal area, it is a fact that there are massive volumes of collections of documents that demand case-by-case reading for decision-making, which represents an extensive period for the punctual solution of each case and its correlation to similar cases.


The application of NLP techniques allied to ML models tends to offer celerity in the processing of this corpus, consequently optimizing the identification of subsidies for the treatment of judicial processes (Dias Canedo et al., 2021; Alrumayyan and Al-Yahya, 2022), such as identification of similar opinions that may guide and link similar cases (Loutsaris and Charalabidis, 2020), summarization of legal texts

^a <https://orcid.org/0000-0003-1070-9403>


^b <https://orcid.org/0000-0002-7329-4447>


^c <https://orcid.org/0000-0001-8978-139X>


^d <https://orcid.org/0000-0002-7423-6695>

^e <https://orcid.org/0000-0003-1070-9403>

^f <https://orcid.org/0000-0002-2159-339X>

^g <https://orcid.org/0000-0001-7100-7304>

^h <https://orcid.org/0000-0001-7100-7304>

ⁱ <https://orcid.org/0000-0001-6433-1587>

of great complexity and high volume (Finegan-Dollak and Radev, 2016); sorting, reading, understanding arguments in summaries, evaluating evidence, applying laws, identifying relevant cases, and drafting decisions for analysis by legal experts (Park and Ko, 2020); extracting effect sentences from legal cases for optimizing jurisprudence search in document collections (Mandal et al., 2021).

Besides the necessary NLP techniques, in the case of this article, its application is in a jurisprudence search system.

Jurisprudence search systems have access to documentary collections and search for similarities in sources of judicial decisions. The result is a set of similar legal situations that can serve as a basis for various legal activities. Summarization of cases is, therefore, essential.

The objective of this paper is to present a case study of jurisprudence search by a Brazilian federal public administration agency obtained from the adoption of NLP methods. It reviews and analyzes the concepts and progress of NLP, analyzes the NLP pipeline created for the system in focus, and presents the stage of development in the Jurisprudence Search system.

2 NATURAL LANGUAGE PROCESSING

NLP comprises an interdisciplinary study area involving Computer Science, Linguistics, Statistics, Logic, and Philosophy, among many others. Its studies date back to the '50s, specifically, Warren Weaver's (Weaver1949) research during the conception of a project of automatic translation of documents. This project, inspired by the work of Alan Turing (Turing, 1950), focused on developing similar methods for translating documents between different languages (Somers, 2012).

Weaver's findings have stimulated the evolution of research in NLP, basically under two types of approaches: i) rule-based: applied to the development of robust systems that require extensive individual effort from linguists in their construction, although they are of simplified maintenance, based on modifications in the translation rules and; ii) statistics-based approach: applied to the development of systems in a shorter period after the collection and cleaning of bilingual data, but complex for adjustments after the start of operation. It is the dominant approach among researchers in the area (Somers, 2012).

Nowadays, the application of NLP techniques relies heavily on textual interpretation due to the wide availability of digital information in this format. Text-

tual interpretation systems help retrieve, categorize, filter, and extract information from texts and are typified as information retrieval systems, textual categorization systems, and data extraction systems (Russell et al., 2010; Loutsaris and Charalabidis, 2020).

2.1 Classical Approaches to NLP

Classical approaches to NLP comprise a set of stages in which the language analysis process is decomposed according to the theoretical linguistic distinctions drawn between syntax, semantics, and pragmatics (Dale, 2010).

It seeks, through analysis and practical actions, to make a computer able to perform six stages of understanding communication: i) phonology: the study of the sounds that make up words; ii) morphological analysis (tokenization): fragmentation of an input text to determine its components, the words, punctuations, numbers, and signs (Palmer, 2010); iii) lexical analysis (lemmatization): relation of morphological variants to their lemmas, canonical forms or form in which they are found in dictionaries, and their meanings (Hippisley, 2010); iv) syntactic analysis: evaluation of the grammar of the language used and representation of the analyzed sentence (parsing) in the form of a grammar; v) semantic analysis: extraction of the meaning of a statement and its representation in a semantic network, and vi) pragmatic analysis: discourse processing for intentionality analysis (Murphy, 2003; Dale, 2010).

2.2 Empirical and Statistical Approaches to NLP

In the scope of empirical and statistical approaches, NLP is used to decide the meaning of a word, its category, its syntactic structure, and the semantic scope around it. Thus, various models and techniques are adopted for this purpose. Statistical models are heavily used for building machine learning systems. Among the main ones are: i) artificial neural networks (ANN): computational systems inspired by biological neural networks able to learn to perform tasks from examples; ii) decision trees: predictive models run over a vector of input values to return a unique output value; iii) support-vector machine (SVM): framework of methods for supervised learning, used for classification and regression and; iv) Bayesian networks: a graphical model of the probability distribution related to a set of variables within the universe of a problem (Mitchell, 1997; Russell et al., 2010).

The techniques generally adopted are: i) word sense disambiguation (WSD): the computational

identification of the meaning of words in context (Navigli, 2009); ii) corpora creation: collections of texts used for learning linguistic models (Xiao, 2008), iii) part-of-speech (POS) tagging: the process of tagging each word in a given sentence with its correct part of speech (Güngör, 2010); iv) treebank annotation: corpora that present tree-structured annotations (graph theory) representing syntactic, semantic, and intersentential relationships (Hajičová et al., 2010) and; v) alignment: automatic parallel text alignment for translation validation purposes (Wu, 2010).

2.3 Related Work

From the linguistic perspective, (Wang, 2019; Jiang and Lu, 2020) states that language comprises the following linguistic levels: phonetics, lexicon, grammar, semantics, discourse, and pragmatics. For the language studies cited, NLP applications can be subdivided into these sections: machine translation, sound recognition, sound synthesis, automatic information retrieval, term database, optical character recognition, human-machine dialogue, and others.

(Loutsaris and Charalabidis, 2020) presents among the possibilities of NLP to assign predefined category labels to new documents, understand the meaning of natural language, and label a word in a sentence or phrase to its appropriate part of speech type.

(Kumar et al., 2022) emphasizes the importance of Natural language understanding (NLU) in understanding human communication because, in textual documents, the annotations used for machine learning are punctual. In real-world communication, because of interactions, the frequency of annotations is significant as marking part of speech, generating sentences, or answering questions. We used frequency-enriched datasets to compare the performance of (IC-NER)) and proposed two changes in domain generalization approaches: domain masks for generalization (DMG) and optimal transport (OT).

The applications of statistical techniques and machine learning are quite diverse. In his research on the topic of neuroscience, (Sarmashghi et al., 2022) presents a study of neural coding using existing Machine Learning (ML) approaches, particularly deep network architectures, and the methods to integrate them with statistical models. For both the simulation and real data analyses, 70% of data were devoted to training, 10% to validation, and 20% to testing with the use of mini-lot gradient descent (GD) as the learning algorithm to update the model parameters. The research demonstrates that the classical statistical methods and supervised machine learning algorithms have

complementary strengths and can be used together to address the limitations of each method on their own.

(Finegan-Dollak and Radev, 2016) presents the use of sentence simplification, compression, and disaggregation for summarization applied to creating sophisticated document summaries in the legal and medical fields. The proposal is to have shorter sentences of the original document reducing the size by about 20%. Due to the texts' complexity, the results were not satisfactory, demonstrating that the techniques applied need to be improved for the areas in question.

(Park and Ko, 2020) presents the use of Machine learning (ML) in the Legal and Economic area in the Chinese context, based on regression modeling for testing legal models. The authors apply three ML models: Train-Test Cycle, Regularization, and Cross-Validation to the Logit model. The authors state that although NLP is reliably applied in the legal field for classification, reading and understanding arguments in briefs, evaluating evidence, applying relevant laws and cases to a factual situation, and drafting a decision. However, it has not yet reached a maturity that allows it to replace lawyers' cognitive power and legal reasoning skills.

Reading a summary of legal cases speeds up the attorney's work in searching for jurisprudence. (Mandal et al., 2021) presents a neural sequence tagging model for extracting catchphrase from legal cases of Supreme Court of India. Cross validation approach was used to train and evaluate all supervised methods. For identification of catchphrase was identified by scoring candidate sentences, modeling the task as a sequence labeling task, use of document context information with sequence markers. As a result, the authors identified that generic extraction methods do not work well in extracting from legal documents, that including the document context improves the performance of the extraction model, and that the variation using noun-phrases outperforms the two variations using n-grams.

3 METHODOLOGY

The present research is characterized, from the point of view of its nature, as applied research; from the point of view of the way of approaching the problem, as qualitative research; from the point of view of the objectives, as exploratory research and; from the point of view of the technical procedures, as a case study considering that, for exemplification, the search for jurisprudence in processes of economic defense in the pharmaceutical sector will be presented (Gil, 1989).

The objective of the case study is to verify the existence of antitrust practices in the elicited domain and its relation to the activities performed by a Brazilian federal public administration, focused on the investigation, judgment, and analysis of such practices.

To this end, a methodological path was established based on three stages:

- i. Building the corpus;
- ii. Running the NLP pipeline;
- iii. Consultation of the results in the Jurisprudence Search System (BJ System).

Figure 1 illustrates the architecture of the system.



Figure 1: Jurisprudence search system architecture.

A description of these steps are presented in the following subsections.

3.1 Construction of the Documentary Corpus

This step foresees the identification and organization of relevant documents for the construction of a corpus to be submitted to NLP and ML methods to meet the objectives of the case study. Relevant documents are those produced and maintained by the Brazilian federal public administration agency to register decisions, technical notes, opinions, and others related to the institution’s performance in the prevention, judgment, and analysis of antitrust practices. The documents must be available for consultation in electronic format.

3.2 NLP Pipeline

Figure 2 contains the three steps of the NLP Pipeline: Cleaning, pre-processing and modeling.

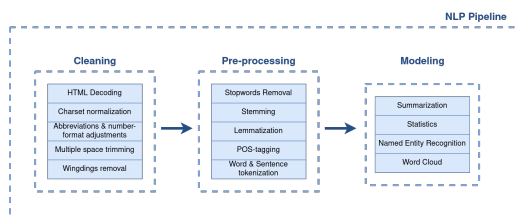


Figure 2: NLP Pipeline.

The cleaning step comprises the application of techniques regarding text cleaning and normalization,

removing abbreviations and normalizing numeric formats, and removing inappropriate characters originating from HTML texts.

Text pre-processing is a fundamental step for NLP, and statistics for correct data loading are applied. This step comprises a few steps:

- Stopwords Removal: Stopwords are words that help to understand the meaning of a sentence but that do not carry in themselves any significance. Words like “a”, “que”, “em” are present in the Portuguese stopwords lists and are removed in preprocessing because they have a high occurrence and do not add to the meaning of the text.
- Stemming and Lemmatization: this techniques are applied to normalize words by removing their inflections(Hippisley, 2010).
- POS-Tagging: Part-of-Speech (POS) tagging is an NLP process that categorizes words from their grammatical class(Güngör, 2010).
- Tokenization: in the legal context, due to specificities in the text, such as abbreviations, Roman numerals, and article and legislation citations, conventional tokenizers have their performance affected. Therefore, the construction of specific tokenizers is planned, taking into account the characteristics of the text of the local government’s documents.

The modeling step consists of statistical analysis of the text and application of the text summarization, Named Entity Recognition, and WordCloud models. (Alshammari and Alanazi, 2020)

4 RESULTS AND DISCUSSIONS

Text summarization is a process that generates a document summary by identifying its most important sentences. It was implemented from the *ensemble* of a set of summarization techniques (Luhn, 1958; Haghghi and Vanderwende, 2009), where the output of these models are combined with choosing the most relevant sentences from the document. Ensemble learning, unlike other methods, selects a set of hypotheses from the hypothesis space, combines their predictions, and reduces the correlation between possible errors in hypothesis classification (Russell et al., 2010).

Most studies employ one of four summarization architectures: Sentence Extraction and Summarization; Feature Extraction and Classification or Classification-based Sentence Selection; Abstract Sentence Compression and Compression; and Language Modeling (Rezazadegan et al., 2022).

Named Entity Recognition is an NLP task that identifies and categorizes real-world entities present in (Grishman and Sundheim, 1995) texts. The names of people, organization names, places, citations to laws, and other documents are identified. A *corpus* was built for training, and a Deep Learning model specific to the documents of the municipality was trained.

Word Cloud is a visual representation of text, where keywords are highlighted from their frequency in the *corpus*. This visualization is implemented in the system for each document only, but for the context of this paper, visualization across multiple documents was implemented using Word Cloud (Mueller et al., 2018) software.

Statistical methods are widely applied in NLP. In the specific case of BJ, the feature extraction step for building summarization are used calculation of Distributions (relative and absolute frequency together with IDF) (Navigli, 2009), cutting processes in FreqDistrib using elbow techniques (Shi et al., 2021) and TF Prioritization techniques (pre-calculations of IDF marks) (Rahmah et al., 2019).

To conduct this case study, two versions of the Jurisprudence Search (BJ) System were used to achieve specific results, being:

Version 1.0 (production environment):

- Construction of the documentary corpus of jurisprudence;
- Retrieval of the corpus necessary for the intended scope of this research.

Version 1.2 (development environment):

- Pipeline execution.

The results obtained in each step of the proposed methodology is shown in the following subsections.

4.1 Construction of the Documentary Corpus

The corpus was built using BJ v1 (Dias Canedo et al., 2021), which can provide advanced search filters, with the option of conditionals, search with specific characters/terms, by proximity or Boolean operators, search by relevance, phonetic search with spell checker and autosuggestion. As a search result, the system presents resources for word highlighting, paging and sorting, controlled vocabulary synonyms, term-stopwords definition, and document standardization. In addition, the system allows the indexing of various file extensions, such as PDF with OCR.

For the case study, we established a cutout around the pharmaceutical industry. The keywords “medica-

ments”, “pharmaceutical”, “medicines”, and the logical operators available in the system were used in the BJ system search in the filter resource. The addition of the term “drugs” was evaluated in the search, but since there was no relevant impact on the search results since the results referred to the context of illicit drugs, the term was discarded.

The system finds documents related to veterinary medicines from the search with the chosen keywords. To adjust for these cases, the logical operator “NOT” was used to exclude the words “animal” and “veterinarians” from the search, resulting in the search `(pharmaceutical* OR medicine* OR medicament*) NOT (veterinarian* OR animal*)`.

For a demonstration of the pipeline results, the document number SEI 1090146 was taken as a base, where the results of the Summarization, and NER models are available in the development environment, Figures 3 and 4 respectively.

4.2 Running the NLP Pipeline and Querying the Results in the BJ System

Following all the steps described in the 3.2 section of the methodology, the pipeline running process occurs transparently within the BJ System. The results are stored in databases and made available for query by the BJ system through an API.

4.2.1 Results after Data Cleaning

For the data cleaning step provided in the pipeline, the BJ System removes stopwords and punctuation.

The BJ system also implements in the cleaning step the removal of abbreviations by replacing them with their corresponding fully spelled ones.

The removal of plurals is implemented in a specific way in the BJ system for handling exceptions not handled by commonly adopted Python libraries.

4.2.2 Results after Pre-Processing

For the pre-processing step, foreseen in the pipeline, the BJ System implements the lemmatization process in a step called “morphosyntactic tagging.”

Another implementation, also performed in this step, refers to the segmentation of representative sentences of the semantic set to define propositions.

4.2.3 Results after Modeling

The query about the pipeline results started on 06/01/2022 and returned 535 documents from the

jurisprudence of the Federal Government’s Departments. Of these, 276 were issued in the last five years.

The result as follows: 62 documents expedited in the year 2018, 61 documents expedited in the year 2019, 80 documents expedited in the year 2020, 67 documents expedited in the year 2021, and 6 documents expedited in the year 2022.

The organizes and distribution of process categories as three are characterized as merger reviewers, 70 are characterized as ordinary mergers, 140 are characterized as summary mergers, one is characterized as consultation, 21 are characterized as administrative inquiries, six are characterized as preparatory proceedings, 32 are characterized as administrative proceedings, two are characterized as voluntary appeals, and one is characterized as cease-and-desist application.

The results related to summarization are customizable according to the number of sentences or percentage of the text informed by the system user. This parameter is sent to the API, which selects the most relevant fragments from the quantity informed. Figure 3 illustrates the results of selecting the most relevant sentences.

Hypera operates in manufacturing allopathic drugs for human use, as does BI.

As previously informed, this Transaction is limited to acquiring the production capacity of an industrial plant currently owned by BI. It does not involve acquiring drugs, intellectual property, or distribution assets. Thus, the Applicants emphasize that Hypera and BI will continue to act as independent competitors in the drug market.

It should be noted that Hypera acquired the rights to the Buscopan brand in Brazil [8] from the BI Group in 2019 in a transaction approved by the Agency in 2020 [9].

According to the Applicants, the plant acquired in this Operation is responsible for producing the drug Buscopan in Brazil for supply to Hypera itself, as established in the acquisition of the Buscopan brand. According to the Applicants, the production of Buscopan still occurs in BI's plant solely for compliance with regulatory steps common to the pharmaceutical industry. In this context, Hypera intends to internalize the production of Buscopan since, currently, the BI Group continues to manufacture and supply both the active ingredient and the ready-made drug to Hypera.

The figure below summarizes the production chain of Buscopan currently:

Source: Petitioners.

The Applicants emphasized that this Transaction boils down to the consolidation of the production of the drug Buscopan in Hypera, which already holds the ownership of Buscopan in Brazil.

Recently, the Hypera Group acquired Solana Agro Pecuaría Ltda from the BI Group, responsible for duboisia production, the primary commercial source of scopolamine (Buscopan's active ingredient) [10]. The Applicants emphasized that the present Operation will not result in vertical integration between the plant acquired for the production of Buscopan and duboisia since the active ingredient scopolamine is an intermediate step in the production chain and continues to be performed only by the BI Group. Hypera clarified that, currently, its economic group does not operate in the manufacturing of active principles.

Figure 3: Summarization output from the BJ System.

Figure 4 contains the results related to Named Entity Recognition (NER), where the entities recognized in the texts are highlighted in different colors, using: red for locations, green for organizations, gray for values, and yellow for jurisprudence regulation documents. The entities identified are stored in the database by the model during the execution of the pipeline. At the moment of the user’s request in the BJ system, they are retrieved and presented visually on the screen.

In the context of economic defense, NER is a tool with great potential since it is relevant to the recovery of entities and the discovery of knowledge, emphasizing organizations mentioned in legal documents. Despite the great value in information retrieval provided by NER, the names of organizations

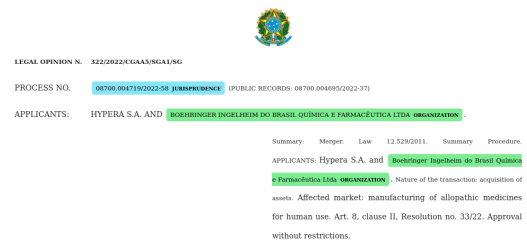


Figure 4: Named Entities recognized by the BJ System.

by themselves need to add more value to the legal context. Therefore, by integrating the retrieval of entity names with the knowledge discovery made possible by the APIs of knowledge bases such as WikiData and the Agency’s database, it is possible to extract, besides the names of the entities, information such as The Brazil National Registry of Legal entities Number (CNPJ), Corporate Name, Organizational Structure, the National Classification of Economic Activities (CNAE), among other types of data relevant to Agency’s target audience.



Figure 5: Word Cloud generated by the system.

Figure 5 shows that besides the keywords referring to the business areas of the Brazilian federal public administration agency, there were also found those that refer to the cut proposed in this article, highlighting: health insurance, hospital, medicine, medicine distribution, patent.

5 FINAL REMARKS

The objective of this paper was to verify the existence of antitrust practices in the pharmaceutical field and their relation with the activities performed by the Brazilian federal public administration agency. Us-

ing the conception and execution of a mixed methodology, encompassing NLP techniques and contemporary machine learning models, articulated in a technological architecture to support a jurisprudence search system under development (Dias Canedo et al., 2021).

Section 2 presents background on NLP, covering the classical, empirical, and statistical approaches, methods, and techniques found in the literature, accompanied by practical examples of its application in related research identified in the last five years. Section 3 presented the three-stage methodology designed for the research, covering the construction of the document corpus, the execution of the NLP pipeline, and the query of results in the BJ system. Section 4 presents the results obtained during the execution of the methodology and preliminary discussions about them.

In compliance with the objective of the case study, it was possible to identify the performance of the agency around the domain elicited, as well as indications of the existence of antitrust practices, since the 276 documents retrieved from the BJ system relate directly to routine processes executed by the agency, either in the sense of investigation, trial or analysis of the business practices. Details about this processes known to merger review, ordinary merger, summary merger, consultation, administrative inquiry, preparatory proceeding, administrative proceeding, voluntary appeals and cease-and-desist application could be found at (BRASIL, Ministério da Justiça e Segurança Pública. Conselho Administrativo de Defesa Econômica, 2021; CADE, 2021; Brasil, 2011).

Given the exploratory nature of the research described in this paper, the content analysis of the recovered documents is the object of a future publication. Using it for a better understanding of the flow of processes in progress in the agency and the relationship that the documents establish between themselves since they can characterize progressive outputs of the processes and sub-processes performed by the organization.

One of the differentials of this project rests on the construction of a domain ontology to support the disambiguation of terms and consequently to optimize ML processing in the BJ System.

ACKNOWLEDGMENTS

This work is supported in part by CNPq - Brazilian National Research Council (Grant 310941/2022-9 PQ-1D), in part by FAPDF - Brazilian Federal District Research Support Foundation (Grant 625/2022

SISTeR City), in part by the University of Brasilia (Grant 7129 UnB COPEI), in part by the General Attorney of the Union (Grant AGU 697.935/2019), in part by the Administrative Council for Economic Defense (Grant CADE 08700.000047/2019-14), and in part by the General Attorney's Office for the National Treasury (Grant PGFN 23106.148934/2019-67).

REFERENCES

- Alrumayyan, N. and Al-Yahya, M. (2022). Neural embeddings for the elicitation of jurisprudence principles: The case of arabic legal texts. *Applied Sciences*, 12(9).
- Alshammari, N. and Alanazi, S. (2020). An arabic dataset for disease named entity recognition with multi-annotation schemes. *Data*, 5(3).
- Brasil (2011). Lei nº 12.529, de 30 de novembro de 2011. *Diário Oficial da República Federativa do Brasil*.
- BRASIL, Ministério da Justiça e Segurança Pública. Conselho Administrativo de Defesa Econômica (2021). *CADE MECUM: Coletânea de normativos brasileiros de defesa da concorrência*. Conselho Administrativo de Defesa Econômica, Brasília: CADE. CDD 341.3787.
- CADE, C. A. d. D. E. (2021). *Regimento interno CADE*. CADE, Brasília, 5a ed. edition.
- Dale, R. (2010). Classical approaches to natural language processing. In *Handbook of Natural Language Processing, Second Edition*, pages 3–8. CRC Press - Taylor and Francis Group, New York, NY, USA.
- Dias Canedo, E., Aymoré Martins, V., Coelho Ribeiro, V., dos Reis, V. E., Carvalho Chaves, L. A., Machado Gravina, R., Alberto Moreira Dias, F., Lopes de Mendonça, F. L., Orozco, A. L. S., Balañuk, R., and de Sousa, R. T. (2021). Development and evaluation of an intelligence and learning system in jurisprudence text mining in the field of competition defense. *Applied Sciences*, 11(23).
- Finegan-Dollak, C. and Radev, D. R. (2016). Sentence simplification, compression, and disaggregation for summarization of sophisticated documents. *Journal of the Association for Information Science and Technology*, 67(10):2437–2453.
- Gil, A. C. (1989). *Métodos e Técnicas de Pesquisa Social*. Atlas, São Paulo, 2nd edition.
- Grishman, R. and Sundheim, B. (1995). Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.
- Güngör, T. (2010). Part-of-speech tagging. In *Handbook of Natural Language Processing, Second Edition*, pages 205–236. CRC Press - Taylor and Francis Group, New York, NY, USA.
- Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American*

- Chapter of the Association for Computational Linguistics*, pages 362–370. Boulder, Colorado. Association for Computational Linguistics.
- Hajičová, E., Abeillé, A., Hajič, J., Mírovský, J., and Uresová, Z. (2010). Treebank annotation. In *Handbook of Natural Language Processing, Second Edition*, pages 167–188. CRC Press - Taylor and Francis Group, New York, NY, USA.
- Hippisley, A. (2010). Lexical Analysis. In *Handbook of Natural Language Processing*, pages 31–58. Nitin Indurkha and Fred J. Damerau, New York: Chapman & Hall /CRC Press, 2nd. edition.
- (ISO), I. O. F. S. and (IEC), I. E. C. (2021). ISO/IEC TR 24030:2021(en), Information technology — Artificial intelligence (AI) — Use cases.
- (ISO), I. O. F. S. and (IEC), I. E. C. (2022). ISO/IEC 22989:2022(en), Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.
- Jiang, K. and Lu, X. (2020). Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review. In *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, pages 210–214, Chongqing City, China. IEEE.
- Kumar, M., Rumshisky, A., and Gupta, R. (2022). Chasing the tail with domain generalization: A case study on frequency-enriched datasets. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–11, Online only. Association for Computational Linguistics.
- Loutsaris, M. A. and Charalabidis, Y. (2020). Legal informatics from the aspect of interoperability: a review of systems, tools and ontologies. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, page 731–737, Athens Greece. ACM.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.
- Mandal, A., Ghosh, K., Ghosh, S., and Mandal, S. (2021). A sequence labeling model for catchphrase identification from legal case documents. *Artificial Intelligence and Law*, 30.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, New York.
- Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R., Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I., vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong, L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). amueller/word.cloud: Wordcloud 1.5.0.
- Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press.
- Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.
- Palmer, D. (2010). Text preprocessing. In *Handbook of Natural Language Processing*, page 22. Nitin Indurkha and Fred J. Damerau, New York, NY, USA, 2nd. edition.
- Park, S. and Ko, H. (2020). Machine learning and law and economics: A preliminary overview. *Asian Journal of Law and Economics*, 11(2):20200034.
- Rahmah, A., Santoso, H. B., and Hasibuan, Z. A. (2019). Exploring technology-enhanced learning key terms using tf-idf weighting. In *2019 Fourth International Conference on Informatics and Computing (ICIC)*, pages 1–4.
- Rezazadegan, D., Berkovsky, S., Quiroz, J. C., Kocaballi, A. B., Wang, Y., Laranjo, L., and Coiera, E. W. (2022). Symbolic and statistical learning approaches to speech summarization: A scoping review. *Comput. Speech Lang.*, 72:101305.
- Russell, S. J., Norvig, P., and Davis, E. (2010). *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, 3rd ed edition.
- Sarmashghi, M., Jadhav, S. P., and Eden, U. T. (2022). Integrating statistical and machine learning approaches for neural classification. *IEEE Access*, 10:119106–119118.
- Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., and Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J. Wirel. Commun. Netw.*, 2021(1):31.
- Somers, H. (2012). Machine Translation: History, Development, and Limitations. In *The Oxford Handbook of Translation Studies*, pages 1–9. Oxford University Press Inc., New York, NY, USA, kirsten malmkjær and kevin windle edition.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX(236):433–460. Number: 236.
- Wang, Y. (2019). Natural language processing and applications in machine learning. *Modern Chinese*, 5:187–191.
- Wu, D. (2010). Alignment. In *Handbook of Natural Language Processing, Second Edition*, pages 367–408. CRC Press - Taylor and Francis Group, New York, NY, USA.
- Xiao, R. (2008). Well-known and influential corpora. In *Corpus Linguistics: An International Handbook*, volume 1, pages 383–457. Mouton de Gruyter, Berlin, Germany, a. lüdeling and m. kyto edition.