

A Legal Framework for Natural Language Processing Model Training in Portugal

Rúben Almeida, Evelin Amorim

INESC TEC, FCUP-Universidade do Porto

ruben.f.almeida@inesctec.pt, evelin.f.amorim@inesctec.pt

Abstract

Recent advances in deep learning have promoted the advent of many computational systems capable of performing intelligent actions that, until then, were restricted to the human intellect. In the particular case of human languages, these advances allowed the introduction of applications like ChatGPT that are capable of generating coherent text without being explicitly programmed to do so. Instead, these models use large volumes of textual data to learn meaningful representations of human languages. Associated with these advances, concerns about copyright and data privacy infringements caused by these applications have emerged. Despite these concerns, the pace at which new natural language processing applications continued to be developed largely outperformed the introduction of new regulations. Today, communication barriers between legal experts and computer scientists motivate many unintentional legal infringements during the development of such applications. In this paper, a multidisciplinary team intends to bridge this communication gap and promote more compliant Portuguese NLP research by presenting a series of everyday NLP use cases, while highlighting the Portuguese legislation that may arise during its development.

Keywords: Portuguese NLP, Legal NLP, PLN Português

1. Introduction

In recent years, deep-learning-based methods have permitted great advances in Computer Science (CS) fields previously known for their computational complexity (Bishop and Bishop, 2023). One of those fields was natural language processing (NLP); the CS field focused on machine understanding of human languages and the generation of coherent human text (Deng and Liu, 2018).

Long before the introduction of state-of-the-art (SOTA) generative large language models (LLMs) like Chat-GPT (OpenAI, 2022), NLP relied on simpler approaches based on rules/word counts (Jurafsky and Martin, 2014) to deliver human-interpretable models capable of addressing simple NLP tasks. The evolution towards deep learning permitted NLP to quit the controlled atmosphere of university labs and embrace mainstream societal adoption at the expense of costly training processes and vast volumes of textual data.

The capabilities revealed by SOTA NLP models were accompanied by ethical and legal concerns among prominent NLP researchers¹, big-tech CEOs², politicians³, and economists⁴ who

appealed for regulation and higher ethical standards during the development of NLP solutions as a way of restricting the usage of such capabilities to induce harm in society.

However, the pace at which new LLMs are currently being developed largely surpasses the pace at which new regulations are introduced. This phenomenon, together with the traditional communication barriers between the formality of legal expertise and the dynamic world of CS, opens the door to legal vacuums, promoting undesirable copyright⁵ and privacy infringements⁶.

In this paper, we intend to bridge this gap by listing the legal concerns that may arise during everyday NLP challenges. We focus on the Portuguese legal system to present a study targeting both computer scientists and legal experts, with the ultimate goal of promoting awareness about the topic in one of the most peripherally and underdeveloped⁷ countries of the European Union (EU). Despite our focus on Portugal, we believe researchers in other EU countries may find the concepts introduced in this paper useful, due to the European nature of many of the legislation listed in this paper.

The paper is structured as follows: Section 2 de-

¹<https://twitter.com/ylecun/status/1733481002234679685>

²<https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/>

³<https://www.ft.com/content/9339d104-7b0c-42b8-9316-72226dd4e4c0>

⁴<https://www.weforum.org/agenda/2017/03/taxing-robots-wont-work-says-yanis-v>

aroufakis/

⁵<https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

⁶<https://www.forbes.com/sites/emmacott/2023/09/01/openai-hit-with-new-lawsuit-over-chatgpt-training-data/?sh=2f3d1b856d84>

⁷https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index

scribes the existing literature about the topic. The lack of research, focused exclusively on the Portuguese case, forced us to broaden our scope towards other EU countries. Section 3 provides a quick overview of the current SOTA for Portuguese NLP, with a special emphasis on explaining the importance of Brazilian NLP research for Portuguese NLP. In Section 4, we briefly introduce the Portuguese legal system, performing a high-level description of the legislation that may engage with the development of NLP solutions. Section 5 lists common licensing agreements used by NLP researchers to publish their works. Finally, in Section 6, we introduce some use cases representing everyday challenges faced by NLP researchers highlighting the legal concerns associated with them.

2. Related Work

The novelty of this subject translates into a lack of literature about the topic. The absence would be even worse if we focused exclusively on the Portuguese legal landscape. Therefore, in the following section, we broaden our scope outside Portugal towards other EU countries.

2.1. Portuguese

The existing literature concerning the legal implications of artificial intelligence (AI) systems in Portugal is constrained to regional conferences and journals in the Portuguese language. We highlight the pioneering works of (Guimarães et al., 2021) that compiled summarized versions of master theses developed by law students about different CS topics, including Big Data (Silva Costa, 2021).

Recently, (Barbosa, 2023) introduced the analysis concerning data privacy during LLM development. Despite focusing on the Portuguese legal framework, the writer relies heavily on EU regulations to provide a brief overview of the recent developments surrounding the topic. In particular, it is demonstrated that the General Data Protection Regulation (GDPR) alone is incapable of regulating LLMs like ChatGPT. The author mentions the importance of the AI Act to establish a legal framework founded on the concept of **risk** to deal with the inaccuracies of these models.

Focusing on copyright issues, we highlight the works of (Nobre, 2012), which provide an extensive listing of what is protected and what is excluded from copyright protection.

Lastly, it is worth mentioning the interesting considerations of (Pereira, 2020) regarding copyright eligibility for AI outputs. In Portugal, the copyrights for AI results are free, unlike in other countries like the United Kingdom, where the research team who

made them inherits the rights⁸.

2.2. European

At the European level, we highlight the works of (Eckart de Castilho et al., 2018) and (Kelli et al., 2020) focusing on copyright. Both works discuss whether an LLM trained in copyrighted protected data inherits the same license as its training data. Both authors concluded there is a lack of clarity on the topic, but it all depends on how much copyright data is used. If a model can reproduce data that is protected by copyright, it is called a derivation of the dataset and has the same copyright as the dataset used.

Nevertheless, most of the European literature about the topic is produced by enterprises. We highlight the efforts of ML6⁹, KPMG¹⁰ or EY¹¹ to promote the debate about the topic. These blog posts often have anonymous authors who are not checked by others. This makes it hard to know if they are accurate and up-to-date.

3. Portuguese NLP: Quick Overview

Portuguese is the official language of 260 million people spread across five continents. In the context of NLP, Portuguese is considered a mid-resourced language (Joshi et al., 2021). Meaning that “they have a large amount of unlabeled data...and are only challenged by lesser amount of labeled data” (Joshi et al., 2021). Despite this classification, the vast majority of these resources are Brazilian Portuguese, produced by Brazilian research teams.

Similar to other mid-resource languages, the shorter investment produces a reduced number of Portuguese LLMs. Until recently, BERTimbau (Souza et al., 2020), a Brazilian Portuguese version of BERT, was the only Portuguese LLM. Progressively, more complex architectures have emerged: Albertina PT (Rodrigues et al., 2023) or the generative models Sabiá (Pires et al., 2023), Gervásio (Santos et al., 2024), and Glória (Lopes et al., 2024).

⁸https://pec.ac.uk/blog_entries/copyright-protection-in-ai-generated-works/

⁹<https://www.ml6.eu/blogpost/navigation-ethical-considerations-developing-and-deploying-large-language-models-llms-responsibly>

¹⁰<https://kpmg.com/ie/en/home/insights/2024/01/eu-artificial-intelligence-act-art-int.html>

¹¹https://www.ey.com/en_ch/forensic-integrity-services/the-eu-ai-act-what-it-means-for-your-business

3.1. Leveraging Brazilian Portuguese Resources

Despite the major phonological, morphological, lexical, syntactical and semantic differences between the numerous Portuguese varieties around the globe (Scherre and Duarte, 2016), prompt-engineering¹² and fine-tuning¹³ of Brazilian LLMs help European Portuguese researchers achieve SOTA results (Almeida et al., 2024).

4. Portuguese Legal System

Portugal, as an EU member, must ensure its legal system is in harmony with EU law. In this section, we list the legislation that engages with NLP development, while describing how EU legislation impacts the Portuguese legal system.

4.1. Scientific Exceptions

Many of the legislation listed in the following subsections establish exceptions to scientific work. For that reason, we believe it is paramount to clarify what is considered **scientific work** by the Portuguese law.

The (European Parliament and Council of the European Union, 2019) directive introduces a comprehensive definition of scientific work founded on the concept of profit: "research organisation means a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State" (European Parliament and Council of the European Union, 2019).

4.2. National Legislation

Portuguese legislation is divided into two major sets of laws: the civil code and the penal code. The penal code lists those actions that constitute a crime and require the existence of **willful misconduct** to be taken into consideration. While in the civil code, the penalties restrained themselves to some kind of compensation, the penal code establishes tougher penalties like prison time. Committing a crime is a severe action that usually implies not only a penal condemnation but also a parallel civilian compensation to repair any harm provoked in society.

¹²https://en.wikipedia.org/wiki/Prompt_engineering

¹³[https://en.wikipedia.org/wiki/Fine-tuning_\(deep_learning\)](https://en.wikipedia.org/wiki/Fine-tuning_(deep_learning))

Regarding NLP, most of the legislation that applies is restricted to the civil code; however, in recent years, with the mass adoption of CS in society, many penal considerations have arisen focusing on copyright infringement, privacy violations, or cybercrime.

Finally, before enumerating the Portuguese legislation directly relevant to NLP, it is essential to clarify the distinctions between public, semi-public, and private crimes. Public crimes warrant the intervention of public prosecutors to safeguard the interests of the victim. Semi-public crimes trigger public prosecution only upon the filing of a complaint with the police. In cases of private crimes, typically of lesser severity, public prosecution does not intervene, and a complaint must also be lodged with the police. In instances where the penal code does not explicitly specify the classification of the crime, it defaults to being considered public¹⁴.

Sensitive Data Protection: Article 35 of the Portuguese constitution (República Portuguesa, 1976) strictly prohibits the digital processing of sensitive ethnic, political, sexual, or religious data without explicit consent and for scientifically motivated purposes that may positively impact society. Consequently, NLP researchers are advised to refrain from operating on such sensitive data.

Right to Expression: Article 79 of the Portuguese civil code (República Portuguesa, 1966) affirms the right of individuals to freely express themselves without fear of their words being recorded or utilized by a third party. This legislation outlines exceptions for scientific research. Without these provisions, the use of tweets to train NLP models would be deemed illegal.

4.3. European Legislation

The great investment the EU has been making to regulate AI has had a great effect in peripheral countries like Portugal, helping reduce inequalities in access to technology between richer and poorer countries within the EU.

Before enumerating EU regulations that engage directly with the development of NLP solutions, it is important to clarify some terminology regarding EU law:

Regulation: Regulations come into force automatically upon approval by the European Parliament and European Council. Requiring a 65% majority for acceptance, regulations are always accompanied by a transition period to allow EU citizens to adapt to the new legal framework.

¹⁴<https://www.ministeriopublico.pt/perguntas-frequentes/crime>

Directive: Directives represent EU laws that member states must adopt within a specified transition period, integrating them into their own legal systems. Despite penalties for countries failing to timely adopt directives, Portugal has earned a reputation for delays in transposing these directives. A pertinent example is the 2019/790 directive, where the transposition was delayed by over two years¹⁵.

Below, we briefly introduce the most relevant EU legislation that engages with NLP development:

General Data Protection (Regulation): Enacted in 2016, this EU regulation (European Parliament and Council of the European Union, 2016) sets the foundational standards for online data protection. Emphasizing transparency and consent, it mandates that all corporate operations involving personal data must be pre-informed and explicitly consented to by users. The regulation applies specifically to private data - information that, if disclosed, can identify the individual to whom it pertains. It also regulates cross-border data transfers, stipulating that EU data protection laws accompany the data wherever it goes. Outside the EU, in countries lacking an **adequacy decision**¹⁶, such as Brazil, additional measures may be necessary to ensure equivalent levels of protection as those mandated within the EU. This regulation was integrated into Portuguese legislation in 2021, complemented by the introduction of the **Portuguese Charter on Human Rights in the Digital Age** (Assembleia da República, 2021). Concerning NLP, explicit consent is required for data collection and usage in training NLP models, with researchers urged to minimize the use of private data whenever possible. While scientific research exceptions exist, the legislature advises their avoidance, favoring maximal consent.

Copyright in the Digital Single Market (Directive): The 2019/790 EU directive (European Parliament and Council of the European Union, 2019) on copyrights introduced a **right to [text] mine**, framing the legal context for NLP development. It broadened exceptions, permitting scientists to use copyrighted data for NLP model training if not pursued for profit. Transposed into Portuguese legislation in 2023 (Presidência do Conselho de Ministros, 2023), this directive provides a legal framework for NLP endeavors.

Database Sui Generis Protection (Directive): Directive 96/97EC (European Parliament and Council

of the European Union, 1996) established copyright protection for databases, granting 15 years of copyright protection to those who compile independent works into structured units. This regulation was revisited during the approval of the 2019/790 directive, with exceptions allowing copyright infringements for scientific endeavors also applying to databases.

AI Act (Regulation): Recently approved, this regulation (European Commission, 2021) aims to be the world’s first comprehensive AI legal framework, positioning the EU at the forefront of AI legislation. It sets rules for AI systems, including NLP, based on their societal risk levels. Additionally, it mandates transparency mechanisms during the training of LLMs exceeding 10^{25} FLOPS. Given that most NLP research does not involve sensitive data or yield automated decisions significantly impacting Europeans’ daily lives, it is categorized as **minimal-risk**, exempting researchers from additional considerations. The AI Act is in the final stages of approval and is expected to be fully implemented within two years.

5. NLP Licensing System

In Table 1, we list the licenses of many NLP resources.

Model Name	Year	License
BERT	2018	Apache 2.0
GPT-2	2019	MIT
Bloom	2022	Custom License
Falcon	2023	Apache 2.0
Llama 2	2023	Custom License
Mistral	2023	Apache 2.0
Phi-2	2023	MIT
Gemma	2024	Custom License
BERTimbau	2020	MIT
Albertina PT	2023	MIT
Sabiá	2023	Llama 1 License
Gervásio	2024	MIT
Glória	2024	ClueWeb22 License
Wikitext	2016	CC-BY-SA 3.0
CNN-Dailymail	2017	Apache 2.0
Flores	2019	CC-BY-4.0
OSCAR	2023	CC0-1.0

Table 1: Licenses for different SOTA NLP resources. The first entries cover non-Portuguese LLMs, while the second set focuses on Portuguese architecture. The last set of resources are commonly used NLP datasets.

The results reveal that many LLMs adopt an Apache 2.0 or MIT license. The datasets identified tend to adopt Common Crawl licenses. It is worth mentioning the case of the Portuguese LLM Glória (Lopes et al., 2024), where the usage of the

¹⁵<https://www.cuatrecasas.com/pt/portugal/art/transposicao-da-diretiva-do-mercado-unico-digital-1>

¹⁶https://www.edpb.europa.eu/sme-data-protection-guide/international-data-transfers_en

clueweb22 dataset (Overwijk et al., 2022) as part of the training corpus required Glória’s authors to adopt this license as well.

The information provided in this section is complemented by the extensive analysis made by the choose a license platform¹⁷.

6. Use Cases

In this section, we cover three use cases that represent everyday NLP tasks. In the following subsections, we assume the perspective of a Portuguese NLP operating under EU law. The green color used in the flowcharts represents the yes/permitted case; in contrast, the red color represents the no/not permitted case.

6.1. Load Brazilian Portuguese Dataset From HuggingFace

This use case describes a standard practice in Portuguese NLP, where Brazilian datasets are used to train new NLP models. In Figure 1 we provide a flowchart summarizing the legal questions that may arise during the process.

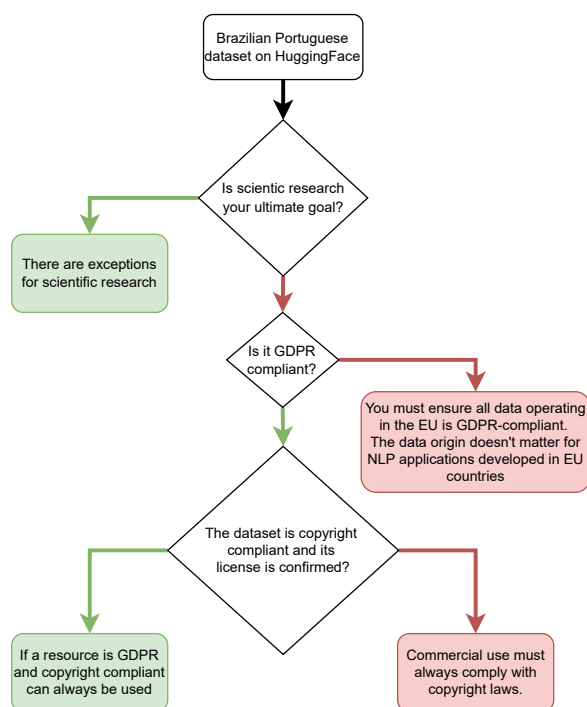


Figure 1: Flowchart summarizing the legal questions associated with the loading of a Non-EU dataset

It is worth mentioning that the geographical origin of the dataset has no impact on the overall legal assessment of the resource.

¹⁷<https://choosealicense.com/appendix/>

6.2. Crawl Portuguese Websites to Produce a Large NLP Corpus

The usage of web crawling techniques is paramount for LLM training. SOTA LLMs require a vast amount of data, whose scale is only comparable to the amount of information existing on the web. In Figure 2, we describe the legal considerations researchers should pay attention to while crawling websites.

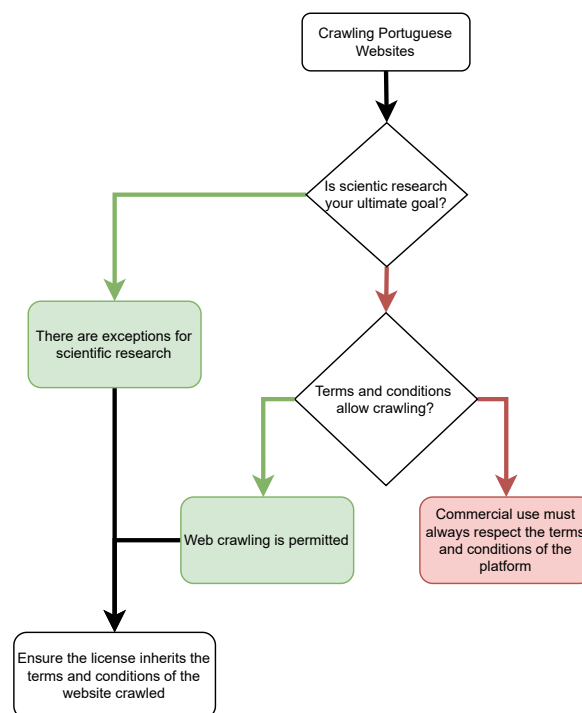


Figure 2: Flowchart highlighting the legal considerations web crawling may arise.

6.3. Use Tweets to Produce Political Profiles: The Facebook-Cambridge Analytica case

This use case draws inspiration from the Facebook–Cambridge Analytica data scandal¹⁸. It aims to encapsulate the legal considerations that may emerge during the development of NLP models, particularly in scenarios involving sensitive data such as political profiling. Figure 3 provides a summarized overview of these legal aspects.

7. Conclusion & Future Work

In this study, we have presented the inaugural legal framework for NLP development in Portugal. The limited awareness of this subject in one of the EU’s less affluent nations exposes vulnerabilities to

¹⁸<https://shorturl.at/uxK47>

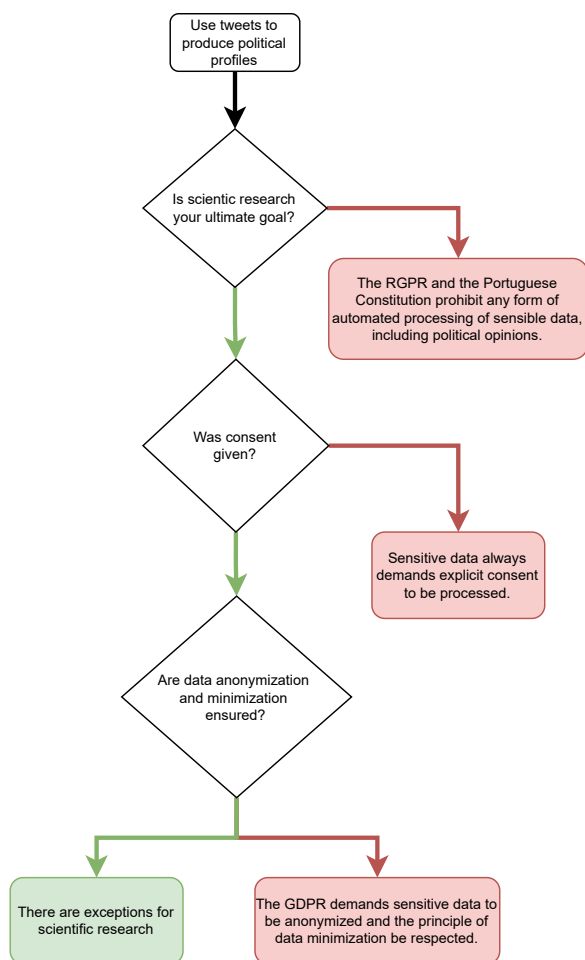


Figure 3: Flowchart concerning the legal issues of processing sensitive data.

legal vacuums and potential infringements of copyright and data privacy. We have tried to introduce key NLP concepts in a way that is accessible to both computer scientists and legal experts. The three use cases outlined serve to summarize the principal insights gained from this research. Utilizing flowcharts to illustrate these scenarios aims to accelerate the procedure of obtaining data and enhance overall comprehension of the research.

Our future work will focus on expanding the scope of use cases, matching to the same structured approach, while exploring additional topics. Specifically, we aim to incorporate insights from the recently introduced regulation on AI, the European AI Act, to further enrich the legal framework for NLP development.

8. Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 |

<https://doi.org/10.54499/LA/P/0063/2020> The author also would like to acknowledge the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

9. Bibliographical References

Rúben Almeida, Ricardo Campos, Alípio Jorge, and Sérgio Nunes. 2024. Indexing portuguese nlp resources with pt-pump-up. *arXiv preprint arXiv:2401.15400*.

Assembleia da República. 2021. *Carta Portuguesa de Direitos Humanos na Era Digital*. Diário da República Eletrónico. Lei n.º 27/2021 de 17 de maio.

Mafalda Barbosa. 2023. *Proteção de dados e inteligência artificial – perigos e soluções (também a propósito do chatgpt)*. *Revista Do Direito Da Responsabilidade*, 5(1):796–833.

Christopher M Bishop and Hugh Bishop. 2023. *Deep Learning*. Springer Nature.

Li Deng and Yang Liu. 2018. *Deep Learning in Natural Language Processing*. Springer, Singapore.

Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou, and Iryna Gurevych. 2018. A legal perspective on training models for natural language processing.

European Commission. 2021. *COM(2021) 206 final*. European Commission Document. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ESTABLISHING HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE REGULATION) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS.

European Parliament and Council of the European Union. 1996. *DIRECTIVE 96/9/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. Official Journal of the European Union. Directive on the legal protection of databases.

European Parliament and Council of the European Union. 2016. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. Official Journal of the European Union. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

- European Parliament and Council of the European Union. 2019. [DIRECTIVE \(EU\) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL](#). Official Journal of the European Union. Directive on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.
- Maria Raquel Guimarães, Rute Teixeira Pedro, Maria Regina Gomes Redinha, Fernanda de Araujo Meirelles Magalhães, et al. 2021. Direito digital.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. [The state and fate of linguistic diversity and inclusion in the nlp world](#).
- Dan Jurafsky and James H Martin. 2014. [Speech and Language Processing : an Introduction to Natural Language processing, Computational linguistics, and Speech Recognition](#). Dorling Kindersley Pvt, Ltd, India.
- Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramunas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värvi, Pavel Stranák, et al. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. In *Selected Papers from the CLARIN Annual Conference 2019*, pages 53–65. Linköping University Electronic Press.
- Ricardo Lopes and João Magalhães and David Semedo. 2024. [Glória – A Generative and Open Large Language Model for Portuguese](#).
- Teresa Nobre. 2012. [Direito de Autor e Direitos Conexos](#). University of Lisbon.
- OpenAI. 2022. [Introducing chatgpt](#).
- Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362.
- Alexandre Pereira. 2020. [A PROTEÇÃO JURÍDICA DO SOFTWARE EXECUTADO POR ROBOTS \(E DAS OBRAS GERADAS POR IA\)*](#), page 239–254. Universidade De Coimbra, Coimbra.
- Pires, Ramon and Abonizio, Hugo and Almeida, Thales Sales and Nogueira, Rodrigo. 2023. [Sabiá: Portuguese large language models](#). Springer.
- Presidência do Conselho de Ministros. 2023. [Decreto-Lei n.º 47/2023](#). Diário da República.
- Decreto-Lei transpondo a Diretiva (UE) 2019/790 sobre direitos de autor e direitos conexos no mercado único digital.
- Erasmus Marcos Ramos. 1998. A influência do bürgerliches gesetzbuch alemão na parte geral do novo código civil português. *Revista da Faculdade de Direito da UFRGS*, (15).
- República Portuguesa. 1966. [Código Civil](#). Diário do Governo. Article 79.º: Direito à imagem.
- República Portuguesa. 1976. [Constituição da República Portuguesa](#). Official Journal of the Portuguese Republic. Article 35: Utilização da informática.
- João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. [Advancing neural encoding of portuguese with transformer albertina pt*](#).
- Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. [Advancing generative ai for portuguese with open decoder gervásio pt*](#).
- Maria Marta Pereira Scherre and Maria Eugênia Lammoglia Duarte. 2016. Main current processes of morphosyntactic variation. *The Handbook of Portuguese Linguistics*, pages 526–544.
- Inês Silva Costa. 2021. [A proteção da pessoa na era dos big data: a opacidade do algoritmo e as decisões automatizadas](#). *Revista Electrónica De Direito*, 24(1):33–82.
- Souza, Fábio and Nogueira, Rodrigo and Lotufo, Roberto. 2020. [BERTimbau: pretrained BERT models for Brazilian Portuguese](#). Springer.