Universidade Estadual de Campinas
Instituto de Computação

Raysa Masson Benatti

Revealing Gender Biases in Court Decisions with Natural Language Processing

Revelando Vieses de Gênero em Decisões Judiciais com Processamento de Linguagem Natural

CAMPINAS
2023

Raysa Masson Benatti

Revealing Gender Biases in Court Decisions with Natural Language Processing

Revelando Vieses de Gênero em Decisões Judiciais com Processamento de Linguagem Natural

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientadora: Prof.ª Esther Luna Colombini
Co-supervisor/Coorientadora: Prof.ª Sandra Eliza Fontes de Avila

Este exemplar corresponde à versão final da Dissertação defendida por Raysa Masson Benatti e orientada pela Prof.ª Esther Luna Colombini.

CAMPINAS
2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

B431r

Benatti, Raysa Masson, 1992-
Revealing gender biases in court decisions with natural language processing / Raysa Masson Benatti. – Campinas, SP : [s.n.], 2023.

Orientador: Esther Luna Colombini.
Coorientador: Sandra Eliza Fontes de Avila.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Viés de gênero. 2. Processamento de linguagem natural (Computação). 3. Computação social. 4. Texto legal. 5. Inteligência artificial. 6. Computação semântica. I. Colombini, Esther Luna, 1980-. II. Avila, Sandra Eliza Fontes de, 1982-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações Complementares

Título em outro idioma: Revelando vieses de gênero em decisões judiciais com processamento de linguagem natural
Palavras-chave em inglês:
Gender bias
Natural language processing (Computer science)
Social computing
Legal text
Artificial intelligence
Semantic computing
Área de concentração: Ciência da Computação
Titulação: Mestra em Ciência da Computação
Banca examinadora:
Esther Luna Colombini [Orientador]
Luanna Tomaz de Souza
Rodrigo Frassetto Nogueira
Data de defesa: 24-04-2023
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-1227-1647
- Currículo Lattes do autor: http://lattes.cnpq.br/3849136766711583

Universidade Estadual de Campinas
Instituto de Computação

Raysa Masson Benatti

Revealing Gender Biases in Court Decisions with Natural Language Processing

Revelando Vieses de Gênero em Decisões Judiciais com Processamento de Linguagem Natural

Banca Examinadora:

- Prof.ª Dra. Esther Luna Colombini
  IC/Unicamp
- Prof.ª Dra. Luanna Tomaz de Souza
  ICJ/UFPA
- Dr. Rodrigo Frassetto Nogueira
  NeuralMind

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 24 de abril de 2023

This work is dedicated to my mother, Silvana.

Acknowledgements

This piece, as any scientific endeavor, is a work of many hearts and hands.

First and foremost, I thank my mother, Silvana, whose unrestricted love brought me here and continues to reverberate all around. I am lucky to have had her support over my entire academic journey and learned from her larger-than-life generosity, compassion, and wisdom. Knowing that she would have been the proudest audience of this work gave me much of the strength needed to accomplish it.

I thank my unbelievably amazing supervisors, Esther Luna Colombini and Sandra Avila, whose kindness, competence, and work ethics I constantly carry as an inspiration. I am grateful for how they placed their trust in me even — and especially — when I couldn't do it myself, and for showing me what outstanding supervision looks like.

I thank Fabiana Severi for her essential role in enriching the path of so many legal and social science scholars in our country, one of which I was privileged to be. I thank her and Camila Villarroel for kindly accepting to be the best research collaborators for which this work could have asked.

I am glad to be a part of the Recod.ai lab, which turned out to be the perfect environment for this work to be developed. I thank its members and partners and hold a special feeling of gratitude for some of the dear friends I made here along the way: Marcela, for her exceptional supportiveness and presence during good and bad times; Gabriel, who enthusiastically said hi to me on my first day here, for our shared sense of humor and energetic chats; Diego, for his companionship and remarkable dancing and photography skills; Víctor, for his vivacity and for always being up for anything; João Phillipe, for the readiness and availability to share his skills and provide valuable input for this work; Rafael, for being a heart and soul of this lab and a great coordinator for our gatherings; Matheus, for his soothing spirit and joyful presence; Wladimir, for his refreshing irreverence; Sadeeq, for always being so kind, open, and understanding towards our blabbing in Portuguese; Levy, for always having a word of encouragement for me. Many other people from this lab have also been a part of this journey in different ways, and I am grateful for all of them.

I am grateful to have been a fellow in the BEAAMO research group at the UC Berkeley, which broadened my perspective as a researcher and gave me the opportunity to know and work with brilliant people like professor Rediet Abebe, Mírian Silva, Tainá Turella, Rodrigo Dornelles, and other members of the team. I am glad that I got to share such a memorable time with them.

I thank my sisters, Biba and Talita, and my father, Paulo, for being my home. A special thanks goes out to Biba, who kindly put her talent in service of this work by drawing all the beautiful diagrams.

Privileged as I am to be surrounded by many amazing people, I also thank my friends Diego, Marcos, and Carol, for never failing to listen to me no matter how boring the content; Gabi, Elisa, and Maíra, for being my rock; Augusto, for trusting me his craziest

and best ideas (and for the willingness to proofread this work).

I thank Helena Branquinho, a special and sensitive professional without whom I would not have made it through this stage.

I thank professors Luanna Tomaz de Souza, Rodrigo Frassetto Nogueira, Nádia Félix Felipe da Silva, and Roberto de Alencar Lotufo, for making themselves available to share their knowledge by enriching this work with their evaluation.

Lastly, I thank all the support provided by the Institute of Computing and its workers, as well as every institution which supported me financially during any period of this master's work: CAPES[1]; the University of Campinas and its support fund for teaching, research, and outreach (FAEPEX); the University of California, Berkeley.

Thank you all so much for allowing me to stand on your shoulders.

[1]This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* — Brazil (CAPES) — Finance Code 001.

Resumo

Dados oriundos das ciências sociais são comumente produzidos em forma de texto digital, o que motiva seu uso como fonte para métodos de processamento de linguagem natural. Pesquisadores e profissionais vêm desenvolvendo e utilizando técnicas de inteligência artificial para coletar, processar e analisar documentos no campo jurídico, especialmente em tarefas como sumarização e classificação de textos. Neste cenário, identificamos um potencial subexplorado do uso de processamento de linguagem natural para lidar com questões de direitos humanos, no contexto da inteligência artificial voltada ao bem social. Métodos qualitativos e quantitativos das ciências sociais têm sido utilizados para estudar questões como a presença de estereótipos de gênero em instituições legais; contudo, abordagens baseadas em processamento de linguagem natural poderiam auxiliar a ampliar a escala de alcance de tais tarefas. Neste trabalho, apresentamos um protocolo de automatização da detecção de estereótipos de gênero institucionais em cortes brasileiras, que inclui: (a) um pipeline de coleta, anotação, e preparo de textos extraídos de decisões judiciais emitidas pelo Tribunal de Justiça de São Paulo em casos de violência doméstica e alienação parental, resultando em dois datasets; (b) um protocolo experimental de classificação binária supervisionada sobre as decisões, executado com redes derivadas do modelo BERTimbau; (c) métodos para avaliação e validação desse protocolo. Mostramos que técnicas baseadas em semântica, como classificação com mecanismos de atenção, são satisfatoriamente adequadas para aprender a identificar automaticamente vieses de gênero em decisões judiciais; no entanto, restam questões de validação, escalabilidade e explicabilidade a serem exploradas. Também elaboramos diretrizes éticas e legais para o uso e disponibilização de decisões judiciais como dado de entrada para modelos de aprendizado automático. Nossa abordagem pode auxiliar especialistas de áreas como Direito, Estudos de Gênero e Políticas Públicas a explorar novas possibilidades de análise em seu domínio — além de fornecer insights sobre o uso de técnicas e ferramentas de processamento de linguagem natural.

Abstract

Data derived from the realm of the social sciences is often produced in digital text form, which motivates its use as a source for natural language processing methods. Researchers and practitioners have developed and relied on artificial intelligence techniques to collect, process, and analyze documents in the legal field, especially for tasks such as text summarization and classification. In this scenario, we identify an underexplored potential of natural language processing used to delve into human rights issues in the context of artificial intelligence for social good. Qualitative and quantitative social science methods have been used to study matters such as institutional gender biasing in legal settings; however, natural language processing-based approaches can help analyze the issue on a larger scale. In this work, we present a protocol to address the automatic detection of institutional gender biasing in Brazilian courts, which comprises: (a) a pipeline of collection, annotation, and preparation of text extracted from court decisions issued by the São Paulo state Court of Justice in cases of domestic violence and parental alienation, which resulted in two datasets; (b) an experimental protocol of supervised binary classification over the decisions, performed with BERTimbau-based models; (c) methods for evaluating and validating such protocol. We show that semantics-based techniques, like attention-based classification, are satisfactorily adequate to learn to reveal gender biases in judicial decisions automatically; however, validation, scalability, and explainability issues remain to be addressed. We also design legal and ethical guidelines on the use and availability of court decisions as a data input for automatic learning models. Our approach might help experts from fields such as Law, Gender Studies, and Public Policy to explore new analysis possibilities in their domain. It also provides insights into natural language processing techniques and tools.

Contents

Chapter 1

Introduction

Natural language processing techniques have been proposed to address issues in many domains. Certain fields are more prone to use and produce texts containing relevant data for analysis, such as the social sciences. The legal field, in particular, is the focus of interest in this work.

As observed in different artificial intelligence domains, natural language processing has benefited from increasingly available digital data. In the last decades, public institutions have substituted physical documents and procedures for digital ones; Brazilian courts have followed this trend, especially after the promulgation of Federal Law 11419/2006 [15], which fosters the use of digital documents in judicial processes. Our interest relies on exploring judicial decisions, largely available in federal and state courts' websites — except for special cases that must remain secret.

One aspect that might be extracted from such decisions raises concerns: the presence of gender biases or stereotypes encrusted in legal reasoning, especially in cases of gender-based violence (GBV). There is evidence that court rulings can bear those biases, as explained in Section 1.2; however, applying computational techniques to address it seems to be underexplored, as discussed in Sections 1.1 and 1.3. This is the gap on which relies the motivation behind this work. We show that techniques such as supervised classification with attention-based networks, described in Section 2.2, can be explored in that sense.

Our **underlying hypotheses** are: **(a) gender biases and stereotypes can be detected in judicial decisions on a large scale**, and **(b) natural language processing offers suitable approaches to detect them**. Therefore, we have designed, developed, and tested a natural language processing protocol to evaluate those hypotheses. Our protocol can (a) be helpful for domain experts who wish to collect and analyze such data on a large scale and (b) provide new possibilities for the application and adaptation of supervised classification tasks over legal texts.

Our approach includes structuring data and metadata and developing an attention-based deep learning solution for its classification. Therefore, we provide a multidisciplinary solution: on the one hand, a new natural language processing protocol and its tools; on the other, the possibility for domain experts to find new answers to their inquiries.

1.1 Artificial Intelligence, Law and Jurimetrics

The term *jurimetrics* was introduced in 1949 by lawyer Lee Loevinger, who claimed that the Law field should embrace scientific (roughly meaning quantitative) approaches to remain respectable and aligned to the demands of the twentieth-century [65]. Although understanding traditional approaches as unscientific is questionable in many ways, quantitative analysis has been playing an essential role in the legal field and other social sciences.

Jurimetrics is the application of quantitative methods to Law [26]. Such methods usually come from math and statistics. Still, they can also include computer science ones — such as artificial intelligence techniques and, specifically, natural language processing, considering that data in the field is mostly text.

In fact, the use of computational approaches in Law — beyond conventional methods from statistics — has been recognized and referred to as *computational legal science* [63]. The intersection between artificial intelligence and the legal field has also been a target of interest in the last decades. Its main results and trends have been presented biennially since 1987 at the International Conference on AI and Law (ICAIL) [10], organized by the International Association for Artificial Intelligence and Law (IAAIL) in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI).

In Brazil, there is also evidence of an increasing interest in the field [70]. Research associations include the Brazilian Association of Jurimetrics (ABJ)[1], Lawgorithm[2], and Direito Tech[3]. Private law practitioners rely on jurimetrics and legal business intelligence services, often provided by specialized companies referred to as *lawtechs* or *legaltechs*, whose services tend to focus on predictive analysis of outcomes, estimation of risks and cost estimation, and supporting legal strategies [36].

Brazilian courts have also put artificial intelligence systems in place to automate tasks such as document classification [71, 13]. Increasing procedural efficiency and reducing costs are motivations behind their use [41, 37]. Other common tasks in the legal field that can be performed using natural language processing techniques include document reviewing [23, 52] and legal writing or summarization [52].

In this scenario, we identify an unexplored potential for using computational methods in the context of human rights and artificial intelligence for social good. In Section 2, we detail how computational approaches are suitable for analysis in the legal field. Data availability is not an issue since using digital documents in judicial processes has been fostered in Brazil since 2006 [26]; Brazilian court websites provide digital court rulings texts and their metadata. We focus on a specific class of cases related to gender violence.

1.2 The Issue of Gender

Stereotyping assumes one's characteristics or roles due to his or her belonging to a certain group; when associating a feature with a group and assuming its members share this feature disregarding their individual traits, we are stereotyping them. Therefore, a

[1] `abj.org.br`
[2] `lawgorithm.com.br`
[3] `direito.tech`

stereotype is a generalized view or preconception about a group [30]. A gender stereotype exists when such a view is related to the gender of its target.

Humans stereotype each other for many reasons: to maximize simplicity and predictability, to assign difference, to script identities — in general, to make sense of the world by reducing its complexity [30]. Stereotypes can reflect statistical evidence about a group, and they are not necessarily negative; however, some might be noxious.

Gender stereotypes, in particular, tend to be especially harmful towards women and represent a "challenge in combating sexism, which is often perpetuated through stereotypes", according to Cook and Cusack [30]. The authors describe how such generalizations might help degrade women, diminish their dignity, disproportionately add to their burden, and hamper their access to rights or justified benefits.

In that sense, illegitimate gender stereotyping is perceived as a pervasive human rights violation [33]. The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) [101], signed and ratified by Brazil, expresses that state parties must take adequate measures to eliminate prejudices and practices based on stereotyped concepts of gender roles. According to the Convention, authorities and institutions, including tribunals, must eradicate discrimination against women.

However, institutions themselves are often the venue in which harmful gender stereotyping occurs and unfolds into destructive consequences. Legislative processes, court rulings, and the Law itself tend to reflect social, political, and economic relations present in society; therefore, despite their neutrality rhetoric, they frequently reinforce gender discrimination practices [8].

Several studies have addressed how judicial proceedings issue gender stereotyping acts and some consequences of this [8, 81, 45, 34, 77, 104]. Particularly in Brazil, Federal Law 11340/2006 (*Lei Maria da Penha*) [16] creates legal mechanisms, including proceeding rules, aiming to prevent and repress violence against women, according to guidelines provided by the country's Federal Constitution [14] and the CEDAW. However, there is evidence that Brazilian courts often disregard some of its provisions while relying on noxious stereotypes, resulting in inappropriate institutional responses to women affected by gender violence [34, 77].

Such studies are mostly based on traditional methods from the social sciences (e.g., content analysis). In general, data of interest — usually decisions and other physical or digital documents issued by courts — is collected manually or (rarely) through web scraping. Quantitative analysis is limited to tens or no more than a few hundred documents and is performed by a person or group. To the best of our knowledge, no academic work has ever addressed such issues using recent natural language processing techniques in Portuguese, despite their abundance and availability of digital legal documents.

In that context, natural language processing tools might help expand possibilities of analysis of such documents — after all, language itself might contain traces of stereotyping [66, 90]. The protocol we describe enables the collection and extraction of patterns from a larger volume of texts since it allows the automation of processes currently handled by humans. Some of the domain inquiries that remain unanswered and that such an approach can help clarify include: detecting correlations between gender biases in court rulings and their metadata (date, location, type of court); confronting such results

with other data (e.g., gender violence statistics); assessing the quality of decisions on a large scale; systematizing outcomes from decisions and verifying how they correlate to institutional gender stereotyping.

We argue that this work can help legal practitioners and researchers answer such questions and, overall, analyze the presence and implications of gender biases in courts. It also provides a methodology that can be used to apply automatic text classification techniques in the social sciences.

1.3 Natural Language Processing and Gender Biases

Recent work addressing gender biases in the field of natural language processing tends to focus on either mitigating gender biases in natural language processing systems, detecting the gender of a text's author, or detecting hostile language, including hate speech towards women [97].

Artificial intelligence researchers recognize that systems that learn from data are prone to reinforce or amplify harmful stereotypes since the data is extracted from the real world [98, 31]. Acknowledging such weakness and making efforts to mitigate it are relevant concerns for two reasons: scientifically, biased systems should generalize results poorly; ethically, biased systems might harm people in many ways [11]. Many studies have explored methods to detect and mitigate gender biases in natural language processing systems and models [98, 31, 88, 89, 106, 6, 12].

The field also explores the concept of gender in artificial intelligence to detect or predict demographic attributes. For example, Burger et al. [17] propose classification methods to identify the gender of Twitter users based on their content.

Automatic hostile speech detection is another context that gathers interest in language, artificial intelligence, and gender studies since such speech is often directed toward gender minorities. Anzovino et al. [4] propose a labeled data corpus and supervised machine learning techniques to identify and classify misogynistic language on Twitter. Schmidt and Wiegand [91] survey natural language processing techniques used to detect online hate speech and show that sentiment analysis, among other approaches, is suitable for the task. Deep learning approaches have been adopted to address veiled toxic speech through unsupervised methods [47, 51]. This demonstrates that gender biases can indeed be automatically detected in language; however, since the language used in legal documents does not tend to follow hate speech patterns, such approaches are not directly applicable to this work.

We also stress that when addressing gender issues in natural language processing, one might be aware of ethical considerations. Larson [62] discusses theoretical and ethical guidelines for researchers and practitioners. Although this is not the main object of this work, our analysis is guided by such considerations. In that sense, our views on gender and gender-based stereotypes are aligned with the theory of gender performativity presented by Judith Butler in 1990 and explored in related work since [18], according to which "language is a part of gender performativity, and (...) a key part of how we transmit and maintain stereotypes, (re)produce meaning, and navigate systems of power" [42].

1.4 Contributions

The main contributions of this work are:

1. Two datasets of court decisions issued by the São Paulo state Court of Justice in gender-based violence cases, with annotation (partial and complete, respectively for Datasets 1 and 2), their metadata on a range of legal attributes, their documentation, and the description of collection, processing, and annotation protocols;

2. An experimental pipeline for automatic detection of gender biases in court decisions issued in Brazilian Portuguese, which can be reproduced by domain experts with some technical training;

3. Legal and ethical guidelines on the use and availability of datasets made of court documents, published by the author and contributors as part of this work in the article *Should I disclose my dataset? Caveats between reproducibility and individual data rights*, published at the Natural Legal Language Processing Workshop, at the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) [69].

1.5 Outline

The remaining of this text is organized as follows. In Chapter 2, we briefly provide a timeline of natural language processing research, delve into the main techniques of interest in this work, and present some of the studies developed in natural language processing applied to the legal domain. In Chapter 3, we describe our methodology, which comprises considerations of the data, experimental pipeline, and validation methods. In Chapter 4, we describe and discuss experimental results. Finally, in Chapter 5, we elaborate on our findings and prospects for future work.

Chapter 2

Background

Natural Language Processing (NLP) is an interdisciplinary field of study belonging to both Linguistics and Computer Science; it is concerned with automating human language-related processes, such as comprehension of what is said (or written) and how this content relates to the world [72]. This usually means enabling (or improving) human-machine communication, text or speech processing, and language generation. Its tasks include conversational agents building, machine translation, question answering, information retrieval, text summarization, topic modeling, and word sense disambiguation [58, 20]. Disambiguation is needed since natural languages are, by definition, ambiguous; therefore, NLP systems are required to clarify not only word sense but also word category, syntactic structure, and semantic scope when possible [72].

Early NLP research took place in the 1950s when artificial intelligence moved towards academic consolidation. Then, automatic translation was an important topic and was handled through simple approaches such as dictionary-based word-for-word processing and parsing [57]. Despite the enthusiasm of researchers, some obstacles, such as semantic-related issues and technical limitations, were challenging to surmount [57].

Later on, during the 60s and 70s, there were attempts to incorporate some world knowledge in NLP systems, and identifying language function and meaning became a trend; therefore, this era was more semantics-oriented [57]. It was followed by a grammatico-logical phase in the 80s when computational grammar theory developed into an active field, and the lexicon's role grew in importance [57].

As data availability (speech and text data included) and processing capabilities increased, probabilistic approaches arose in the field. New possibilities emerged for exploring semantic issues such as detecting key concepts in a text, disclosing discourse structure, and integrating NLP in multi-modal systems [57]. Remarkable techniques developed in this context include Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), mainly used for topic modeling and other semantic classification tasks.

Information extraction techniques also leaped with the expansion of possibilities. We stress sentiment analysis, sometimes referred to as emotion recognition, opinion mining, or polarity detection (there are subtle differences between terms, but they refer to the same group of methods). Such techniques evolved from canonical statistical approaches (e.g., term frequency count and keyword spotting) to more sophisticated ones, including lexical affinity, Bayesian inference-based techniques, graphs-based techniques, and Web

ontology [19].

Traditional machine learning algorithms, such as naïve Bayes classifiers, hidden Markov models, decision trees, and support vector machines, have been widely used to address NLP issues for a while [80]. However, over the last decade, deep and reinforcement learning techniques were proposed more frequently, partly due to computational power increase and parallelization [80, 94]. One of the promises of deep learning is to perform challenging tasks while avoiding the need for extensive feature engineering and prior knowledge [25]. Although prior domain knowledge remains necessary in developing, implementing, and using deep learning models, such techniques reduce the need for feature engineering.

It is also realistic to rely on deep learning to enhance the automatic understanding of a language (through language modeling, morphology, parsing, and semantics), allowing better performance of canonical tasks [80]. Canonical tasks boosted by deep learning include information retrieval and extraction, text classification, text generation, summarization, question answering, and machine translation [80]. Deep learning innovations worth stressing in the field include the Transformer network architecture, based on attention mechanisms [102], and the language representation model BERT (Bidirectional Encoder Representations from Transformers) [43] — along with its Portuguese-trained version BERTimbau [96], which is especially relevant to this work.

In short, the current NLP area can be broadly defined as "primarily a data-driven field using statistical and probabilistic computations along with machine learning" [80]. The task that concerns us the most is automatic supervised text classification with attention-based networks, detailed in Section 2.2. Section 2.3 presents an overview of related work addressing tasks from the legal domain with natural language processing methods and tools. Now, we delve into different forms of text data representation.

2.1 Data Representation

Text data can be digitally represented in many ways. While it is not our goal to describe every detail of each of them, discerning a general picture is essential to understand what happens behind NLP techniques. We stress three primary forms of text representation, presented in order of complexity: one-hot encoding, bag-of-words-based representations, and word embeddings.

One-hot encoding represents each token[1] as a vector of ones and zeros. In general, such a vector will be filled with zeros except in an index position; since tokens must be distinguishable, each token vector has a different index. The larger the vocabulary is, the more vectors are needed to represent it, which generates a representation made of sparse matrices. Overall, it has limited applications.

Bag-of-words-based representations tend to be more efficient. Such approaches convert the text into a bag of words, i.e., a set of tokens; no sequential information is preserved. Instead, the analysis is based on the frequency of tokens. Bag-of-words representations include **term-document matrices** and **term-term matrices**.

[1] In our context, a token is a defined document unit. Text preprocessing usually involves some form of tokenization to extract those units, which might be solitary alphabetic characters, words, word roots, n-grams.

A term-document matrix A presents each term in a row and each corpus document in a column. Each matrix cell $a_{i,j} \in A$ contains the frequency in which term i appears in document j. Such a structure helps identify similar words based on the documents they appear to the most since related words occur in the same documents.

A term-term matrix B presents target terms in each row and context terms in each column. Each cell $b_{i,j} \in B$ contains the frequency in which a target term i co-occurs with a context term j in some corpus. The idea behind it is to detect similar words based on their context. A context can be, for example, a set of words that appear near a target term.

Additionally, instead of raw frequencies, such matrices can have their cells filled with weighted frequencies. The most common scheme to compute weighted frequencies is **TF-IDF**. TF stands for *term frequency*; $TF(t, d)$ represents the number of times a given term t appears in a given document d. IDF stands for *inverse document frequency*. Given a corpus D of N documents, $IDF(t, D)$ is defined as $\ln \frac{N}{DF(t)}$, in which DF stands for *document frequency*, and $DF(t)$ is the number of documents containing the given term t [54]. The final value of $TF\text{-}IDF(t, d, D)$ is given as $TF(t, d) \cdot IDF(t, D)$.

The intuition behind it is that the inverse document frequency should reflect how relevant a given term is in a corpus. Texts are filled with common words that carry little meaning, such as prepositions and articles, showing a high DF value. If $DF(t)$ is high enough to equal the total number N of documents, IDF will remain as $\ln(1) = 0$. On the other hand, rare terms — which tend to be more meaningful in a corpus — will make for a low $DF(t)$ value and a high IDF. Thus, the value of $TF\text{-}IDF(t, d, D)$ reflects the occurrence frequency of a term t in a document d, weighted by how relevant t is in the corpus D.

Although bag-of-words-based representations are largely used in NLP techniques due to their simplicity and robustness, **word embeddings** perform better in semantics-dependent tasks. Such tasks include detecting polysemy and identifying non-trivial similarities between elements of a text. For instance, the two sentences "Ana eats an apple" and "Maria tastes a banana" have no words in common but share a similar meaning that traditional text representations cannot grasp.

In word embeddings, words are represented as dense vectors of features; a model — usually a neural network — learns those features, being trained over a large corpus. Then, it is possible to extract associations between words based on their features. The model might learn, for example, that "king" is to "man" as "queen" is to "woman". Word2vec [75] is a largely used architecture to produce word embeddings. In the Portuguese language, which concerns us the most, Hartmann et al. [53] trained and evaluated word embedding models using fastText, GloVe, wang2vec, and word2vec.

More recently, the development of attention-based architectures [102], such as the BERT language representation model, incorporated context-based learned embedding vectors to represent textual data — which is particularly suitable for context-dependent tasks, such as bias detection. We now delve into a description of these architectures and the network we use in this work to address supervised classification for gender biases detection: BERTimbau.

2.2 Transformers-based Architectures

Different NLP approaches can help reveal gender biases in language [97]. Since identifying them while disregarding semantic issues would be challenging, semantics-centered techniques seem to be adequate — which is especially relevant considering that legal documents do not always follow patterns of explicit biases.

Our main goal is to distinguish between biased and unbiased decisions, which can be unfolded into a task of canonical automatic classification. Text classification can be performed in many ways: through unsupervised or supervised training, using different classifiers, and relying on distinct forms of data representation. In this work, we adopt a supervised classification approach with attention-based networks, which we now delve into.

As a behavioral and cognitive concept, attention is a process that allows one to focus on parts of information that are more relevant according to some criteria; computational systems based on this idea have existed for at least the last three decades [40], and especially after the 2000s [49]. A seminal deep learning model of attention was proposed in 2014 to enhance machine translation by introducing an extension to the traditional encoder–decoder model [5].

The first architecture entirely based on attention mechanisms, the Transformer, was described in 2017 [102]. The intuition behind it is establishing dependencies between the input and output of sequential data without the need for recurrence or convolution, which had been the main approaches up to that point. It showed state-of-the-art performance for machine translation, also generalizing well for other tasks.

Attention mechanisms remained at the core of the state-of-the-art performance of a range of natural language processing tasks. A step further in this direction was accomplished in 2019 with the introduction of BERT (Bidirectional Encoder Representations from Transformers) [43], a pre-trained language representation model based on the Transformer architecture previously described. In 2020, the main BERT-based pre-trained model for Brazilian Portuguese was released, improving state-of-the-art performance in this language in tasks of sentence textual similarity, recognizing textual entailment, and named entity recognition [96]. Since then, it has been the primary attention-based tool for natural language processing tasks in Brazilian Portuguese — and the one we use in this work to perform the task of binary classification over our data.

2.3 Related Work

In general, natural language processing applied to the legal domain focuses on one (or more) tasks: text classification and/or sentiment analysis; text summarization; topic modeling; information retrieval; named entity recognition. Some studies propose new frameworks or methods to address a specific demand, while others look at the issue in a broader sense by comparing approaches or refining research questions.

Techniques also vary, ranging from traditional **parsing approaches** to extract information from the text to **word embeddings approaches** and **neural networks classifiers**. Most studies use **conventional machine learning** techniques for classification, especially in sentiment analysis. **Topic modeling** is also present in some studies, mainly

based on traditional techniques (LSA and LDA). Some use or fine-tune **attention-based networks** to perform a task of choice.

McCarty [73] proposes a statistical parser to automatically extract judicial opinions from 111 federal civil case documents in United States appellate courts. This work's approach is based on a definite clause grammar for semantic interpretation, with around 700 rules. Although parsing strategies pose limitations, the author proposes a deep semantic interpretation and shows a way to compute it. The ultimate goal from such results would be to provide semantic interpretations of judicial opinions, allowing to summarize main information from a given case — thus making a law practitioner's job easier.

Parsing approaches are also studied by Wyner et al. [105], who surveyed works in which context-free grammars were applied, or could be applied, to identify arguments in legal cases. Such works use databases of less than 100 cases and usually test such grammar over one or more extracted arguments. By relying on concepts from argumentation theory, the idea is to evaluate how suitable this technique is to distinguish between different kinds of sentences automatically. Although some results reveal to be satisfactory, the authors stress that advanced machine learning techniques could improve some issues — for example, manual knowledge labeling, which is a costly task.

Most researchers approach opinion mining as a classification problem and use conventional machine-learning approaches to address it. For example, Moens et al. [76] present experiments on the automatic detection of arguments from texts, including legal ones. A labeled database of around 3,700 sentences extracted from different domains trained a naïve Bayes classifier using features involving lexical, syntactic, semantic, and discourse properties. This allowed them to separate arguments from mere rhetorical sentences, but accuracy measures showed that legal texts were more challenging to classify than others. The authors hypothesize that such a limitation could be attributed to the lack of data and its ambiguity. However, they showed that the chosen features are relevant to detect arguments automatically.

Conrad and Schilder's work [28] surveys approach to perform opinion mining in legal blogs and proposes a language modeling method to address both subjectivity and polarity analysis over the material. They built a corpus with 200 blog entries, with around 1000 sentences classified as positive/negative (polarity) and opinion/non-opinion (subjectivity). A naïve Bayes classifier with language model smoothing exhibited the best accuracy measures to perform the task. Still, the results were not exceptional, and the authors stress that more sophisticated approaches could achieve better accuracy.

Polarity analysis is also addressed by Gómez et al. [50], who apply a pointwise mutual information-based approach on a corpus of judicial sentences to detect polarity in text, i.e., whether their content is positive, negative, or neutral-charged. Besides measuring the accuracy of results, such polarity is also confronted with emotions in the content — sadness, happiness, and neutrality —, as identified by humans. They show compatibility between the automatically detected polarities and the human-labeled emotions.

Liu and Chen [64] also use a supervised classification model; their motivation task, however, differs from most. They propose a method to predict pending judgments using legal documents by extracting features from precedents and classifying them. They use a precedent corpus made of approximately 1,400 documents. In their work, sentiment

analysis is not the goal but a classification feature. They also apply pointwise mutual information to establish semantic associations between terms and compute a sentiment score, which leads to more accurate results. Besides revealing the importance of sentiment as a classification feature, they stress that performance could be better with a more extensive and domain-specific vocabulary.

Still in sentiment analysis, Aires and Morais [2, 38] describe a database annotation method to perform this task over decisions issued by the Brazilian Federal Court. Their method consists of labeling such documents based on the decision's outcome, allowing supervised machine learning systems to classify them. The motivation behind it is to enhance the performance of jurisprudence search engines. Although they do not implement their framework, their work suggests that using sentiment analysis to extract judicial decision outcomes is feasible. Their method, like ours, depends on data annotation.

Text summarization is another task tackled by conventional machine learning. Kanapala et al. [59] approach it as a binary optimization problem, thus proposing a gravitational search algorithm to produce summaries of legal judgments. They trained and evaluated their method over the legal track of the FIRE-2014 (Forum of Information Retrieval Evaluation) dataset, containing around 1000 supreme court judgments. Traditional evaluation metrics for text summarization — ROGUE-1 and -2 — were used to compare their technique to other artificial intelligence algorithms, such as a genetic algorithm, LSA, and particle swarm optimization; their approach performed better than the others. This work illustrates the demand for text summarization in the legal field and demonstrates how it can be addressed differently.

Merchant and Pande [74] also address text summarization in this domain. They use LSA to capture concepts from legal documents and build a summarization system — which shows how LSA can tackle tasks other than classic topic modeling. This approach uses concepts extracted from the text to produce its summarization. To implement their method, they also developed a Python web crawler using Selenium and Beautiful Soup libraries to scrape legal judgments issued by the Indian judiciary system — which is an interesting approach to legal data collection.

We also identify researchers who use neural networks-based classifiers. Undavia et al. [100], for instance, present a system based on neural networks to classify legal court documents. This work aims to classify 8,419 US Supreme Court opinions, extracted from the Washington University School of Law Supreme Court Database (SCDB), into 15 legal categories. They use 300-dimensional pre-trained word2vec vectors. Combined with convolutional neural network architecture, they produce more accurate results than classic machine learning approaches — with the benefit of demanding less data preprocessing. The authors believe that specific word embeddings from the legal domain could produce even better results than general ones.

Silva et al. [95] describe a method to classify documents issued by the Brazilian Supreme Court based on their legal category. They use a convolutional neural network proposed by Conneau et al. [27] over a dataset of 6,814 decisions. This model performs the task with 90.35% accuracy — however, they do not present other models to compare. Ferreira [46] fulfills this gap: the author tackles a similar goal over the same dataset, showing that a neural LSTM (long short-term memory) model achieves the highest ac-

curacy (94%). In both cases, what motivates the work is the possibility of automating a task currently performed by humans.

Ahmad et al. [1] also use a deep learning model to perform text classification in this domain; however, instead of classifying documents in some categories, their goal is to classify sentences in legal documents based on their rhetorical role. Around 6,153 sentences, extracted from the open-source repository Veteran Claims, were used to train a BiLSTM (bidirectional LSTM) model with 300-dimensional GloVe word embeddings — which achieved better accuracy than other embeddings.

Chalkidis and Kampas [21] go beyond: besides studying deep learning techniques for text classification, information extraction, and information retrieval in the legal field, they present Law2Vec. It is a word2vec-based word embedding model trained over a large corpus of legislation from different countries, comprised of around 120,000 documents with 492M tokens. They analyze results qualitatively and stress that, although their approach produced satisfactory semantic relations, an even larger database could increase the model's semantic representation. It remains an additional option for word embeddings in English, focused on the legal domain.

Named entity recognition (NER) is also a relevant natural language processing task in the legal field. In the Brazilian context, Albuquerque et al. [3] developed UlyssesNER-Br, a corpus of legislative documents evaluated according to this goal. The motivation to perform NER automatically relates to information retrieval, which facilitates navigation for Law practitioners and researchers — especially in large legal systems such as the Brazilian one.

BERT-based models are also used in the legal context. Elwany et al. [44], for instance, describe how to fine-tune BERT on a large corpus of legal agreements. They use it to perform supervised classification over such documents and show that applying the pre-trained BERT model improves results. Sun et al. [98] show that BERT can also be used to tackle aspect-based sentiment analysis; however, their work is not specific to the legal domain. Still, they show consistent results: their approach was evaluated over a dataset of around 5,200 sentences.

In Brazil, different projects have trained some BERT models over legal data to generate domain-specific embeddings. For instance, Polo et al. [85] trained the original BERTimbau model over datasets of Brazilian legal publications, legal documents from the São Paulo state Court of Justice, and procedural updates from different courts. Their resulting model, called BERTikal, was used as a feature extractor to perform supervised classification of procedural updates, showing fair results. Viegas [103] developed JurisBERT, a BERT model which was pre-trained over 410 MB of raw text of a diversity of legal documents; their model was then fine-tuned and used to perform a task of semantic text similarity. Such approaches show that domain-specific models can perform well on domain-specific tasks; however, their use is often limited to the task for which they have been trained, which is why we did not use any of them in our work.

Finally, some works have addressed the automatic detection of gender biases in legal contexts. Pinto et al. [83] propose a project to develop a linguistic model and a tool to perform such a task over a (manually annotated) corpus of legal sentences published by the Portuguese Ministry of Justice on gender-based violence cases. While their approach

is similar to the one we propose, they have not published results yet or settled on a methodology. On the other hand, Sexton et al. [93] show results on using different supervised classification models to detect gender biases in Fijian court documents issued in the context of gender violence cases. Their dataset has 13,384 court documents, of which 809 were annotated — the same strategy we use in this work. In their case, however, they evaluated performance on different models: a support-vector machine, convolutional neural network architectures, and BERT-based architectures. They all showed promising results, but the authors stress challenges such as managing overfitting — due to the low availability of annotated data —, having experiments hampered by limitations on computational processing, and dealing with data heterogeneity. As shown in Chapters 4 and 5, there are overlaps between their conclusions and the ones we present.

Overall, the literature review shows that, despite the increase of technical possibilities in NLP, long-established methods still prevail over new architectures in the legal domain — even in recent studies. We found no evidence of delivered research linking natural language processing tools to automatically detecting gender biases in legal documents in the Portuguese language.

Chapter 3

Methodology

This chapter describes the methodology we followed to accomplish the work, which comprises:

1. Data (Section 3.1): choice, collection, processing, annotation, and ethical considerations regarding the data;
2. Experimental pipeline (Section 3.2): choice of techniques, modeling, and implementation of experiments;
3. Evaluation and validation methods (Section 3.3): choice of evaluation metrics and tools and a validation pipeline.

Figure 3.1 presents a high-level view on our methodology. It starts with protocols of collection, annotation, and preparation of two datasets of Brazilian court decisions, whose texts are cleaned and transformed in chunks. This step aims at making the data's content and size adequate to feed the models which perform the task of classifying them.

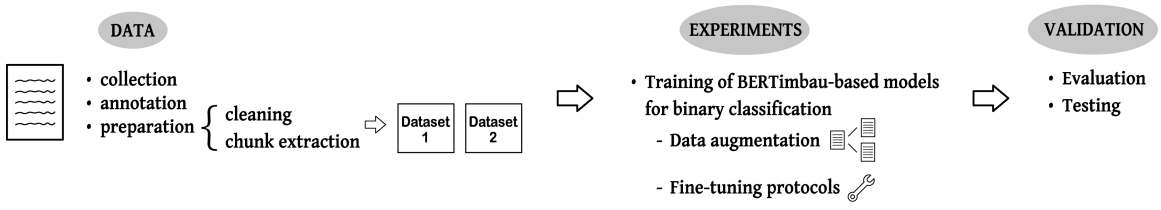


Figure 3.1: High-level view on the methodology.

Classification is performed in the experimental phase. We run a set of experiments over BERTimbau-based models with different degrees of data augmentation, with the goal of training them to differentiate between biased and non-biased chunks of labeled text. In this phase, we apply different fine-tuning protocols over the pre-trained networks, using our own data to adjust their parameters.

The training and validation sets are used to teach and evaluate the models; for evaluation, we use metrics of loss and balanced accuracy. Besides evaluating model performance

on the validation set, the validation methods include a testing pipeline. While our test sets are too small to pose statistically significant validation results, we deliver a pipeline that uses the best versions of the trained models to assign a label on the whole texts of court decisions that are compatible with our framework. It could therefore be used in enriched versions of our datasets or in new ones.

3.1 Data

This section details the choice, collection, annotation, and processing of the data used in our work. We also provide context on the courts from which the data is extracted and ethical considerations on its use and availability. At each phase description, we provide a diagram of the corresponding files and how they relate to each other.

3.1.1 Choice of Datasets

All of the decisions used as input for our investigation were issued in the second instance of the São Paulo state Court of Justice (TJSP, *Tribunal de Justiça de São Paulo*), one of the 27 Brazilian common state courts (one for each of the country's federative units). Its jurisdiction reaches criminal and civil state-level disputes in virtually all but elections-, military-, and labor-related matters, which fall under the competence of special courts.

Gender biasing in legal settings can take place differently, given the pervasiveness of gender-related stereotyping in culture and social institutions. In court, decisions in which gender stereotypes play a role as part of the motivation seem to emerge regularly in cases of domestic violence [45, 77], custody and other family disputes [55, 104], health care and reproductive rights [56, 81], and rape [81, 34]. Therefore, to analyze such biases in a large scale, sentences issued in these contexts often provide the content under investigation. In the Brazilian justice system, they usually fall under the jurisdiction of state common courts, such as TJSP.

We built and performed experiments over two datasets of decisions issued by the court: Dataset 1, which comprises gender violence-related criminal cases, and Dataset 2, which comprises parental alienation-related civil and criminal cases. In both datasets, search criteria and instances were selected by domain experts for purposes other than this work; however, due to their relevance for our investigation of interest (and with consent and collaboration from the experts), they were also used in our context. The ways in which gender biases play a role in each of the datasets and details on how we made annotations on such biases are described in Section 3.1.2.

Besides the data availability derived from the work of the experts, other criteria behind the choice for the state of São Paulo include: (a) proximity and familiarity with the local court, given that the author, advisors, and collaborators of this work are all based and affiliated in the state (and some of us have worked with the institution); (b) data volume, given that TJSP has the highest amount of legal cases among all of the courts in the country (more than 28 million as of 2022 [29]); (c) ease of collection, since the court's official website and search engines allow for data scraping, and auxiliary tools are avail-

able[1]. We recognize that, by not including data from other courts, we are unable to assess our protocol's performance and limits in a more diverse range of regional particularities.

Reproducibility and data protection issues While we recognize the importance of making datasets publicly available for scientific reproducibility purposes, the court documents used in this work can contain sensitive personal information on the subjects involved, which imposes the need for legal and ethical compliance on its publicization. Details on related issues can be found in previous work by Benatti et al. [69], according to whose proposed guidelines we decided to disclose the datasets by demand with a deed of the undertaking. Instructions on access and use can be found at the official dataset page, hosted at Zenodo [68]. Our codes, on the other hand, are fully available at the project page on GitHub[2]. We argue that this structure of publicization, along with the detailed methodology description provided in the work, makes up for a fair balance between scientific reproducibility and compliance with the informational self-determination of individuals, an elemental dimension of their human rights.

3.1.2 Biases

A core element of the data annotation process — which determines what the models learn from the input texts — is the definition of bias. As explained in Section 1.2, stereotyping is the assumption of one's characteristics or roles due to his or her belonging to a specific group; therefore, gender stereotypes take place when such assumptions are related to one's gender[3].

There are several examples on institutional gender biasing and their harmful consequences for the groups affected by it. In health care, for instance, access to legal abortion-related care can be delayed for younger and single women or women whose pregnancies resulted from violence perpetrated by someone close to them [48]. In legal systems, gender stereotypes can hamper access to proper institutional response in several ways: in cases of sexual violence, for example, the victim's behavior, personal history, and relationship with perpetrator(s) often play a role in how state agents perceive her testimony and other evidentiary elements [32].

Regarding the São Paulo state Court of Justice, for instance, qualitative investigations have shown tendencies of undervaluing victim's testimonies in cases of rape when she does not fulfill the ideal of an "*honest woman*" [35]; an analysis of more than 1,500 cases of domestic violence judged by the court between 2009 and 2018 revealed several biases to be stated by judges, prosecutors, and attorneys to determine whether the violence under analysis had been gender-motivated — for example, physical features or the relationship of the subjects involved [77].

[1]While there are scraping tools for data produced in other courts, each website and search engine has its standard, which hampered the possibility of using other sources.

[2]`https://github.com/ra-ysa/gender_law_nlp`

[3]While we do not delve into definitions of gender — which are better explained by other fields of science —, we recognize the existence of different gender identities and expressions, which unfolds in such stereotyping taking place in diverse forms. For instance, one could be stereotyped due to their assigned gender, their gender identity, their perceived gender, and so on.

Moyses and Severi [77] stress how the recognition of gender-based violence and discrimination should not depend on proof of intention in that sense by the perpetrator(s) but instead can be determined by results, according to the CEDAW. Therefore, a statement issued by a judge is biased if it is not based on evidence, results, or legal statutes but on his or her perception of how gender-weighted features of the subjects involved play a role in the case. Such perceptions often influence if — and to which extent — institutional response will be given to a victim[4].

Gender biases also play a role in decisions regarding family disputes. Severi and Villarroel [104] show how the scientifically unsound concept of parental alienation[5] is used in court against women who report sexual abuse and other forms of family-perpetrated violence on their children. Stereotypes that play a role in such cases usually involve questioning the woman's nurturing capabilities and/or the child's behavior, often based on underlying conservative values on family and relationships.

For Dataset 1 (domestic violence cases), statements that we considered being biased include:

- Statements on the **relationship dynamics** between victim(s) and alleged perpetrator(s). Examples: stressing that aggression was mutual; stressing that the victim went back to, or did not break up with, the perpetrator; describing the relationship as "troubled"; stressing that the aggression was an isolated incident in the context of the relationship;

- Statements on individual gender-weighted features of the **victim** or another **woman** featured in the case. Examples: understanding that the victim's behavior gave cause to the aggression; diminishing the woman's testimony;

- Statements on individual features of the alleged **aggressor**. Examples: describing the defendant's personality as either "moderate" or "twisted" and "prone to crime". While these stereotypes are not gender-weighted per se, they reveal a tendency to address the violence claims when the defendant is perceived as a dangerous person, and dismiss them otherwise;

- **General** statements on legally and/or scientifically unsound conservative values, gender perceptions, and/or the victimhood of women in domestic violence cases. Examples: arguing for preserving the family and protecting "societal values"; claiming women's fragility as a natural feature; deriding on women's fear of reporting their aggressors.

[4]The standard institutional responses for crimes in our legal system include imprisonment and fines. While discussing the effectiveness of such responses goes beyond the scope of this work, we acknowledge that they are not necessarily appropriate instruments to eliminate structural gender violence, and the author stresses her position as a prison abolitionist. Still, punishment issued in court decisions is a proxy of how the state responds to conflict in society; in that sense, biased-motivated decisions for gender-based aggression — resulting in unjust outcomes — represent an inadequate institutional response to the victim(s) and the collectivity, given that no other measures are sought to address the issue.

[5]Brazilian law defines parental alienation as "the interference in the child's or adolescent's psychological development, perpetrated or induced by one of the birth parents, by the grandparents, or by whom has authority, custody, or supervision over the minor, with the goal of repudiating a birth parent or causing damage to the establishment or preservation of the bonds between them" (Law n. 12318/2010, article 2).

For Dataset 2 (parental alienation cases), statements that we considered being biased include:

- Statements on the **relationship dynamics** between mother and the alleged perpetrator. Examples: describing the relationship as "troubled"; stressing that claims of aggression were mutual;
- Statements on individual gender-weighted features of the **mother**. Examples: describing the woman as "prone to emotional outbursts", "egoistic", "self-centered", "arrogant", or "unarticulated";
- Statements on individual features of the alleged **aggressor**. Examples: describing the defendant's reputation as "unblemished" or "prestigious"; describing the defendant as a "good father"; stressing the positive perceptions of the defendant's community on his personality and behavior;
- **General** statements on legally and/or scientifically unsound conservative values, gender perceptions, and/or the child's behavior. Examples: arguing in favor of traditional family settings for proper children's development; diminishing statements expressed by the child; assuming what an expected "abused child behavior" would look like.

We also annotated the target of each biased sentence. While this attribute was not used in our pipeline, it can be helpful in future work. Those include:

- `vitima`: victim;
- `reu`: defendant;
- `test`: witness;
- `mae`: mother;
- `mul`: woman (individually — some specific woman that does not fall under previous categories);
- `abs_mul`: the collectivity of women;
- `abs_reu`: the collectivity of defendants;
- `abs_cri`: the collectivity of children;
- `soc`: society as a whole, abstractly.

3.1.3 Dataset 1: Domestic Violence Cases

The first dataset we use in this work comprises 1,604 decisions issued by TJSP between 2012 and 2019 in domestic violence-related cases. The list of legal cases was manually gathered by domain experts who used the court's official search engine[6], filtering for bodily injury (and associated offenses) cases in the context of domestic violence; only non-confidential and fully digital cases were selected.

Metadata was added to each decision both from automatic extraction and manual annotation. Although most of the metadata was left out of our experimental pipeline (since this work focuses on the biases only), they contain information that could be explored in future research. Additionally, for some attributes, categories can be clustered based on similarity to reduce the dimensionality of the domain.

Extraction of data and metadata

We extracted corresponding PDF files from the court's website using the `pseudoscraper_lesao.R` script from the list of legal case numbers. It is not an actual scraper since the scraping itself is performed, step by step, by the following external functions, provided in the `tjsp` package for R[7]:

- (1) `baixar_cposg`: receives legal case number(s) and directory name as input; returns and saves corresponding HTML file(s) in directory;
- (2) `tjsp_ler_dados_cposg`: receives directory or vector of HTML files as input; returns table with their corresponding metadata;
- (3) `tjsp_baixar_acordaos_cposg`: receives legal case number(s) and directory name as input; returns table with their corresponding metadata and PDF file, which is saved in the directory.

The `pseudoscraper_lesao.R` script is then left to process that information and return it in a single, tidy JSON file. The primary need for processing arises from the difference in the data returned by functions (2) and (3), as shown in Listings 3.1 and 3.2. A description of each attribute is provided in Table 3.1[8].

```
[
  {
    "processo": "00000029220168260556",
    "cd_processo": "RI003UK2H0000",
    "area": "Criminal",
    "assunto": "DIREITO PENAL - Lesão Corporal - Decorrente de Violência
    Doméstica",
    "classe": " Apelação Criminal\n",
    "orgao_julgador": " 6ª Câmara de Direito Criminal\n",
    "origem": "Comarca de Itápolis / Foro de Itápolis / 2ª Vara",
```

[6] https://esaj.tjsp.jus.br/cjsg/consultaCompleta.do

[7] https://github.com/jjesusfilho/tjsp

[8] We stress that these are not official descriptions provided by the court but rather inferred ones.

```
    "outros_numeros": "\n                                    \n
              \n                                          04/2016\n
                    \n                              ",
    "relator": "MACHADO DE ANDRADE",
    "revisor": "JOSÉ RAUL GAVIÃO DE ALMEIDA",
    "secao": " Direito Criminal\n",
    "volume_apenso": "1 / 0"
  }
]
```

Listing 3.1: JSON file example as produced by function (2).

```
[
  {
    "processo": "0000002-92.2016.8.26.0556",
    "data_jugalmento": "2017-06-12",
    "doc_texto": "Acordão Finalizado",
    "decisao": "Acórdão Dr. Machado de Andrade",
    "doc_num": "24",
    "url": "https://esaj.tjsp.jus.br/pastadigital/getPDF.do?nuSeqRecurso
   =00000&nuProcesso=0000002-92.2016.8.26.0556&cdDocumento=40652303&
   conferenciaDocEdigOriginal=false&nmAlias=SG5TJ&origemDocumento=M&
   nuPagina=0&numInicial=251&tpOrigem=2&flOrigem=S&deTipoDocDigital=Ac%
   F3rd%E3os+Eletr%F4nicos&cdProcesso=RI003UK2H0000&cdFormatoDoc=5&
   cdForo=990&idDocumento=40652303-251-0&numFinal=251&sigiloExterno=N"
  }
]
```

Listing 3.2: JSON file example as produced by function (3).

In addition to these original attributes, our script adds the following to the output JSON file for unique identification (therefore the viability of database operations) purposes:

- `cd_documento` (document code): code associated with the document, extracted from the URL;
- `id_documento` (document identification): another number associated with the document, also extracted from the URL;
- `nome_arquivo` (file name): final name designed to the PDF and plain text files of each decision. It comprises the decision date and the legal case number, which makes each file name unique for this dataset.

Having extracted all PDF files and set up the final JSON file, we produce a plain text file with the content of each decision; its specifics are further described in Section 3.1.5. Figure 3.2 shows the structure of files used in this phase.

Annotation

Since relying on minimally supervised approaches, we identified the need to partially annotate the data for information not provided in the extraction phase. For this dataset,

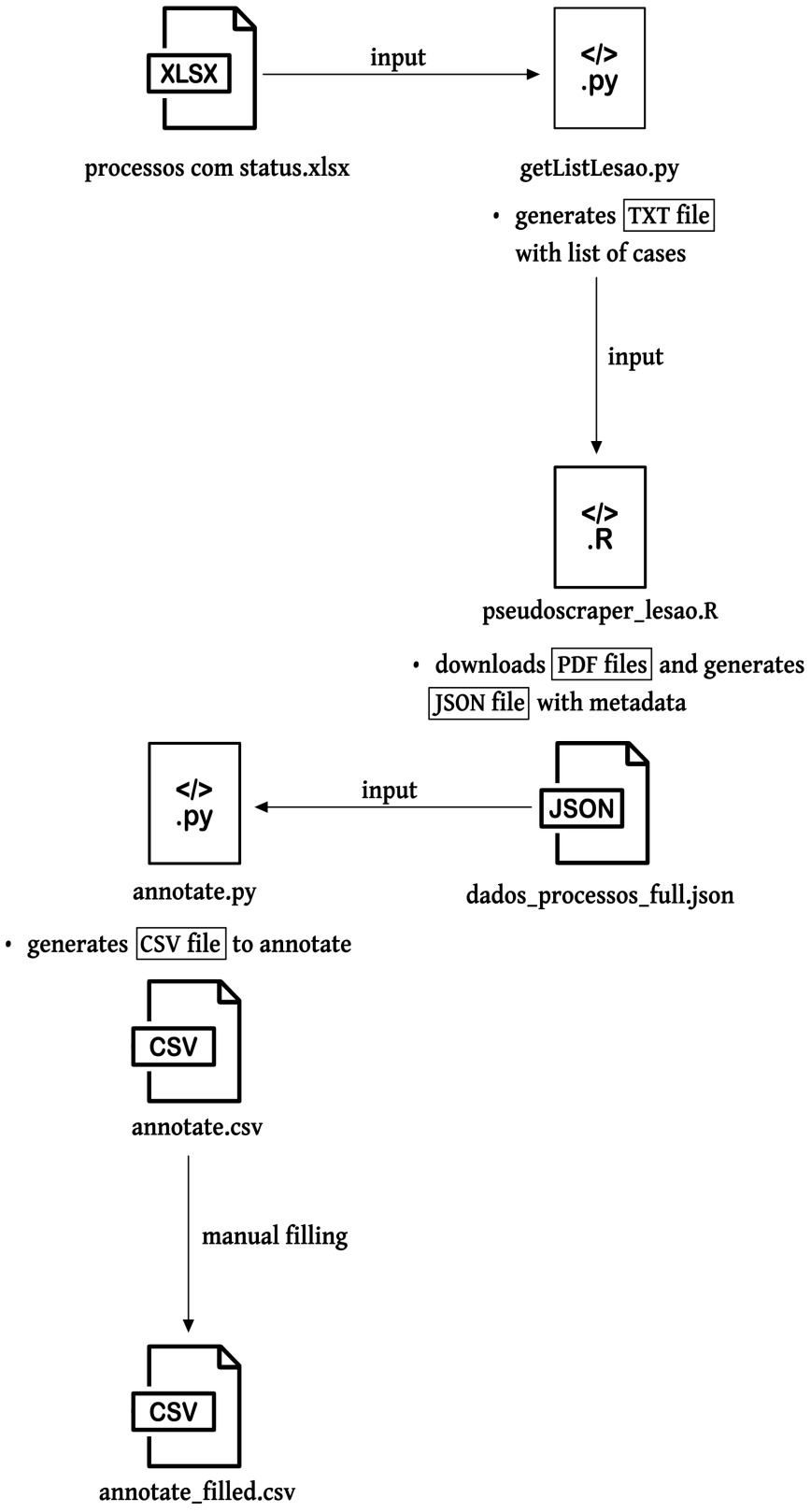


Figure 3.2: Structure of files used to create Dataset 1.

Table 3.1: Data attributes returned by functions (2) and (3).

Original Attribute Name	Meaning	Description
`processo`	legal case number	Main number associated with a legal case
`cd_processo`	legal case code	Code associated with a legal case
`area`	area	Main legal field to which the case belongs (e.g., Criminal Law)
`assunto`	theme	Main topic, inside the field, to which the case belongs (e.g., bodily injury under domestic violence)
`classe`	class	Type of legal procedure that led to the decision (e.g., appeal)
`orgao_julgador`	issuing body	Court body which issued the decision (e.g., 6$^{\text{th}}$ Criminal Law Chamber)
`origem`	origin	Court of first instance where the case started (e.g., judicial district of Itápolis)
`outros_numeros`	other numbers	Other numbers associated with the case (e.g., year)
`relator`	judge-rapporteur	Judge assigned to write the case report for the other judges of the issuing body
`revisor`	reviewer	Judge assigned to review the case report
`secao`	section	Court department to which the case belongs. Possible values: Criminal Law, Private Law, Public Law
`volume_apenso`	volume (folder) / attached cases	Amount of folders and attached cases linked to the original case
`ultima_carga`	last withdrawal	Last update on withdrawal of the case files from the court registry (only appears in 2 cases)
`data_julgamento`	decision date	Date when the decision was issued
`doc_texto`	document text	Piece of text with brief information on the document (e.g., "*acórdão finalizado*" (finished decision))
`decisao`	decision	An indication to where the decision can be found (e.g., digital support) (often empty)
`doc_num`	document number	Internal number associated with the document
`url`	url	Direct link to the document (uniform resource locator)

we randomly selected N documents for annotation, in which N is the integer part of 10% of the number of documents — therefore, for Dataset 1, $N = 160$. Table 3.2 summarizes the added attributes and their domains, followed by a dictionary of values and descriptions of annotation protocols.

Table 3.2: Data attributes added to 10% of the documents in Dataset 1.

Attribute name	Description	Domain[(a)]
`apelante`	identification of the appellant party (anonymized if natural person)	Any combination of name initials; `mpsp`
`apelante_genero`	gender of the appellant	`masc; fem; masc_trans; fem_trans`
`apelado`	identification of the appealed party (anonymized if natural person)	Same as `apelante`
`crime`	legal code(s) of crime(s) under analysis in the case	`cp129p6; cp129p9; cp147; cp150p1; cp330; cp331; cp345; ct306; lcp21; lcp65`
`vitima`	victim(s) main relationship with the defendant	`comp; esposa; namo; ex; fam_ex; rel_ex; filha; ent; irma; irmao; sob; cnh; mae; pai; tia; amiga`
`vitima_genero`	gender of the victim(s)	Same as `apelante_genero`
`pena_original`	time of prison punishment, in months, issued against the defendant in first instance	$[0, 23.5]$
`requer`	main request(s) made by the appellant	`abs; cond; abrand; desclass; cond_sem_qual; afast_altern; maj; conc_mat`
`requer_subsid`	subsidiary request(s) made by the appellant	`abrand; desclass; afast_sursis`
`requer_motivo`	main reason(s) claimed by the appellant	`provas; aut_mater; insig; atip; aus_dolo; leg_def; conf; cp129p4; inimputab; fato; jur; vit; antec; n_antec`
`mp_pj`	position stated by the Public Prosecutor's Office	`s; n; parcial; prej`
`resultado`	final decision on the merits[(b)] of the appeal	`s; n; parcial`
`resultado_razoes`	main reason(s) stated by the court to motivate the result	`provas; aut_mater; fund_legal; bis_in_idem; jur; vit; conf; n_antec; imputab; leg_def; circ; presc; prej`
`pena_atual`	penalty issued against the defendant after the appeal	$[0, 15.17]$; `idem; sursis; sem_sursis; abrand_reg; sem_serv; prej`
`vies`	biased statement(s) identified in the decision	See Section 3.1.2
`vies_alvo`	target(s) of the biased statement(s)	`vitima; reu; test; abs_mul; abs_reu; soc;` See Section 3.1.2

(a) An empty value is part of the domain for all the attributes. It was omitted from the table to avoid redundancy.

(b) Discussions on appeal admissibility and other preliminary issues were not considered, except when they motivated acquittal (e.g., in case of statute of limitations).

- Gender:
  - `masc`, `fem`, `masc_trans`, and `fem_trans` mean, respectively, cisgender masculine, cisgender feminine, transgender masculine, and transgender feminine. While we acknowledge the existence of other genders, their labels are not used in official court records to the best of our knowledge. We assigned gender labels considering: (a) the usual gender attributed to the name of the subject; (b) pronouns used in the decision to refer to the subject; (c) gender descriptions stated in the document. Gender self-identification would have been a primary criterion if stated in the documents, which is not the case.
- Appellant / Appealed parties:
  - In most of the documents, `mpsp` (*Ministério Público do Estado de São Paulo* — state of São Paulo Prosecutor's Office) is the appealed party since, in domestic violence cases, it is the plaintiff by default, and court decisions tend to accept its claims. The appellant is usually the person accused of the crime — and convicted in the first instance —, here identified by initials only. Sometimes, the opposite happens, and the prosecutor appeals against the defendant (e.g., when the first instance grants acquittal); in that case, we use the initials of the appealed person's name in the `apelado` field, and `mpsp` as `apelante`. Very rarely, the court addresses appeals from both the defendant and the prosecutor in a single decision; in that case, we annotate both parties as `apelante` and `apelado`, but the other attributes are labeled considering the defendant's appeal only.
- Crime:
  - `cp129p6`: unintentional bodily injury (Criminal Code, article 129, paragraph 6);
  - `cp129p9`: intentional bodily injury perpetrated in the context of domestic relationships (Criminal Code, article 129, paragraph 9);
  - `cp147`: intimidation (Criminal Code, article 147);
  - `cp150p1`: aggravated trespassing (Criminal Code, article 150, paragraph 1);
  - `cp330`: defiance of the lawful authority of public servants (Criminal Code, article 330);
  - `cp331`: contempt of the work of public servants (Criminal Code, article 331);
  - `cp345`: taking the law into one's own hands (Criminal Code, article 345);
  - `ct306`: driving under the influence (Traffic Code, article 306);
  - `lcp21`: assault (Misdemeanors Act, article 21);
  - `lcp65`: harassment (Misdemeanors Act, article 65[9]).

[9]This article was revoked in 2021 since a new related definition was included in the Criminal Code (stalking, article 147-A); however, it was valid when the facts brought to court and figuring in our dataset happened.

- Victim:
  - `comp`: partner (*companheira(o)*, sometimes *amásia(o)*);
  - `esposa`: wife;
  - `namo`: girlfriend or boyfriend (*namorada(o)*);
  - `ex`: ex-partner, ex-wife/husband, or ex-girlfriend/boyfriend;
  - `fam_ex`: someone belonging to the ex's family;
  - `rel_ex`: someone related to the ex by bonds other than family (e.g., friend or current partner);
  - `filha`: daughter;
  - `ent`: stepdaughter or stepson (*enteada(o)*);
  - `irma`: sister;
  - `irmao`: brother;
  - `sob`: niece or nephew (*sobrinha(o)*);
  - `cnh`: sister-in-law or brother-in-law (*cunhada(o)*);
  - `mae`: mother;
  - `pai`: father;
  - `tia`: aunt;
  - `amiga`: female friend.

  Descriptions of both female and masculine genders were included when either (a) the abbreviation chosen for labeling the category allows for any gender to be included or (b) a case with a male victim of that category appeared in the dataset. We note, however, that the majority of victims are women.

  Relationship status is always stated as it was when the facts happened. When the document provides conflicting information on the relationship between the victim(s) and defendant, we annotate it as informed by the victim(s); if s/he provided conflicting testimonials in different phases of the case, we interpreted the available information and circumstances to decide on a label. If the victim and defendant were legally married but factually separated, we label this attribute as `ex`. If the victim and defendant have a non-clarified companionship bond, the default label is `comp`.

- Penalty:
  - If annotated with a number, the attributes `pena_original` and `pena_atual` state for how long, in months, the punishment of liberty restraint is imposed to last. Decimal parts are computed considering a 30-day month. We do not differentiate between types of prison/jail, nor annotate conditions of imprisonment and other penalties that might have been imposed, such as fines. An amount of zero means acquittal. The upper limit of the domain is established

according to the longest penalty found in the annotated dataset, even if the crime under analysis can entail a longer prison time.

Penalty issued after the appeal (`pena_atual`) can have the same imprisonment length as the original but softened by other conditions, which justifies adding information in that attribute. Its domain of textual labels is:

- `idem`: same imprisonment length as first instance;
- `sursis`: grant of *sursis* (suspended sentence);
- `sem_sursis`: dismissal of *sursis*;
- `abrand_reg`: some form of mitigation of penalty other than length (*abrandamento de regime*);
- `sem_serv`: dismissal or mitigation of community service order (*sem prestação de serviços à comunidade*).

- Requests:
  - `abs`: acquittal (*absolvição*);
  - `cond`: conviction (*condenação*);
  - `abrand`: some form of mitigation of penalty (*abrandamento*);
  - `desclass`: criminal downgrading to a less severe offense (*desclassificação*);
  - `cond_sem_agr`: conviction without the aggravation motive stated in the Criminal Code, article 61 IIf[10] (*condenação sem agravante*);
  - `afast_altern`: dismissal of alternative punishment (*afastamento de pena alternativa*);
  - `maj`: increase of punishment time (*majoração*);
  - `conc_mat`: admission of charge stacking (*concurso material*);
  - `afast_sursis`: dismissal of *sursis* (*afastamento de sursis*).

- Reasoning:
  - `provas`: evidence; this label is used to state an argument of absence, insufficiency, or any inadequacy of evidence to support a conviction;
  - `aut_mater`: used if attribution and materiality of the crime are well established (*autoria e materialidade*);
  - `insig`: criminal pettiness (*insignificância*);
  - `atip`: used to argue that whatever happened cannot be defined as a criminal action (*atipicidade*);
  - `aus_dolo`: absence of intention (*ausência de dolo*);

[10]This article states the aggravation of the punishment to any crime if it is perpetrated (a) under an abuse of authority, or (b) in the context of domestic relationships — if those circumstances are not already stated in the description of the crime itself.

  - `leg_def`: lawful self-defense (*legítima defesa*);
  - `conf`: confession; admission of guilt (*confissão*);
  - `cp129p4`: the existence of moral motivations behind the crime or intense emotions of the perpetrator following unjust provocation made by the victim, as stated in Criminal Code, article 129, paragraph 4;
  - `inimputab`: unimputability (*inimputabilidade*);
  - `imputab`: imputability (*imputabilidade*);
  - `n_antec`: absence of criminal records (*não antecedentes*);
  - `antec`: presence of criminal records (*antecedentes*);
  - `fato`: fact, i.e., anything related to factual elements of the case;
  - `vit`: victim (*vítima*), i.e., any argument related to a deed from the victim at some point during the legal procedures (e.g., retraction of allegations);
  - `fund_legal`: legal ground (*fundamento legal*), i.e. anything directly linked to a legal statement;
  - `bis_in_idem`: double jeopardy;
  - `jur`: analogous to `fund_legal`, but linked to a court precedent instead (*jurisprudência*);
  - `circ`: circumstances (*circunstâncias*), unspecifically;
  - `presc`: statute of limitations (*prescrição*).

- Prosecutor's position (`mp_pj`):
  - The Prosecutor's Office is granted the right to provide an opinion in some court cases as *custos legis* (warden of the law). Such a right derives from an interpretation of its constitutional definition as guardian of social interest (CF, article 127); there is no explicit legal provision behind it. In fact, some argue that such a deed would be unconstitutional under certain conditions since the prosecution is an interested party in many cases [82]. Regardless, having this statement given in court is common practice, and the attribute `mp_pj` represents its content: `s` if in favor of the appeal (*sim*), `n` if against it (*não*), and `parcial` if partially in favor. The same labels are used to state the final decision (attribute `resultado` (result)).
  - Rarely, the first instance prosecutor (`mp` — *Ministério Público*[11]) and the second instance prosecutor (`pj` — *Procuradoria de Justiça*) state two distinct opinions; in that case, they were both annotated in the same field.

[11] *Ministério Público* is the prosecution institution as a whole, but, in this context, refers to the first instance division. In Brazil, generally, *promotor de justiça* is the first instance prosecutor and *procurador de justiça* is the second instance prosecutor. Both of them belong to the (in our case, state level) Prosecutor's Office (*Ministério Público*), but when *Ministério Público* and *Procuradoria de Justiça* are used as distinct elements, the former refers to the first instance and the latter to the second instance divisions.

- Extra considerations:
  - The label `prej` is used when the analysis for an attribute was impaired (*prejudicada*) due to limitations from the case itself;
  - Empty values were used when the corresponding attribute does not exist in the case (e.g., when prosecution appeals, it is common to omit their reasoning from the decision report since it usually repeats the arguments from the original petition);
  - While this dataset consists mostly of court answers to strict sense appeals (i.e., on the merits), six out of the 160 annotated documents answer to an appeal on formal and/or preliminary issues (*embargos*). In those cases, all attributes were left empty since such procedural matters are beyond our scope;
  - All decisions described here result from a trade-off between precision and simplicity of the annotation; different contexts of use might entail different degrees of annotation diversity. We also acknowledge that the annotation process carries intrinsic biases from the researches, which we try to mitigate by (a) describing such process thoroughly, and (b) using domain knowledge as a reference behind each decision.

3.1.4 Dataset 2: Parental Alienation Cases

The second dataset was built from a list of legal cases gathered by domain experts in the context of Severi and Villarroel's work on parental alienation and women's access to justice [92]. They selected pertinent civil and criminal first- and second-instance cases from the four state courts in the Brazilian southeast (TJSP, TJRJ, TJMG, TJES)[12], using their official online search engines with keyword filtering (*"alienação parental"* — parental alienation).

In their work, experts manually downloaded available PDF files. Access to full-content PDF files of decisions was possible for TJSP and TJMG only; TJRJ provided access to, at most, decision abstracts, due to secrecy established over the selected cases. Many cases from TJSP and TJMG were also under secrecy; their documents were available due to a sluggish response of these courts to the Brazilian General Data Protection Act (LGPD, *Lei Geral de Proteção de Dados*). Finally, TJES's website did not provide any information on the cases, which caused them to be kept out of the analysis.

From reading the material and theoretical references, experts settled for a set of metadata to annotate, which was carried out in an Excel file `ap_dados_sudeste.xlxs`. They provided us with this file, which allowed us to perform the following steps to consolidate Dataset 2:

1. standardize column names for all courts;
2. select a subset of cases based on allegations of sexual violence against minors[13];

[12] Respectively, state Courts of Justice of São Paulo, Rio de Janeiro, Minas Gerais, and Espírito Santo.

[13] This decision was based on domain knowledge according to which gender biases under investigation are more frequent in this subset of cases.

3. generate a list of process numbers for cases of interest in a TXT file;
4. generate a CSV file with metadata on this subset only;
5. clean the CSV file, making its cells more computer-friendly (e.g., by stripping or replacing spaces, getting rid of special characters, sorting rows by date, and converting every character to lowercase ones);
6. add columns `vies` and `vies_alvo`;
7. generate a CSV file to be filled with annotation on bias and bias targets;
8. annotate information on bias and bias targets.

Up to item 7, this pipeline was followed for all available annotated data, i.e., for TJMG, TJSP (first and second instance), and TJRJ (second instance) lists of cases. For consistency, however, we performed annotation, analyses, and experiments over data from the second instance of TJSP only.

To download the PDF of selected decisions, we built a `pseudoscraper_ap.R` based on the one used for Dataset 1. For Dataset 2, however, we did not extract metadata from the court's website, since experts had already annotated relevant ones. For 27 cases, we had to do a manual search in the court's website[14] due to scraping limitations; details of which was found in this search are described in Table 3.3.

By the end of the preparation process, Dataset 2 had 49 annotated court decisions issued between 2012 and 2019. Their attributes are described in Table 3.4, which is followed by a dictionary. Figure 3.3 shows the structure of files regarded in the steps described above. We take note of the following special cases:

- Cases 0005806-30.2003.8.26.0028 and 0000998-54.2015.8.26.0450 had more than one PDF file scraped, i.e., more than one associated decision, although experts had annotated one for each. In that case, we kept only the already annotated decisions;
- Cases 0002505-97.2009.8.26.0470, 0314134-38.2009.8.26.0100, and 2009691-82.2015.8.26.0000 were originally annotated regarding procedural decisions even though merit ones existed and were more easily accessible[15]. In these cases, we stored PDF of merit decisions only, while also replacing the corresponding metadata when needed.

Dictionary of attributes

Since annotation for Dataset 2 was entirely built by experts except for bias-related attributes, the domain of each attribute is more detailed, exhaustive, and redundant than in Dataset 1. We kept the original annotation but stress the recommendation for gathering similar values depending on the context of use.

[14] `https://esaj.tjsp.jus.br/cposg/open.do`

[15] For instance, the scraping tool extracted a merit decision only, and/or procedural ones were not available in the court's website.

Table 3.3: Manual search results for cases whose corresponding PDF files could not be extracted through our web scraping tool.

Case number	Retrieved
0024979-66.2007.8.26.0071	No merit decisions available
0041784-65.2004.8.26.0050	
0063816-59.2007.8.26.0050	
0143437-56.2010.8.26.0000	No information available due to secrecy
0008238-76.2008.8.26.0309	
0111201-72.2006.8.26.0006	
0021378-89.2006.8.26.0361	
0036948-48.2004.8.26.0309	No files
0027955-46.2012.8.26.0564	Merit decision available and manually downloaded
1013369-08.2014.8.26.0114	
0314134-38.2009.8.26.0100	
2070734-54.2014.8.26.0000	
2032611-79.2017.8.26.0000	
0002184-85.2010.8.26.0063	
2152311-49.2017.8.26.0000	
2009691-82.2015.8.26.0000	
0000998-54.2015.8.26.0450	
0010265-93.2014.8.26.0156	
2159812-25.2015.8.26.0000	
2029596-68.2018.8.26.0000	
2258473-05.2016.8.26.0000	
2124564-90.2018.8.26.0000	
0056445-78.2006.8.26.0050	
0040998-79.2008.8.26.0050	
0002505-97.2009.8.26.0470	
0010963-26.2012.8.26.0009	

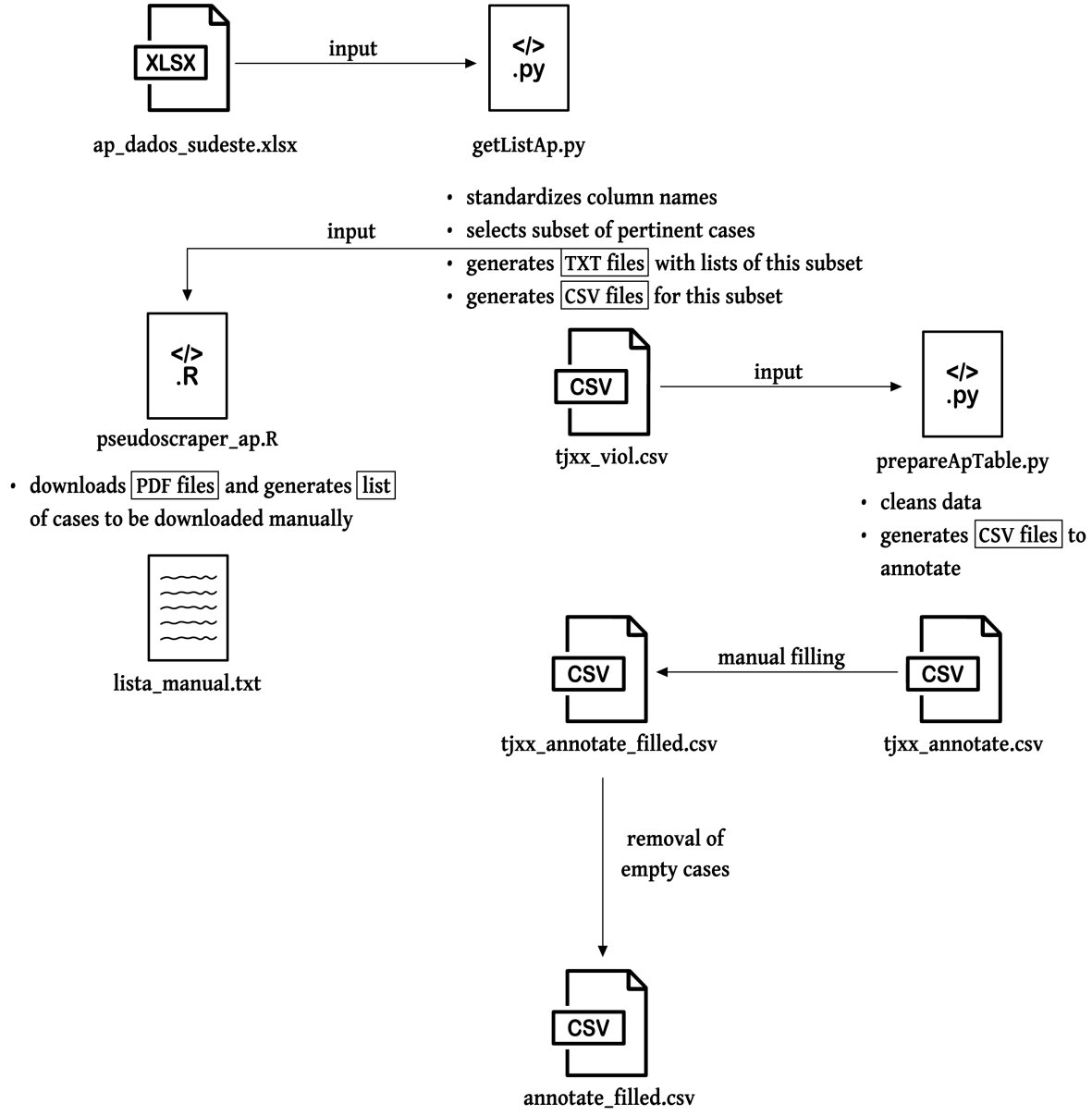


Figure 3.3: Structure of files used to create Dataset 2.

Table 3.4: Data attributes chosen and/or validated by domain experts to label decisions on parental alienation.

Attribute name	Description	Domain[(a)]
`processo`	legal case number	Any number in the format xxxxxxx-xx.xxxx.8.26.xxxx
`relator`	judge-rapporteur	Any judge assigned to operate in TJSP at second instance level
`orgao_julgador`	issuing body	Any second instance court body belonging to TJSP
`data_julgamento`	decision date	Any date in the format yyyy-mm-dd
`tipo_recurso`	type of appeal	See dictionary
`colegialidade`	collegiality degree under which the decision was issued	`acordao` (at least three judges) `decisao_monocratica` (one judge)
`inteiro_teor`	availability of decision's full content	`available`[(b)]
`assunto`	theme	See dictionary
`alegou_ap`	who claimed parental alienation	See dictionary
`acusado_ap`	who was accused of parental alienation	See dictionary
`viol_mulher`	claim(s) of violence against woman	See dictionary
`viol_menor`	claim(s) of violence against minor	See dictionary
`acusado_viol`	who was accused of violence against minor	See dictionary
`resultado_viol`	result on violence allegations	`sim` (yes); `nao` (no); `indicios` (signs)
`prova_viol`	evidence used to decide on claims of violence	See dictionary
`resultado_ap`	result on parental alienation allegations	See dictionary
`prova_ap`	evidence used to decide on claims of parental alienation	See dictionary
`vies`	biased statement(s) identified in the decision	`prej`[(c)]; See Section 3.1.2
`vies_alvo`	target(s) of the biased statement(s)	`vitima; mae; mul; soc; abs_mul; abs_reu; abs_cri; prej`[(c)]; See Section 3.1.2

(a) An empty value is part of the domain for all the attributes. It was omitted from the table to avoid redundancy;
(b) Originally, the contents of all selected second instance decisions from TJSP were available, and we did not change annotation made by experts unless when explicitly stated — which is why the domain for this attribute in our dataset has only one value;
(c) The entry `prej` was used when a PDF file for the decision was unavailable, preventing proper assessment of biases.

- `tipo_recurso`:
  - Criminal merit appeals: `apelacao_criminal`, `habeas_corpus_criminal`[16];
  - Civil merit appeals: `apelacao_civel`, `agravo_de_instrumento`;
  - Criminal appeals on procedural and/or formal issues: `embargos_de_declaracao_criminal`, `recurso_em_sentido_estrito`, `carta_testemunhavel`;
  - Civil appeals on procedural and/or formal issues: `embargos_de_declaracao_civel`, `embargos_infringentes`, `embargos_infringentes_e_de_nulidade`, `agravo_regimental_civel`;
- `assunto`:
  - (`acao_de_`) (case regarding): `atentado_ao_pudor`: assault; `visita`: visitation; `violencia_domestica`: domestic violence; `estupro`: rape; `guarda`: custody; `dissolucao`: dissolution; `danos_morais`: non-material damages; `suprimento_de_consentimento`: consent supply; `guarda_e_visita`: custody and visitation; `alimentos_e_dissolucao`: alimony and dissolution; `alienacao_parental`: parental alienation; `divorcio`: divorce; `ameaca`: menacing; `maus_tratos`: maltreatment; `destituicao_do_poder_familiar`: loss of parental authority; `doacao`: donation; `alimentos_e_guarda`: alimony and custody; `busca_e_apreensao`: search and seizure; `danos_morais_e_materiais`: material and non-material damages;
- `alegou_ap`:
  - `genitor`: birth father; `genitora`: birth mother; `ex-companheiro_pai_que_nao_e_genitor`: former partner / non-birth father; `ambos`: both;
- `acusado_ap`:
  - `genitor`: birth father; `genitora`: birth mother; `ambos`: both; `agravada`: appealed party; `perita`: (female) court expert; `avo_materna`: maternal grandmother; `avos_paternos`: paternal grandparents; `atual_companheiro_da_genitora`: current birth mother's partner; `genitora_e_sogra`: birth mother and mother-in-law;
- `viol_mulher`:
  - `agressao`: physical offense; `lesao_corporal`: bodily injury; `existencia_de_medida_protetiva`: presence of restraining order; `ameaca_e_agressao`: menacing and physical offense;
- `viol_menor`:

[16] In Brazilian legal system, the *habeas corpus* is not an appeal but rather a cause per se; detailing such a technicality, however, is beyond the scope of this work.

  - `abuso_sexual`: sexual abuse; `ameaca_e_abuso_sexual`: menacing and sexual abuse; `maus_tratos_e_abuso_sexual`: maltreatment and sexual abuse; `acusacao_anterior_de_abuso_sexual`: former complaint of sexual abuse; `lesao_corporal`: bodily injury; `agressao`: physical offense;

- `acusado_viol`:

  - `genitor`: birth father; `madrasta`: stepmother; `companheiro_da_genitora`: birth mother's partner; `ex-companheiro_da_genitora`: former birth mother's partner; `companheira_do_genitor`: birth father's partner; `pai_adotivo`: adoptive father; `filho_da_companheira_do_genitor`: birth father's partner's son; `rapazes_que_moram_com_a_genitora`: men who live with the birth mother; `esposo_da_avo_materna_e_pai_da_genitora`: maternal grandmother's husband and birth mother's father; `ambos`: both;

- `prova_viol`:

  - `in_dubio_pro_reo`: in dubio pro reo; `estudo_psicossocial`: psychosocial assessment; `exame_iml`: forensic exam; `pericia`: expert examination; `estudo_psicologico`: psychological assessment; `exame`: exam; `necessidade_de_instrucao_probatoria`: evidence collection needed; `arquivamento_do_inquerito_policial`: criminal investigation shelved; `rejeicao_da_denuncia`: complaint rejected; `processo_penal_arquivado`: criminal procedure shelved; `nao_houve_oferecimento_da_denuncia`: complaint not presented; `condenacao_criminal`: criminal conviction; `conselho_tutelar`: child protection services;

- `resultado_ap`:

  - `alienacao_parental_evidenciada`: evidence of parental alienation; `sindrome_da_alienacao_parental_evidenciada`: evidence of parental alienation syndrome; `nao_ocorrencia`: no parental alienation; `nao_ocorrencia_sindrome`: no parental alienation syndrome; `indicios_de_alienacao_parental`: signs of parental alienation; `necessidade_de_instrucao_probatoria`: evidence collection needed; `materia_estranha_ao_processo`: non-pertinent issue;
    `existencia_de_acao_declaratoria_de_alienacao_parental`: parental alienation formerly acknowledged; `citacao_de_jurisprudencia_pelo_tribunal`: court mentioned precedents;

- `prova_ap`:

  - `estudo_psicossocial`: psychosocial assessment; `estudo_psicologico`: psychological assessment; `pericia`: expert examination; `prova_emprestada`: evidence from another case; `em_outro_processo`: idem.

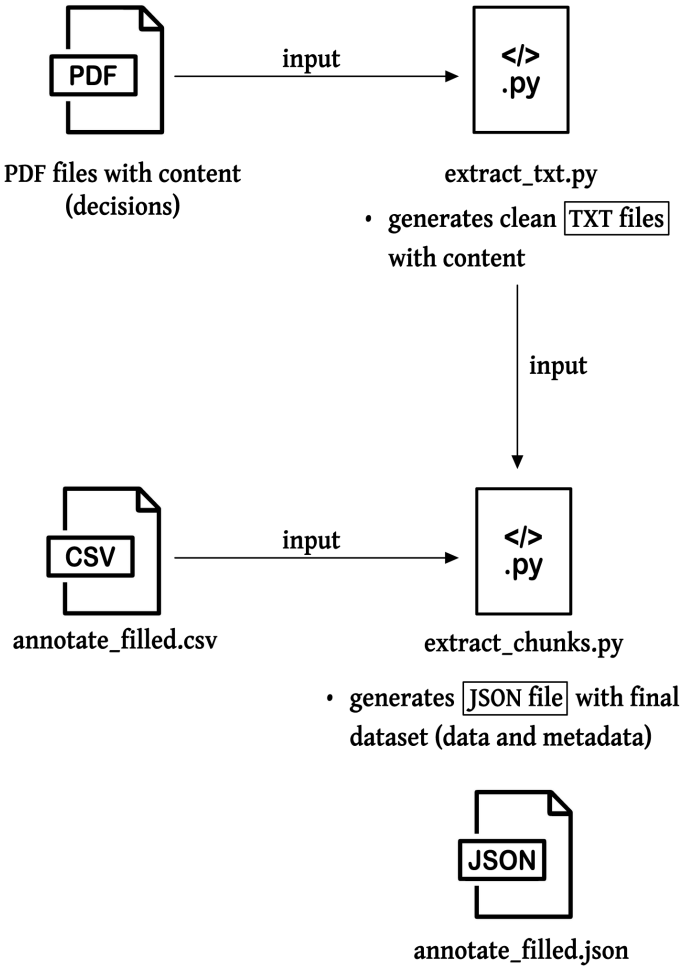


Figure 3.4: Structure of files used to prepare Datasets 1 and 2.

3.1.5 Data Preparation

Text to be used as input to the models went through a preparation phase that involved (a) cleaning and (b) chunk extraction, which we describe in the following sections. Figure 3.4 shows the structure of files used in this phase, which can be briefly described as follows:

- `extract_txt.py`: with PDF files as input, this script executes functions of cleaning and plain text generation;
- `extract_chunks.py`: with the plain texts and the annotation-filled CSV file as inputs, this script generates a JSON file for each dataset with its data and metadata.

Text cleaning

Plain text extracted from the PDF files comes with some noisy elements. Although attention-based models do not require them to be resolved, some of these elements, in

our case, were known to be irrelevant, such as headers, electronic signatures, special characters, and some punctuation marks. The `clean_text` function inside `extract_txt.py` performs the task of removing such elements and solving the following cases, which are specific to our data:

- Codes and numbers were replaced by the tags *[CODE]* and *[NUMBER]* — which were then removed;
- Patterns such as *(...)ABb* and *(...)aB* were replaced by *(...)A Bb* and *(...)a B*, therefore splitting words that came mistakenly aggregated from the PDF file;
- Patterns such as *JUDICIÁRIOTRIBUNAL*, *JUSTIÇAPODER*, *NOMERELATOR(A)*, *PALAVRAACORDAM*, and *PALAVRANUMBER* were fixed to their correct forms *JUDICIÁRIO TRIBUNAL*, *JUSTIÇA PODER*, *NOME RELATOR(A)*, *PALAVRA ACORDAM*, and *PALAVRA NUMBER*.

Headers, signatures, extra spaces, new lines, and other noisy elements were removed from the text. As of punctuation marks, we kept exclamation points, question marks, hyphens, commas, semicolons, and what we call "legal" periods — i.e., those that precede an uppercase letter.

Figure 3.5 shows an example extracted from Dataset 1, in which a plain text with noisy elements is generated from the original PDF, and is then cleaned to generate a final plain text instance.

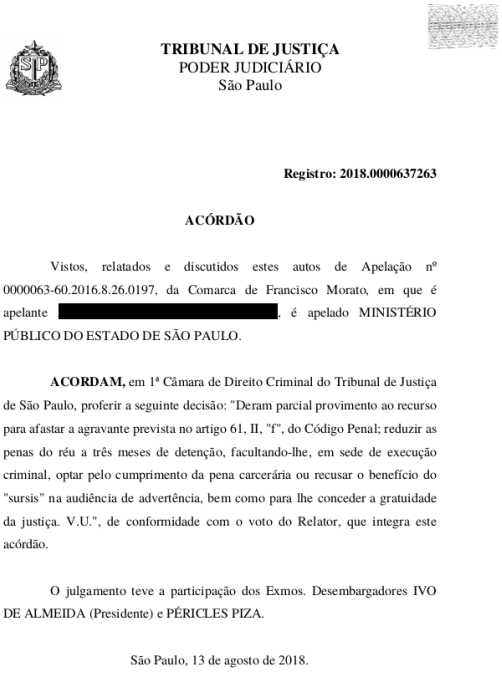
Chunk extraction[17]

Having the plain, clean text corresponding to each annotated decision is still not enough to feed our models of interest due to (a) its size and (b) its content. Attention-based networks typically require input text not to be larger than 512 tokens [102, 96]. There are techniques to deal with longer texts, such as the Long-Document Transformer [9]; however, applying them to our data would be challenging to the point of going beyond the scope of this work, considering that our texts are written in Portuguese and are often even longer than the sizes accepted by such models.

Additionally, court decisions display a significant amount of content that would be likely meaningless for automatic learning. Depending on the task for which the model is being trained, choosing specific parts of the content increases the odds of the learning happening. For instance, the biases that interest us tend to appear in the middle of the text amidst a broader argumentation context; other information, such as the verdict itself, is typically found in the first and/or last paragraphs.

To overcome these issues, we applied a protocol of chunk extraction over the data. We define a chunk as an excerpt from a text — with no particular size but expected to be necessarily smaller than the whole content and ideally have a word count below 512 (also considering that tokenization might increase word count since a single word is typically unfolded in more than one token). The size of a chunk is defined by the number

[17] Not to be confused with *chunking* [61, 24].



TRIBUNAL DE JUSTIÇA
PODER JUDICIÁRIO
São Paulo

Registro: 2018.0000637263

ACÓRDÃO

Vistos, relatados e discutidos estes autos de Apelação nº 0000063-60.2016.8.26.0197, da Comarca de Francisco Morato, em que é apelante [redacted], é apelado MINISTÉRIO PÚBLICO DO ESTADO DE SÃO PAULO.

ACORDAM, em 1ª Câmara de Direito Criminal do Tribunal de Justiça de São Paulo, proferir a seguinte decisão: "Deram parcial provimento ao recurso para afastar a agravante prevista no artigo 61, II, "f", do Código Penal; reduzir as penas do réu a três meses de detenção, facultando-lhe, em sede de execução criminal, optar pelo cumprimento da pena carcerária ou recusar o benefício do "sursis" na audiência de advertência, bem como para lhe conceder a gratuidade da justiça. V.U.", de conformidade com o voto do Relator, que integra este acórdão.

O julgamento teve a participação dos Exmos. Desembargadores IVO DE ALMEIDA (Presidente) e PÉRICLES PIZA.

São Paulo, 13 de agosto de 2018.

(a) PDF

```
TRIBUNAL DE JUSTIÇAPODER JUDICIÁRIOSão PauloRegistro:
2018.0000637263ACÓRDÃOVistos, relatados e discutidos estes
autos de Apelação nº 0000063-60.2016.8.26.0197, da Comarca de
Francisco Morato, em que é apelante [redacted]
[redacted], é apelado MINISTÉRIO PÚBLICO DO ESTADO DE SÃO
PAULO.ACORDAM, em 1ª Câmara de Direito Criminal do Tribunal
de Justiça de São Paulo, proferir a seguinte decisão: "Deram
parcial provimento ao recurso para afastar a agravante
prevista no artigo 61, II, "f", do Código Penal; reduzir as
penas do réu a três meses de detenção, facultando-lhe, em
sede de execução criminal, optar pelo cumprimento da pena
carcerária ou recusar o benefício do "sursis" na audiência de
advertência, bem como para lhe conceder a gratuidade da
justiça. V.U.", de conformidade com o voto do Relator, que
integra este acórdão.O julgamento teve a participação dos
Exmos. Desembargadores IVO DE ALMEIDA (Presidente) e PÉRICLES
PIZA.São Paulo, 13 de agosto de 2018. MÁRIO DEVIENNE FERRAZ
RELATORAssinatura Eletrônica

PODER
JUDICIÁRIO
```

(b) Noisy plain text

```
Registro ACÓRDÃO Vistos, relatados e discutidos estes autos
de Apelação n -, da Comarca de Francisco Morato, em que é
apelante [redacted], é apelado MINISTÉRIO
PÚBLICO DO ESTADO DE SÃO PAULO. ACORDAM, em Câmara de Direito
Criminal do Tribunal de Justiça de São Paulo, proferir a
seguinte decisão Deram parcial provimento ao recurso para
afastar a agravante prevista no artigo , II, f, do Código
Penal; reduzir as penas do réu a três meses de detenção,
facultando-lhe, em sede de execução criminal, optar pelo
cumprimento da pena carcerária ou recusar o benefício do
sursis na audiência de advertência, bem como para lhe
conceder a gratuidade da justiça. VU, de conformidade com o
voto do Relator, que integra este acórdão O julgamento teve a
participação dos Exmos. Desembargadores IVO DE ALMEIDA
Presidente e PÉRICLES PIZA. São Paulo, de agosto de . MÁRIO
DEVIENNE FERRAZ RELATOR  PODER JUDICIÁRIO TRIBUNAL DE JUSTIÇA
DO ESTADO DE SÃO PAULO Apelação n- Comarca de Francisco
Morato Apelação n - Vara de Francisco Morato Apelante [redacted]
[redacted] Apelado Ministério Público do Estado de São
Paulo Voto n Inconformado com a decisão do MM Juiz de Direito
da Vara da Comarca de Francisco Morato, que o condenou como
incurso no artigo , , do Código Penal, a três meses e quinze
dias de detenção, em regime prisional aberto, concedido o
sursis, pelo prazo de dois anos, mediante condições, por ter,
no dia de setembro de , por volta de h min, na Avenida
[redacted], naquela cidade,
ofendido a integridade corporal de sua ex-companheira [redacted]
[redacted], provocando-lhe lesão corporal de natureza
leve, o réu [redacted] apelou em busca da
absolvição por insuficiência de provas ou quanto ao dolo,
alternativamente pretendendo o benefício da justiça gratuita,
a exclusão da agravante reconhecida, a redução das penas e o
afastamento do sursis fixado na sentença. Regularmente
processado o recurso, pelo desprovimento opinou a douta
Procuradoria de Justiça É a síntese do necessário A
absolvição é meta impossível de ser alcançada, em face do que
```

(c) Clean plain text

Figure 3.5: Excerpts of the PDF file, the noisy and clean text from an instance of Dataset 1. Black stripes were manually added to preserve the identity of private subjects figured in the decision.

of sentences it contains; a sentence is delimited by the presence of punctuation marks that suggest the completion of content (question marks, exclamation points, semicolons, or periods). Each chunk belongs to one of the following categories:

- **Introductory chunk:** extracted from the beginning of the text, defined as its first N sentences;

- **End chunk:** extracted from the end of the text, defined as its last N sentences;
- **Random chunk:** extracted from a random position of the text, defined by a random sentence plus its context (the N sentences above and the N sentences below it);
- **Full text chunk:** extracted in the context of the validation pipeline (explained in Section 3.3), which requires the whole content of each decision to be split in chunks;
- **Bias chunk:** defined by a sentence that contains bias as annotated plus its context (the N sentences above and the N sentences below it).

Having annotated the data for attributes of interest, we can take advantage of knowing where each piece of information is most likely to be found, dismissing insignificant parts of the content[18]. Therefore, in the training phase, each decision is represented by a chunk, or set of chunks, which make sense — according to a domain expertise-related decision — for the task being performed.

For each decision, we extracted the following standard sets of chunks, which are incorporated in the JSON files corresponding to the datasets:

- One introductory chunk of size `N_SENTENCES_INTRO` (default value = 5);
- One end chunk of size `N_SENTENCES_END` (default value = 10);
- Ten random chunks of size `N_SENTENCES_RANDOM` (default value = 3);
- `N_CHUNKS` full text chunks of size `len(sentences) // N_CHUNKS`, in which `len(sentences)` is the amount of sentences in the decision and `N_CHUNKS` is defined by the user (default = 32);
- A variable amount of bias-related chunks, according to the following guidelines:
  - For biased decisions, each annotated biased sentence is used as the seed to generate context chunks of size $N = 1, 2$, and 3, making a total of 3× `occurrences of each biased sentence` bias-related chunks for each of these decisions;
  - For non-biased decisions, random sentences are used as the seed to generate context chunks of size $N = 2$ and 3, making a total of 2 bias-related chunks for each of these decisions.

    We chose to generate more bias-related chunks for the biased texts to improve the balance between bias and no-bias classes — since the original amount of biased decisions are around 18% for Dataset 1 and 26% for Dataset 2. Occasionally, annotated biased sentences appear close to each other in the original text, which causes some context chunks to have the same content.

[18] We recognize that such process carries intrinsic biases from the researchers.

3.2 Experimental Pipeline

We developed an experimental pipeline of supervised learning for the task of binary classification over the annotated portion of each one of our datasets. The classification was performed over the bias attribute only; other attributes could be explored as learning inputs in future work, as discussed in Section 5.2.

Figure 3.6 illustrates our experimental pipeline. The original annotated texts, stored in a JSON file `annotation_filled.json`, are encoded with the BERTimbau tokenizer; the dataset is then split in proportions of 72:18:10 for training, validating, and testing, respectively. Training and validation portions are fed into the classification model while testing instances are left for the validation pipeline, described in Section 3.3. Details on the parameters of each experiment are presented in Chapter 4.

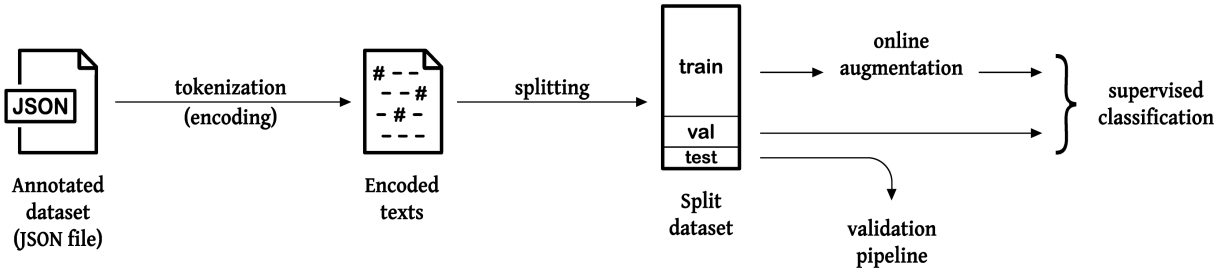


Figure 3.6: A representation of the experimental pipeline.

3.3 Evaluation and Validation Methods

Each experiment was evaluated based on loss metrics and balanced accuracy for training and validation sets over 20 epochs. We also produced confusion matrices to help visualize model performance on the epochs with the lowest loss value.

The validation of our protocol over the test set was hampered by the low availability of data, since only 10% of the annotated portion of each dataset was set aside for testing — whose results, therefore, are not statistically significant in our context. However, the validation pipeline can be explored in future work with larger amounts of annotated data, besides serving as a tool for final users who are interested in using our model over full, non-annotated decisions.

In this phase, for each dataset, we chose the version of the trained model which showed the best balanced accuracy value in the validation set, over all experiments described in Chapter 4. The whole content of each decision is split in `N_CHUNKS` of size `len(sentences) // N_CHUNKS)`, as explained in Section 3.1.5; for a given decision, if any of its chunks is classified as biased by the model, all of its chunks are given the same classification. This protocol considers that, when not in the learning phase, detecting bias in one portion of a decision is equivalent to detecting the whole decision as a biased one.

Chapter 4

Experiments, Results, and Discussion

This chapter presents details on the experimental pipeline, their main results — which are summarized in Table 4.1 —, and discussions. The complete output logs for each experiment, including all resulting graphs and confusion matrices, are available at the project's GitHub page.

4.1 Data Augmentation

Data augmentation, the creation of synthetic data to be used as input in automatic learning models, is a possible approach to overcome the issue of low data availability [7]. It becomes then a powerful ally in our context of partial data annotation — given that augmenting data is usually cheaper than annotating it, especially when annotation is too domain-dependent, which is the case.

Synthetic text can be derived from original ones through different techniques, of which we chose synonym replacement. It consists of changing a word for a synonym, thus (theoretically) preserving the original meaning and allowing the model to learn from a more diverse range of data. We performed online (during training) synonym replacement according to the following steps for each input text from the training set:

- for every word of the text aside from stop words[1], we flip a coin of `weight = 0, 0.3, 0.7` or `1.0` to decide if it will be changed for a synonym;
- in case the change happens,
  - if the input text is labeled as biased, the word is replaced by (a) a synonym extracted from a domain-specific synonym dictionary `BIAS_SYN_DICT`, which we built from scratch based on the most bias-associated words in the annotated biased chunks, or (b) a synonym extracted from a general dictionary[2], in case the word to be replaced does not exist in `BIAS_SYN_DICT`. Otherwise,

[1] Stop words are those with less semantic significance, usually the ones that appear frequently in text — such as articles and prepositions. To filter them out of synonym replacement, we used the Natural Language Toolkit corpus of Portuguese stop words (`https://www.nltk.org/howto/portuguese_en.html`).

[2] We used the Brazilian Portuguese synonym dictionary from OpenWordnet-PT [39].

– the word is replaced by a synonym extracted from a general dictionary.

Noticeably, there is a trade-off between the augmentation weight (expected to correlate to model learning performance) and the processing cost of the experiment, as shown in Table 4.1.

4.2 Model and Parameters

The task of binary classification on bias for Datasets 1 and 2 was learned by the BERTimbau model [96]. While originally trained for masked-language modeling[3], the model can be used as a classifier through its Hugging Face interface[4]. We imported the `bert-base-portuguese-cased` version of the model as an `AutoModelForSequenceClassification`.

While the original BERTimbau embeddings were preserved (freezed) during learning, we fine-tuned some of the model's parameters with our inputs. For each dataset and augmentation weight, two fine-tuning protocols were used:

1. Baseline protocol (`BertBaseline` class): the whole original network is preserved (freezed) except for the last layer, where the actual classifier is;

2. Deep fine-tuning protocol (`BertFineTuner` class): we preserve (freeze) all but the last `N_L = 5` layers of the network, over which the fine-tuning is performed. The value of `N_L` was chosen empirically after preliminary experiments showed the optimal value to be between 4 to 6 since overfit increases significantly for $\texttt{N\_L} \geq 7$. Processing costs also increase prohibitively for higher `N_L` values.

Having two datasets, four augmentation weights, and two fine-tuning protocols, we performed 16 final training experiments. In all of them, the following parameters were used: (a) a batch size of 32 instances; (b) 20 epochs of training; and (c) a loss-based optimization with PyTorch's `AdamW` optimizer and `CosineAnnealingLR` scheduler.

4.3 Findings

Our main experimental results are summarized in Table 4.1. It shows, for each dataset and fine-tuning protocol, the best balanced accuracy for training and validation sets, as well as the balanced accuracies observed in the last epoch of each experiment. For each dataset, we chose the trained model with the best balanced accuracy value in the validation set to be used in the testing pipeline.

Data augmentation helped make up for the low availability of annotated data. In most experiments, values of balanced accuracy increase with the augmentation weight while overfitting decreases. In the deep fine-tuning protocol, an augmentation weight of 0.3 increased accuracy significantly, especially in Dataset 1. Therefore, combining this strategy with partial data annotation is helpful to achieve a reasonable trade-off between

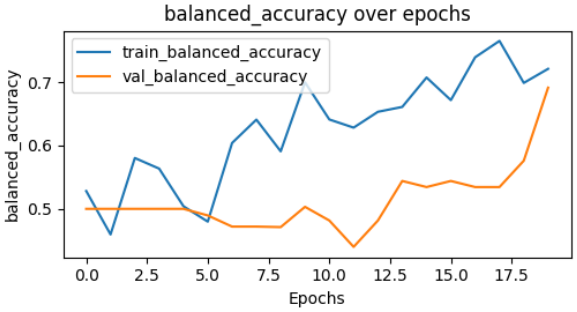
[3]See `https://github.com/neuralmind-ai/portuguese-bert`.
[4]See `https://huggingface.co/neuralmind`.

Fine-tuning protocol	Augmentation weight	Best balanced accuracy (epoch)	Last balanced accuracy (epoch 19)	Processing time (s)
Baseline	0	0.7654 (T) (17) 0.6915 (V) (19)	0.7213 (T) 0.6915 (V)	132.3014
	0.3	0.7492 (T) (16) 0.7131 (V) (19)	0.7299 (T) 0.7131 (V)	190.1523
	0.7	0.7554 (T) (19) 0.7332 (V) (16)	0.7554 (T) 0.7035 (V)	257.3917
	1.0	0.7295 (T) (19) 0.7452 (V) (16)	0.7295 (T) 0.7252 (V)	257.3917
Deep	0	1.0000 (T) (10) 0.8574 (V) (19)	1.0000 (T) 0.8574 (V)	211.4069
	0.3	1.0000 (T) (10) 0.8886 (V) (7)	1.0000 (T) 0.8670 (V)	272.4788
	0.7	1.0000 (T) (13) 0.8670 (V) (12)	1.0000 (T) 0.6691 (V)	334.9803
	1.0	1.0000 (T) (14) 0.8574 (V) (8)	1.0000 (T) 0.8053 (V)	384.0748
Baseline	0	0.7450 (T) (14) 0.8393 (V) (19)	0.7000 (T) 0.8393 (V)	81.7429
	0.3	0.7374 (T) 16) 0.8571 (V) (19)	0.7000 (T) 0.8571 (V)	118.5765
	0.7	0.7459 (T) (18) 0.8571 (V) (17)	0.6933 (T) 0.8571 (V)	159.8615
	1.0	0.7247 (T) (16) 0.8790 (V) (19)	0.6789 (T) 0.8790 (V)	190.218
Deep	0	1.0000 (T) (8) 0.8790 (V) (3)	1.0000 (T) 0.8214 (V)	126.0416
	0.3	1.0000 (T) (9) 0.9405 (V) (5)	1.0000 (T) 0.8393 (V)	162.2621
	0.7	1.0000 (T) (11) 0.9405 (V) (9)	1.0000 (T) 0.8750 (V)	204.3155
	1.0	1.0000 (T) (11) 0.9583 (V) (11)	1.0000 (T) 0.8750 (V)	233.6886

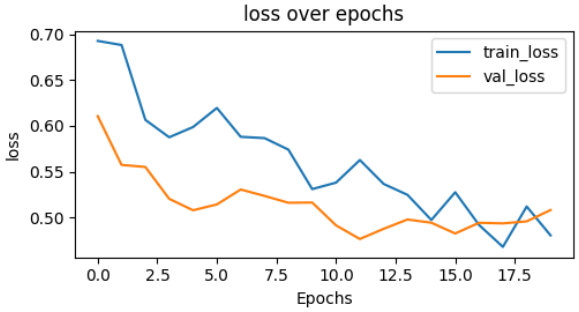
Table 4.1: Summary of the main experimental results. Cells in orange (warm color) represent results for Dataset 1; cells in blue (cold color) represent results for Dataset 2. Label (T) stands for *training* and (V) for *validation*. In case the same best value is observed in more than one epoch, we report the one where it appears first.

the cost of building a quality dataset and getting good performance in the task that we want a model to learn. Figure 4.1 shows how coherence between training and validation sets seems to increase with augmentation weights (in this case, for Dataset 1), even in the simplest fine-tuning protocol.

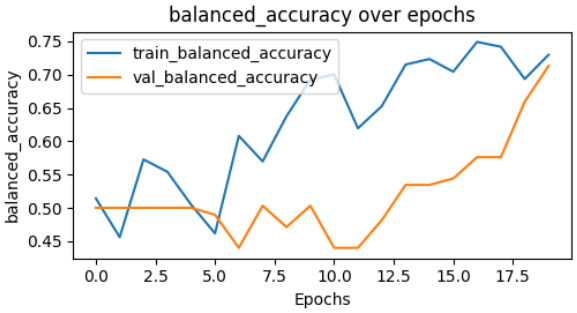
Differences between results for Datasets 1 and 2 also reveal the influence of the amount of data on model performance. When in baseline fine-tuning protocol, for instance, the



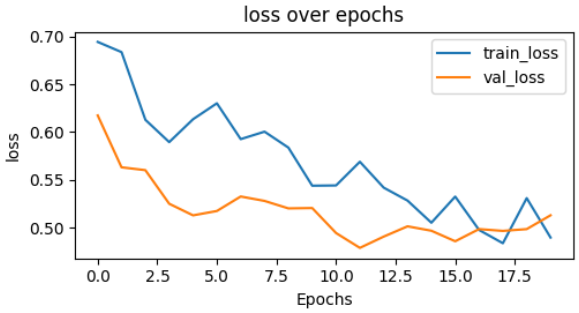
(a) augmentation weight = 0



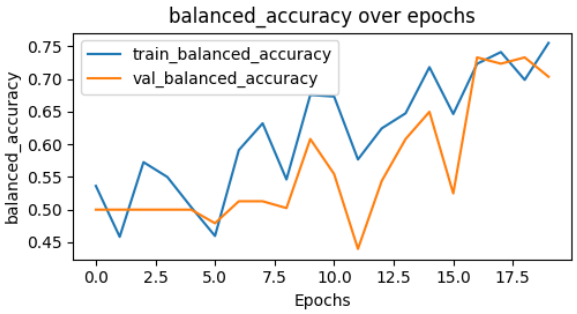
(b) augmentation weight = 0



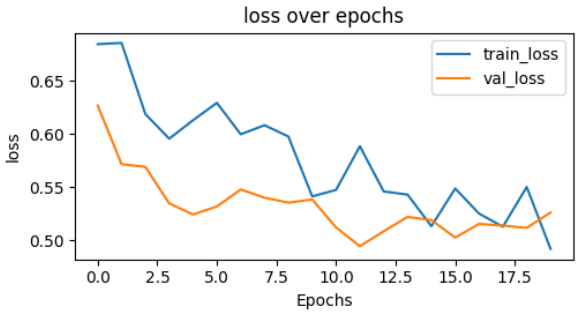
(c) augmentation weight = 0.3



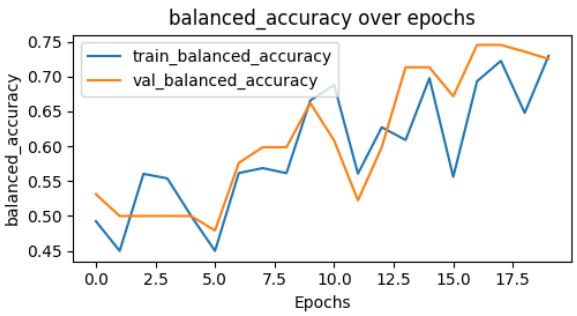
(d) augmentation weight = 0.3



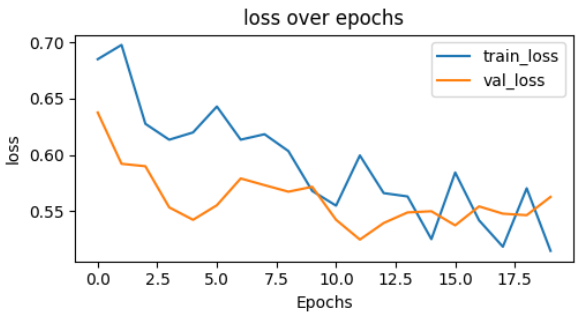
(e) augmentation weight = 0.7



(f) augmentation weight = 0.7



(g) augmentation weight = 1.0

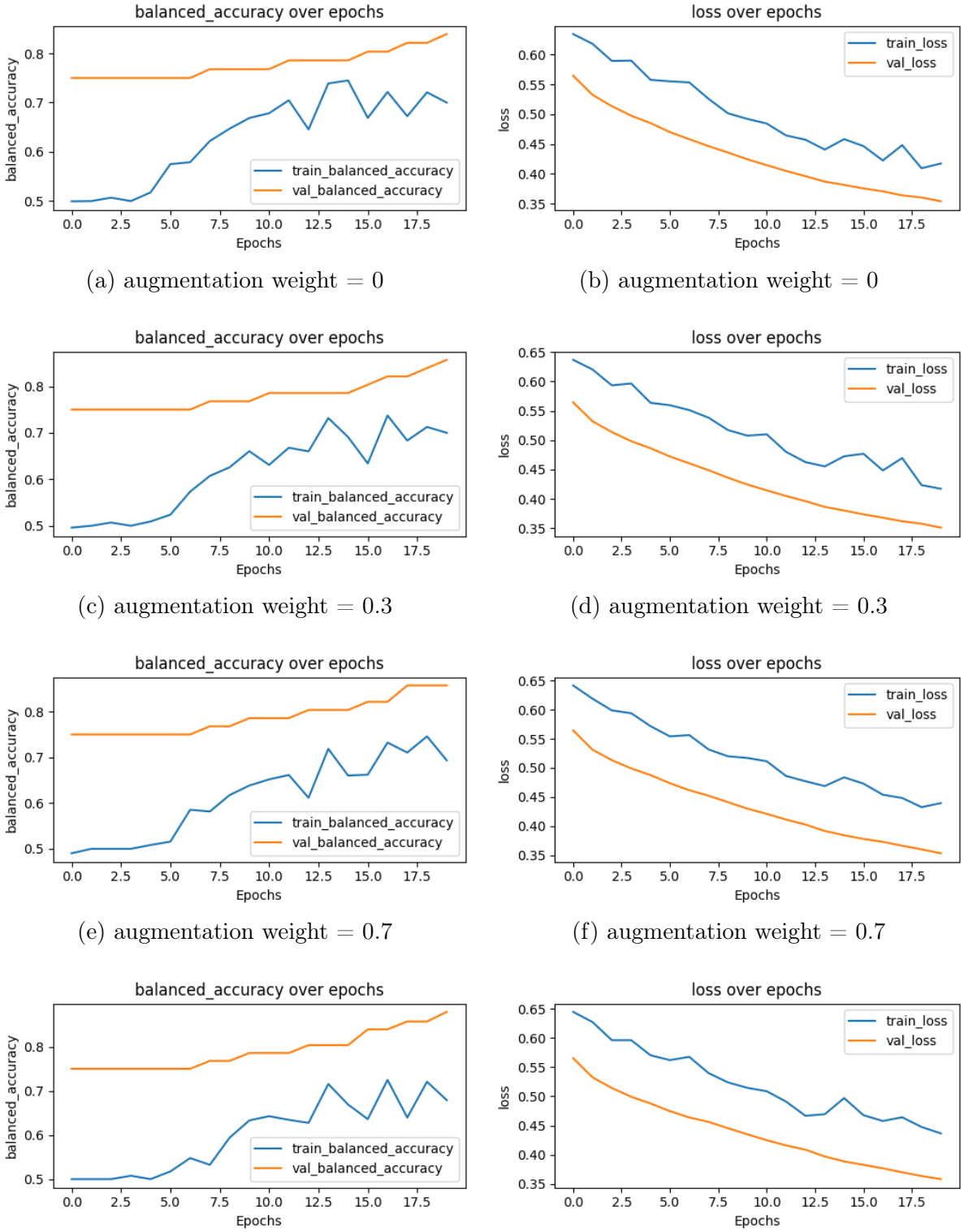


(h) augmentation weight = 1.0

Figure 4.1: Graphs of balanced accuracy and loss over epochs for Dataset 1 (baseline fine-tuning protocol). Each line represents results for a different augmentation weight.

model underfits when learning from Dataset 2, which is more than three times smaller than Dataset 1 — therefore, it would probably require more training epochs to achieve

better results, as shown in Figure 4.2.



(a) augmentation weight = 0

(b) augmentation weight = 0

(c) augmentation weight = 0.3

(d) augmentation weight = 0.3

(e) augmentation weight = 0.7

(f) augmentation weight = 0.7

(g) augmentation weight = 1.0

(h) augmentation weight = 1.0

Figure 4.2: Graphs of balanced accuracy and loss over epochs for Dataset 2 (baseline fine-tuning protocol). Each line represents results for a different augmentation weight.

When in the deep fine-tuning protocol, the model overfits when learning from both

datasets, but it happens slower in Dataset 2 — and with better balanced accuracy values in the validation set when compared to Dataset 1. Figures 4.3 and 4.4, which show graphs of loss and balanced accuracy over epochs for both datasets when in deep fine-tuning, help us visualize the tendency for overfitting.

Overall, overfit is more prevalent in experiments that used the deep fine-tuning protocol over the baseline ones; they also showed better evaluation metrics and less confusion between classes. For instance, Tables 4.2 and 4.3 show confusion matrices of results over the validation set of Dataset 2 at each fine-tuning protocol, using the maximum augmentation weight. While deep fine-tuning protocol showed a slight increase in the number of false negatives, overall classification was more accurate, with a significant decrease of false positives.

Table 4.2: Confusion matrix for results over the validation set of Dataset 2 (baseline fine-tuning protocol, augmentation weight = 1.0).

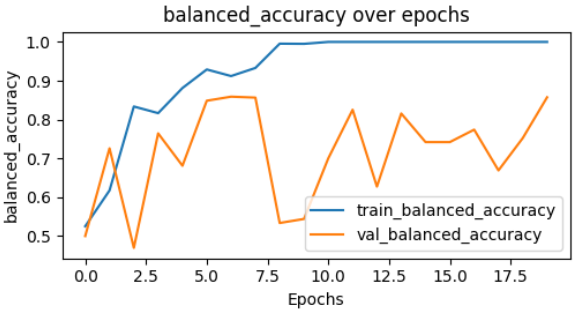
		Predicted class	
		Non-biased (%)	Biased (%)
Actual class	Non-biased	21.05	15.79
	Biased	2.63	60.53

Table 4.3: Confusion matrix for results over the validation set of Dataset 2 (deep fine-tuning protocol, augmentation weight = 1.0).

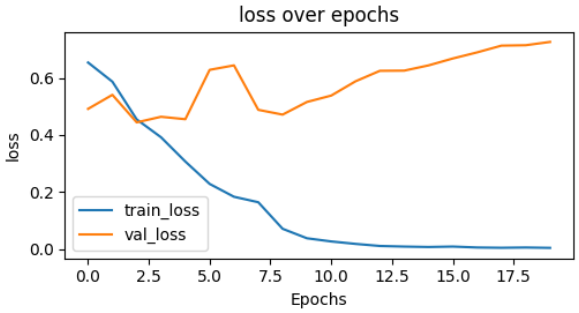
		Predicted class	
		Non-biased (%)	Biased (%)
Actual class	Non-biased	34.21	2.63
	Biased	5.26	57.89

Using an augmentation weight above zero, combined with the deep fine-tuning protocol, seems to be the best approach regarding model performance between the ones we tested; however, in future work, it should be enhanced with strategies to mitigate overfitting.

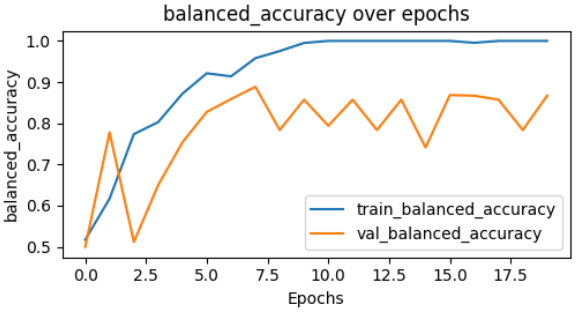
Although our approach seems to make sense from an automatic learning perspective, lack of proper validation prevents us from assessing the generalization capabilities of the models. Future work, as discussed in Section 5.2, could address this issue with larger datasets — which could include collecting new data and/or enriching Datasets 1 and 2 with more annotated instances. Adapting the protocol to be more annotation independent would also allow for exploring other validation possibilities.



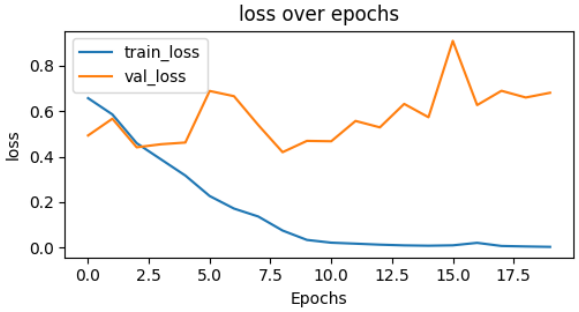
(a) augmentation weight = 0



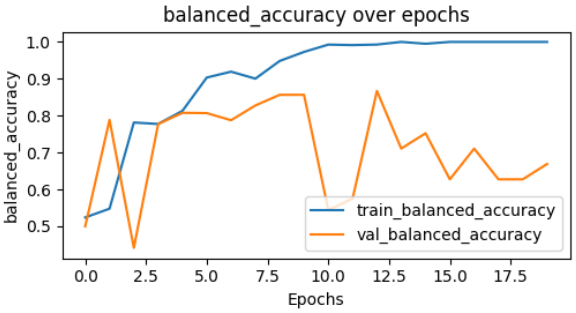
(b) augmentation weight = 0



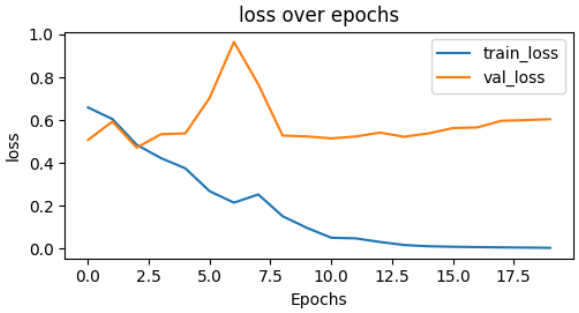
(c) augmentation weight = 0.3



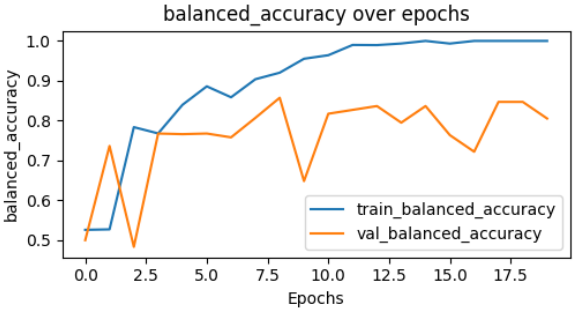
(d) augmentation weight = 0.3



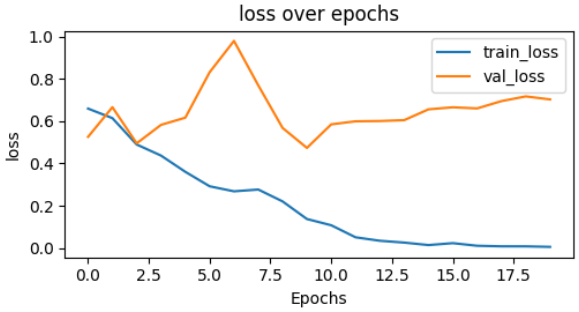
(e) augmentation weight = 0.7



(f) augmentation weight = 0.7

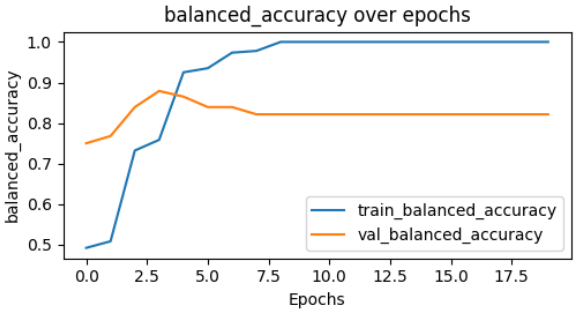


(g) augmentation weight = 1.0

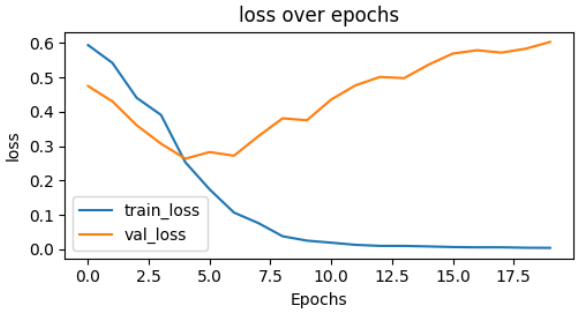


(h) augmentation weight = 1.0

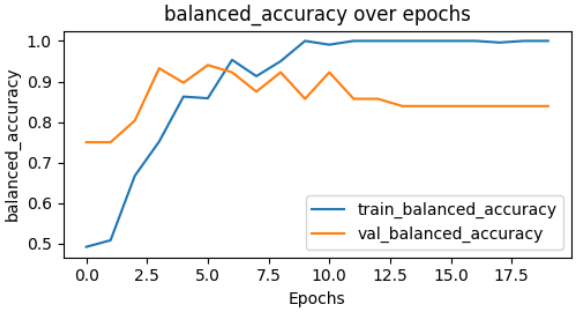
Figure 4.3: Graphs of balanced accuracy and loss over epochs for Dataset 1 (deep fine-tuning protocol). Each line represents results for a different augmentation weight.



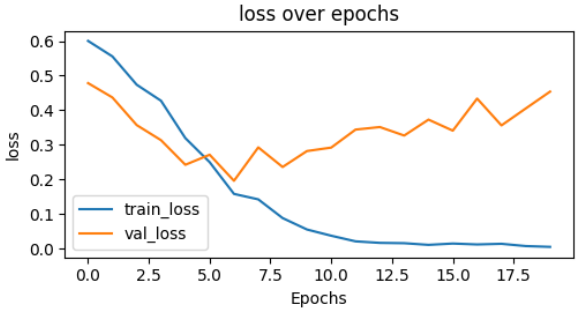
(a) augmentation weight = 0



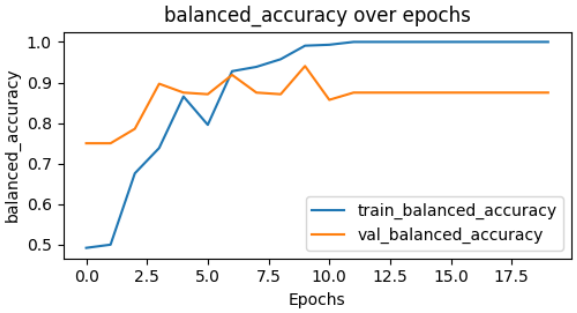
(b) augmentation weight = 0



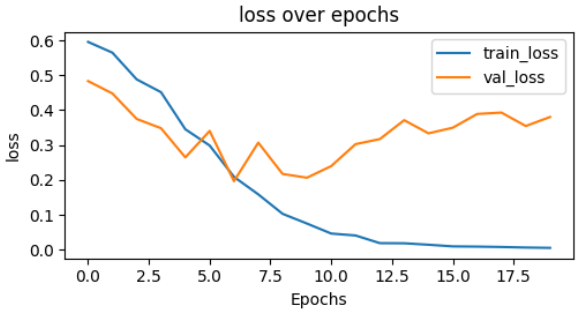
(c) augmentation weight = 0.3



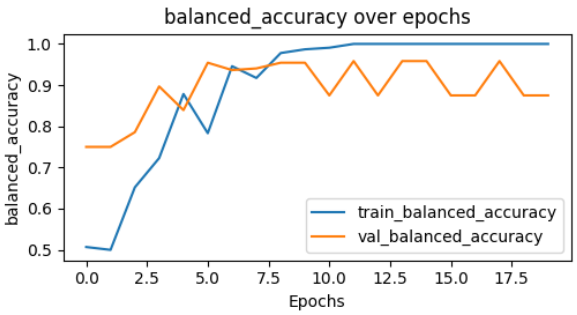
(d) augmentation weight = 0.3



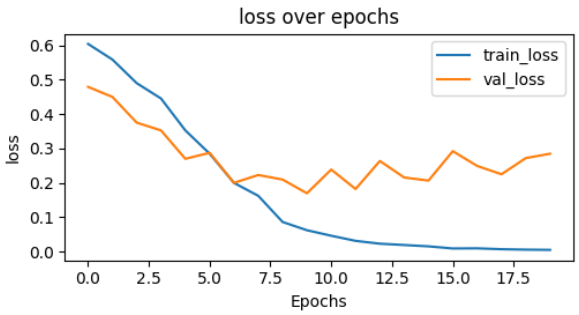
(e) augmentation weight = 0.7



(f) augmentation weight = 0.7



(g) augmentation weight = 1.0



(h) augmentation weight = 1.0

Figure 4.4: Graphs of balanced accuracy and loss over epochs for Dataset 2 (deep fine-tuning protocol). Each line represents results for a different augmentation weight.

Chapter 5

Conclusion

This works presents an attention-based natural language processing binary classification protocol to address the issue of automatic gender bias detection in Brazilian court decisions delivered in the context of gender-based violence cases. Our framework comprises:

1. The collection, partial annotation, and preparation of data — which, in our case, was extracted from the São Paulo state Court of Justice and made up for two datasets built with the help of domain experts;

2. The usage of an experimental pipeline based on BERTimbau, a pre-trained BERT model for the Brazilian Portuguese language;

3. The evaluation of such pipeline and a validation protocol.

Automatic detection of gender biases in court decisions would allow domain experts to address some of their research inquiries on the matter and enrich diagnoses on how such harmful practice is institutionally perpetrated. The underlying hypotheses behind this project are that (a) gender biases and stereotypes can be detected in judicial decisions on a large scale, and (b) natural language processing offers suitable approaches to detect them. While there are caveats behind the answer for each one of them and the protocol we developed needs improvement, we consider our results to corroborate both hypotheses.

Data was collected automatically due to the availability of scraping tools, combined with input from domain experts — which was crucial throughout the whole work. However, our approach has scalability issues, especially for Dataset 2, since the tools only sometimes worked as expected and had to be adapted for our instances and complemented with manual interventions.

Annotating our data also required domain knowledge, which hampers the possibility of annotating full large datasets — after all, that would defeat the purpose of using automatic strategies to facilitate the human work of analyzing each decision. Still, domain knowledge remains an ally rather than an obstacle since it allowed us to build the dataset from scratch, mindfully annotate it, choose and calibrate adequate models, create a validation pipeline for the protocol, and thoroughly document and be aware of the references behind our decisions.

5.1 Limits

In addition to the scalability issues in data collection, other caveats should be considered regarding our approach. While some could be addressed in future work (as discussed in Section 5.2), others are intrinsic to our conceptual and experimental choices. We stress the following:

- **Use by domain experts:** While domain experts could use and/or adapt our protocol as a tool to help them address research inquiries and build diagnoses on public policies — which is a major motivation behind this work —, some technical training would be required, given that our framework does not include a final, user-friendly, graphic interface-based product;
- **Improvements:** While our protocol has shown fair results and indicates a promising approach, we do not vouch for its indiscriminate use, especially not before improvements are made on the automatic learning process and its explainability — an issue in which we dig deeper in Section 5.2;
- **Validation:** Validation of our protocol over the test sets was hampered by the scarcity of annotated data, causing testing results to be statistically insignificant. Therefore, although experimental results are fair and we present a usable validation pipeline, we did not properly evaluate its generalization capability;
- **Need for human assessment**: Our protocol should not be used without human assessment during the process, nor its decisions should be trusted without proper human (and preferably domain-based) evaluation. It can be a helpful diagnostic tool in combination with the richer tools and abilities provided by human experts — especially for the analysis of individual cases rather than populations of instances;
- **Gender and bias definitions:** Our definitions of gender and biases are intrinsically limited by the references we have had access to, as well as our own interpretations and perceptions on such references — even if logical, well-based, and scrutinizable, which are the qualities that make them acceptably scientific. Building a tool to learn gender biases from court decisions automatically requires some degree of a discretization of concepts to be performed, and we should be aware that there is a trade-off between discretizing concepts and acknowledging their nuances, which might be lost in the process;
- **Annotation dependency:** Our protocol is based on a supervised learning approach requiring domain-specific data annotation. This costly process could be addressed in future work, as explained in Section 5.2.

5.2 Future work

Although we propose a complete pipeline for data collection and automatic gender bias detection in court decisions issued in Brazilian Portuguese in gender-based violence cases, many issues remain to be addressed and could be explored in future work. Those include:

- **Datasets:** Our approach could be applied to, validated in, and/or expanded for other datasets of court decisions featuring gender issues. Besides enhancing the scalability features of our protocol of collection, documents issued by other courts, in different time frames, or a more diverse range of cases and attributes (including the ones for which we provided annotation protocols) could be explored in that sense;

- **Use by domain experts:** Since our pipeline requires technical training, future work could improve its usability — and, therefore, its reach power;

- **Modeling:** A more diverse range of models could be explored for automatic bias detection. They might include domain-specific fine-tuned models, approaches based on feature extraction, and approaches based on traditional models rather than attention-based ones. Examining such options could improve performance results and enrich our understanding of the task;

  - **Use of other large language models:** The release of pre-trained large language models in the past months — such as GPT-4 [79] and LLaMA [99], as well as comparable options trained in languages other than English, such as Sabiá for Brazilian Portuguese [84] — redefined standards for state-of-the-art performance in many natural language processing tasks. The possibilities offered by them for our investigation could be explored in future research;

- **Annotation:** Dependency on domain-specific annotation, which causes low annotated data availability, can be addressed differently. Annotating more data improves availability, but it is costly; data augmentation is a cheaper, feasible option, which we chose in this project. Future work could explore automatic annotation protocols and/or unsupervised techniques to make the pipeline more annotation independent;

- **Explainability:** Explainability is an essential dimension of automatic learning models — mainly when its outcomes are intended to support decision-making processes in Law and Public Policy settings — that could also be included in our pipeline. Being able to explain or interpret[1] results provided by such models make them more scrutinizable and, therefore, trustable. The abundant literature on the relevance, definitions, and techniques regarding explainability [86, 67, 87, 60, 78, 22] could be explored in future work.

5.3 Ethics statement

The main purpose of our contributions is to provide an approach for researchers and practitioners who are interested in investigating gender biases and related features in court decisions issued in Brazilian Portuguese. We foresee our protocols and guidelines being helpful for them to, among others:

[1] Some authors distinguish between "explainability" and "interpretability"; exploring such conceptual differences, however, goes beyond the scope of this work.

- Decide whether, and to which extent, to disclose datasets made of court documents, especially in gender-based violence and other human rights violations-related cases;
- Collect, process, and annotate court documents as a data source for automatic learning models, by either using our protocol or deriving similar ones;
- Explore the information provided by our datasets to investigate institutional gender biasing in Brazilian courts, especially from the state of São Paulo, as well as other features associated with the metadata and annotation we provided;
- Use, expand, and assess our experimental pipeline and our baseline testing protocol to detect gender biases in court decisions on a large scale, thus unlocking helpful diagnostic information on the matter.

Despite the positive impacts that our work might induce, we must acknowledge that distorted and/or unpredicted interpretations and uses derived from it can arise, which could lead to unwanted outcomes. These include but are not limited to:

- Breach of the terms of the deed of undertaking to which one must abide to access our datasets — which, although entails liability, carries the risks associated with wrongfully using and/or disclosing their content, as explained by the author in previous work [69];
- Bypassing of human assessment and previous domain-informed knowledge when using and evaluating our tools and their derived results, which could lead to misdiagnosis of the issues we propose to address. Examples include:
  - dismissing other sources of institutional gender biasing in justice systems;
  - wrongfully pointing specific individuals or court chambers as bias perpretrators;
  - over or underestimating occurrences of institutional gender biasing in Brazilian courts.

We try to mitigate unwelcome derivations of our work by thoroughly describing its processes, methods, caveats, and intended implications, also believing that foreseeing associated risks within reason helps us understand the limits and possibilities offered by our approach.

Bibliography

[1] Syed Rameel Ahmad, Deborah Harris, and Ibrahim Sahibzada. Understanding legal documents: Classification of rhetorical role of sentences using deep learning and natural language processing. In *14th IEEE International Conference on Semantic Computing*, pages 464–467, San Diego, 2020. IEEE.

[2] Miguel Teixeira Jacobina Aires. A técnica de análise de sentimentos aplicada às certidões de julgamento do Supremo Tribunal Federal, 2019.

[3] Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In Vládia Pinheiro, Pablo Gamallo, Raquel Amaro, Carolina Scarton, Fernando Batista, Diego Silva, Catarina Magro, and Hugo Pinto, editors, *Computational Processing of the Portuguese Language*, pages 3–14. Springer International Publishing, 2022.

[4] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems (NLDB 2018). Lecture Notes in Computer Science*, volume 10859 LNCS, pages 57–64, 2018.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014.

[6] Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the 2nd Workshop on Gender Bias in Natural Language Processing at COLING 2020*, 2020.

[7] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, 55(7):1–39, dec 2022.

[8] Lidia Casas Becerra, Juan Pablo González Jansana, and María Soledad Molina. Estereotipos de género en sentencias del Tribunal Constitucional. *Anuario de Derecho Público UDP*, pages 250–272, 2012.

[9] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, 2020.

[10] Trevor Bench-Capon, Katie Atkinson, Adam Z. Wyner, Michał Araszkiewicz, Kevin Ashley, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald P. Loui, L. Thorne McCarty, Henry Prakken, Frank Schilder, Erich Schweighofer, Paul Thompson, Alex Tyrrell, Bart Verheij, and Douglas N. Walton. A History of AI and Law in 50 papers: 25 Years of the International Conference on AI and Law. *Artificial Intelligence and Law*, 20:215–319, 2012.

[11] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[12] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating Gender Bias in BERT, 2020.

[13] Fernanda Bragança and Laurinda Fátima da F. P. G. Bragança. Revolução 4.0 no Poder Judiciário: levantamento do uso de inteligência artificial nos tribunais brasileiros. *Revista da Seção Judiciária do Rio de Janeiro*, 23(46):65–76, 2019.

[14] Brasil – Casa Civil da Presidência da República. Brazilian Federal Constitution (1988), 1988. Brasília.

[15] Brasil – Casa Civil da Presidência da República. Brazilian Access to Information Act (Law n. 12527, November 18, 2011), 2006. Brasília.

[16] Brasil – Secretaria Geral da Presidência da República. Law n. 11340, August 7, 2006, 2006. Brasília.

[17] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, 2011. Association for Computational Linguistics.

[18] Judith Butler. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, 1990.

[19] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

[20] Erik Cambria and Bebo White. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57, 2014.

[21] Ilias Chalkidis and Dimitrios Kampas. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198, 2019.

[22] H. Chefer, S. Gur, and L. Wolf. Transformer Interpretability Beyond Attention Visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society.

[23] João Victor de Assis Brasil Ribeiro Coelho. Aplicações e Implicações da Inteligência Artificial no Direito, 2017.

[24] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[25] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[26] Bruna Armonas Colombo, Pedro Buck, and Vinicius Miana Bezerra. Challenges when using jurimetrics in Brazil-A survey of courts. *Future Internet*, 9(4), 2017.

[27] Alexis Conneau, Holger Schwenk, Yann Le Cun, and Loïc Barrault. Very Deep Convolutional Networks for Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017.

[28] Jack G Conrad and Frank Schilder. Opinion Mining in Legal Blogs. *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 231–236, 2007.

[29] Conselho Nacional de Justiça. Resolução Nº 121 de 05/10/2010, 2010.

[30] Rebecca J. Cook and Simone Cusack. *Gender Stereotyping: Transnational Legal Perspectives*. University of Pennsylvania Press, 2010.

[31] Marta R. Costa-jussà. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496, 2019.

[32] D.G. Coulouris. Violência, gênero e impunidade: a construção da verdade nos casos de estupro. *Anais do XVII Encontro Regional de História -– O lugar da História*, 2004.

[33] Simone Cusack. Gender stereotyping as a human rights violation. Technical report, Office of the United Nations High Commissioner for Human Rights, 2013.

[34] Gabriela Perissinotto de Almeida. Estereótipos de gênero sobre mulheres vítimas de estupro: uma abordagem a partir do viés de gênero e dos estudos de teóricas feministas do direito, 2017.

[35] Gabriela Perissinotto de Almeida and Sérgio Nojiri. Como os juízes decidem os casos de estupro? Analisando sentenças sob a perspectiva de vieses e estereótipos de gênero. *Revista Brasileira de Políticas Públicas*, 8(2), 2018.

[36] Mariana Dionísio De Andrade, Beatriz De Castro Rosa, and Eduardo Régis Girão de Castro Pinto. Legal tech: analytics, inteligência artificial e as novas perspectivas para a prática da advocacia privada. *Revista Direito GV*, 16(1):1–22, 2020.

[37] Mariana Dionísio de Andrade, Eduardo Régis Girão de Castro Pinto, Isabela Braga de Lima, and Alex Renan De Sousa Galvão. Inteligência Artificial para o rastreamento de ações com repercussão geral: o Projeto Victor e a realização do princípio da razoável duração do processo. *Revista Eletrônica de Direito Processual*, 21(1):312–335, 2020.

[38] Guilherme Ramos de Morais. Inteligência artificial aplicada ao direito: análise de sentimento em julgamentos de mandados de segurança no Supremo Tribunal Federal, 2019.

[39] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, 2012. The COLING 2012 Organizing Committee. Published also as Techreport.

[40] Alana de Santana Correia and Esther Luna Colombini. Attention, please! A survey of Neural Attention Models in Deep Learning. *CoRR*, abs/2103.16775, 2021.

[41] Weslei Gomes de Sousa. Inteligência artificial e celeridade processual no Judiciário: mito, realidade ou necessidade?, 2020.

[42] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA, 2022. Association for Computing Machinery.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[44] Emad Elwany, Dave Moore, and Gaurav Oberoi. BERT goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, 2019.

[45] Gema Fernández Rodríguez de Liévana. Los Estereotipos de Género en los Procedimientos Judiciales por Violencia de Género: El Papel del Comité CEDAW en la Eliminación de la Discriminación y de la Estereotipación. *Oñati Socio-legal Series*, 5(2):498–519, 2015.

[46] Marcelo Herton Pereira Ferreira. Classificação de peças processuais jurídicas: Inteligência Artificial no Direito, 2018.

[47] Anjalie Field and Yulia Tsvetkov. Unsupervised Discovery of Implicit Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 596–608. Association for Computational Linguistics, 2020.

[48] Sandra Costa Fonseca, Rosa Maria Soares Madeira Domingues, Maria do Carmo Leal, Estela M. L. Aquino, and Greice M. S. Menezes. Aborto legal no Brasil: revisão sistemática da produção científica, 2008-2018. *Cadernos de Saúde Pública*, 36(Cad. Saúde Pública, 2020 36 suppl 1):e00189718, 2020.

[49] Simone Frintrop, Erich Rome, and Henrik I. Christensen. Computational Visual Attention Systems and Their Cognitive Foundations: A Survey. *ACM Trans. Appl. Percept.*, 7(1), jan 2010.

[50] A. Hector F. Gómez, B. Franco O. Guaman, Jorge Benitez, Luis Roberto Jacome Galarza, Victor Hernandez Del Salto, Daniel Sanchez Guerrero, and Gabriel Garcia Torres. Semantic analysis of judicial sentences based on text polarity. *Iberian Conference on Information Systems and Technologies*, 2016-July:5–8, 2016.

[51] Xiaochuang Han and Yulia Tsvetkov. Fortifying Toxic Speech Detectors Against Veiled Toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7732–7739. Association for Computational Linguistics, 2020.

[52] Brian Seamus Haney. Applied Natural Language Processing for Law Practice. *Intellectual Property & Technology Forum at Boston College Law School*, 2020.

[53] Nathan S. Hartmann, Erick Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *Proceedings of Symposium in Information and Human Language Technology*, pages 122–131, 2017.

[54] Lukáš Havrlant and Vladik Kreinovich. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1):27–36, 2017.

[55] Inter-American Court of Human Rights. Case of Atala Riffo and Daughters v. Chile, 2012.

[56] Inter-American Court of Human Rights. Case of I.V. v. Bolivia, 2016.

[57] K Sparck Jones. Natural Language Processing: A Historical Review. In Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer, editors, *Current issues in computational linguistics: in honour of Don Walker*. Springer, 1994.

[58] Dan Jurafsky and James H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

[59] Ambedkar Kanapala, Srikanth Jannu, and Rajendra Pamula. Summarization of legal judgments using gravitational search algorithm. *Neural Computing and Applications*, 31(12):8631–8639, 2019.

[60] Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online, 2021. Association for Computational Linguistics.

[61] Taku Kudo and Yuji Matsumoto. Chunking with support vector machines. In *Second meeting of the North American chapter of the Association for Computational Linguistics*, 2001.

[62] Brian Larson. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, 2017. Association for Computational Linguistics.

[63] Nicola Lettieri, Antonio Altamura, Rosalba Giugno, Alfonso Guarino, Delfina Malandrino, Alfredo Pulvirenti, Francesco Vicidomini, and Rocco Zaccagnino. Ex Machina: Analytical platforms, Law and the Challenges of Computational Legal Science. *Future Internet*, 10(5), 2018.

[64] Yi Hung Liu and Yen Liang Chen. A two-phase sentiment analysis approach for judgement prediction. *Journal of Information Science*, 44(5):594–607, 2018.

[65] Lee Loevinger. Jurimetrics–The Next Step Forward. *Minnesota Law Review*, 33(5):455–493, 1949.

[66] June Luchjenbroers and Michelle Aldridge. Conceptual manipulation by metaphors and frames: Dealing with rape victims in legal discourse. *Text and Talk*, 27(3):339–359, 2007.

[67] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[68] Raysa M. Benatti. Revealing Gender Biases in (TJSP) Court Decisions with Natural Language Processing. Available at `https://doi.org/10.5281/zenodo.7794781`, 2023.

[69] Raysa M. Benatti, Camila M. L. Villarroel, Sandra Avila, Esther L. Colombini, and Fabiana Severi. Should I disclose my dataset? Caveats between reproducibility and individual data rights. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 228–237, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics.

[70] Marcos Maia and Cicero Aparecido Bezerra. Análise bibliométrica dos artigos científicos de jurimetria publicados no Brasil. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, 18:e020018, 2020.

[71] Mamede Said Maia Filho and Tainá Aguiar Junquilho. Projeto Victor: perspectivas de aplicação da Inteligência Artificial ao Direito. *Revista de Direitos e Garantias Fundamentais*, 19(3):219–238, 2018.

[72] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[73] L. Thorne McCarty. Deep semantic interpretations of legal texts. *Proceedings of the International Conference on Artificial Intelligence and Law*, pages 217–224, 2007.

[74] Kaiz Merchant and Yash Pande. NLP Based Latent Semantic Analysis for Legal Text Summarization. *International Conference on Advances in Computing, Communications and Informatics*, pages 1803–1807, 2018.

[75] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.

[76] Marie-Francine Moens, Erik Boiy, Chris Reed, and Raquel Mochales Palau. Automatic Detection of Arguments in Legal Texts. *ICAIL: International Conference on Artificial Intelligence and Law*, pages 225–230, 2007.

[77] Juliana Fontana Moyses and Fabiana Cristina Severi. Os estereótipos de gênero nas interpretações do critério "violência baseada no gênero" da Lei Maria da Penha pelo TJ/SP. *42º Encontro Anual da ANPOCS - GT 21 - Os juristas na sociedade: conflitos políticos e sentidos do direito*, 2018.

[78] Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks, 2019.

[79] OpenAI. GPT-4 Technical Report, 2023.

[80] Daniel Otter, Julian Medina, and Jugal Kalita. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–21, 04 2020.

[81] María Angélica Peñas Defago. Estereotipos de género: la perpetuación del poder sexista en los tribunales argentinos. *Revista Estudos Feministas*, 23(1):35–51, 2015.

[82] Sonia Maria Demeda Groisman Piardi. A (In)Constitucionalidade do *Custos Legis* nas Ações Penais Públicas em Segundo Grau. *Atuação: Revista Jurídica do Ministério Público Catarinense*, 9(20):9–42, 2012.

[83] Alexandra Guedes Pinto, Henrique Lopes Cardoso, Isabel Margarida Duarte, Catarina Vaz Warrot, and Rui Sousa-Silva. Biased Language Detection in Court Decisions. In Cesar Analide, Paulo Novais, David Camacho, and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2020*, pages 402–410. Springer International Publishing, 2020.

[84] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese Large Language Models, 2023.

[85] Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J. Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. LegalNLP – Natural Language Processing methods for the Brazilian Legal Language, 2021.

[86] Eliott Remmer. Explainability Methods for Transformer-based Artificial Neural Networks: a Comparative Analysis. Master's Programme, Machine Learning, KTH Royal Institute of Technology, Stockholm, Sweden, 2022.

[87] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[88] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, 2017. Association for Computational Linguistics.

[89] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, 2018. Association for Computational Linguistics.

[90] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *The 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[91] Anna Schmidt and Michael Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, 2017. Association for Computational Linguistics.

[92] Fabiana Cristina Severi and Camila Maria de Lima Villarroel. Análise jurisprudencial dos tribunais da região sudeste sobre a aplicação do instituto: (síndrome da) alienação parental. *Pensar – Revista de Ciências Jurídicas*, 26(2):1–14, 2021.

[93] Chris Sexton and Greg Tozzi. Detecting Evidence of Gender Discrimination in Fijian Court Documents, 2020.

[94] Akanksha Rai Sharma and Pranav Kaushik. Literature survey of statistical, deep and reinforcement learning in natural language processing. *IEEE International Conference on Computing, Communication and Automation*, pages 350–354, 2017.

[95] Nilton Correia Da Silva, Fabricio Ataides Braz, Teófilo Emídio De Campos, André Bernardes Soares Guedes, Danilo Barros Mendes, Davi Alves Bezerra, Davi Benevides Gusmão, Felipe Borges de Souza Chaves, Gabriel Gomes Ziegler, Lucas Hiroshi Horinouchi, Marcelo Herton Pereira Ferreira, Pedro Henrique Gonçalves Inazawa, Victor Hugo Dias Coelho, Ricardo Vieira De Carvalho Fernandes, Fabiano Hartmann Peixoto, Mamede Said Maia Filho, Bernardo Pablo Sukiennik, Lahis da Silva Rosa, Roberta Zumblick Martins Da Silva, Tainá Aguiar Junquilho, and Gustavo H. T. A. Carvalho. Document type classification for Brazil's supreme court using a Convolutional Neural Network. In *Proceedings of the Tenth International Conference on Forensic Computer Science and Cyber Law*, São Paulo, 2018. Brazil Chapter of the HTCIA.

[96] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, pages 403–417. Springer International Publishing, 2020.

[97] Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing, 2021.

[98] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In *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)*, pages 380–385, Minneapolis, 2019. Association for Computational Linguistics.

[99] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023.

[100] Samir Undavia, Adam Meyers, and John E. Ortega. A comparative study of classifying legal documents with neural networks. *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems*, 15:515–522, 2018.

[101] United Nations General Assembly. Convention on the Elimination of All Forms of Discrimination against Women, 1979. New York.

[102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[103] Charles Felipe Oliveira Viegas. JurisBERT: Transformer-based model for embedding legal texts, 2022.

[104] Camila Maria de Lima Villarroel. Acesso à justiça para as mulheres e (Síndrome da) Alienação Parental: Análise Jurisprudencial dos Tribunais da Região Sudeste. Technical report, Universidade de São Paulo, Ribeirão Preto, 2020.

[105] Adam Wyner, Raquel Mochales-Palau, Marie Francine Moens, and David Milward. Approaches to text mining arguments from legal cases. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6036 LNAI:60–79, 2010.

[106] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, 2018. Association for Computational Linguistics.

Appendix A

List of legal statutes mentioned in this work

1. CF (*Constituição Federal*): Brazilian Federal Constitution (1988);
2. CP (*Código Penal*): Brazilian Criminal Code (Decree-Law n. 2848, December 7, 1940);
3. LCP (*Lei das Contravenções Penais*): Brazilian Misdemeanors Act (Decree-Law n. 3688, October 3, 1941);
4. CT (*Código de Trânsito*): Brazilian Traffic Code (Law n. 9503, September 23, 1997);
5. LAP (*Lei da Alienação Parental*): Brazilian Law on Parental Alienation (Law n. 12318, August 26, 2010);
6. Brazilian Law n. 11419/2006 (December 19, 2006);
7. LGPD (*Lei Geral de Proteção de Dados*): Brazilian General Data Protection Act (Law n. 13709, August 14, 2018) — also available in English (unofficial translation);
8. Brazilian Law n. 11340/2006 (*Lei Maria da Penha*) (August 7, 2006);
9. CEDAW: United Nations General Assembly's Convention on the Elimination of All Forms of Discrimination against Women (1979).