Gender Bias Detection in Court Decisions: A Brazilian Case Study

Raysa Benatti
raysa.benatti@uni-tuebingen.de
University of Tübingen
Germany
Universidade Estadual de Campinas (UNICAMP)
Brazil

Fabiana Severi
fabianaseveri@usp.br
Faculdade de Direito de Ribeirão Preto, Universidade de São Paulo (USP)
Brazil

Sandra Avila
sandra@ic.unicamp.br
Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
Brazil

Esther Luna Colombini
esther@ic.unicamp.br
Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
Brazil

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. While increasing procedural efficiency is often the primary motivation behind natural language processing in the field, several works have proposed solutions for human rights-related issues, such as assessment of public policy and institutional social settings. One such issue is the presence of gender biases in court decisions, which has been largely studied in social sciences fields; biased institutional responses to gender-based violence are a violation of international human rights dispositions since they prevent gender minorities from accessing rights and hamper their dignity. Natural language processing-based approaches can help detect these biases on a larger scale. Still, the development and use of such tools require researchers and practitioners to be mindful of legal and ethical aspects concerning data sharing and use, reproducibility, domain expertise, and value-charged choices. In this work, we (a) present an experimental framework developed to automatically detect gender biases in court decisions issued in Brazilian Portuguese and (b) describe and elaborate on features we identify to be critical in such a technology, given its proposed use as a support tool for research and assessment of court activity.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → **Information extraction**; **Supervised learning by classification**.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658937

KEYWORDS

gender bias, natural language processing, social computing, legal text

ACM Reference Format:
Raysa Benatti, Fabiana Severi, Sandra Avila, and Esther Luna Colombini. 2024. Gender Bias Detection in Court Decisions: A Brazilian Case Study. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3630106.3658937

1 INTRODUCTION

Natural language processing (NLP) techniques have been proposed to address issues in many domains. Specific fields, such as the social sciences, are more prone to use and produce texts containing relevant data for analysis. The legal field, in particular, has been the focus of interest of many practitioners and researchers who propose techniques to perform tasks such as document classification [56, 62], information extraction [48], and text summarization [31, 41].

The increase of digital data availability in such domains plays a significant role in using NLP to address some of their inquiries. In recent decades, public institutions have replaced physical documents and procedures with digital ones in many jurisdictions. We stress the Brazilian case: being the most populated Latin American country, it has a substantial court system[1], with large databases of judicial documents and an engaged community focused on developing computational approaches for the legal field. In that sense, it has emerged as a legal data hotspot.

Although increasing procedural efficiency is the primary motivation behind most artificial intelligence-based solutions in legal systems, they can also be explored to address other issues. The possibility of analyzing content on a larger scale offers new methods of investigation and expands the range of research questions about social institutions to be explored. In that context, NLP can be framed as a support tool for assessing court activity.

One aspect that might be extracted from court decisions raises concerns: the presence of gender biases or stereotypes encrusted in legal reasoning, especially in cases of gender-based violence. There is evidence that court rulings can bear those biases, and NLP approaches can help detect them on a larger scale; however, despite

[1]The country has one lawyer for each batch of around 150 people [22] and approximately 80 million active legal cases [40].

their technical promises and accomplishments, legal and ethical considerations must be carried out when designing, developing, and using such approaches.

As a case study to support this argument, we introduce an experimental framework of data construction and text classification designed to automatically detect gender biases in court decisions issued in Brazilian Portuguese in the context of gender-based violence-related cases. The pipeline includes structuring data and metadata and developing an attention-based deep-learning solution for its classification. Therefore, it provides a methodological possibility for domain experts to find new answers to their inquiries. We describe the methodology used for developing the framework, the experiments and their main results, and a baseline evaluation protocol.

From the development of the framework, we identify issues to be addressed if it were used as a standard diagnostic auxiliary tool by domain experts and other stakeholders. Critical aspects of being mindful of such technology include data sharing and use, reproducibility, domain expertise, and value-charged choices carried out during the process.

In summary, the contributions of our paper are threefold:

(1) We propose a framework for detecting gender biases in court decisions, comprising an experimental pipeline of binary classification on the presence of gender biases in court decisions issued in Brazilian Portuguese, which can be reproduced by domain experts with some technical training;
(2) We introduce two datasets of court decisions issued by the São Paulo state Court of Justice (Brazil) in gender-based violence cases, DVC (Domestic Violence Cases) and PAC (Parental Alienation Cases), with annotation (partial and complete, respectively), their metadata on a range of legal attributes, their documentation, and the description of collection, processing, and annotation protocols;
(3) We highlight and describe critical features that should be present in computational technologies proposed as support tools for assessing court activity in gender issues, in particular, and in human rights issues, in general.

The remaining of this text is organized as follows. In Sec. 1.1, we explain the motivation behind addressing gender stereotypes in court decisions while presenting concepts related to gender biasing. In Sec. 2, we briefly present part of the literature that also addresses the automatic detection of gender biases in the legal domain. In Sec. 3, we describe the case study data and framework: the methodology followed to build them, a baseline validation protocol, and the main results observed from the experimental pipeline. In Sec. 4, we propose a discussion based on what we identify as critical technical, legal, and ethical aspects to be addressed for this kind of technology to fulfill its purposes. In Sec. 5, we elaborate on our findings and prospects for future directions. Finally, we present an ethics statement.

1.1 Institutional Gender Bias

Stereotyping assumes one's characteristics or roles due to belonging to a particular 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 stereotype is a generalized view or preconception about a group [13]. 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 [13]. 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 [13]. 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 a pervasive human rights violation [16]. The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) [63], having 189 parties as of 2024[2], expresses that state parties must take adequate measures to eliminate prejudices and practices based on stereotyped concepts of gender roles; 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 reflect social, political, and economic relations present in society; therefore, despite their neutrality rhetoric, they frequently reinforce gender discrimination practices [6]. Several studies have addressed how judicial proceedings issue gender stereotyping acts and some consequences of this [6, 17, 25, 43, 45, 53]. Particularly in Brazil, Federal Law 11340/2006 (*Lei Maria da Penha*) [10] creates legal mechanisms, including proceeding rules, aiming to prevent and repress violence against women, according to guidelines provided by the country's Federal Constitution [9] 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 [17, 43].

Studies providing that kind of evidence 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 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. 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 [37, 52].

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. It provides ways for legal practitioners and researchers to answer domain questions and analyze the presence and implications of gender biases in courts. It also contributes as a methodology that can be used to apply automatic text classification techniques in the social sciences.

[2]See this United Nations Treaty Collection page for a complete and updated list of signatures and ratifications, accessions, or successions.

We argue, however, for a cautious approach when designing and implementing this kind of technology.

2 RELATED WORK

Previous work has addressed the issue of automatic detection of gender biases in legal contexts, from which we stress the following ones.

Pinto et al. [49] proposed 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 or settled on a methodology. On the other hand, Sexton and Tozzi [55] showed results on using 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 used in our framework. 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. There are overlaps between their results and the ones we present; they do not mention non-technical challenges or ethical constraints that might have been present.

Sevim et al. [54] reconstructed the corpora used to train Law2Vec, a legal domain-specific word embedding model, to assess gender biases present in legal documents from different sources. While their work focuses on legislation rather than court decisions, they provide technical insights for evaluating language biases in the legal domain and mention ethical aspects concerning the task — such as the potential of unfair outcomes when informed by biased applications.

Baker Gillis [3], on its turn, proposed an approach focused on determining the presence of gender bias within the US judicial system, primarily based on case law. From a dataset of over 6.7 million decisions, the author proposes new ways for automating the creation of biases-related word lists and uses clustering algorithms to group the documents; their main contribution relies on stressing that consistent definitions of biases are essential to achieve consistent results. That conclusion is aligned with what we observed while developing our framework and the beyond-technical aspects that we identify as critical in developing such technologies, as discussed in Sec. 4 — particularly regarding the importance of domain expertise.

3 FRAMEWORK

This section describes our methodology for developing the framework under study, the data, and the main results observed from the proposed experimental pipeline.

Fig. 1 summarizes the methodology. It starts with protocols of collection, annotation, and preparation of two datasets of Brazilian court decisions, whose texts are cleaned and transformed into chunks. This step aims to adequate the data's content and size to feed the models that classify them.

Classification is performed in the experimental phase. We ran a set of experiments over BERTimbau-based models [58], BERT-based pre-trained models for Brazilian Portuguese, with different degrees of data augmentation, to train them to differentiate between biased and non-biased chunks of labeled text. In this phase, we applied different fine-tuning protocols over the pre-trained networks, using our own data to adjust their parameters.

We used the training and validation sets to teach and evaluate the models. For evaluation, we used loss metrics and balanced accuracy. Besides evaluating model performance on the validation set, the validation methods include a baseline testing pipeline. While our test sets are too small to pose statistically significant validation results, we ran a pipeline that uses the best versions of the trained models to label all the texts of court decisions compatible with the framework. It could, therefore, be used in enriched versions of our datasets or new ones.

Complete documentation of technical aspects of the framework, data, and codes can be found in the project's open repository[3]. The datasets [38] can be downloaded and used under conditions as discussed in Sec. 4.

3.1 Data

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 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 in diverse ways, 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 [25, 43], custody and other family disputes [29, 53], health care and reproductive rights [30, 45], and rape [17, 45]. Therefore, to analyze such biases on 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: DVC, which comprises 1,604 decisions issued between 2012 and 2019 in domestic violence-related cases, and PAC, which comprises 49 decisions issued between 2012 and 2019 in parental alienation-related civil and criminal cases. In both datasets, domain experts selected search criteria and instances.

Besides the data selection derived from the work of the experts, other criteria behind the choice for the state of São Paulo include: (a) 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 [12]); (b) ease of collection, since the court's official website and search engines allow for data scraping, and auxiliary tools

[3]Available at https://github.com/ra-ysa/gender_law_nlp.

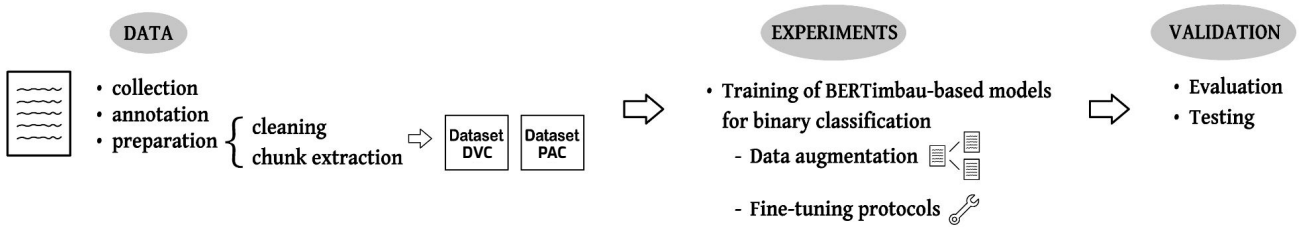

Figure 1: High-level view on the methodology. It comprises three blocks: the first one, Data, includes collection, annotation, and preparation with cleaning and chunk extraction, generating Domestic Violence Cases (DVC) and Parental Alienation Cases (PAC) datasets; they are the input of the second block, Experiments, containing training of BERTimbau-based models for binary classification, with data augmentation and fine-tuning protocols. Finally, the third block, Validation, includes evaluation and testing.

are available[4]; (c) proximity and familiarity with the local court, given that the authors and collaborators of this work are all based and affiliated in the state (and some of us have previously worked with the institution). By not including data from other courts, we acknowledge that we cannot assess the protocol's performance and limits in a more diverse range of regional particularities.

3.1.1 Data Annotation. Since relying on minimally supervised approaches, we identified the need to partially annotate the data for information not provided in the extraction phase. Metadata and other features, from automatic extraction to manual annotation, were added to each decision. Although most features were left out of the experimental pipeline (which focused 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. For DVC, we randomly selected N documents for manual annotation, in which N is the integer part of 10% of the number of documents — therefore, $N = 160$. PAC was annotated entirely, given its limited size.

Three people carried out the process of annotation: (1) the first author has a background in Law and Computer Science; (2) the second author is an expert in Law, human rights, and related gender issues; and (3) a domain researcher from the same field. Therefore, theoretical references — mainly based on Cook and Cusack's work on gender stereotyping in legal contexts [13, 16] —, combined with previous domain expertise, provided the foundations on which annotation decisions were based.

For each annotated document, an attribute `vies` (bias) contains the statement(s) in which some bias is identified; for model training and classification purposes, they are considered positive cases. Such identification was performed by (1), following guidelines and posterior qualitative validation from (2) and (3). To make the best out of the annotation labor — since it was being made for identification of biases anyway —, we also systematized further judicial, less interpretative information contained in the decisions, such as legal codes of the crimes under investigation, features of the parties, decision outcome[5], and others. Such additional information was primarily annotated by (1) for DVC and (3) for PAC. A complete list of the annotated attributes and their domains, as well as a dictionary of values and descriptions of annotation protocols, can be found in Appendix.

Biases. A core element of the data annotation process — which determines what the models learn from the input texts — is the definition of bias. 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[6].

There are several examples of institutional gender biases and their harmful consequences for the groups affected by them. 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 [26]. 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 [14].

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*" [18]; 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 [43].

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

[5]We note that biased language can exist in overall valid decisions with legitimate outcomes. However, detecting such biases is an essential task on its own since they taint the legitimacy of what should be an unbiased, soundly motivated institutional response. Assessing correlations between the presence of gender biases and decision outcomes was beyond the scope of this work; future research could explore such endeavors.

[6]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, or their perceived gender.

Moyses [43] stresses 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.

Gender biases also play a role in decisions regarding family disputes. Severi and Villarroel [53] show how the scientifically unsound concept of parental alienation[7] 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.

3.1.2 Data Preparation. Text to be used as input to the models went through a preparation phase that involved (a) cleaning and (b) chunk extraction[8]. The digital decisions are issued in PDF files; plain text extracted from them comes with some noisy elements. Although attention-based models do not require noise 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. They were, therefore, removed.

Having plain, clean text corresponding to each annotated decision is still insufficient to feed the models of interest due to (a) its size and (b) its content. Attention-based networks typically require input text not larger than 512 tokens [58, 65]. There are techniques to deal with longer texts, such as the Long-Document Transformer [7]; however, applying them to our data would be challenging to the point of going beyond the scope of the 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 significant content that would likely be 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 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).

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. 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.

3.2 Experimental Design

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.

Fig. 2 illustrates our experimental pipeline. The original annotated texts, stored in a JSON file, 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.

3.2.1 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 [5]. 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. Augmentation also partially made up for the uneven class distribution of the data: the original amount of biased decisions is around 18% for DVC and 26% for PAC, which were adjusted for 45% and 212%, respectively.

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[9], we flip a coin of `weight` $= \{0, 0.3, 0.7, 1.0\}$ 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[10], in case the word to be replaced does not exist in `BIAS_SYN_DICT`;
  - otherwise, the word is replaced by a synonym extracted from a general dictionary.

[7] 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 who has authority, custody, or supervision over the minor, to repudiate a birth parent or causing damage to the establishment or preservation of the bonds between them" (Law n. 12318/2010, article 2).

[8] Not to be confused with *chunking* [32].

[9] 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).

[10] We used the Brazilian Portuguese synonym dictionary from OpenWordnet-PT [19].

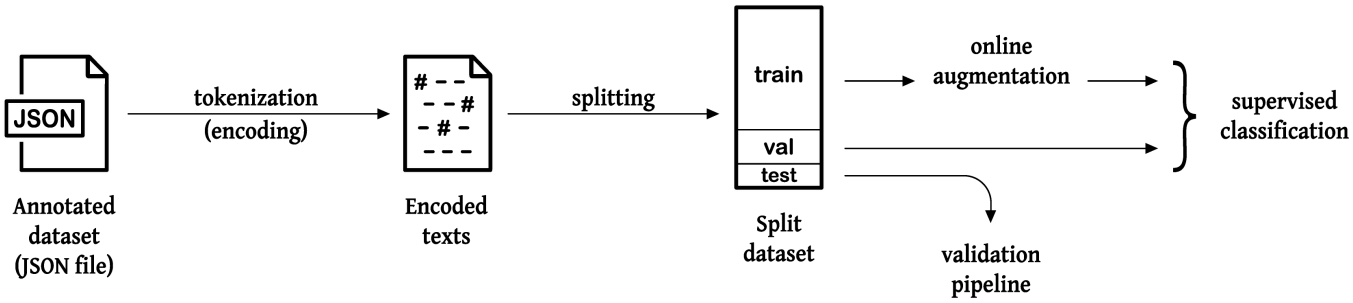

Figure 2: A representation of the experimental pipeline. It starts with a JSON file, the annotated dataset, which is tokenized (encoded). The encoded texts are split and become the split dataset, made of portions for training, validation, and testing. The training set is augmented online and, along with the validation set, is fed into a supervised classification process; the test set is fed into a validation pipeline.

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

3.2.2 Model and Parameters. The binary classification task on the bias for DVC and PAC was learned by the BERTimbau model [58]. While originally trained for masked-language modeling[11], the model can be used as a classifier through its Hugging Face interface[12]. We imported the `bert-base-portuguese-cased` version of the model as an `AutoModelForSequenceClassification`.

While the original BERTimbau embeddings were preserved (frozen) 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 (frozen) 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 and 6 since overfit increases significantly for `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.

3.3 Evaluation and Validation Methods

The low availability of data hampered the validation of our protocol over the test set 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 baseline tool for final users interested in using our model over full, non-annotated decisions.

[11]See https://github.com/neuralmind-ai/portuguese-bert.
[12]See https://huggingface.co/neuralmind.

In this phase, we chose the version of the trained model for each dataset that showed the best-balanced accuracy value in the validation set over all experiments. We split the whole content of each decision into chunks; for a given decision, if any of its chunks are 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 biased.

We used confusion matrices to help visualize model performance on the epochs with the lowest loss value.

3.4 Main Findings

Our main experimental results are summarized in Tables 1 and 2. They show, for each dataset and fine-tuning protocol, the best-balanced accuracy for training (label 'T') and validation (label 'V') sets, as well as the first epoch in which it was observed. 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 DVC. Therefore, combining this strategy with partial data annotation helps achieve a reasonable trade-off between the cost of building a quality dataset and getting good performance in the task that we want a model to learn.

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 3 and 4 show confusion matrices of results over the validation set of PAC at each fine-tuning protocol, using the maximum augmentation weight. While the deep fine-tuning protocol slightly increased false negatives, overall classification was more accurate, significantly decreasing false positives.

Using an augmentation weight above zero, combined with the deep fine-tuning protocol, is 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 makes sense from an automatic learning perspective, the lack of robust validation prevents us from assessing

Table 1: Summarized results for DVC. 'T' stands for training; 'V' stands for validation.

Fine-tuning protocol	Augmentation `weight`	Best-balanced accuracy (%) (epoch)
Baseline	0	76.54 (T) (17), 69.15 (V) (19)
	0.3	74.92 (T) (16), 71.31 (V) (19)
	0.7	75.54 (T) (19), 73.32 (V) (16)
	1.0	72.95 (T) (19), 74.52 (V) (16)
Deep	0	100.00 (T) (10), 85.74 (V) (19)
	0.3	100.00 (T) (10), 88.86 (V) (7)
	0.7	100.00 (T) (13), 86.70 (V) (12)
	1.0	100.00 (T) (14), 85.74 (V) (8)

Table 2: Summarized results for PAC. 'T' stands for training; 'V' stands for validation.

Fine-tuning protocol	Augmentation `weight`	Best-balanced accuracy (%) (epoch)
Baseline	0	74.50 (T) (14), 83.93 (V) (19)
	0.3	73.74 (T) (16), 85.71 (V) (19)
	0.7	74.59 (T) (18), 85.71 (V) (17)
	1.0	72.47 (T) (16), 87.90 (V) (19)
Deep	0	100.00 (T) (8), 87.90 (V) (3)
	0.3	100.00 (T) (9), 94.05 (V) (5)
	0.7	100.00 (T) (11), 94.05 (V) (9)
	1.0	100.00 (T) (11), 95.83 (V) (11)

Table 3: Confusion matrix for results over the validation set of PAC (*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: Confusion matrix for results over the validation set of PAC (*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

the generalization capabilities of the models. As discussed in Sec. 5.1, future directions could address this issue with larger datasets — which could include collecting new data and/or enriching DVC and PAC with more annotated instances. Adapting the protocol to be more annotation-independent would allow for exploring other validation possibilities.

4 DISCUSSION

Computer-enhanced information extraction provides possibilities of automating tasks previously performed by humans, increasing the investigation scale. In the context of social institutions, they can be support tools for public policy diagnoses and decision-making, assessment of institutional activity, and social science research.

In that sense, frameworks like the one we present can potentially fulfill roles in social change, as proposed by Abebe et al. [1]. Detecting human rights violations in court decisions, such as harmful gender biases, helps measure the problem, diagnose how it manifests, and understand how we specify it. It is also an effort towards the call to "study institutions up", a concept previously described in the anthropology field, now reframed as a power-aware research focus in machine learning [4, 42].

As in any technical intervention, clarifying its limits is essential. The development of our framework highlighted critical aspects about which one must be mindful when proposing computational tools to support decision-making in social settings. While some could be addressed in future directions (as discussed in Sec. 5.1), others are intrinsic to conceptual and experimental choices, and we argue that they should be contemplated in designing, implementing, and using such technologies. The following paragraphs are dedicated to describing them.

Data sharing and reproducibility. Reproducibility is a critical quality of modern research [2, 28, 36], given its role in scientific scrutiny, fraud prevention and detection, and strengthening of research communities, which upholds the purpose of science as an endeavor of public interest. In computer science research, the gold reproducibility standard can be attained by publishing linked and executable code and data along with results, according to Peng [47].

In this context, data sharing and quality assessment emerge as an object of concern [8, 27]. Data collecting, cleaning, labeling, and/or

processing are often part of the experimental pipeline in machine learning research, which justifies interest in making them available for peers and stakeholders. However, the use and availability of datasets produced by social institutions can pose ethical and legal constraints that researchers and practitioners must consider.

Court decisions, for instance, often contain sensitive personal information[13] on the subjects involved; while secrecy and/or anonymization is generally expected in those cases, it is not always properly performed, and documents with restricted personal data can end up publicly available — and well-meaning researchers can be held accountable for propagating it.

Even setting data sensitivity aside, other restrictions may apply. Some jurisdictions impose specific constraints on data use and disclosure — for example, when it concerns minors, issues of social or public interest, private life matters, and others. Publicizing documents containing this kind of information can pose legal liability or be ethically debatable, given that it amplifies risks for the subjects involved. Those risks include violation of privacy and intimacy rights, exposure of confidential information, and exposure of any information that might jeopardize the safety or integrity of the subject(s) involved in a legal case. Such violations can be particularly harmful in human rights-related disputes, which often figure socially vulnerable groups. In gender violence claims, for instance, decisions frequently contain descriptions of family and relationship dynamics, information on the health and sex life of the parties, identification of persons (including minors) and communities, and other delicate data.

To guarantee acceptable levels of scientific reproducibility while maintaining the informational self-determination of individuals — an elemental dimension of their human rights —, researchers and practitioners should comply with legal and ethical guidelines for data use and availability; they can include disclosure by demand with a deed of undertaking, anonymization, and other mitigation measures [15, 39, 64]. In this work, we chose the first option; data usage and constraints instructions can be found in the project's public repository. 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 data restriction issues. Researchers should evaluate which risks and mitigation choices might apply to their context to decide on the extent of data disclosure considering available resources, aiming at preserving scientific reproducibility while respecting ethical and legal restrictions.

Domain expertise. Domain expertise in machine learning has been an ongoing topic of scientific interest. Several researchers have addressed discussions on the matter; while the development of data-intensive tools, such as deep neural networks and large language models, brought possibilities of reducing the need for prior knowledge to deliver solutions, some argue that human expertise remains essential in the machine learning loop [33, 35], and can enhance the quality of the results [21, 46].

[13]According to the European General Data Protection Regulation [24] and similar provisions, personal data is sensitive when it concerns the racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health or sex life, or personal genetic or biometric information.

The importance of human expertise is particularly visible when the data available for learning is not abundant, well processed, properly documented, or overall does not meet quality, safety, and ethics standards for the task being performed — which is often the case in real-world problems. The development of our framework, for instance, required the input of domain experts in several steps of the data construction pipeline: selection of cases, collection, annotation, cleaning, chunking, and documentation. Expertise helped us determine which groups of cases were pertinent for the task, which metadata was necessary, how to identify biased decisions, which parts of the text were relevant, and what should be registered for reproducibility by other researchers and practitioners.

The benefits of integrating domain knowledge in technical solutions go beyond the data construction. Processes such as defining and formalizing which problems should be tackled, how they should be modeled, assessing the quality of the computer-enhanced solutions, and comparing them to the available ones can all benefit from prior human expertise. For instance, previous work on text classification from domestic violence online posts showed how domain-specific embeddings produce more informative results than generic ones [59].

We argue that protocols to support decision-making in institutional contexts should not be used without human assessment, nor should their decisions be trusted without proper human (and preferably domain-based) evaluation. When designed as a diagnostic tool, their purposes are fulfilled when combined with the knowledge and abilities provided by human experts — especially for analyzing individual cases rather than populations of instances. Expertise can also enrich context-specific validation strategies, as part of a broader participatory design [60]. We stress, however, that the participation of domain experts in the loop should be meaningfully integrated into the process; "participation washing" [57] should be avoided.

Value-charged choices. Values are pervasive in every scientific endeavor — not only in pre- and post-scientific activities, such as the definition of problems and application of results but also at the core of scientific reasoning. The acceptance and rejection of hypotheses require scientists to make value judgments [51]; scientific investigations in which errors can cause non-epistemic[14] consequences require non-epistemic values to be considered in methodological choices, data characterization, and interpretation of results [23].

Particularly in machine learning research and practice, value-charged decisions play a role in different process stages, including data construction (comprising selection, processing, annotation, and availability), choice of models and parameters, and selection of evaluation metrics. When using natural language processing to address gender issues, for instance, one's views on gender — and gender-based stereotypes, if pertinent — can influence how these steps will be performed.

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 of such references — even if logical, well-based, and scrutinizable, which are the qualities that make

[14]Social, ethical, and political aspects are examples of non-epistemic values in science [23].

them acceptably scientific. Building a tool to learn gender biases from court decisions requires some degree of a discretization of concepts, and we should be aware of the trade-off between discretizing concepts and acknowledging their nuances, which might be lost in the process.

Larson [34] discusses theoretical and ethical guidelines to consider when dealing with gender-related concepts in natural language processing research. In that sense, the views on gender and gender-based stereotypes imprinted in our framework are aligned with the theory of gender performativity presented by Judith Butler in 1990 and explored in related work since [11], 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" [20].

5 FINAL REMARKS

This work 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 of two datasets, DVC and PAC, 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 baseline validation protocol.

We also described critical features concerning data sharing and use, reproducibility, domain expertise, and value-charged choices that should be considered in the design and implementation of computational technologies proposed as support tools for the assessment of court activity, especially in human rights-related issues, such as the identification of gender biases.

Automatic detection of gender biases in court decisions allows domain experts to address some of their research inquiries 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; in that sense, the model we propose can be used and understood as proof of concept.

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 PAC, 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.

Overall, 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 to the automatic learning process and the critical features described in Sec. 4. The following section describes limitations that could be addressed in future endeavors.

5.1 Future Directions

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 directions. 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, further work could integrate other forms of participation and improve its usability — and, therefore, its reach power;
- **Modeling:** A more diverse range of models can 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 the GPT series [44] and LLaMA [61], as well as comparable options trained in languages other than English, such as Sabiá for Brazilian Portuguese [50] — 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;
- **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, a more robust evaluation of its generalization capability remains to be developed — yet another dimension in which more domain expertise participatory efforts should be integrated;

- **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 directions could explore automatic annotation protocols and/or unsupervised techniques to make the pipeline more annotation-independent.

ETHICS STATEMENT

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

- 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 biases 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 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;
- Bypassing human assessment and previous domain-informed knowledge when using and evaluating our tools and their derived results could lead to misdiagnosis of the issues we propose to address. Examples include:
  - dismissing other sources of institutional gender biases in justice systems;
  - wrongfully pointing specific individuals or court chambers as bias perpetrators;
  - over or underestimating occurrences of institutional gender biases 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.

ACKNOWLEDGMENTS

This work was funded by CAPES/Brazil and FAEPEX/Unicamp, having been developed mostly at the Recod.ai lab in the University of Campinas, Brazil. The Recod.ai lab is supported by projects from FAPESP/São Paulo, CNPq/Brazil, and CAPES. R. Benatti is currently funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645. F. Severi is supported by the University of São Paulo's Law School of Ribeirão Preto. S. Avila is partially funded by CNPq PQ-2 grant 316489/2023-9, FAPESP 2013/08293-7, 2020/09838-0, 2023/12086-9, and H.IAAC (Artificial Intelligence and Cognitive Architectures Hub). E. Colombini is partially funded by CNPq PQ-2 grant 315468/2021-1 and H.IAAC. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Raysa Benatti. We also thank Camila Maria de Lima Villarroel, Luanna Tomaz de Souza, Rodrigo Frassetto Nogueira, and Konstantin Genin for helpful discussions and feedback. Finally, we are grateful for the reviewers of this work, who enriched it with their evaluation and comments.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 252–260. https://doi.org/10.1145/3351095.3372871

[2] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533 (2016), 452–454. https://doi.org/10.1038/533452a

[3] Noa Baker Gillis. 2021. Sexism in the Judiciary: The Importance of Bias Definition in NLP and In Our Courts. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster (Eds.). Association for Computational Linguistics, Online, 45–54. https://doi.org/10.18653/v1/2021.gebnlp-1.6

[4] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/3351095.3372859

[5] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. *Comput. Surveys* 55, 7 (2022), 1–39. https://doi.org/10.1145/3544558

[6] Lidia Casas Becerra, Juan Pablo González Jansana, and María Soledad Molina. 2012. Estereotipos de género en sentencias del Tribunal Constitucional. *Anuario de Derecho Público UDP* 1 (2012), 250–272. http://derecho.udp.cl/wp-content/uploads/2016/08/13_Casas_Gonzalez.pdf

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. https://doi.org/10.48550/arXiv.2004.05150

[8] Hendrik Blockeel and Joaquin Vanschoren. 2007. Experiment Databases: Towards an Improved Experimental Methodology in Machine Learning. In *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, Berlin, Heidelberg, 6–17. https://link.springer.com/chapter/10.1007/978-3-540-74976-9_5

[9] Brasil — Casa Civil da Presidência da República. 1988. Brazilian Federal Constitution (1988). http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm Brasília.

[10] Brasil — Secretaria Geral da Presidência da República. 2006. Law n. 11340, August 7, 2006. http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm Brasília.

[11] Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, NY.

[12] Conselho Nacional de Justiça. 2010. Resolução Nº 121 de 05/10/2010. https://atos.cnj.jus.br/atos/detalhar/atos-normativos?documento=92

[13] Rebecca J. Cook and Simone Cusack. 2010. *Gender Stereotyping: Transnational Legal Perspectives*. University of Pennsylvania Press, Pennsylvania, USA. http://www.jstor.org/stable/j.ctt3fhmhd

[14] Daniella Georges Coulouris. 2004. Violência, gênero e impunidade: a construção da verdade nos casos de estupro.

[15] Gergely Márk Csányi, Dániel Nagy, Renátó Vági, János Pál Vadász, and Tamás Orosz. 2021. Challenges and Open Problems of Legal Document Anonymization. *Symmetry* 13, 8 (2021). https://doi.org/10.3390/sym13081490

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

[17] Gabriela Perissinotto de Almeida. 2017. *Estereótipos de gênero sobre mulheres vítimas de estupro: uma abordagem a partir do viés de gênero e dos estudos de teóricas feministas do direito.* Master's thesis. Universidade de São Paulo. https://doi.org/10.11606/D.107.2019.tde-05022019-093155

[18] Gabriela Perissinotto de Almeida and Sérgio Nojiri. 2018. 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), 825–853. https://doi.org/10.5102/rbpp.v8i2.5291

[19] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In *Proceedings of COLING 2012: Demonstration Papers*. The COLING 2012 Organizing Committee, Mumbai, India, 353–360. https://aclanthology.org/C12-3

[20] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2083–2102. https://doi.org/10.1145/3531146.3534627

[21] Michelangelo Diligenti, Soumali Roychowdhury, and Marco Gori. 2017. Integrating Prior Knowledge into Deep Learning. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, Cancun, Mexico, 920–923. https://doi.org/10.1109/ICMLA.2017.00-37

[22] OAB Ordem dos Advogados do Brasil Conselho Federal (Federal Council of the Brazilian Bar Association). 2024. Institucional / Quadro da Advocacia / Quantitativo Total. https://www.oab.org.br/institucionalconselhofederal/quadroadvogados

[23] Heather Douglas. 2000. Inductive Risk and Values in Science. *Philosophy of Science* 67, 4 (2000), 559–579. http://www.jstor.org/stable/188707

[24] European Parliament, Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). http://data.europa.eu/eli/reg/2016/679/2016-05-04

[25] Gema Fernández Rodríguez de Liévana. 2015. 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 (2015), 498–519. http://ssrn.com/abstract=2611539

[26] Sandra Costa Fonseca, Rosa Maria Soares Madeira Domingues, Maria do Carmo Leal, Estela M. L. Aquino, and Greice M. S. Menezes. 2020. Aborto legal no Brasil: revisão sistemática da produção científica, 2008-2018. *Cadernos de Saúde Pública* 36, Cad. Saúde Pública, 2020 36 suppl 1 (2020), e00189718. https://doi.org/10.1590/0102-311X00189718

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. https://doi.org/10.1145/3458723

[28] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8, 341 (2016), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027 arXiv:https://www.science.org/doi/pdf/10.1126/scitranslmed.aaf5027

[29] Inter-American Court of Human Rights. 2012. Case of Atala Riffo and Daughters v. Chile. https://corteidh.or.cr/docs/casos/articulos/seriec_239_ing.pdf

[30] Inter-American Court of Human Rights. 2016. Case of I.V. v. Bolivia. https://www.corteidh.or.cr/docs/casos/articulos/seriec_329_ing.pdf

[31] Ambedkar Kanapala, Srikanth Jannu, and Rajendra Pamula. 2019. Summarization of legal judgments using gravitational search algorithm. *Neural Computing and Applications* 31, 12 (2019), 8631–8639. https://doi.org/10.1007/s00521-019-04177-x

[32] Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/N01-1025

[33] Pawan Kumar and Manmohan Sharma. 2022. Data, Machine Learning, and Human Domain Experts: None Is Better than Their Collaboration. *International Journal of Human–Computer Interaction* 38, 14 (2022), 1307–1320. https://doi.org/10.1080/10447318.2021.2002040

[34] Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, 30–40. https://doi.org/10.18653/v1/w17-1601

[35] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* 1, 2 (2023), 100017. https://doi.org/10.1016/j.metrad.2023.100017

[36] Joseph Loscalzo. 2012. Irreproducible Experimental Results: Causes, (Mis)interpretations, and Consequences. *Circulation* 125, 10 (2012), 1211–1214. https://doi.org/10.1161/CIRCULATIONAHA.112.098244 arXiv:https://www.ahajournals.org/doi/pdf/10.1161/CIRCULATIONAHA.112.098244

[37] June Luchjenbroers and Michelle Aldridge. 2007. Conceptual manipulation by metaphors and frames: Dealing with rape victims in legal discourse. *Text and Talk* 27, 3 (2007), 339–359. https://doi.org/10.1515/TEXT.2007.014

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

[39] Raysa M. Benatti, Camila M. L. Villarroel, Sandra Avila, Esther L. Colombini, and Fabiana Severi. 2022. Should I disclose my dataset? Caveats between reproducibility and individual data rights. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 228–237. https://aclanthology.org/2022.nllp-1.20

[40] Consultor Jurídico (ConJur) Magazine. 2022. Brazil Justice Yearbook 2022. https://anuario.conjur.com.br/pt-BR/profiles/78592e4622f1-anuario-da-justica/editions/brazil-justice-yearbook-2022

[41] Kaiz Merchant and Yash Pande. 2018. NLP Based Latent Semantic Analysis for Legal Text Summarization. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, Bangalore, India, 1803–1807. https://doi.org/10.1109/ICACCI.2018.8554831

[42] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 34 (jan 2022), 14 pages. https://doi.org/10.1145/3492853

[43] Juliana Fontana Moyses. 2018. *Os enquadramentos da violência contra as mulheres no componente estrutural da Lei Maria da Penha: análise de conteúdo de decisões de 2ª instância do TJ/SP sobre 'violência baseada no gênero'.* Master's thesis. Universidade de São Paulo. https://doi.org/10.11606/D.107.2019.tde-29052019-154919

[44] OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774

[45] María Angélica Peñas Defago. 2015. Estereotipos de género: la perpetuación del poder sexista en los tribunales argentinos. *Revista Estudos Feministas* 23, 1 (2015), 35–51. https://www.redalyc.org/articulo.oa?id=38135331003

[46] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. https://doi.org/10.48550/arXiv.2302.12813

[47] Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (2011), 1226–1227. https://doi.org/10.1126/science.1213847

[48] Jayr Pereira, Andre Assumpcao, Julio Trecenti, Luiz Airosa, Caio Lente, Jhonatan Cléto, Guilherme Dobins, Rodrigo Nogueira, Luis Mitchell, and Roberto Lotufo. 2024. INACIA: Integrating Large Language Models in Brazilian Audit Courts: Opportunities and Challenges. https://arxiv.org/abs/2401.05273

[49] Alexandra Guedes Pinto, Henrique Lopes Cardoso, Isabel Margarida Duarte, Catarina Vaz Warrot, and Rui Sousa-Silva. 2020. Biased Language Detection in Court Decisions. In *Intelligent Data Engineering and Automated Learning – IDEAL 2020*, Cesar Analide, Paulo Novais, David Camacho, and Hujun Yin (Eds.). Springer International Publishing, Cham, 402–410. https://doi.org/10.1007/978-3-030-62365-4_38

[50] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems*. Springer, 226–240. https://doi.org/10.1007/978-3-031-45392-2_15

[51] Richard Rudner. 1953. The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science* 20, 1 (1953), 1–6. http://www.jstor.org/stable/185617

[52] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5477–5490. https://doi.org/10.18653/v1/2020.acl-main.486

[53] Fabiana Cristina Severi and Camila Maria de Lima Villarroel. 2021. Análise jurisprudencial dos tribunais da região sudeste sobre a aplicação do instituto: (síndrome da) alienação parental. *Pensar — Revista de Ciências Jurídicas* 26, 2 (2021), 1–14. https://doi.org/10.5020/2317-2150.2021.11443

[54] Nurullah Sevim, Furkan Şahinuç, and Aykut Koç. 2023. Gender bias in legal corpora and debiasing it. *Natural Language Engineering* 29, 2 (2023), 449–482. https://doi.org/10.1017/S1351324922000122

[55] Chris Sexton and Greg Tozzi. 2020. Detecting Evidence of Gender Discrimination in Fijian Court Documents. https://icaad.ngo/wp-content/uploads/2020/09/w266_final_project.pdf

[56] 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. 2018. 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. Brazil Chapter of the HTCIA, São Paulo, 7–11. https://doi.org/10.5769/C2018001

[57] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 6 pages. https://doi.org/10.1145/3551624.3555285

[58] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Ricardo Cerri and Ronaldo C. Prati (Eds.). Springer International Publishing, Cham, 403–417. https://link.springer.com/chapter/10.1007/978-3-030-61377-8_28

[59] Sudha Subramani, Sandra Michalska, Hua Wang, Jiahua Du, Yanchun Zhang, and Haroon Shakeel. 2019. Deep Learning for Multi-Class Identification From Domestic Violence Online Posts. *IEEE Access* 7 (2019), 46210–46224. https://doi.org/10.1109/ACCESS.2019.2908827

[60] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (, Seoul, Republic of Korea,) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 667–678. https://doi.org/10.1145/3531146.3533132

[61] 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. 2023. LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971

[62] Samir Undavia, Adam Meyers, and John E. Ortega. 2018. A comparative study of classifying legal documents with neural networks. *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems* 15 (2018), 515–522. https://doi.org/10.15439/2018F227

[63] United Nations General Assembly. 1979. Convention on the Elimination of All Forms of Discrimination against Women. https://www.un.org/womenwatch/daw/cedaw/text/econvention.htm New York.

[64] Marc van Opijnen, Ginevra Peruginelli, Eleni Kefali, and Monica Palmirani. 2017. *On-Line Publication of Court Decisions in the EU: Report of the Policy Group of the Project 'Building on the European Case Law Identifier'*. Technical Report. https://doi.org/10.2139/ssrn.3088495

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

A DVC DATASET: DOMESTIC VIOLENCE CASES

Table 5 summarizes the annotated attributes and their domains, followed by a dictionary of values and descriptions of annotation protocols.

A.1 Dictionary of attributes

- 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[15]).

- 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;

[15] 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.

  - 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[16] (*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*);
  - 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 (Federal Constitution, 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. 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)).

[16]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.

  - Rarely, the first instance prosecutor (`mp` — *Ministério Público*[17]) and the second instance prosecutor (`pj` — *Procuradoria de Justiça*) state two distinct opinions; in that case, they were both annotated in the same field.

- 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.

B PAC DATASET: PARENTAL ALIENATION CASES

Table 6 summarizes the annotated attributes and their domains, followed by a dictionary of values and descriptions of annotation protocols.

B.1 Dictionary of attributes

Since annotation for PAC was previously made by one of the experts in the context of another work (except for bias-related attributes), the domain of each attribute is more detailed, exhaustive, and redundant than in DVC. We kept the original annotation but stress the recommendation for gathering similar values depending on the context of use.

- `tipo_recurso`:
  - Criminal merit appeals: `apelacao_criminal`, `habeas_corpus_criminal`[18];
  - 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`:
  - `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

[17] *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.

[18] 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.

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.

C BIASES

For DVC (domestic violence cases), biased statements 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.

For PAC (parental alienation cases), biased statements 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.

Table 5: Data attributes annotated to 10% of the documents in DVC.

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 on biases
vies_alvo	target(s) of the biased statement(s)	vitima; reu; test; abs_mul; abs_reu; soc

(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).

Table 6: Data attributes annotated to the documents in PAC.

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	See section on biases; also includes `prej`[c]
`vies_alvo`	target(s) of the biased statement(s)	`vitima`; `mae`; `mul`; `soc`; `abs_mul`; `abs_reu`; `abs_cri`; `prej`[c]

(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.