



# Revealing Gender Biases in Court Decisions with Natural Language Processing

Candidate: Raysa Masson Benatti

Supervisors: Esther Luna Colombini and Sandra Avila

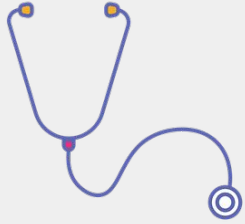
April 24<sup>th</sup>, 2023

# Stereotypes

Disregarding individual traits

- Why do we stereotype people?
- Is stereotyping bad?

# Institutional gender stereotyping



## health care

- age
- relationship status
- ...

Fonseca et al., 2020



## legal systems

- behavior
- personal history
- relationship with perpetrator(s)
- ...

Coulouris, 2004;  
Almeida et al., 2018;  
Moyses et al., 2018

# Institutional gender stereotyping

Atala Riffo and Daughters v. Chile

Inter-American Court of Human Rights, 2012

In this context, the [Supreme] Court [of Chile] concluded that:

(...)

iii) Ms. Atala “put her own interests before those of her daughters when she chose to begin to live with a same sex partner, at the same home where she raised and cared for her daughters, separately from the girls' father” and iv) “the potential confusion over sexual roles that could be caused in them by the absence from the home of a male father and his replacement by another person of the female gender poses a risk to the integral development of the children from which they must be protected.”

# Institutional gender stereotyping



perpetuates sexism

helps diminish the dignity of  
gender minorities and add to their  
burden

hampers gender minorities access  
to rights or justified benefits

So, we should investigate it, right?

# Institutional gender stereotyping

Social scientists have been doing just that

**but**

some questions remain difficult to address

- In which cases do court rulings express gender biases? How often does it happen?
- Can we find correlations between gender biases in court rulings and their metadata (date, location, type of court, etc.)?
- Can we systematize outcomes from court decisions? Can we verify how they correlate to institutional gender stereotyping?
- Given all this information, how can we improve policies towards better institutional response to gender violence?

# Hypotheses

1. Gender biases can be detected in judicial decisions on a large scale
2. Natural Language Processing offers suitable approaches to detect them

**Goal:** to build an NLP framework to classify Brazilian court decisions based on the presence of gender biases

# Contributions

- Two **datasets** of court decisions issued by the São Paulo state Court of Justice (TJSP)  
+ metadata, documentation, and protocols of collection, processing, and annotation
- An **experimental pipeline** for automatic detection of gender biases in court decisions issued in Brazilian Portuguese
- Legal and ethical **guidelines** on the **use and availability** of datasets made of court documents

Benatti et al. @ NLLP Workshop 2022, *Should I disclose my dataset? Caveats between reproducibility and individual data rights*



# Related work: NLP in the legal domain

Parsing approaches

Argument detection

Traditional machine learning techniques

Argument detection, sentiment analysis, document classification, text summarization

Topic modeling

Topic structure detection, text summarization, document clustering

Neural network classifiers

Document classification, subjectivity analysis

Word embeddings

Sentiment analysis, document classification  
**Sexton et al. (2020):** detection of gender biases in Fijian court documents

# Background: NLP paradigms and data representation

## frequency-based statistical approaches

### bag-of-words

this  
is  
really  
really  
cool

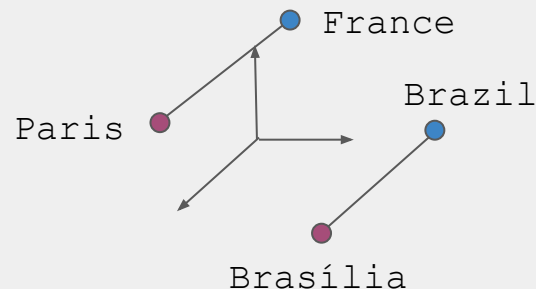
is really  
cool really  
this

simple and robust

no sequence

## contextual information

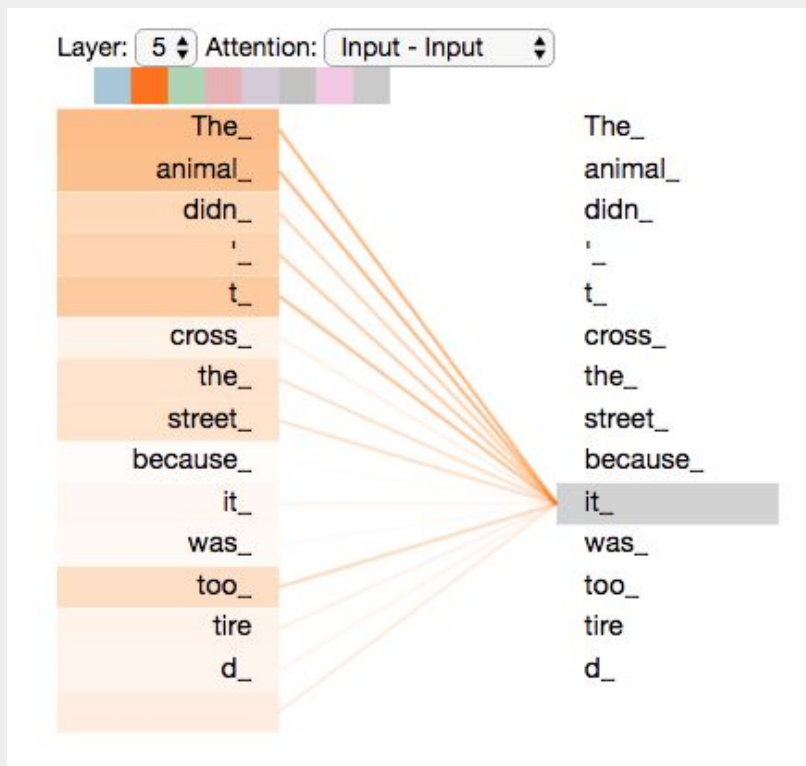
### word embeddings



semantic  
associations

# Background: Self-attention mechanism

**Attention:** focus on relevant parts of information



# Background: Transformers-based architectures

**Transformer** (2017): dependencies between input and output  
(no more need for recurrence or convolution on sequential data)



**BERT** (2019) (Bidirectional Encoder Representations from Transformers)



**BERTimbau** (2020) (BERT-based pre-trained model for Brazilian Portuguese)

# Methodology

## DATA



- collection
- annotation
- preparation

{ cleaning  
chunk extraction



## EXPERIMENTS

- Training of BERTimbau-based models for binary classification

Classes:

1 (biased)

0 (non-biased)

- Data augmentation



- Fine-tuning protocols



## VALIDATION


- Evaluation
- Testing

# Methodology: data



- **Dataset 1:** 1,604 decisions (2012-2019), criminal cases, domestic violence and related offenses  
Annotation on 160 cases
- **Dataset 2:** 49 annotated decisions (2012-2019), civil and criminal cases, parental alienation
- Selection of cases by domain experts
- Collection: scraping tools (with caveats)

# Methodology: data

 **TRIBUNAL DE JUSTIÇA**  
**PODER JUDICIÁRIO**  
São Paulo

Registro: 2018.0000637263

**ACÓRDÃO**

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

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

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

São Paulo, 13 de agosto de 2018.

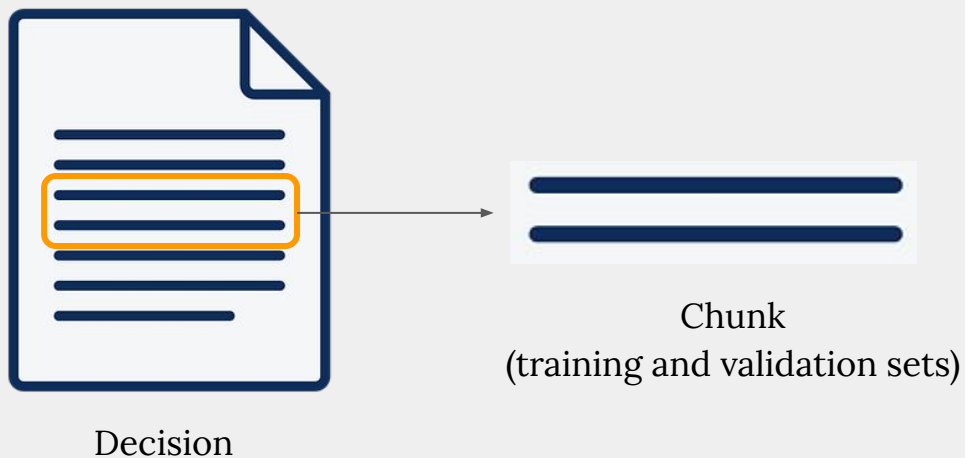
TRIBUNAL DE JUSTIÇA PODER JUDICIÁRIO São Paulo Registro: 2018.0000637263 ACÓRDÃO Vistos, relatados e discutidos estes autos de Apelação nº 0000063-60.2016.8.26.0197, da Comarca de Francisco Morato, em que é apelante [REDACTED], é apelado MINISTÉRIO PÚBLICO DO ESTADO DE SÃO PAULO. ACORDAM, em 1ª Câmara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

PODER  
JUDICIÁRIO

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

Text cleaning

# Methodology: data



- **Chunk:** seed sentence + context (N sentences above and below)
- For biased decisions, seed sentence = biased statement(s) and N = 1, 2, and 3
- For non-biased decisions, seed sentence = random sentence and N = 2 and 3

## # of chunks:

D1: 120 biased  
264 non-biased  
D2: 153 biased  
72 non-biased

Chunk extraction



# Methodology: data

context sentences

"Cabe recordar que, em crimes de natureza do aqui tratado, a palavra da vítima é de fundamental importância, sobretudo, quando apoiada pelas demais provas constantes nos autos, como ocorreu in casu. Somase a isso o fato de inexistirem incongruências em suas declarações

seed sentence

Da análise das provas, portanto, infere-se que estão suficientemente demonstradas a autoria e a materialidade do delito, em contexto de Violência Doméstica, atribuído ao apelante

Tampouco seria caso de aplicação do princípio da insignificância, como sugeriu o patrono do apelante em suas razões, haja vista o bem jurídico tutelado. A conduta que atinge a integridade física da pessoa não pode ser considerada de mínima ofensividade, desprovida de"

context sentences

Chunk extraction

# Data annotation: General attributes

Dataset 1 (domestic violence)	Dataset 2 (parental alienation)
<ul style="list-style-type: none"><li>• appellant and appealed parties;</li><li>• gender of the appellant;</li><li>• legal code(s) of crime(s) under analysis;</li><li>• victim(s) main relationship with defendant;</li><li>• gender of the victim(s);</li><li>• time of punishment of imprisonment in first and second instances;</li><li>• main request(s);</li><li>• subsidiary request(s);</li><li>• main reason(s) claimed by the appellant;</li><li>• Public Prosecutor's main statement;</li><li>• final decision on the merits;</li><li>• main reason(s) for decision;</li><li>• <b>biased statement(s);</b></li><li>• target(s) of biased statement(s)</li></ul>	<ul style="list-style-type: none"><li>• judge-rapporteur;</li><li>• issuing body;</li><li>• decision date;</li><li>• type of appeal;</li><li>• collegiality degree;</li><li>• availability of full content;</li><li>• theme;</li><li>• person who claimed alienation;</li><li>• person who was accused of alienation;</li><li>• claim(s) of violence against woman;</li><li>• claim(s) of violence against minor;</li><li>• who was accused of violence against minor;</li><li>• result on violence allegations;</li><li>• evidence (violence);</li><li>• result on alienation allegations;</li><li>• evidence (alienation);</li><li>• <b>biased statement(s);</b></li><li>• target(s) of biased statement(s)</li></ul>

# Data annotation: Biases

- CEDAW: Convention on the Elimination of All Forms of Discrimination Against Women (UN General Assembly, 1979)

**Article 2.** States Parties condemn discrimination against women in all its forms, agree to pursue by all appropriate means and without delay a policy of eliminating discrimination against women and, to this end, undertake:

(...)

(d) To refrain from engaging in any act or practice of discrimination against women and to ensure that public authorities and institutions shall act in conformity with this obligation;

- Moyses and Severi (2018): results > intention
- **Bias:** stereotype-based **motivation**

# Data annotation: Biases

*Na segunda fase, por considerar posterior reconciliação entre réu e vítima, a primariedade do acusado, a confissão parcial e ausência de novos fatos desabonadores ao acusado, reconheço tais circunstâncias como atenuantes*

Veja-se que desde o relatório da autoridade policial, há a menção da prática de agressões mútuas e não foram prestados depoimentos por parte de testemunhas presenciais para se saber quem deu início à contenda.

acontecimentos, na fase extrajudicial. De resto, as lesões que suportou são leves, o que, pondere-se, indica alguma moderação por parte do réu. Por último, cuida destacar que o episódio deu-se de forma isolada enquanto durou o relacionamento, ao que não se pode deixar de conferir o devido significado. Diante do quadro apresentado, impõe-se a aplicação do brocardo *in dubio pro reo*.

havendo a possibilidade, conforme bem observado pela magistrada *a quo*, terem sido começadas pela ofendida, o que desencadeou a atitude do réu de segurar os seus braços, com o intuito de fazer cessar a abordagem ou, até mesmo, de afastá-la.

Na gravação, a vítima - de forma um pouco confusa, talvez até em razão de responder simultaneamente a perguntas relacionadas a dois processos distintos - asseverou que o acusado

O regime para cumprimento da pena imposta foi o inicial semiaberto eis que adequado face ao Princípio da Suficiência, e por se tratar de réu reincidente demonstrando ter personalidade voltada para a prática de crimes, o que faz merecer maior reprovabilidade de sua conduta e uma terapêutica penal mais rigorosa.

restando, portanto, inteiramente isolada a versão judicial ofertada por [REDACTED] - no claro intuito de favorecer o réu, ladainha sempre repetida pelas mulheres que têm medo dos seus parceiros -, uma vez que um simples empurrão não seria hábil para causar os ferimentos encontrados na pessoa da vítima pelos peritos.

# Data annotation: Biases

Destarte, o egoísmo da requerida não pode prevalecer, já que o pseudoindividualismo em nada contribui para a criação e formação da prole, que necessita de ambos os pais para que venha ter o necessário *a posteriori*.

*aventada. O diagnóstico diferencial afirma o perfil de personalidade e inclui a autal condição de estabilidade conjugal e profissional que obteve após a separação (incomparáveis ao perfil de molestadores);*

Todavia, não se ocupou de convencer acerca da demora na revelação das agressões (cerca de um ano), o que se mostrou fora dos padrões da normalidade, inclusive porque os fatos somente vieram à baila depois de o seu genitor, por iniciativa da mãe, ter sido posto para fora de casa, e por outras razões, via ação cautelar de separação de corpos cujas cópias estão no feito.

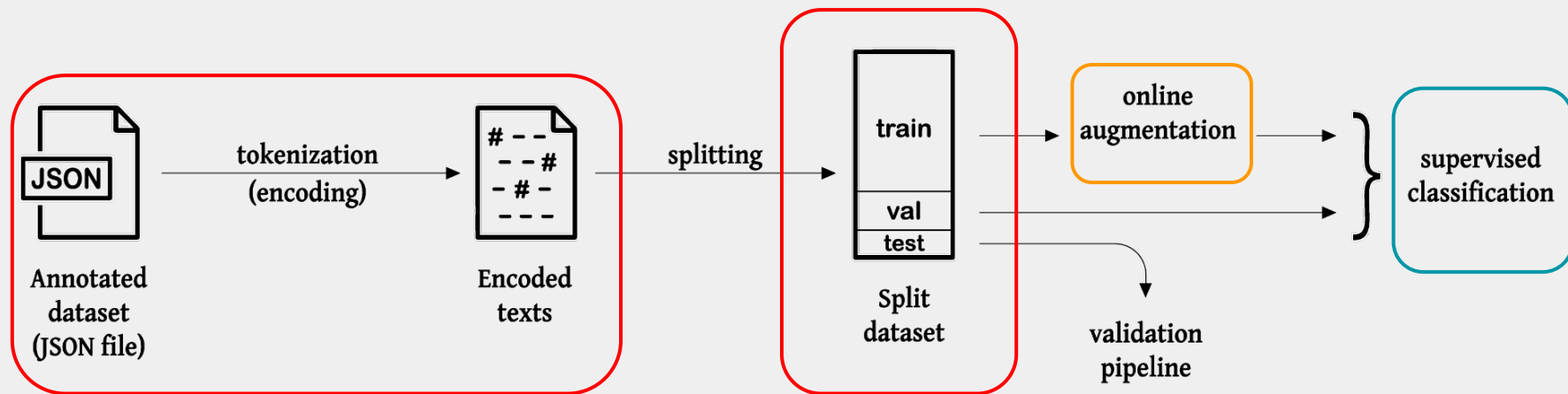
Como se sabe, a criança molestada sexualmente não age dessa forma.

Minha decisão funda-se na constatação de que o convívio dos filhos com ambos os pais é importante elemento de seu desenvolvimento saudável e que a forma arbitrada preserva a segurança da menor.

*b) o perfil psicológico da mãe indica tendência a reagir de modo intenso, que se acompanha de ansiedade disruptiva (não-consciente), que permite sejam desencadeados comportamentos com explosões emocionais. Esta tendência, por sua vez, contamina a percepção que tem sobre os fatos e experiências, particularmente as afetivas. De cuja expressão se depreende na maneira distorcida com que interpreta aqueles e na ação intempestiva que brota daí.*

De se convir que é especial a natureza dos atos descritos na exordial, peculiares a alguém que deveria exibir outro perfil psicológico, não aquele inicialmente descrito pelas testemunhas de acusação (bom pai, insuspeito).

# Methodology: experimental pipeline



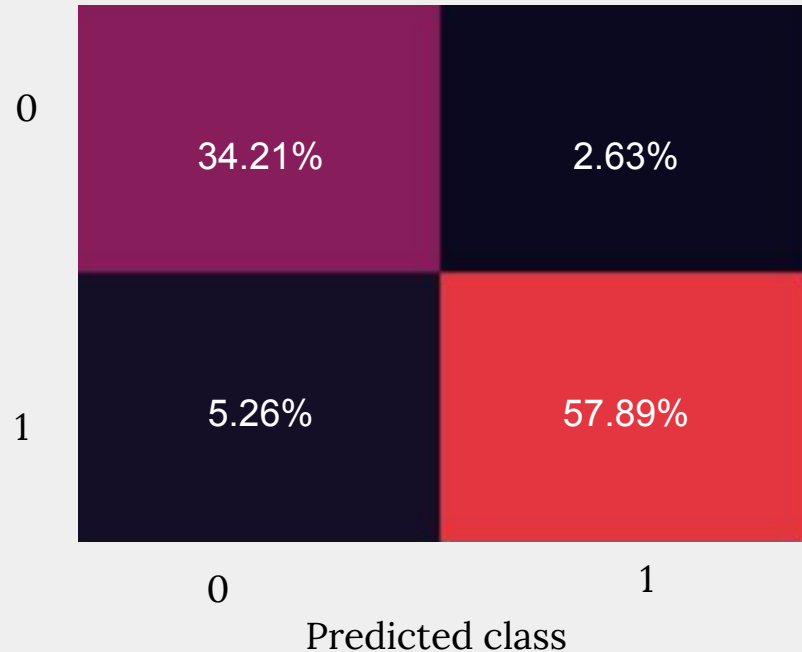
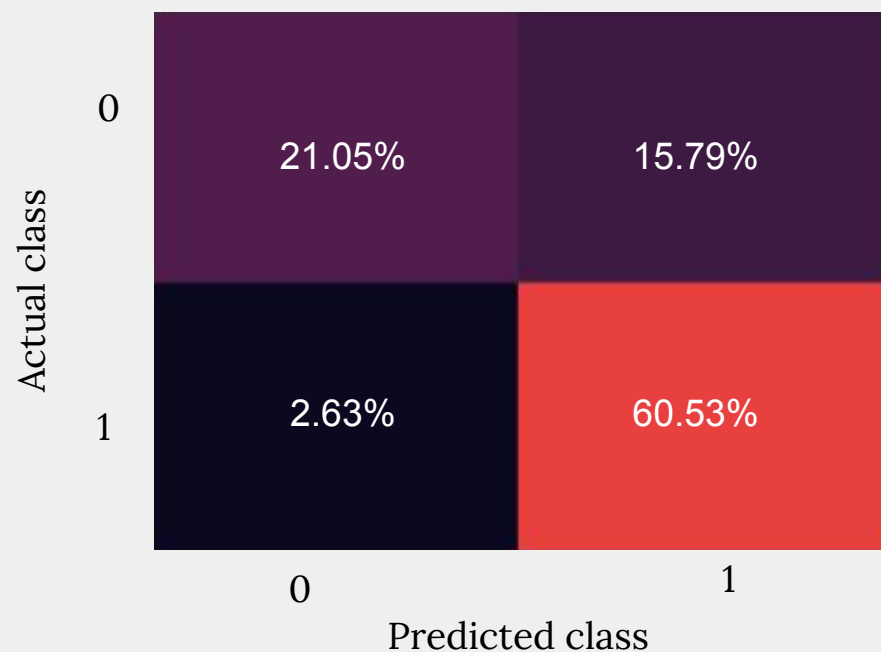
- **Augmentation**: synonym replacement with different probabilities (weights: 0, 0.3, 0.7, 1.0)
- **Fine-tuning** protocols:
  - baseline (classification layer only)
  - "deep" (all but last 5 layers freezed)

# Methodology: validation

- **Evaluation** metrics over training and validation sets:  
loss, balanced accuracy
- **Testing** baseline protocol:
  - Input is the whole decision, chunked
  - If any chunk is biased, then the decision is biased

# Results

"Deep" fine-tuning provides better accuracies than baseline fine-tuning...



(validation sets, Dataset 2 (parental alienation), augmentation weight = 1.0)

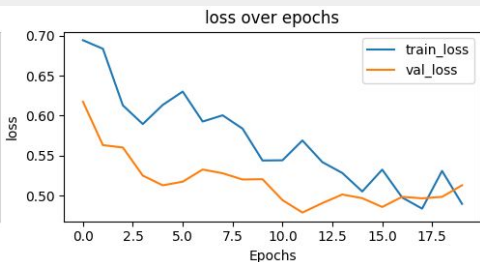


# Results

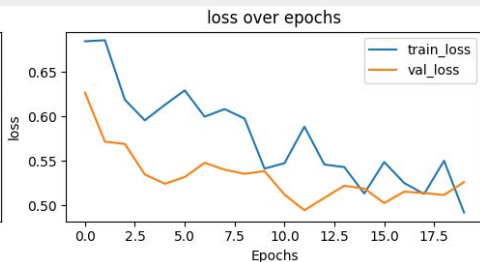
...but it overfits more



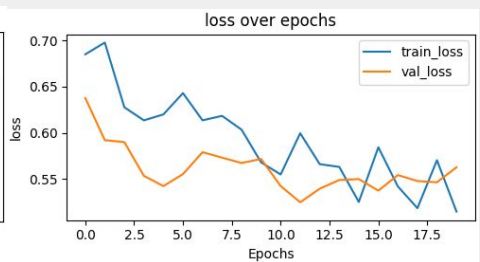
aug. weight = 0



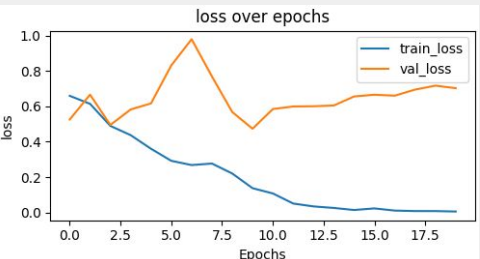
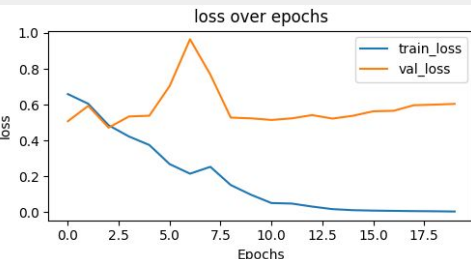
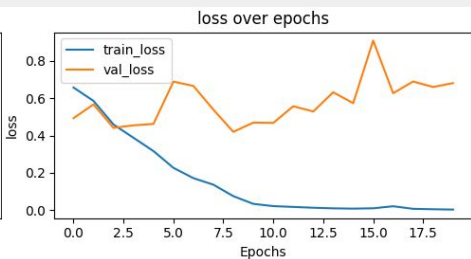
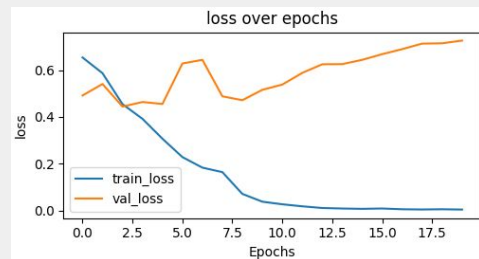
aug. weight = 0.3



aug. weight = 0.7



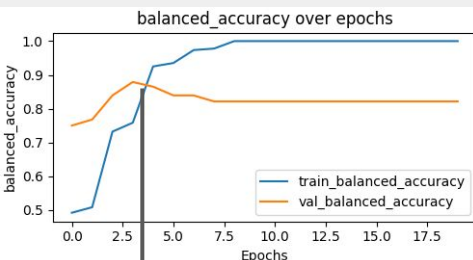
aug. weight = 1



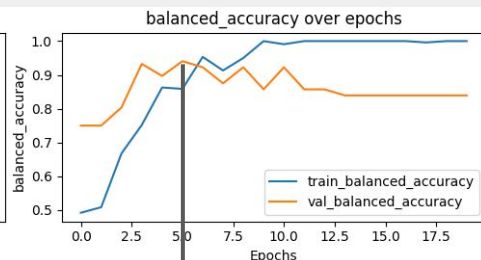
(loss over epochs for Dataset 1 (domestic violence))

# Results

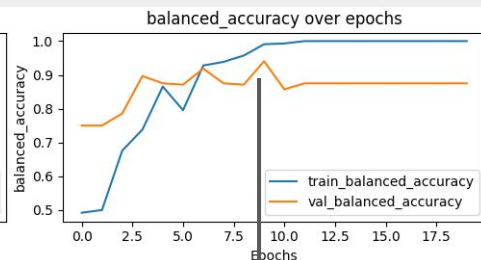
## Data augmentation: successful strategy



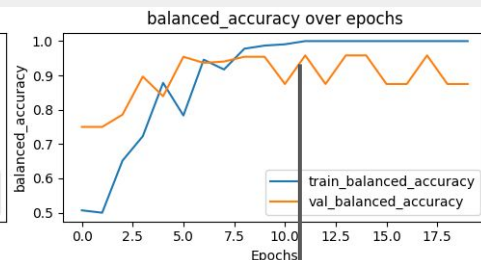
0.8790



0.9405



0.9405



0.9583

(best balanced accuracies in validation sets for Dataset 2 (parental alienation), deep fine-tuning protocol)

# Limits

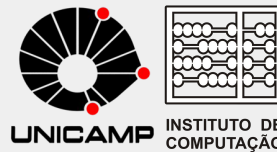
- Use by domain experts (training required)
- Need for human assessment
- Lack of proper validation (assessment of generalization capabilities)
- Gender and bias definitions
- Annotation dependency

# Future work

- More data
- More annotated data (and/or more annotation independence)
- Improvements on modeling and validation
- Use of our annotated attributes
- Explainability issues

Thank  
you!

Supported by:



raysa.benatti@gmail.com