Identificação e classificação de trechos racistas e sexistas em processos judiciais

Yasmin Francisquetti Barnes[1], **Marcelo Finger**[1]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)
05508-090 – São Paulo – SP – Brasil

yasminfrancisquetti@usp.br, mfinger@ime.usp.br

Abstract. *This study describes the development of an artificial intelligence tool to identify and classify racist and sexist discourses from legal agents. Focusing on sexual violence lawsuits at the São Paulo State Court of Justice (TJSP), the research aims to highlight biases that corroborate secondary victimization. The computational solution relies on a Natural Language Processing model based on the Transformer architecture, utilizing a BERT-derived model pre-trained for Brazilian Portuguese. The model is trained with textual data categorized by a gender stereotype ontology, and its performance is evaluated through 5-fold cross-validation using accuracy, precision, recall, and F-measure.*

Resumo. *Este estudo descreve o desenvolvimento de uma ferramenta de inteligência artificial para identificar e classificar discursos racistas e sexistas proferidos por agentes jurídicos. Focado em processos de violência sexual do Tribunal de Justiça de São Paulo (TJSP), a pesquisa visa evidenciar vieses que corroboram a vitimização secundária. A solução computacional baseia-se em um modelo de Processamento de Linguagem Natural com arquitetura Transformer, utilizando um modelo derivado do BERT pré-treinado para o português brasileiro. O modelo é treinado com dados textuais categorizados por uma ontologia de estereótipos de gênero e seu desempenho é avaliado por validação cruzada ($k = 5$) com métricas de acurácia, precisão, revocação e medida-F.*

1. Introdução

Em 2024, foram registrados 83.114 casos de estupro no Brasil, representando o maior número de ocorrências entre os 5 últimos anos [Ministério da Justiça e Segurança Pública 2025] e, dentre esses casos, 86% das vítimas são do sexo feminino. Entretanto, historicamente, o estupro em si raramente constitui uma violação isolada: após o crime, a vítima normalmente é submetida a diversas manifestações de violência simbólica, tanto institucionais quanto sociais [Siqueira 2016, Sousa 2017].

No âmbito judicial, o momento após o ataque é crucial para a manifestação da vítima e o estabelecimento do processo jurídico que tratará o caso. No entanto, é justamente durante a busca pela justiça que um novo ciclo de violência se inicia, uma vez que o Sistema de Justiça Criminal está condenado, em suas raízes, a reproduzir comportamentos derivados do patriarcalismo presente na sociedade [Siqueira 2016]. Essa experiência, caracterizada pelo tratamento inadequado, insensível ou deslegitimador por

parte das instituições judiciais, é conhecida pela Criminologia como vitimização secundária [Conselho Nacional do Ministério Público sd]. Nela, a denúncia da vítima e suas reivindicações são tratadas com negligência e de maneira desumana.

Nesse contexto, a análise manual do elevado volume de processos em andamento para identificar padrões discursivos potencialmente enviesados torna-se inviável, demandando soluções computacionais capazes de automatizar a triagem de dados textuais em larga escala. Para esse fim, o Processamento de Linguagem Natural (PLN) oferece ferramentas consolidadas para a detecção de discursos nocivos em português brasileiro, incluindo linguagem ofensiva, misoginia e discurso de ódio. Trabalhos de referência nessa tarefa, como os de Pelle e Moreira [Pelle e Moreira 2017], Leite et al. [Leite et al. 2020] e Vargas et al. [Vargas et al. 2022], empregam desde modelos clássicos até arquiteturas baseadas em Transformers, como o BERTimbau. A aplicação dessas técnicas ao domínio jurídico viabiliza o mapeamento sistemático de padrões discursivos em grandes bases processuais, fornecendo subsídios para políticas públicas e práticas formativas.

Tendo isso em vista, o presente estudo busca a elaboração de um modelo de rede neural que auxiliará na classificação dos discursos proferidos por agentes do direito em processos de violência sexual, identificando discurso nocivo e desumano nos âmbitos de gênero e raça.

2. Metodologia

Este estudo apresenta o desenvolvimento de um modelo de processamento de linguagem natural, baseado em rede neural, treinado a partir de um *corpus* composto por processos judiciais de casos de estupro em formato textual, contendo trechos anotados com manifestações de viés. Os processos que formam esse *corpus* estão sob segredo de justiça do Tribunal de Justiça do Estado de São Paulo (TJSP) e seu acesso foi autorizado para a pesquisa, mediante a assinatura de Termo de Confidencialidade e Sigilo por todos os pesquisadores e colaboradores envolvidos no projeto. O processo de anotação é realizado por uma equipe interdisciplinar, que identifica trechos de acordo com categoria, conotação e alvo pré-definidos, embasados e adaptados da ontologia estruturada por Marques [Marques 2021].

A categorização dos trechos identificados como portadores de estereótipo é estruturada em eixos principais e subcategorias. Os itens avaliados englobam: o conceito do crime de estupro (incluindo noções muitas vezes equivocadas de consentimento, menosprezo ao crime, violência e reação da vítima); discussões sobre provas; conduta pré-fato (abrangendo conduta sexual e vestimenta); e a vida pregressa (englobando contexto familiar, profissão e sexualidade). Além disso, mapeiam-se a descredibilização da vítima (por meio de alegadas inconsistências no depoimento ou falsa denúncia por vingança), o estado mental, a vulnerabilidade (envolvendo entorpecimento/embriaguez e grau de consciência) e a responsabilidade parental, que em geral ocorre atribuindo-se à mãe uma negligência pela ocorrência da violência.

Após a categorização, estabelece-se a conotação do trecho, dividida entre qualificadora (fala de cunho positivo com base na categorização padrão da sociedade) ou desqualificadora (cunho negativo). Por fim, define-se o alvo da fala, que engloba as possibilidades de vítima, pessoa agressora, mulher, homem, pessoa genérica, ou um conceito/ideia. Em paralelo com a anotação, realiza-se uma contínua análise dessas rotulações para a

melhor definição dos parâmetros de entrada e saída do modelo.

O *corpus* em construção é composto por cerca de 300 processos do TJSP. Atualmente, a base encontra-se em uma etapa de aprimoramento, passando por uma nova extração automatizada via portal e-SAJ com o intuito de maximizar a captura de informações. No estágio atual da pesquisa, a etapa de anotação encontra-se em fase piloto, contando com quatro processos integralmente rotulados por pesquisadoras especialistas da área do Direito e três em fase de processamento. Cada trecho é avaliado de maneira independente por, em média, dois anotadores (raramente três). Devido ao volume baixíssimo de trechos anotados nesta etapa preliminar, e considerando que a ontologia definitiva ainda será aplicada sistematicamente aos novos processos extraídos, a apresentação da distribuição estatística das categorias e a mensuração da concordância interanotadores serão realizadas apenas após a consolidação da base. Pretende-se aplicar o Coeficiente Kappa de Cohen [Cohen 1960] caso o protocolo de anotação seja consolidado com dois anotadores, ou uma métrica equivalente caso haja variação no *corpus*.

Em relação à arquitetura, estabelece-se um modelo de rede neural *Transformer*, baseado em mecanismos de atenção (*self-attention*). Em específico, implementa-se um modelo da família BERT [Devlin et al. 2019]. Visando o processamento de textos jurídicos em português brasileiro, adota-se inicialmente o modelo pré-treinado BERTimbau [Souza et al. 2020] como *baseline*. Adicionalmente, se viável no decorrer da pesquisa, será realizada uma avaliação comparativa de desempenho com outros modelos recentes para o português, como o ModernBERT-pt-BR [Wu e Garcia 2025] e o NorBERTo [Pellicer e Rinaldo 2024]. A partir dessa base, desenvolve-se o classificador automático.

Para a aplicação do modelo, adota-se o método de validação k-fold *cross-validation* com $k = 5$. O modelo passa por 5 turnos, sendo treinado com 80% do *corpus* disponível e testado com os 20% restantes, mantendo a separação o mais balanceada possível entre as categorias. Para avaliar a pertinência da rede, seu funcionamento é medido globalmente e por categoria através das métricas de acurácia, precisão, revocação e medida-F. Para os fins da pesquisa, classifica-se como "positivo" o evento no qual realmente existe uma expressão nociva no discurso, e "negativo" como um discurso neutro.

A estrutura geral da metodologia proposta, abrangendo desde a etapa de coleta de dados até a avaliação das métricas do modelo, é apresentada na Figura 1.

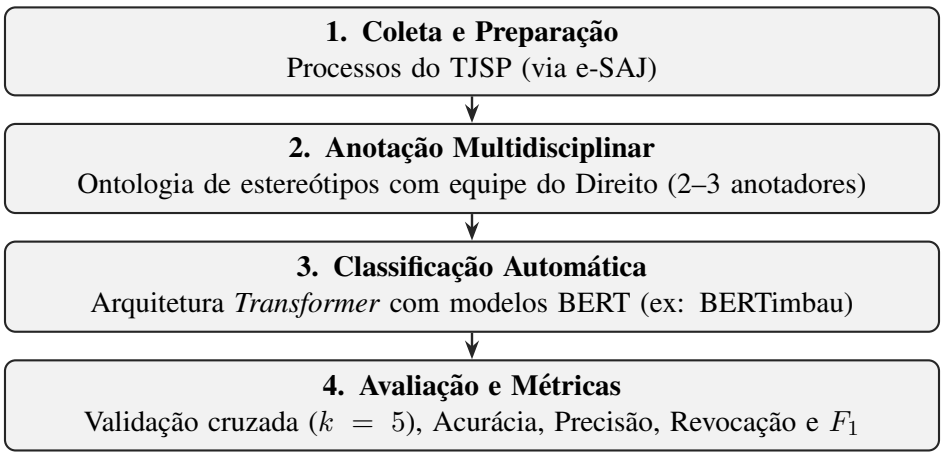

Figura 1. Fluxo metodológico geral da solução computacional proposta.

3. Resultados Esperados

Como resultado da investigação de possíveis manifestações de racismo e sexismo em discursos proferidos por agentes do sistema de justiça, esperam-se impactos de relevância tanto científica quanto social. Do ponto de vista social, a identificação desses padrões contribuirá para a compreensão empírica dos mecanismos institucionais associados à vitimização secundária em casos de violência sexual.

Do ponto de vista científico, espera-se contribuir para o desenvolvimento de métodos computacionais aplicados à análise de linguagem no domínio jurídico, particularmente no contexto do processamento de linguagem natural voltado à identificação de discurso nocivo. Ferramentas dessa natureza auxiliarão pesquisadores e instituições na análise de grandes volumes de documentos judiciais, permitindo investigações que seriam inviáveis por meio de análise manual.

Dessa forma, a consolidação do modelo de rede neural para a classificação de discursos racistas e sexistas em processos judiciais fornecerá um instrumento que auxilie na análise sistemática desses documentos, contribuindo efetivamente para estudos sobre viés institucional e para o aprimoramento de práticas operacionais no sistema de justiça.

4. Considerações Finais

Em suma, este estudo propõe uma abordagem interdisciplinar para a identificação de vieses racistas e sexistas no judiciário brasileiro, integrando técnicas de Processamento de Linguagem Natural à análise do discurso jurídico. A estruturação de um classificador automático baseado na arquitetura *Transformer*, alimentado por processos de crimes de violência sexual do TJSP, representa um avanço prático para a detecção em larga escala de falas associadas à vitimização secundária.

Ainda que a lida com dados sensíveis e sob segredo de justiça imponha desafios inerentes de acesso e processamento, a metodologia de anotação ancorada em ontologias de estereótipos confere rigor à ferramenta em desenvolvimento. Ao aplicar o aprendizado de máquina à revisão processual, a pesquisa oferece um método quantitativo e estruturado para observar a linguagem jurídica. Dessa forma, o projeto viabiliza a extração de métricas concretas sobre a presença de vieses nos autos, fornecendo recursos computacionais que facilitam a análise empírica desses documentos.

Referências

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Conselho Nacional do Ministério Público (s.d.). Vitimização. `https://www.cnmp.mp.br/defesadasvitimas/vitimas/vitimizacao`. Acesso em: 3 ago. 2026.

Devlin, J., Chang, M., Lee, K., e Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Leite, J. A., Silva, D., Bontcheva, K., e Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: Task and dataset with multilabel annotation. Em *Proceedings of the 4th Workshop on Online Abuse and Harms (ALW)*, pages 1–13.

Marques, C. S. (2021). Uma ontologia para permitir a análise da presença e influência do estereótipo de gênero em processos jurídicos de casos de estupro. Trabalho de conclusão de curso (bacharelado em ciência da computação), Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Ministério da Justiça e Segurança Pública (2025). Mapa de segurança pública 2025. Relatório técnico, Ministério da Justiça e Segurança Pública, Brasília, DF.

Pelle, R. d. e Moreira, V. (2017). Offensive comments in the Brazilian web: a dataset and baseline results. Em *Anais do VI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 510–519, Porto Alegre, RS, Brasil. SBC.

Pellicer, L. F. A. O. e Rinaldo, G. (2024). NorBERTo: A ModernBERT model trained for Portuguese with 331 billion tokens corpus. Em *Proceedings of the International Conference on the Computational Processing of Portuguese (PROPOR)*.

Siqueira, C. K. B. (2016). A liberdade sexual da mulher na prática judicial: análise da aplicação de estereótipos de gênero em processos de estupro. Dissertação de mestrado, Faculdade de Direito, Universidade Federal do Ceará, Fortaleza.

Sousa, R. F. d. (2017). Cultura do estupro: prática e incitação à violência sexual contra mulheres. *Estudos Feministas*, 25(1):9–29.

Souza, F., Nogueira, R., e Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems: 9th Brazilian Conference (BRACIS)*, pages 403–417. Springer International Publishing.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., e Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., e Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Wu, W. B. T. L. e Garcia, L. P. F. (2025). ModBERTBr: A ModernBERT-based model for Brazilian Portuguese. Em *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*.