# Desenvolvimento de um etiquetador de expressões sexistas baseado em BERT

Caio Túlio de Deus Andrade
Orientador: Marcelo Finger

26 de Maio de 2021

## 1 Introdução

No seu livro de 1996, "Estupro, Crime ou Cortesia?"[1] A Professora Sylvia Pimentel da Faculdade de direito da PUC-SP traz à tona uma série de fatores que conectam os processos judiciais de violência contra a mulher a estereótipos de gênero identificados nas falas dos entes jurídicos participantes das decisões do processo judicial. O livro resultou de uma pesquisa sócio-jurídica iniciada em 95 elaborada pela professora Sylvia Pimentel e outras autoras na qual estas fizeram uma análise qualitativa de decisões judiciais arquivadas em Tribunais de Justiça em 5 grandes cidades Brasileiras representando cada uma das regiões do Brasil: São Paulo (Sudeste), Manaus (Norte), Salvador (Nordeste), Goiânia (Centro oeste) e Porto Alegre (Sul).

Em 2019, nasce o projeto "Crime de Estupro No Sistema de Justiça Brasileiro - Abordagem sociojurídica de genêro". O projeto, coordenado pela professora Sylvia Pimentel novamente, visa atualizar as conclusões qualitativas da pesquisa de 96 e, devido ao incremento em poder computacional e quantidade de processos digitalizados nos últimos 20 anos, visa também trazer uma discussão quantitativa além de qualitativa ao objeto de pesquisa.

Como observado por Pymentel em 96, os estereótipos nas decisões dos entes jurídicos podem ser identificáveis no texto do processo. O foco deste trabalho de conclusão de curso será, trabalhando no enfoque quantitativo do projeto de pesquisa, utilizar técnicas de Aprendizado de Máquina em Processamento de Linguagem Natural (PLN) para automaticamente detectar trechos dos processos que possuem falas estereotipadas.

Como estamos procurando vieses subjetivos no texto, essa informação não está contida explicitamente no texto. Para tanto, um processo de etiquetagem é necessário para incluir a informação no texto, de forma que o modelo seja capaz de entender a presença da informação alvo que ele deve aprender. No entanto, o processo de etiquetagem é caro por se tratar de trabalho braçal e intelectual humano. Dessa forma, temos potencialmente acesso a um volume significativo de dados não etiquetados e um processo custoso de etiquetagem. Nesse tipo de

contexto, a técnica de aprendizado ativo[2] se destaca: utilizando essa técnica, seremos capazes de construir um conjunto de dados mais representativo e que, por sua vez, possibilita que o modelo atinja um nível satisfatório de performance com maior rapidez.

Os processos judiciais disponibilizados abertamente pelo TJ-SP [https://www.tjsp.jus.br/] serão anotados por uma equipe de advogadas, usando o INCEpTION [3]: uma ferramenta de anotação que permite adicionar informação explicitamente a um texto. O intuito em utilizar uma equipe para anotar em conjunto um processo é obter um julgamento mais sólido possível sobre a presença ou não de estereótipos na fala dos entes jurídicos.

Um exmeplo de uso do INCEpTION seria analisar um trecho de uma reportagem. Digamos que um pesquisador está interessado em destacar termos relacionados a organizações (ORG), locais (LOC) e Predicados relacionados a indivíduos (PERDeriv) no texto. Todos outros termos relevantes, mas sem categoria definida podem ser chamados de Outros(OTH). O nome dessas etiquetas é completamente arbitrário e é definido pelo anotador. Este texto anotado por um indivíduo utilizando o INCEpTION tem a seguinte aparência após ser anotado:

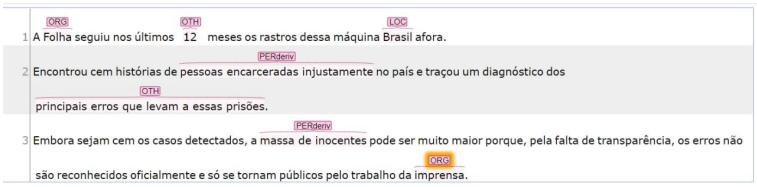


Figure 1: Exemplo de texto anotado pelo INCEpTION [3]

O dado estruturado retornado pela ferramenta (um arquivo de texto em formato JSON, XML, CONLL dentre outros possíveis formatos) pode, então, ser alimentado a um modelo de aprendizado de máquina.

Abaixo, a saída simplificada do INCEpTION ao anotar o trecho da reportagem. Um algoritmo de aprendizado supervisionado, então, poderia potencialmente aprender a reconhecer essas entidades nomeadas.

```
<?xml version="1.0" encoding="UTF-8"?>
    <Document>
    <!-- Metadados -->
            <ner.type.NamedEntity begin="0" end="1" value="null"
                identifier="null">A</ner.type.NamedEntity>
            <ner.type.NamedEntity begin="2" end="7" value="ORG"
                identifier="null">Folha</ner.type.NamedEntity>

            <!-- ... -->
             <segmentation.type.Token begin="75" end="84"
                 >Encontrou</segmentation.type.Token>

            <segmentation.type.Token begin="85" end="88"
```

```
            >cem</segmentation.type.Token>

        <segmentation.type.Token begin="89" end="98"
            >historias</segmentation.type.Token>

        <segmentation.type.Token begin="99" end="101"
            >de</segmentation.type.Token>

        <ner.type.NamedEntity begin="102" end="135" value="PERderiv"
            identifier="null">
           <segmentation.type.Token begin="102" end="109"
               >pessoas</segmentation.type.Token>

           <segmentation.type.Token begin="110" end="122"
               >encarceradas</segmentation.type.Token>

           <segmentation.type.Token begin="123" end="135"
               >injustamente</segmentation.type.Token>
        </ner.type.NamedEntity>


        <segmentation.type.Token begin="68" end="73"
            >afora</segmentation.type.Token>
        <segmentation.type.Token begin="73" end="74"
            >.</segmentation.type.Token>
    </segmentation.type.Sentence>
</Document>
```

O texto anotado apresenta etiquetas que marcam o início e o fim de regiões de interesse (no caso deste trabalho, falas contendo julgamentos preconceituosos). O modelo de linguagem natural deve, então, ser capaz de aprender a posicionar corretamente tags de início e fim no texto de forma a delimitar de forma satisfatória uma região de interesse.

No campo de PLN, modelos que resolvem esta tarefa geralmente são modelos que aprendem um modelo de linguagem: um modelo de linguagem é uma formalização de um aspecto linguístico. Os modelos se encaixam em três possíveis categorias: lógico/gramatical, probabilístico ou conexionista. Cada uma dessas classes foca em uma característica diferente do texto para construir o modelo: em modelos lógicos/gramaticais, o foco é simbólico; Em modelos probabilísticos, a modelagem é construída sobre a estatística da ocorrência de palavras em um dado corpus. Por fim, a abordagem conexionista tem como foco a relação entre as palavras, e para tanto utiliza uma abordagem neural, assinalando pesos para relacionar e definir palavras. Essa técnica é chamada de word embedding [6]. Neste trabalho, focaremos em modelos de linguagem conexionistas, ou neurais.

Na literatura, as soluções prevalentes para construção de modelos neurais de linguagem e, portanto, tarefas de marcação, envolviam Redes Neurais Recorrentes, mais em específico, a arquitetura de Redes Long Short Term Memory (LSTM [4]). Essas duas arquiteturas parseiam o texto palavra a palavra, lendo

sequencialmente uma palavra por vez. Este tipo de abordagem apresenta dois problemas:

1. Treino lento: Como o processo de aprendizado da rede neural é necessariamente sequencial, o uso de paralelismo é reduzido.

2. Impossibilidade de aprender dependência de palavras da direita para a esquerda e da esquerda para a direita ao mesmo tempo. Uma solução intermediária é usar uma rede neural lendo o texto do início ao fim e outra rede lendo do fim ao início (LSTM bidirecional) e combinando o resultado das duas redes.

Em 2018, pesquisadores do Google apresentaram o modelo BERT[5] (Bidirecional Encoder Representation from Transformers), um modelo pré-treinado e verdadeiramente bidirecional que faz uso da arquitetura de transformadores [7] para resolver diversas tarefas em PLN, dentre elas, a tarefa de marcação e, portanto, de modelos de linguagem.

Utilizar um modelo pré-treinado é fundamentalmente uma estratégia de empregar modelos poderosos com baixos tempos de treino. Em um processo comum de treino, devemos encontrar quais são os parâmetros do nosso modelo que maximizam a predição dele em uma tarefa específica. O processo de pré treino consiste em um processo de treinamento em uma ou diversas tarefas genéricas (mas razoavelmente próximas à tarefa de interesse) em máquinas extremamente potentes e com acesso a um grande volume de dados. Dessa maneira, o processo de pré treino nos poupa tempo ao providenciar um modelo razoavelmente acurado.

No entanto, o processo de pré treinamento não é capaz de criar um modelo com grande acurácia para qualquer tarefa, afinal, o processo de pré treinamento é, por essência, um processo generalista. Dessa forma, é importante realizar uma etapa de refinamento do modelo à uma tarefa específica. No caso deste trabalho, o refinamento será responsável por especializar o modelo na tarefa de etiquetagem. Isto posto, o BERT é um modelo capaz de reduzir drasticamente os tempos de treino para obter resultados satisfatórios mesmo empregando uma arquitetura sofisticada e complexa.

Outro ponto da arquitetura do modelo que nos possibilita melhores tempos de treino é o uso de uma arquitetura de Redes Neurais chamada de transformadores [7], que permite o processamento paralelo com muito mais facilidade. Ademais, o BERT é verdadeiramente bidirecional devido ao seu uso do mecanismo de auto-atenção, na qual todas palavras da sentença analisada são relacionadas entre si de maneira densa.

O modelo em questão rapidamente se tornou o padrão da área e, portanto, foi adotado como ferramenta central neste trabalho.

# 2 Objetivo

O objetivo do trabalho de conclusão será desenvolver um etiquetador de expressões sexistas baseado no modelo BERT. Este etiquetador poderá ser uti-

lizado, dentre outros possíveis contextos, para auxiliar a identificar expressões sexistas em decisões judiciais.

# 3 Cronograma

1. Estudar o modelo BERT
2. Estudo de aprendizagem ativa
3. Estudar ferramentas de etiquetagem
4. Preparação de um ambiente para etiquetagem colaborativa
5. Construir um modelo de etiquetagem baseado em BERT para fins didáticos
6. Preparação dos dados
7. Treinamento do modelo
8. Escrita da Monografia
9. Preparação da apresentação

| Atividade | Abr | Mai | Jun | Jul | Ago | Set | Out | Nov | Dez | Jan |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | x | x | x | x | | | | | | |
| 2. | | x | x | x | x | | | | | |
| 3. | | | x | x | | | | | | |
| 4. | | | x | x | x | | | | | |
| 5. | | | | x | | | | | | |
| 6. | | | | x | x | | | | | |
| 7. | | | | | x | x | x | x | x | |
| 8. | | | | | x | x | x | x | x | |
| 9. | | | | | | | | x | x | x |

# References


[1] Silvia Pimentel, Ana Lucia P. Schritzmeyer, Valeria Pandjiarjian, "Estupro: crime ou "cortesia"? : abordagem sociojurídica de gênero" Porto Alegre, S.a. Fabris, 1998.

[2] Settles, Burr; University of Wisconsin–Madison, 2009 - "Active Learning Literature Survey"
http://axon.cs.byu.edu/ martinez/classes/778/Papers/settles.activelearning.pdf

[3] INCEpTION - https://inception-project.github.io/

[4] Hochreiter, Schmidhuber, 1997 - "Long Short Term Memory" - www.bioinf.jku.at/publications/older/2604.pdf

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018 - "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"
https://arxiv.org/abs/1810.04805

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013 - " Efficient Estimation of Word Representations in Vector Space"
https://arxiv.org/abs/1301.3781

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, 2017 - "Attention is all you need"
https://arxiv.org/abs/1706.03762