UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Classificação de decisões judiciais usando modelos de linguagem

André Santos Cavatoni Serra

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

André Santos Cavatoni Serra

Classificação de decisões judiciais usando modelos de linguagem

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dr. Brucce Neves dos Santos

Versão original

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Cavatoni, André
S856m Classificação de decisões judiciais usando modelos de linguagem / André Santos Cavatoni Serra ; orientador Dr. Brucce Neves dos Santos. – São Carlos, 2024.
39 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Processamento Natural de Linguagem. 2. Classificação automática de textos. I. SANTOS, Brucce N., orient. II. Título.

André Santos Cavatoni Serra

Classification of judicial decisions using language models

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Dr. Brucce Neves dos Santos

Original version

São Carlos

2024

AGRADECIMENTOS

Primeiramente, agradeço à minha mãe e ao meu pai (*in memoriam*), por me ensinarem o amor incondicional e como os estudos e o trabalho árduo podem melhorar o mundo. À minha esposa, pelo amor imenso e por todo o apoio e compreensão, sem os quais este trabalho não seria possível. Às minhas filhas, por serem meu motor propulsor e por me ensinarem a ser cada dia melhor através da paciência e da dedicação. Ao meu irmão, por continuar crescendo comigo, mesmo em cidades diferentes. Agradeço à *startup* Quero Meus Direitos por me proporcionar a oportunidade de desenvolver este trabalho. Finalmente, agradeço ao meu orientador por toda a paciência e ensinamentos.

"A máquina ameaça tudo que se conquistou.
Ao pretender estar no espírito e não na obediência."
Rainer M. Rilke

Sonetos a Orfeu 2005, p. 91, parte II, 10.

RESUMO

CAVATONI, André **Classificação de decisões judiciais usando modelos de linguagem**. 2024. 39p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

No Brasil, em 2023, tivemos 84.868.444 processos em andamento, sendo 37.046.875 novos processos, o que resulta em 106 novos processos por minuto. Tais demandas levam à necessidade de automatizar tarefas repetitivas e organizar as informações com o intuito de gerar *insights* para a sociedade. Nesse contexto, este trabalho se propõe a avaliar técnicas de mineração de texto, mais precisamente o refinamento de modelos de linguagem pré-treinados, para a classificação de decisões judiciais brasileiras. O conjunto de dados é composto por movimentações processuais do tipo decisão, classificadas manualmente como: Sentença, Acórdão e Embargo. Cada uma dessas decisões também foi classificada como Procedente, Improcedente ou Extinta. Os modelos foram treinados em duas tarefas: classificação multi-rótulo e classificação multi-classe. Foram avaliados diferentes modelos, dentre eles o modelo BERTimbau. O modelo BERTimbau obteve a melhor performance, tanto na tarefa de classificação multi-classe quanto na tarefa de classificação multi-rótulo, que se mostrou a melhor abordagem para esse problema de classificação.

Palavras-chave: Modelos de Linguagem Pré-treinados. Classificação Multi-rótulo. Classificação Multi-classe. Processamento de Linguagem Natural (PLN). Análise de Sentenças. Inteligência Artificial no Direito.

ABSTRACT

CAVATONI, André **Classification of judicial decisions using language models**. 2024. 39p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

In Brazil, in 2023, there were 84,868,444 ongoing cases, with 37,046,875 new cases, resulting in 106 new cases per minute. These demands highlight the need to automate repetitive tasks and organize information to generate insights for society. In this context, this work aims to evaluate text mining techniques, specifically the fine-tuning of pre-trained language models, for the classification of Brazilian judicial decisions. The dataset consists of procedural actions classified as: *Sentença*, *Acórdão*, and *Embargo*. Each of these decisions was also classified as either *Procedente*, *Improcedente*, or *Extinta*. The models were trained for two tasks: multi-label classification and multi-class classification. Several models were evaluated, among them the BERTimbau model. The BERTimbau model achieved the best performance in both the multi-class classification task and the multi-label classification task, which proved to be the best approach for this classification problem.

Keywords: Pre-trained Language Models. Multi-label Classification. Multi-class Classification. Natural Language Processing (NLP). Sentence Analysis. Artificial Intelligence in Law.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O número de novos processos no Brasil vem aumentando consideravelmente nos últimos anos. Segundo o último relatório de 2023 do Conselho Nacional de Justiça (CNJ) (CNJ, 2023), o número de processos em tramitação no Brasil foi de 84.868.444 e o número de novos processos foi de 37.046.875, o que dá 106 novos processos por minuto. De acordo com o Relatório Justiça em Números 2022, do CNJ (CNJ, 2022), o número total de processos distribuídos no Brasil em 2022 foi de 32.883.738. Esse número representa um aumento de 12,66% em relação a 2021. Na Tabela 1 abaixo, podemos comparar o aumento de processos desde 2020:

ANO	PROCESSOS NOVOS
2020	27.325.451
2021	30.053.083
2022	32.883.738
2023	37.046.875

Tabela 1 – Novos Processos por Ano

Só no Tribunal de Justiça do Estado de São Paulo (TJSP), em 2023, o número de processos em tramitação foi de 25.026.307, representando um montante de 29,48% do total de processos em tramitação no Brasil (Painel CNJ, 2023). O número de processos distribuídos no TJSP em 2023 foi de 7.073.767, representando um montante de 19,09% do total de processos distribuídos no Brasil (Painel CNJ, 2023). Segundo o site do TJSP, o Tribunal de Justiça do Estado de São Paulo é o maior tribunal do mundo em volume de processos [1].

Esse volume e crescimento são um desafio para o Poder Judiciário, que precisa lidar com a demanda crescente de processos e com a necessidade de dar respostas mais rápidas à sociedade. A tecnologia pode ser uma aliada para enfrentar esse desafio. A automação de tarefas repetitivas auxilia no aumento da produtividade dos servidores e de escritórios de advocacia. Além disso, o uso de algoritmos de inteligência artificial pode reduzir o tempo de tramitação dos processos, organizar melhor as informações e tornar mais precisas as análises processuais, tanto dos tribunais quanto dos advogados.

A gestão e análise de dados sobre decisões nos tribunais brasileiros enfrentam obstáculos significativos, incluindo a escassez e a desorganização de informações. Essa situação prejudica não apenas pesquisadores e profissionais do direito, mas também a formulação de políticas públicas eficientes. Por meio da inteligência artificial, é possível automatizar a análise de grandes volumes de dados, identificando padrões em decisões

[1] https://www.tjsp.jus.br/QuemSomos - Acessado em 13 de Abril de 2024.

de processos judiciais brasileiros. Essa abordagem pode oferecer *insights* valiosos para o Estado e organizações, permitindo, por exemplo, mensurar a efetividade jurídica no Brasil.

1.1 Objetivos

Este trabalho tem como objetivo geral avaliar a aplicabilidade de técnicas de mineração de texto, mais precisamente ajuste fino em modelos de linguagem pré-treinados, para classificação de decisões judiciais brasileiras. Para atingir o objetivo geral, foram definidos objetivos específicos:

- Coletar e rotular um conjunto de dados com as movimentações processuais de decisões judiciais.
- Comparar modelos de linguagem pré-treinados, ajustados por *fine-tuning*, para tarefas de classificação de textos multi-rótulo e multi-classe, utilizando publicações judiciais, do tipo decisão, como base para a classificação.

2 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo tem o objetivo de esclarecer os principais fundamentos para o bom entendimento deste trabalho. Primeiramente na Seção 2.1 é descrito com mais detalhes o tema desta pesquisa, incluindo conceitos jurídicos e as classes que foram usadas para rotular os dados. Na Seção 2.2 é descrito o primeiro passo prático que é a obtenção dos dados. Na Seção 2.3 é descrito alguns modelos de língua e sua aplicação nesse projeto. E, por fim, na Seção 2.4 são apresentados os trabalhos relacionados.

2.1 Publicações jurídicas

Publicações jurídicas são textos informativos que visam comunicar qualquer andamento de um processo jurídico, para as partes interessadas. Essas publicações podem ser públicas ou privadas. As publicações privadas somente os advogados das partes, e o juiz do caso, têm acesso e se dão, em sua maioria, quando o processo corre em segredo de justiça. As publicações que serão tratadas neste TCC são públicas e estão disponíveis nos sites dos diários oficiais eletrônicos dos tribunais de justiça dos estados brasileiros. P. Portanto, por se tratar de dados públicos, sua coleta não infringe qualquer regra, em especial à LGPD.

Existem várias classes em que cada publicação jurídica pode se enquadrar dentre elas: decisões, acordo, agravo de instrumento, apelação, audiência, cálculos, dilação, impugnação, etc. Neste trabalho o foco está nas publicações do tipo Decisão. Decisões são publicações que se referem aos documentos oficiais emitidos por autoridades judiciais, como juízes ou tribunais, que resolvem disputas e determinam o resultado de processos judiciais. As Decisões podem ser classificadas em três tipos únicos: Sentença, Acórdão e Embargos.

- Sentença é a decisão judicial que resolve o mérito da causa ou encerra o processo sem resolução de mérito. É proferida por um juiz, na primeira instância do processo. Existem quatro possíveis resultados (classes): “Totalmente procedente”, quando o juiz aceita todos os pedidos da parte autora, como danos morais e honorários sucumbenciais; “Parcialmente procedente”, quando apenas alguns pedidos são aceitos, como conceder honorários sucumbenciais, mas não danos morais; “Improcedente”, quando nenhum pedido da parte autora é aceito; e “Processo extinto”, quando o processo é encerrado por motivos como pedido da parte autora ou erros materiais.
- Acórdão, diferente da sentença que é proferida por um único juiz de primeiro grau, o acórdão é resultado da deliberação de um órgão colegiado, composto por vários juízes ou desembargadores. Normalmente ocorre em segunda instância ou órgão superiores. O acórdão pode ser “Provido”, quando o tribunal acolheu o recurso, ou “Desprovido”, quando o tribunal não acolheu o recurso.

- Embargo é um tipo de recurso cuja finalidade é contestar, esclarecer ou corrigir uma decisão. Utilizado para esclarecer pontos obscuros, omissões ou contradições em decisões judiciais. Os Embargos podem ser de Primeira Instância (Embargos de Sentença) ou de Segunda Instância (Embargos de Acórdão). Os Embargos podem ser classificados em "Acolhidos" ou "Não acolhidos".

Para simplificação da classificação desse trabalho, as classes "Totalmente procedente" e "Parcialmente procedente" foram agrupadas em uma única classe chamada "Procedente". As classes "Provido" e "Desprovido", foram classificadas como "Procedente" e "Improcedente" respectivamente. Da mesma forma as classes "Acolhido" e "Não acolhido" foram classificadas como "Procedente" e "Improcedente", respectivamente. A classificação sugerida segue na Figura 1.

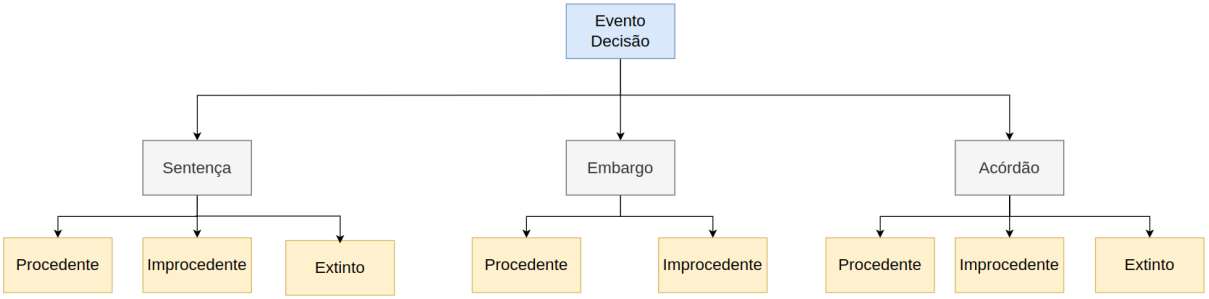

Figura 1 – Árvore de classificação de decisões judiciais

2.2 Classificação

Para a classificação das Decisões serão utilizadas técnicas de mineração de texto. A Mineração de texto é definida como um "conjunto de técnicas e processos para descoberta de conhecimento inovador a partir de grandes coleções textuais" (REZENDE, 2003)

Classificação de textos visa categorizar textos em um ou mais rótulos pré-definidos. A classificação pode ser de três tipos: binária, multi-classe ou Multi-rótulo. A classificação binária é a mais simples, onde cada entrada é classificada entre um de dois rótulos (ex.: verificar se um *e-mail* é *spam* ou não). Na classificação multi-classe o texto é classificado somente com um rótulo, entre vários rótulos disponíveis (ex.: classificar notícias como esporte, política ou tecnologia). Já na classificação multi-rótulo, a entrada textual pode pertencer a vários rótulos simultaneamente (ex.: uma notícia poderia ser rotulada como política, economia e internacional ao mesmo tempo). O foco deste trabalho está na rotulação multi-classe e multi-rótulo.

Dentre as diversas técnicas para realizar classificação de texto, este estudo tem como objetivo investigar modelos de linguagem pré-treinados. Modelos de linguagem são

ferramentas desenvolvidas com técnicas de *deep learning* com o intuito de interpretar, manipular e gerar linguagem natural. Atualmente, os modelos baseados em *Transformer* proposto por (VASWANI *et al.*, 2017), possibilitaram grandes avanços no campo do processamento natural de linguagem - PLN - por realizar processamento paralelo de sequências e captura de dependências de longo alcance entre palavras e subunidades de texto. Com seu mecanismo de atenção, ele aprende a focar em partes relevantes da sequência de entrada ao realizar uma tarefa específica, como tradução ou geração de texto.

Os modelos de linguagem transformam palavras, frases, e até documentos, em representações vetoriais numéricas, de tal forma que possam ser processadas computacionalmente. Essas representações vetoriais são chamadas de *embeddings*. Cada palavra ou frase é mapeada para um ponto no espaço de N dimensões, de tal forma que palavras com significados, ou em contextos semelhantes, são posicionadas próximas umas das outras. Modelos de linguagem modernos usam *embeddings* contextuais, ou seja, a representação vetorial de uma palavra varia de acordo com o contexto em que a mesma está inserida.

Dentre os diversos modelos de linguagem existentes, neste trabalho será explorado o uso de modelos de linguagem pré-treinados, em que para serem utilizados é necessário realizar o *fine-tuning* em uma determinada tarefa, nesse caso será a classificação de textos. O processo de *fine-tuning*, envolve usar dados específicos para refinar um modelo já pré-treinado, o especializando em uma tarefa específica. Essa abordagem é interessante pois permite que o modelo atinja os resultados esperados com muito menos dados de treinamento do que seria necessário se começasse do zero ou utilizando outras abordagens para classificação de texto, além de consumir menos recursos computacionais, se comparado com a etapa de pré-treinamento.

Dentre os modelos de linguagem existentes, este trabalho irá explorar os seguintes modelos de linguagem:

- **BERT** (DEVLIN *et al.*, 2019): modelo de linguagem baseado em *Transformers*. O treinamento desse modelo consistiu em ensiná-lo a identificar contextos de palavras de forma bidirecional, analisando tanto o texto da esquerda para a direita quanto da direita para a esquerda. Essa capacidade de entender o texto de forma bidirecional permite que o BERT capture nuances e subtextos em textos, sendo importante para muitas tarefas de classificação, como determinar o tema de um texto.

- **JurisBERT** (VIEGAS, 2022): especialmente treinado em um grande corpus de documentos jurídicos brasileiros, como leis, decisões judiciais, contratos e outros textos legais. Esse treinamento especializado permite que o modelo compreenda melhor a terminologia, a sintaxe e as nuances da linguagem jurídica, tornando-o uma ferramenta poderosa para aplicações legais.

- **BERTimbau** (SOUZA; NOGUEIRA; LOTUFO, 2020): uma versão do BERT treinada com corpus em português do Brasil, o que o torna mais eficiente para aplicações de Processamento de Linguagem Natural (PLN) nesse idioma.

- **BERTikal** (POLO *et al.*, 2021): parte de um pacote chamado LegalNLP, que contém *embeddings* e modelos pré-treinados para a linguagem jurídica do Brasil. Além de disponibilizar funções que facilitam a manipulação de textos legais, este modelo pode ser visto como uma especialização do BERTimbau para textos jurídicos, uma vez que continuou o pré-treinamento do BERTimbau focado em textos jurídicos.

2.3 Trabalhos relacionados

A classificação de textos jurídicos é um desafio, uma vez que a linguagem jurídica é bem particular. Com o intuito de compreender o estado da arte, foram selecionados artigos focados em classificação de textos e, em sua maioria, em técnicas aplicadas à língua portuguesa do Brasil.

O estudo realizado por POLO *et al.* (2021) descreve um projeto que resultou em modelos de linguagem pré-treinados para a linguagem jurídica do Brasil. Além disso, os autores disponibilizaram um pacote Python com funções que facilitam o uso desses modelos. O trabalho se mostra muito relevante uma vez que há carência de modelos treinados para aplicações no âmbito jurídico brasileiro. Importante destacar que aplicações são demonstradas no artigo como: previsão de status de processos jurídicos e tokenização.

Os autores GOMES LUCAS MOREIRA; JADER MARTINS CAMBOIM DE Sá; PENG (2020), demonstraram o uso de técnicas de mineração de texto focado em classificação de sentenças judiciais em relação à procedência relativa ao pedido feito pelo autor da ação. Entre as técnicas discutidas destaca-se *Bag-of-Words*, TF-IDF e N-gramas, que são usadas de forma combinada para atingir um resultado satisfatório. O artigo relata grande eficiência (por meio das métricas: precisão, *recall* e F1-*score*) dessas técnicas combinadas para classificar sentenças como procedentes, parcialmente procedentes, improcedentes, acordos, e outras.

O autor VIEGAS (2022) propôs o modelo JurisBERT um modelo BERT treinado do zero (não realizaram *fine-tuning*), utilizando um grande corpus, de desenvolvimento próprio, de textos específicos da área jurídica, incluindo: leis, decisões, votos de decisões, tratados legais. A arquitetura do JurisBERT é baseada em *Transformers*, o que permite ao modelo reter contextos bi-direcionais (tanto à direita quanto à esquerda do tokens) de forma eficaz.

O BERTimbau, desenvolvido por SOUZA R.F. NOGUEIRA (2023), é um modelo de linguagem, específico para a língua brasileira, disponível em duas versões sendo elas a versão básica (base) e a versão grande (large). O BERTimbau foi avaliado em três

tarefas de NLP: similaridade textual de sentenças, reconhecimento de implicação textual e reconhecimento de entidades nomeadas, superando a versão multilingue do BERT. Os modelos estão disponíveis para a comunidade em bibliotecas de código aberto.

Uma estratégia baseada em assemble, foi proposta por GUIMARAES (2023) com o objetivo de melhorar a precisão na classificação de questões jurídicas nas seguintes áreas do Direito brasileiro: Direito do Consumidor, Direito de Família e Direito do Trabalho. Questões jurídicas a classificar podem ser: falha na prestação de serviços, publicidade enganosa, cobranças indevidas, divórcio, adoção, jornadas de trabalho, assédio moral no trabalho, etc. Dois sistemas especialistas de processamento de linguagem natural foram empregados nessa abordagem. Foram treinados respectivamente com textos em linguagem "popular" e "não popular", e um classificador para identificá-los, alcançando uma acurácia geral de 96%.

3 METODOLOGIA

Como discutido na seção 2.1, uma publicação jurídica pode ser dividida em diversas categorias, dentre elas esse trabalho explora publicações envolvendo eventos de decisão. Para isso, essa proposta é dividida em cinco etapas, sendo elas: aquisição dos dados e pré-processamento que são abordados na seção 3.1; já a etapa de classificação das publicações é abordada na seção 3.2. Por fim, as etapas de avaliação, ajustes e implantação do modelo são abordados no Capítulo 4. A metodologia proposta é ilustrada na Figura 2.

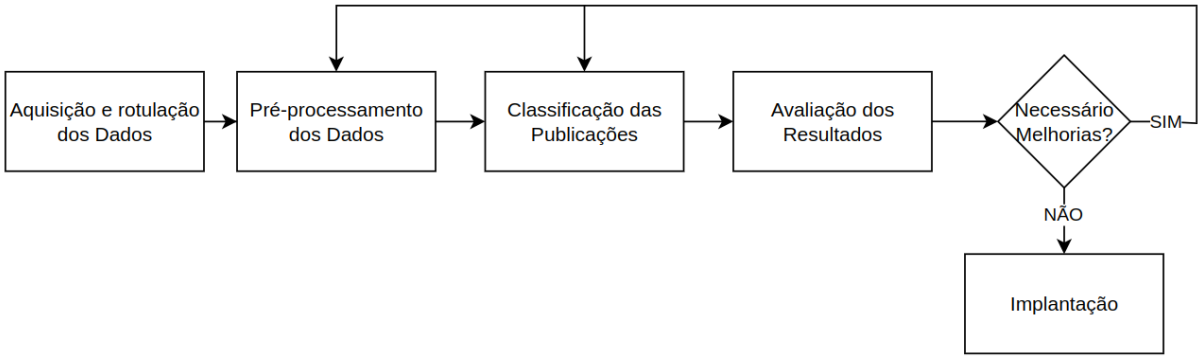

Figura 2 – Fluxograma da metodologia proposta.

3.1 Aquisição, Rotulação dos Dados e Pré-processamento

Os sistemas dos tribunais de justiça do Brasil disponibilizam publicamente as publicações jurídicas, que são capturadas por meio de uma API proprietária, desenvolvida pela empresa Publicações Online[1]. Com o acesso a essa API é possível coletar as publicações do tipo Decisão. Em seguida, o texto dessas publicações de Decisão foram limpos e disponibilizados para rotulação em todas as classes discutidas na seção 2.1. Na Figura 3 está ilustrado esse processo.

[1] <http://www.publicacoesonline.com.br> (acessado em 22/06/2024)

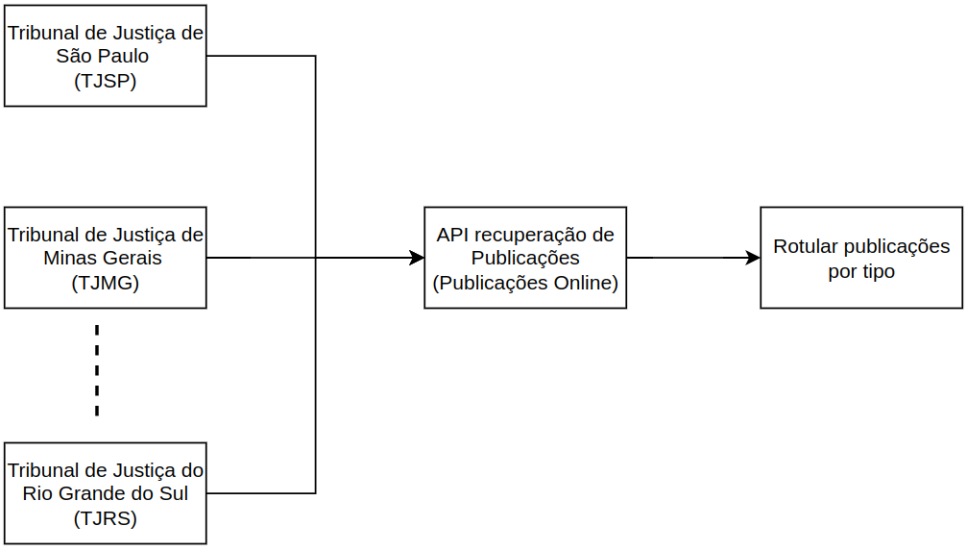

Figura 3 – Aquisição e rotulação dos dados.

3.2 Classificação das publicações

A classificação foi realizada em duas etapas conforme ilustrados na Figura 4.

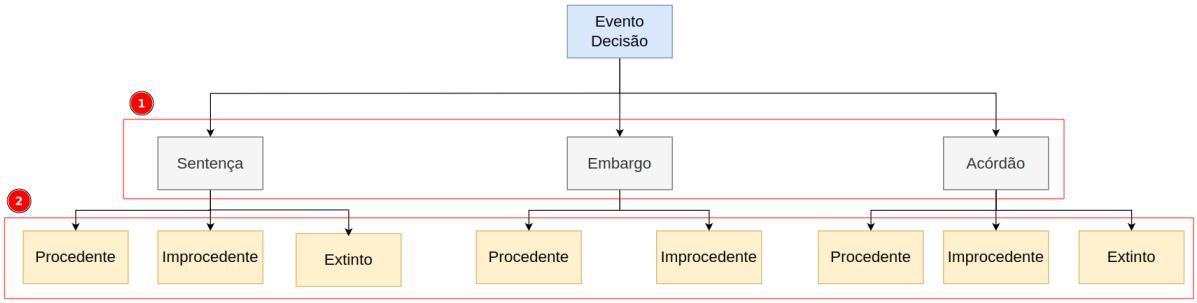

Figura 4 – Etapas classificação de Decisões.

Essas etapas são:

- **Etapa 1:** consiste em classificar um evento de decisão como Sentença, Acórdão ou Embargo.
- **Etapa 2:** dependendo da classificação da etapa anterior, o evento será classificado em: Procedente, Improcedente ou Extinto caso o evento tenha sido classificado como Sentença ou Acórdão. Se o evento foi classificado como Embargo, então nessa etapa ele será classificado como Procedente ou Improcedente.

4 AVALIAÇÃO EXPERIMENTAL

Esse capítulo tem como objetivo descrever a parte prática do desenvolvimento do projeto e os resultados obtidos. Como citado na seção 2.2 este estudo tem como objetivo avaliar 4 modelos de linguagem sendo eles: BERT, BERTimbau, BERTikal e JurisBERT. Esses modelos foram avaliados em duas tarefas distintas, sendo elas a classificação multi-rótulo e a classificação multi-classe.

Na seção 4.1 é descrito o conjunto de dados bem como a distribuição das classes. Na seção 4.2 é descrito a etapa de experimentação passando pela limpeza e divisão dos dados, algoritmos usados e métricas para avaliação do resultado. Finalmente na seção 4.3 são apresentados e discutidos os resultados.

4.1 Conjuntos de Dados

Como mencionado no capítulo 3, o conjunto de dados é formado por textos de publicações jurídicas brasileiras, do tipo decisão. Como ilustrado na Figura 4 , nosso conjunto de dados possui dois níveis de classificação. O primeiro nível dividimos em: Sentença, Embargo e Acórdão. Já no segundo Nível dividimos em: Procedente, Improcedente e Extinto, sendo que a classificação Extinto só se aplica a Sentença e Acórdão.

Os textos são extraídos do site dos tribunais de justiça de todo o país, e foram classificados manualmente, totalizando aproximadamente duzentos textos de cada par de classe. Já com os textos classificados, foram criados dois conjuntos de dados: multi-rótulo e multi-classe. Como citado na seção 2.2, na abordagem multi-rótulo podemos ter vários rótulos para o mesmo texto enquanto na abordagem multi-classe temos apenas um rótulo por texto. A Tabela 2 mostra a distribuição dos rótulos para a abordagem multi-classe e multi-rótulo.

Multi-rótulo		**Multi-classe**	**Quantidade**
Nível 1	**Nível 2**	**Rótulo**	
Sentença	Procedente	Sentença - Procedente	196
Sentença	Improcedente	Sentença - Improcedente	198
Sentença	Extinção	Sentença - Extinção	199
Embargos	Procedente	Embargos - Procedente	199
Embargos	Improcedente	Embargos - Improcedente	200
Acórdão	Procedente	Acórdão - Procedente	177
Acórdão	Improcedente	Acórdão - Improcedente	173
Acórdão	Extinção	Acórdão - Extinção	199
		Total	**1541**

Tabela 2 – Quantidade de rótulos para as abordagens multi-classe e multi-rótulo.

4.2 Experimentação

As abordagens multi-rótulo e a multi-classe foram aplicadas para cada um dos quatro modelos de linguagens pré treinados escolhidos, totalizando 8 experimentos.

4.2.1 Limpeza dos Dados e Tokenização

Em um primeiro momento os dados são carregados e limpos utilizando a função clean da biblioteca LegalNLP. Em seguida os textos são tokenizados, utilizando o tokenizador compatível com cada modelo de linguagem escolhido. A função de tokenização utiliza *padding* e truncamento garantindo assim que todas as sequências tenham o mesmo comprimento.

4.2.2 Divisão dos Dados

A divisão de dados foi realizada utilizando a técnica "*Stratified* K-*Fold*", uma técnica de validação cruzada que garante que a distribuição de classes seja mantida em cada um dos *folds*. A validação cruzada garante que o modelo seja treinado e avaliado em diferentes subconjuntos de dados de tal forma que seja possível avaliar sua capacidade de generalização. O conjunto de dados foi dividido em 5 *folds* sendo 4 para treinamento e 1 para teste, sendo essa uma prática comum ao trabalhar com modelos de linguagem. Os *folds* utilizados para treinamento são divididos em conjunto de treino e validação, sendo a validação 10% do conjunto de treino.

4.2.3 Treinamento

Todos os modelos foram treinados com o mesmo conjunto de hiperparâmetros, sendo eles:

- Taxa de aprendizado (*learning rate*): foi definida como 2e-5, esse é um valor comum ao realizar o *fine-tuning* para evitar grandes mudanças nos pesos do modelo. Assim os ajustes nos pesos são feitos gradualmente, preservando o conhecimento adquirido na fase de pré-treinamento.
- Número de épocas (*epochs*): os modelos foram treinados por 10 épocas e foi selecionado a melhor época com base no conjunto de validação.
- Decaimento de peso (*weight decay*): foi definido como 0.01, sendo utilizado durante a regularização, minimizando risco de *overfitting*.

4.2.4 Avaliação

Após o treinamento, o modelo é avaliado usando o conjunto de teste. As métricas utilizadas foram: precisão, F1-Macro revocação e acurácia, sendo que a métrica F1-Macro,

foi usada para escolher o melhor modelo dentre todas as épocas. Para a abordagem multi-rótulo, além das métricas citadas acima, também utilizamos para comparação as métricas:

- Hamming *Loss*: mede a taxa de erro em um problema multi-rótulo. Contabiliza os rótulos incorretamente previstos, calculando a fração de rótulos classificados de forma errada para todas as instâncias. Essa métrica é calculada pela proporção de rótulos incorretos (falsos positivos + falsos negativos) em relação ao total de rótulos. Assim, quanto menor o valor, melhor é o desempenho do modelo. Vale notar que estaremos utilizando a Hamming *Loss* invertida, para facilitar a comparação com as outras métricas.

- Jaccard *Score*: mede a similaridade entre os conjuntos de classes previstos e verdadeiros. Para cada instância é calculado o número de classes previstas corretamente dividido pelo número total de classes únicas previstas e verdadeiras. Sua variação é de 0 a 1, onde 1 indica que todas as classes previstas coincidem exatamente com as classes verdadeiras, e 0 indica que não houve nenhuma sobreposição entre as classes previstas e as classes verdadeiras.

- *Coverage Error*: mede o número de classes que precisam ser cobertas até que todas as verdadeiras classes de uma instância estejam incluídas nas previsões ordenadas pelo modelo. Para cada instância, as previsões são classificadas em ordem decrescente de confiança, e o *Coverage Error* é o índice do rótulo corretamente previsto mais distante na lista ordenada. O valor ideal é o menor possível, pois indica que o modelo consegue identificar as classes verdadeiras rapidamente nas primeiras predições.

4.3 Resultados e Discussões

Como citado na seção 4.2 os experimentos foram executados uma validação cruzada com 5 *folds*, e para cada *fold* foi selecionada a melhor época com o conjunto de validação, e essa época foi avaliada pelo teste. Na seção 4.3.1 são discutidos os resultados para a classificação multi-classe. Já na seção 4.3.2 são discutidos os resultados para a classificação multi-rótulo. A seção 4.3.3. apresenta a discussão final.

4.3.1 Resultados Multi-classe

Na Tabela 3 e na Figura 5 são apresentados os resultados obtidos para a classificação multi-classe. Para cada uma das métricas o melhor valor está destacado em negrito.

Modelo	Acurácia	F1-Macro	Precisão	Revocação
BERT	0,8539	0,8483	0,8497	0,8485
BERTimbau	**0,8932**	**0,8882**	**0,8916**	**0,8897**
BERTikal	0,8831	0,8762	0,8892	0,8775
JurisBERT	0,8770	0,8724	0,8726	0,8729

Tabela 3 – Resultados multi-classe.

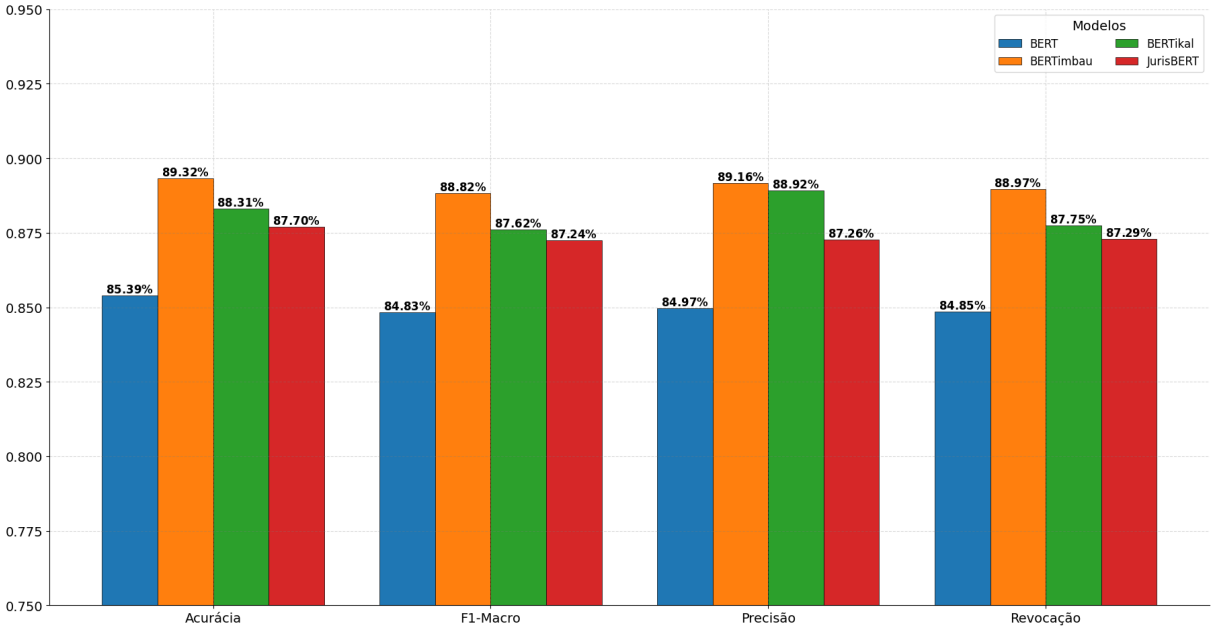

Figura 5 – Gráfico comparativo dos resultados multi-classe.

Com base nos resultados, o melhor modelo é o BERTimbau. A métrica principal analisada é a F1-Macro, pois apresenta o equilíbrio entre precisão e revocação, sendo que o modelo BERTimbau apresentou melhor resultado (0,8882). Este modelo também tem a maior precisão (0,8916), indicando que a maioria das previsões positivas são corretas. Apresenta o maior revocação (0,8897), indicando que identifica bem a classe positiva. Apresenta ainda a maior acurácia (0,8932), o que significa que classifica corretamente a maior parte dos exemplos.

4.3.2 Resultados Multi-rótulo

Na Tabela 4 e na Figura 6 são apresentados o melhor resultado para cada modelo, para cada uma das métricas o melhor valor está destacado em negrito. Verificamos que o modelo BERTimbau obteve o melhor desempenho em todas as métricas.

Modelo	Acurácia	F1-Macro	Precisão	Revocação	Hamming *Loss*	Jaccard *Score*	*Coverage Error*
BERT	0,8799	0,9442	0,9439	0,9446	0,9589	0,9188	2,1851
BERTimbau	**0,8896**	**0,9505**	**0,9507**	**0,9504**	**0,9632**	**0,9264**	**2,1688**
BERTikal	0,8641	0,9366	0,9367	0,9373	0,9536	0,9083	2,2136
JurisBERT	0,8571	0,9356	0,9354	0,9358	0,9524	0,9048	2,1721

Tabela 4 – Resultados multi-rótulo.

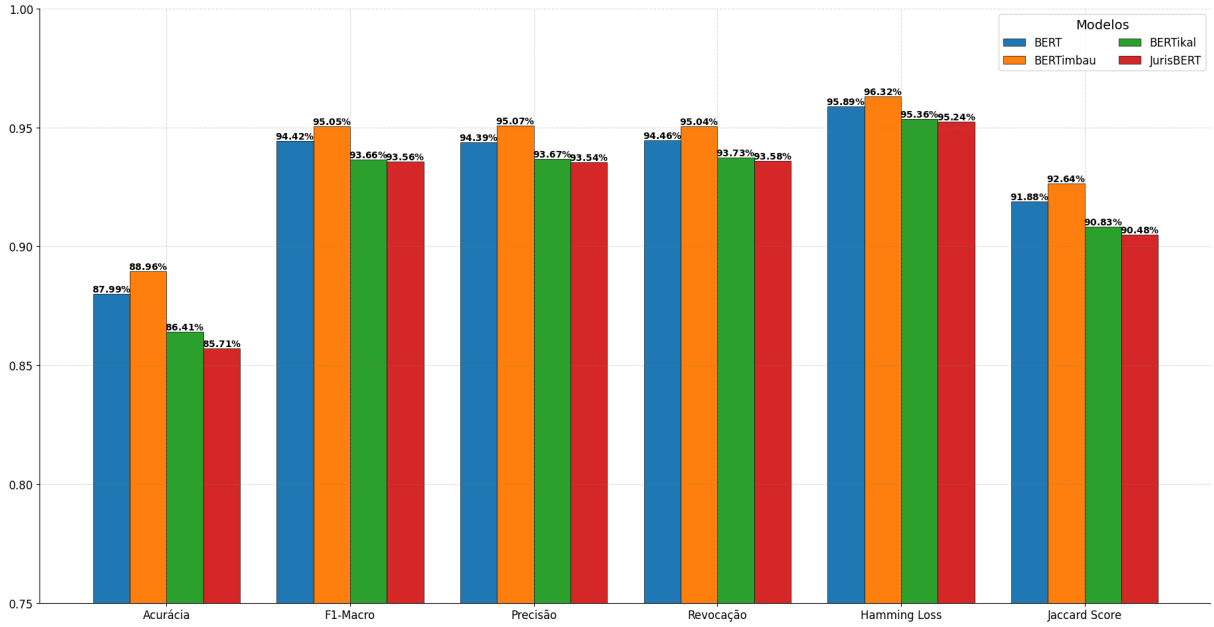

Figura 6 – Gráfico comparativo dos resultados multi-rótulo.

A métrica principal analisada é a F1-Macro, pois apresenta o equilíbrio entre precisão e revocação, sendo que o modelo BERTimbau apresentou melhor resultado (0,9504). Ambos os valores de precisão (0,9506) e revocação (0,9504) são os mais altos entre os modelos, mostrando que ele é tanto preciso quanto eficaz na detecção das classes corretas. Além disso este modelo apresenta o maior Hamming *Loss* (0,9632), o que indica menos predições incorretas (vale relembrar que fizemos Hamming *Loss* invertida para facilitar a comparação com as demais métricas). E tem também o maior Jaccard *Score*, indicando que 92,64% das classes previstas coincidem exatamente com as classes verdadeiras. O modelo também tem a menor *Coverage Error* (2,1688), o que indica que ele está classificando corretamente as classes com menor erro.

4.3.3 Considerações Finais

Avaliando os resultados é possível observar que, para o conjunto de dados escolhido, a abordagem multi-rótulo tem desempenho superior ao multi-classe em termos de F1-Macro. Em ambas as abordagens o modelo BERTimbau apresentou melhor desempenho. Esse resultado é interessante pois tendemos a prever que a abordagem multi-classe, por ser mais simples, dará resultados mais assertivos em comparação com a abordagem multi-

rótulo. Outro ponto interessante é o desempenho superior do modelo BERTimbau a outros modelos pré-treinados com corpus jurídicos brasileiros, mesmo ele tendo sido treinado fora do contexto jurídico.

Como próximos passos o melhor modelo será colocado em produção, com o objetivo de classificar automaticamente decisões judiciais. Os resultados serão monitorados por especialistas. No futuro, utilizaremos um conjunto de dados maior para treinar os classificadores, repetindo os testes para verificar se as conclusões deste estudo se mantêm. Além disso, buscaremos desenvolver classificadores mais robustos, capazes de acompanhar as mudanças na redação dos textos jurídicos.

5 CONCLUSÕES

Neste trabalho foi feita uma comparação entre modelos para classificar publicações de decisões judiciais. Realizamos *fine-tunning* de 4 modelos de linguagem pré-treinados, utilizando como conjunto de dados as abordagens multi-classe e multi-rótulo. Os resultados indicam que a abordagem multi-rótulo aplicada ao modelo BERTimbau obteve melhor desempenho na tarefa de classificação. É importante observar que o modelo BERTibau saiu-se melhor que modelos pré-treinados em corpus jurídicos brasileiros (JurisBERT e BERTikal).

A principal contribuição deste estudo é apresentar um modelo capaz de classificar decisões judiciais de forma efetiva, podendo ser aplicado tanto em escritórios de direito quanto em tribunais para organização das informações e otimizar processos de trabalho. Por exemplo, é possível classificar automaticamente decisões de tribunais gerando indicadores comparativos de como determinada tese jurídica é julgada em cada tribunal brasileiro. O classificador também pode ser aplicado em escritórios, identificando o resultado da decisão automaticamente e, consequentemente, alertando os advogados para que ações sejam tomadas em tempo hábil, de acordo com o resultado.

Apesar das descobertas significativas, o estudo tem algumas limitações. A amostra utilizada é pequena (máximo de 200 exemplos para cada classe), podendo conter amostras enviesadas, que podem levar o modelo a padrões incorretos. Em futuras pesquisas um conjunto de dados maior será utilizado, a fim de obter modelos ainda melhores. Apesar disso, o melhor classificador dessa tese será utilizado em produção para avaliarmos sua eficácia na prática e observar suas tendências de erros a fim de obter *insights* para o re-treino.

Em conclusão, este estudo fornece uma comparação detalhada sobre a melhor abordagem para classificação automática de decisões judiciais brasileiras, e abre caminhos para novas investigações. As implicações práticas deste estudo podem contribuir para organização de informações em tribunais e melhorias dos processos de trabalho em escritórios de direito. O tema continuará a ser explorado para aprofundar o entendimento e contribuir para a inovação tecnológica do direito no Brasil.

REFERÊNCIAS

CNJ. **Justiça em Números 2022. <https://www.cnj.jus.br/portal/images/stories/pdf/justicaemnumeros/justica_em_numeros_2022.pdf>**. 2022. Último acesso em 13 de Janeiro de 2024.

CNJ. **Justiça em Números 2023. <https://www.cnj.jus.br/wp-content/uploads/2023/08/justica-em-numeros-2023.pdf>**. 2023. Último acesso em 13 de Janeiro de 2024.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *In*: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [*S.l.: s.n.*], 2019. p. 4171–4186.

GOMES LUCAS MOREIRA; JADER MARTINS CAMBOIM DE Sá; PENG, Y. Línguas naturais e máquinas artificiais: Aplicação de técnicas de mineração de texto para a classificação de sentenças judiciais brasileiras. *In*: IPEA. **Texto para Discussão**. [*S.l.: s.n.*], 2020.

GUIMARAES, J. P. F. Integrando máquinas de processamento de linguagem natural para otimizar a classificação de textos jurídicos com diferentes padrões linguísticos. *In*: **Lium Concilium**. [*S.l.: s.n.*], 2023. p. 524–536.

Painel CNJ. **Estatísticas do Poder Judiciário. <https://painel-estatistica.stg.cloud.cnj.jus.br/estatisticas.html>**. 2023. Último acesso em 23 de Agosto de 2024.

POLO, F. M. *et al.* Legalnlp-natural language processing methods for the brazilian legal language. *In*: SBC. **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional**. [*S.l.: s.n.*], 2021. p. 763–774.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [*S.l.: s.n.*]: Editora Manole Ltda., 2003.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. *In*: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [*S.l.: s.n.*], 2020.

SOUZA R.F. NOGUEIRA, R. A. L. F. C. Bert models for brazilian portuguese: Pretraining, evaluation and tokenization analysis. *In*: **Applied Soft Computing Journal 149**. [*S.l.: s.n.*], 2023.

VASWANI, A. *et al.* Attention is all you need. *In*: **Advances in Neural Information Processing Systems**. [*S.l.: s.n.*], 2017. p. 5998–6008.

VIEGAS, B. C. C. R. P. I. C. F. O. **JurisBERT: Transformer-based model for embedding legal texts**. 2022. <https://repositorio.ufms.br/handle/123456789/5119>.