UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

REPRESENTAÇÕES *EMBEDDINGS* ORIENTADAS À LINGUAGEM JURÍDICA BRASILEIRA

FABRÍCIO ALMEIDA DO CARMO

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do Prof. Dr. Fábio Manoel França Lobato e Co-orientação do Prof. Dr. Ewaldo Eder Carvalho Santana.

2024

UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

FABRÍCIO ALMEIDA DO CARMO

REPRESENTAÇÕES *EMBEDDINGS* ORIENTADAS À LINGUAGEM JURÍDICA BRASILEIRA

2024

Carmo, Fabrício Almeida do.
Representações Embeddings orientadas à Linguagem Jurídica Brasileira./ Fabrício Almeida do Carmo . – São Luís (MA), 2024.
84p.
Dissertação (Mestrado Profissional em Engenharia da Computação e Sistemas) Universidade Estadual do Maranhão - UEMA, 2024.
Orientador: Prof. Dr. Fábio Manoel França Lobato.
Co-orientador: Prof. Dr. Ewaldo Eder Carvalho Santana.

- Documentos Jurídicos. 2. Word Embeddings .3. Modelos de Linguagem. 4. Classificação de dados . I.Título.

CDU: 341.76 (81)

Fabrício Almeida do Carmo

Representações *Embeddings* Orientadas à Linguagem jurídica Brasileira

Trabalho apresentado ao curso de Mestrado Profissional em Engenharia da Computação e Sistemas na Universidade Estadual do Maranhão como pré-requisito para obtenção do título de Mestre sob orientação do Prof. Dr. Fábio Manoel França Lobato e Co-orientação do Prof. Dr. Ewaldo Eder Carvalho Santana.

Trabalho aprovado. São Luís - MA, 26 de fevereiro de 2024:

Documento assinado digitalmente
gov.br
FABIO MANOEL FRANCA LOBATO
Data: 12/04/2024 19:13:03-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Fábio Manoel França Lobato
(Orientador - UFOPA)

Documento assinado digitalmente
gov.br
EWALDO EDER CARVALHO SANTANA
Data: 12/04/2024 16:19:30-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Ewaldo Eder Carvalho Santana
(Coorientador - UEMA)

Documento assinado digitalmente
gov.br
ANTONIO FERNANDO LAVAREDA JACOB JUNIO
Data: 12/04/2024 15:48:15-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Antonio Fernando Lavareda Jacob Junior
(Examinador interno - UEMA)

CLAUDIO HENRIQUE CARNEIRO SAMPAIO:42774918391
Assinado de forma digital por CLAUDIO HENRIQUE CARNEIRO SAMPAIO:42774918391
DN: c=BR, o=ICP-Brasil, ou=AC SOLUTI Multipla v5, ou=35622406000190, ou=Presencial, ou=Certificado PF A3, cn=CLAUDIO HENRIQUE CARNEIRO SAMPAIO:42774918391
Dados: 2024.04.24 17:41:28 -03'00'

Cláuldio Henrique Carneiro Sampaio
(Examinador externo - TJMA)

Documento assinado digitalmente
gov.br
RICARDO MARCONDES MARCACINI
Data: 22/04/2024 09:53:09-0300
Verifique em https://validar.iti.gov.br

Profº. Dr. Ricardo Marcondes Marcacini
(Examinador externo - USP)

São Luís - MA

2024

AGRADECIMENTOS

Prezados familiares, professores, amigos e colegas, é com imensa satisfação que expresso meus sinceros agradecimentos a todos que contribuíram para a realização deste trabalho. Este momento marca não apenas o término de uma etapa acadêmica, mas também o resultado de esforços e apoios inestimáveis ao longo dessa jornada desafiadora.

Aos meus familiares, expresso minha profunda gratidão pelo constante incentivo e apoio. Vocês foram a base sólida que sustentou cada passo dessa caminhada, fundamental para superar os desafios.

Aos professores, agradeço pela dedicação, orientação e inspiração. Suas habilidades pedagógicas, conhecimento científico profundo e a disposição para compartilhar experiências foram cruciais para o desenvolvimento desta pesquisa. Cada *feedback* construtivo e orientação crítica foram fundamentais para meu crescimento acadêmico e profissional.

Aos meus colegas de curso, compartilho minha gratidão pela colaboração, amizade e apoio. Juntos enfrentamos desafios, superamos obstáculos e celebramos objetivos alcançados. A troca de conhecimento e experiências enriqueceu não apenas meu trabalho, mas também minha formação como indivíduo.

Ao Tribunal de Justiça do Maranhão (TJMA), à Financiadora de Estudos e Projetos (FINEP) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), meus agradecimentos pela oportunidade concedida e pelo suporte financeiro que possibilitou a realização desta pesquisa. Que o compromisso dessas instituições com o desenvolvimento acadêmico continue a oportunizar cada vez mais pessoas.

Esta jornada não teria sido possível sem o apoio de cada um de vocês. Espero que possamos continuar compartilhando sucessos e crescendo juntos no futuro.

RESUMO

O processamento automático de textos jurídicos dispostos em linguagem natural viabiliza a construção de uma gama de aplicações baseadas em inteligência artificial para o setor, tais como: a classificação e agrupamento de processos por assunto, sumarização de documentos, tradução para linguagem cidadã, entre outros. Nesse sentido, o judiciário brasileiro lançou o programa Justiça 4.0 buscando incentivar o desenvolvimento de soluções que ofereçam celeridade nas atividades processuais. Destaca-se que a linguagem técnica é predominante nesse domínio de aplicação, exigindo modelos especializados para o segmento. Frente ao exposto, esse trabalho tem como objetivo a construção de modelos *embeddings* orientados ao âmbito jurídico visando alimentar aplicações na área. Para isso, foram extraídos aproximadamente 5,3 milhões de documentos de instituições de justiça do Brasil das mais variadas esferas como civil, criminal e trabalhista. Os modelos foram avaliados por meio da classificação de petições iniciais e os resultados obtidos se mostraram promissores quando comparados a modelos generalistas da língua portuguesa. Tais achados de pesquisa demostram que modelos treinados com documentos jurídicos compreendem melhor as especificidades da linguagem do segmento e têm o potencial de fomentar novas aplicações para o setor.

Palavras-chave: Justiça 4.0, Processamento de Linguagem Natural, Word embeddings, Ciência de dados, Petições Iniciais.

ABSTRACT

The automatic processing of legal texts arranged in natural language makes it possible to build a range of applications based on artificial intelligence, such as classification and grouping of processes by subject, document summarization, and translation into citizen language. In this sense, the Brazilian judiciary launched the Justice 4.0 program, looking to encourage the development of solutions that offer speed in procedural activities. Notably, technical language is predominant in this application domain, requiring specialized models for the segment. Bearing in mind this context, this work aims to build models *embeddings* oriented to the legal sphere with a view to feeding applications in the area. In this sense, approximately 5.3 million documents were extracted from Brazilian justice institutions from the most varied spheres, such as civil, criminal, and labor. The models were evaluated by classifying initial petitions, and the results obtained were promising when compared to generalist models of the Portuguese language. Such research findings demonstrate that models trained with legal documents better understand the segment's language's specificities and can potentially promote new applications for the sector.

Key-words: Justice 4.0, Natural Language Processing, Word embeddings, Data Science, Legal Domain .

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BNPR	Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatório
BERT	*Bidirectional Encoder Representations from Transformers*
Bi-LSTM	*Bidirectional Long Short-Term Memory*
BoW	*Bag Of Words*
CBoW	*Continuous Bag-of-Words*
BPE	*Byte-Pair Encoding*
CJF	Conselho da Justiça Federal
CNJ	Conselho Nacional de Justiça
CNN	*Convolutional Neural Network*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DL	*Deep Learning*
FT	*Further pre-train*
GRU	*Gated Recurrent Unit*
GPT	*Generative Pre-trained Transformer*
GloVe	*Global Vectors for Word Representation*
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
IRDR	Incidente de Resolução de Demandas Repetitiva
ITD	*Iudicium Textum Dataset*
k-NN	*k-Nearest Neighbors*
LR	*Logistic Regression*
LSTM	*Long Short-Term Memory*
MLM	*Masked Language Model*

ML	*Machine Learning*
NER	*Name Entity Recognition*
NSP	*Next Sentence Prediction*
NILC	Núcleo Interinstitucional de Linguística
NPS	*Next Sentence Prediction* Computacional
PLN	*Processamento de Linguagem Natural*
POS tagging	*Part-of-Speech tagging*
RF	*Random Forest*
RIT	Reconhecimento de Implicação Textual
SC	*Pretraining from Scratch*
STF	Supremo Tribunal Federal
STM	Superior Tribunal Militar
SVM	*Support Vector Machine*
TCU	Tribunal de Contas da União
TF-IDF	*Term Frequency - Inverse Document Frequency*
TJMA	Tribunal de Justiça do Maranhão
TSE	Tribunal Superior Eleitoral
TST	Tribunal Superior do trabalho
TM	*Text Mining*
UMAP	*Uniform Manifold Approximation and Projection for Dimension Reduction*
UEMA	Universidade Estadual do Maranhão

SUMÁRIO

1 INTRODUÇÃO

O presente capítulo apresenta os principais aspectos que nortearam a concepção e a condução deste estudo, organizando-se da seguinte forma: *i)* contextualização: delineando o cenário e a apontando os principais desafios de pesquisa; *ii)* Justificativa: discutindo a relevância do tema estudado e os *iii)* objetivos traçados para o trabalho, finalizando com as contribuições do estudo.

1.1 CONTEXTUALIZAÇÃO

O uso de aplicações baseadas em Inteligência Artificial (IA) e *Big Data* vem apoiando a tomada de decisões em diversos segmentos da sociedade (GARCIA, 2020; HARIRI; FREDERICKS; BOWERS, 2019). No âmbito jurídico, estas soluções podem guiar os profissionais tanto nas atividades administrativas quanto nos trâmites processuais, atuando principalmente sobre o grande volume de dados gerados no dia-a-dia da prestação jurisdicional (PINTO, 2020; PEREIRA; RODRIGUES, 2021; ARAÚJO; GABRIEL; PORTO, 2022). No Brasil, onde o sistema judiciário conta com cerca de 81,4 milhões de processos em tramitação, segundo o último relatório "justiça em números"[1] do Conselho Nacional de Justiça (CNJ), já se entende que a celeridade processual passa necessariamente pela adoção de aplicações que utilizam recursos da IA (MACHADO; COLOMBO, 2021).

Neste contexto, o programa Justiça 4.0[2] é uma das iniciativas do CNJ que fomenta o desenvolvimento de soluções com estas tecnologias, promovendo aquelas que visem a automatização das atividades dos tribunais, otimizando o trabalho dos magistrados, servidores e advogados. Um exemplo dessas iniciativas é a plataforma Codex[3], que trabalha na estruturação de dados jurídicos visando oferecer insumos para diversas finalidades, entre elas, o desenvolvimento de modelos de IA (SOUZA; SALLES, 2022). Outro exemplo notório é a plataforma SINAPSES, responsável tanto pelo estabelecimento de parâmetros (legais e técnicos) para o desenvolvimento e implantação de modelos de IA nos tribunais, como para o armazenamento e distribuição dos mesmos (PEREIRA; RODRIGUES, 2021).

Dentre as aplicações que utilizam os recursos da IA para análises de dados jurídicos destacam-se os que consomem os recursos do Processamento de Linguagem Natural (PLN). PLN é uma área multidisciplinar que envolve a computação e a linguística, contemplando estudos e desenvolvimento de métodos e procedimentos que permitam a compreensão e o processamento automático de textos dispostos na forma natural (SOUSA; FABRO, 2019; HIRSCHBERG; MANNING, 2015). Na seara jurídica, aplicações como classificação de do-

[1] https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/
[2] https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/
[3] https://www.cnj.jus.br/sistemas/plataforma-codex/

cumentos e processos (POLO et al., 2021; BAMBROO; AWASTHI, 2021), reconhecimento de Entidade Nomeada (WANG et al., 2020; BATISTA et al., 2021) e sumarização de texto (VIANNA; MOURA; SILVA, 2023; VIANNA; MOURA, 2022), já oferecem resultados práticos.

Um dos principais desafios no trato com documentos jurídicos está na compreensão das especificidades da linguagem utilizada pelo setor, composta por jargões e termos técnicos (POLO et al., 2021). O tamanho dos textos jurídicos também é um ponto significativo, geralmente composto de textos longos e dotado de formalismo (BAMBROO; AWASTHI, 2021). A exemplos, fazem parte deste conjunto de dados: sentenças, contratos, opiniões legais, petições inicias etc (ZHONG et al., 2020; MOTA et al., 2020). Dessa forma, treinar modelos de IA eficientes para este domínio perpassa pela forma como estes dados são tratados e representados, visando a incorporação de tais peculiaridades no modelo.

Neste sentido, as representações de dados textuais visam criar modelos vetoriais de um determinado texto. Elas são elementos fundamentais em PLN dado sua capacidade de transformar os documentos de entrada em vetores numéricos, preservando informações originais e fornecendo entradas consumíveis por modelos de *Machine Learning* (ML) e *Deep Learning* (DL) (LE-KHAC; HEALY; SMEATON, 2020). Neste contexto, os modelos *embeddings* são amplamente utilizados dado sua capacidade de adicionar informações semânticas no processo representacional (CHALKIDIS; KAMPAS, 2019; MIKOLOV et al., 2013a). A exemplos destes modelos para representações de palavras e textos, destacam-se o *Word2Vec* (MIKOLOV et al., 2013a), o *FastText* (BOJANOWSKI et al., 2017) e, mais recentemente, o *Bidirectional Encoder Representations from Transformers* (*BERT*) (DEVLIN et al., 2019).

Os modelos baseados em *word embeddings* utilizam redes neurais rasas para o aprendizado dos vetores de representação. Tal procedimento permite o mapeamento de palavras e suas respectivas relações no *corpus* de treinamento, produzindo vetores que capturam informações semânticas e sintáticas advindas destas relações contextuais (MIKOLOV et al., 2013a). Por exemplo, as palavras “juiz” e “magistrado” estariam próximas no espaço vetorial devido seu grau de similaridade. Os modelos contextuais, como o *BERT*, são caracterizados pela capacidade de geração dinâmica de vetores de representação, ou seja, dependendo do contexto em que determinada palavra se encontra, um vetor diferente é mapeado (DEVLIN et al., 2019).

Frente ao exposto, esta pesquisa busca fomentar o desenvolvimento de soluções baseadas no processamento de linguagem natural na área na jurídica brasileira por meio da construção de um *framework* experimental com diferentes modelos de representações *embeddings*, treinados a partir de documentos do setor. Tais modelos são avaliados em aplicações finais de PLN, visando um entendimento prático do desempenho dos modelos produzidos.

1.1.0.1 ACORDO DE COOPERAÇÃO TÉCNICA

Buscando o desenvolvimento de soluções que contemplem os recursos de IA para o segmento jurídico, diversos tribunais de justiça do país têm firmado acordos e convênios, assim como projetos de cooperação técnica com universidades e empresas especializadas em tecnologia (SILVA; FILHO, 2020). Nesse contexto, o Tribunal de Justiça do Maranhão (TJMA) e a Universidade Estadual do Maranhão (UEMA) estabeleceram uma parceria para a execução de projetos de pesquisa e desenvolvimento de *softwares* que façam uso dessas tecnologias para automatizar processos no tribunal. A exemplo, uma das principais demandas do tribunal é a análise e identificação automática de precedentes jurídicos. Entende-se por precedente jurídico, uma decisão tomada em um caso concreto, que pode servir como base para julgamentos similares, evitando, assim, tanto retrabalho quanto insegurança jurídica [4]. Dessa forma, ao fazer análises automáticas desses documentos (petições iniciais) impetrados no tribunal são produzidos ganhos significativos no fluxo processual.

Inserido no escopo do Acordo de Cooperação N.º 002/2021, entre TJMA e UEMA, essa pesquisa utiliza de PLN para o processamento de dados jurídicos. O foco está no treinamento de modelos para representação de documentos do setor, dispostos na língua portuguesa, visando maximizar os resultados de tarefas de PLN no tribunal. Exemplificando, modelos treinados nessa pesquisa, denominados BumbaBert por serem baseados na arquitetura do *BERT*, já estão sendo consumidos pela aplicação "Robô Firmina", que visa a classificação de procedentes do tribunal.

1.2 JUSTIFICATIVA

Um dos grandes desafios no desenvolvimento de aplicações baseadas em PLN está na necessidade de um grande volume de dados para o treinamento de representações vetoriais que compreendam características importantes do texto de entrada. Essa dependência se acentua ainda mais quando se trata do treinamento de vetores *embeddings*, que consideram as relações das palavras (*tokens*) no conjunto de dados de treinamento. Ou seja, quanto maior a incidência de tais *tokens*, maior é o aprendizado da representação final. Visando superar essa limitação, diversos autores disponibilizaram modelos pré-treinados em grandes *corpora* de dados e em diferentes idiomas, geralmente extraídos da web, para que esses modelos pudessem ser utilizados em uma gama de aplicações em PLN, como a classificação de textos, a sumarização por assunto, o reconhecimento de entidades nomeadas - *Name Entity Recognition* (NER), entre outras. É importante notar que esses modelos podem ser adaptados por meio de *fine-tuning* para domínios diferentes com menor custo computacional. No entanto, considerando a natureza genérica dos dados utilizados no pré-treinamento,

[4] https://www.tjdft.jus.br/institucional/imprensa/campanhas-e-produtos/direito-facil/edicao-semanal/precedente-x-jurisprudencia-x-sumula

esses modelos apresentam limitações quando aplicados a domínios mais específicos, nos quais o vocabulário utilizado nem sempre está alinhado com a linguagem comum.

No cenário brasileiro, diferentes segmentos já investigam a necessidade de treinar representações específicas. Por exemplo, na área do petróleo e gás, os estudos de Consoli et al. (2020) e Gomes et al. (2021) destacam essa importância, considerando que os termos técnicos do setor divergem do português comum. Na área médica, o estudo proposto por Schneider et al. (2020) corrobora tal necessidade ao comparar seu modelo, o BioBertPt, treinado com documentos biomédicos, com modelos generalistas da língua portuguesa, sendo que o BioBertPt apresentou melhores resultados nos cenários avaliados. Na área jurídica, os estudos de Silveira et al. (2023) e Polo et al. (2021) enfatizam a necessidade de modelos treinados especificamente para este domínio, considerando as especificidades da linguagem utilizada no setor.

Ainda considerando o cenário jurídico brasileiro, observa-se que os desafios se tornam ainda maiores quando se consideram as variações que a língua portuguesa sofre nas diferentes regiões do país (POLO et al., 2021). Este mesmo autor destaca que os regionalismos podem impactar diretamente na forma como as peças jurídicas são desenvolvidas. Destarte, treinar modelos de representações que incorporem tanto as nuances da linguagem jurídica quanto a diversidade linguística do país é necessário.

Outro desafio importante para o treinamento de modelos orientados à linguagem jurídica é a indisponibilidade de dados públicos, estruturados, confiáveis e representativos do segmento. Apesar de iniciativas como as de Sousa e Fabro (2019), Araujo et al. (2020) e Luz de Araujo et al. (2018), que construíram e disponibilizaram *corpus* deste domínio, tais recursos mostram-se insuficientes para suprir essa demanda devido ao seu tamanho diminuto e não representatividade, seja regional ou de subdomínios (*e.g.,* esferas cível, criminal, tributária, previdenciária, militar etc). Dessa forma, faz-se necessário adicionar mais uma etapa nesse processo já complexo: a prospecção e extração de dados/documentos jurídicos em domínios públicos de instituições jurídicas brasileiras. O próprio CNJ, como já mencionado, corrobora esse desafio ao lançar a plataforma Codex, visando essa estruturação de dados. No entanto, tais dados não são disponibilizados para a comunidade acadêmica de modo geral, deixando tal lacuna em aberto.

Mesmo com os desafios e as limitações discutidas, diversas aplicações estão sendo desenvolvidas no âmbito dos tribunais brasileiros. De acordo com um levantamento realizado pelo CNJ, em 2022, houve aumento significativo no número de projetos baseados em IA nos tribunais. Os dados apresentados no "Painel de Projetos de IA no Poder Judiciário" [5] apontam para mais de 100 projetos desenvolvidos ou em desenvolvimento em 53 tribunais

[5] https://paineisanalytics.cnj.jus.br/single/?appid=9e4f18ac-e253-4893-8ca1-b81d8af59ff6&sheet=b8267e5a-1f1f-41a7-90ff-d7a2f4ed34ea&lang=pt-BR&theme=IA_PJ&opt=ctxmenu,currsel&select=language,BR

de vários segmentos da justiça, como mostra a Figura 1. Essa iniciativa no sistema de justiça também reforça a necessidade de se treinar representações que fornecerão insumos para o desenvolvimento de soluções de IA para o setor.

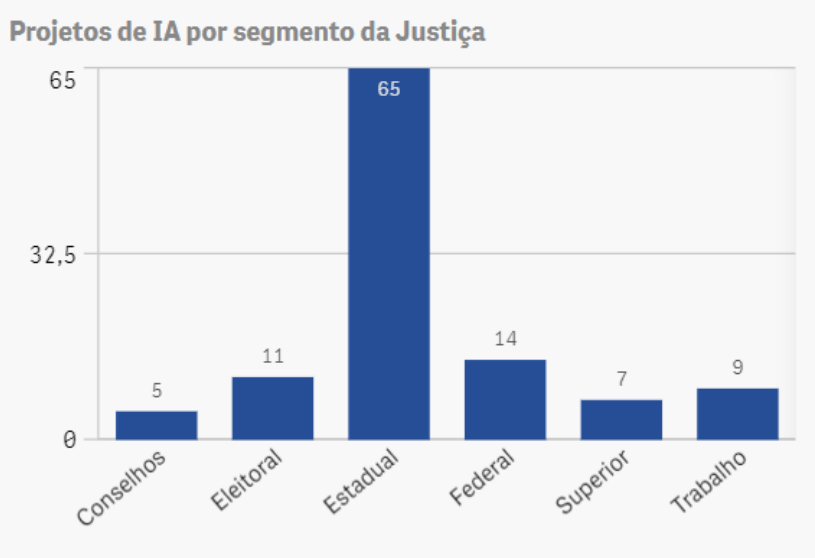

Figura 1 – Projetos de IA na justiça brasileira.

Fonte: CNJ (2022).

Frente ao exposto, esta pesquisa abrange aspectos teóricos e práticos ao realizar o treinamento de representações orientadas à linguagem jurídica brasileira. Tal cobertura oferece ao meio acadêmico e ao segmento da justiça vários ativos importantes, tais como estudos sobre os desafios da linguagem do setor, procedimentos de busca e tratamento de dados, experimentos computacionais para o treinamento e avaliação dos modelos, os próprios modelos treinados, assim como sua avaliação em aplicações finais de PLN.

A relevância do presente trabalho se amplia ao se observar que ele emerge de um acordo de cooperação técnica entre a UEMA e o TJMA. Dessa forma, os modelos construídos estão sendo utilizados como insumos para aplicações PLN no tribunal, produzindo impacto direto no sistema de justiça. A título de exemplo, destaca-se o projeto, construção e implantação da solução denominada "Robô Firmina", que busca a identificação de precedentes judiciais a partir de textos de petições iniciais. A aplicação já está em testes no tribunal e com versões utilizando modelos treinados nesta pesquisa no processo de vetorização. Os modelos baseados na arquitetura do *BERT* foram nomeados como BumbaBert em referência cultural ao estado do Maranhão. Em uma etapa posterior, os modelos construídos no presente estudo também serão compartilhados na plataforma SINAPSES/CNJ para o uso por outras instituições da justiça.

1.3 OBJETIVOS

À luz do contexto apresentado e da lacuna na literatura, o presente trabalho tem como objetivo geral construir e disponibilizar representações *embeddings* que compreendam as especificidades da linguagem jurídica brasileira.

Visando alcançar este objetivo, os seguintes objetivos específicos foram delineados:

1. Construir um *corpus* jurídico contemplando diferentes tipos de documentos, segmentos, provenientes também de diferentes instituições e regiões do país, visando conferir volume e diversidade necessárias;

2. Comparar as representações *embeddings* treinadas no presente estudo com com modelos genéricos e também com outros modelos específicos já disponíveis na literatura;

3. Aplicar e avaliar os modelos treinados em tarefas finais do processamento de linguagem natural como a classificação de petições iniciais, a detecção de possível aplicação de precedentes qualificados, reconhecimento de entidades nomeadas etc.

1.4 CONTRIBUIÇÕES DO ESTUDO

Considerando a ampla cobertura experimental desenvolvida no presente trabalho, vislumbram-se contribuições tanto de cunho acadêmico quanto para a indústria. A seguir são detalhados tais contribuições:

1. *Corpus* jurídico: criação de um *corpus* contendo dados estruturados, confiáveis e representativos do segmento. Este artefato tem como objetivo preencher a lacuna identificada ao longo da pesquisa, proporcionando subsídios essenciais para o desenvolvimento de novas aplicações PLN direcionadas ao setor jurídico. Tais dados serão disponibilizados para a comunidade acadêmica visando incentivar novas pesquisas na área.

2. Modelos Treinados: Os modelos produzidos neste trabalho integram um projeto de pesquisa que visa soluções com IA para o TJMA, representando, dessa forma, um contributo concreto para o segmento. A exemplos, os modelos BumbaBert, nossos modelos baseados na arquitetura do BERT, já estão sendo utilizadas na aplicação "Robô Firmina". Tal aplicação é focada na descoberta e classificação de precedentes jurídicos do tribunal. Além disso, os modelos serão disponibilizados à comunidade acadêmica, visando disseminar o conhecimento na área.

3. Cobertura experimental. Os experimentos realizados fornecerão elementos para a reprodução dos treinamentos e avaliação de modelos, além da aplicação dos mesmos em tarefas finais envolvendo PLN.

4. Publicação de Artigos: Visando reprodutividade e disseminação do conhecimento, os resultados dos experimentos estão sendo publicados em eventos da área:

   - **Artigo 1**: Os resultados dos experimentos preliminares com os modelos *word embeddings* aplicados ao domínio jurídico foram apresentados no XI Workshop de Computação Aplicada em Governo Eletrônico (WCGE 2023). O trabalho, intitulado "*Embeddings* Jurídico: Representações Orientadas à Linguagem Jurídica Brasileira", aborda os procedimentos de treinamento dos modelos, com dados do domínio, e o processo de avaliação do na classificação de dados jurídicos. O artigo correspondente está disponível nos anais do evento, acessível pelo endereço eletrônico: https://sol.sbc.org.br/index.php/wcge/article/view/24876, e no Apêndice A deste documento.
   - **Artigo 2**: O artigo contemplando os resultados dos experimentos com modelos contextuais está em fase de finalização e também será um meio de divulgação desta pesquisa. O trabalho contemplará todo processo de treinamento e avaliação dos modelos denominados BumbaBert, nossos modelos baseados na arquitetura do *BERT* para o segmento jurídico.

1.5 ORGANIZAÇÃO DO TRABALHO

Este documento encontra-se estruturado como segue. O Capítulo 2 apresenta uma visão geral das representações baseadas em *embeddings* utilizadas no estudo. Os Trabalhos Corretos são abordados no Capítulo 3, detalhando as contribuições significativas na literatura existente. No Capítulo 4, são descritos os Materiais e Métodos empregados na condução da pesquisa, descrevendo os passos metodológicos e tecnologias utilizadas. Os Resultados são discutidos no quinto capítulo 5, destacando as análises decorrentes dos experimentos realizados. Finalmente, o Capítulo 6 Considerações Finais, consolidando as principais conclusões da dissertação e delineando possíveis direções para pesquisas futuras.

2 REPRESENTAÇÕES TEXTUAIS BASEADAS EM *EMBEDDINGS*

O campo interdisciplinar de PLN e da Mineração de Textos desempenha um papel significativo na compreensão e extração de informações relevantes a partir de grandes volumes de dados textuais. Neste capítulo, serão abordados conceitos pertinentes ao entendimento deste trabalho, destacando estratégias adotadas no estado da arte e da prática para o processo de representação de dados e extração de informações.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

O campo de PLN pode ser dividido em duas etapas principais: a primeira é a representação do texto de entrada (dados brutos) em formato numérico (vetores ou matriz), e a segunda é o *design* de modelos para processar os dados numéricos a fim de cumprir a tarefa desejada (PATIL et al., 2023). No contexto do PLN, as representações de dados textuais desempenham um papel fundamental, pois têm a capacidade de transformar os documentos de entrada em vetores numéricos, preservando informações originais. Essas representações são essenciais para modelos de ML e DL. Por exemplo, a frequência dos termos na sentença de entrada pode ser computada e mapeada para um vetor numérico como no caso da representação por saco de palavras, mais conhecida por *Bag-Of-Words* (BoW). Entretanto, uma das grandes limitações dessas abordagens simplificadas é que ela não considera as relações entre as palavras, dificultando o entendimento contextualizado do problema. Nessa linha, as representações baseadas em *word embeddings* têm apresentado resultados promissores em diferentes tarefas envolvendo PLN, uma vez que conseguem adicionar informações semânticas no processo representacional (MIKOLOV et al., 2013a).

Os modelos *word embedding* buscam fornecer vetores que compreendam as conexões entre um termo/*token* e seus vizinhos, oferecendo uma projeção de cada palavra no espaço vetorial de forma a expressar as relações semânticas entre as palavras (PATIL et al., 2023). Uma das formas bastante utilizadas para obter esta projeção é treinar redes neurais rasas para prever palavras e seus contextos (MIKOLOV et al., 2013a). As representações *word embeddings* produziram um marco no campo de PLN a partir dos modelos propostos por Mikolov et al. (2013a) e Mikolov et al. (2013b). Esses modelos treinam redes neurais rasas para construir vetores de representações de palavras que incorporam significado semântico, analisando as palavras e suas relações no *corpus* de treinamento. Conhecidos como *Word2Vec*, esses modelos fundamentam trabalhos como o *FastText* (BOJANOWSKI et al., 2017), que introduziu uma abordagem inovadora analisando determinada palavra como um conjunto de n-gramas de caracteres. Dessa forma, a representação vetorial

resultante para a palavra é obtida pela soma das sub-representações, possibilitando a obtenção de *embeddings* para palavras não vistas nos dados de treinamento. A seguir os modelos *Word2Vec*, *FastText* e *BERT* são sucintamente descritos.

2.1.1 MODELOS *WORD2VEC*

Proposto por Mikolov et al. (2013a), o *Word2Vec* é um algoritmo para geração de vetores de palavras, amplamente utilizado na literatura de PLN. Treinado em grandes *córpora*, este modelo de representação tem a capacidade de treinar vetores que compreendem as relações entre as palavras, superando amplamente modelos baseados na contagem de *tokens*, como BoW, cuja representação é focada em computar co-ocorrências dos termos (QADER; AMEEN; AHMED, 2019). A redução da dimensionalidade e o tratamento da esparsidade do vetor também são vantagens do modelo *word embeddings* sobre o BoW.

O *Word2Vec* utiliza duas estratégias de treinamento com redes neurais rasas: o *Continuous Bag-of-Words* (CBoW) e o modelo *Skip-gram* (MIKOLOV et al., 2013b). No CBoW, o algoritmo tenta prever a palavra central (alvo), com base no contexto em que ela está inserida. Já no modelo *Skip-gram* a predição é feita de maneira oposta, as palavras do contexto é que são previstas com base na palavra central. As predições são realizadas utilizando janelas de contextos local, procedimento que seleciona k palavras (w) vizinhas em torno do alvo. A Figura 2 apresenta essas arquiteturas utilizadas pelo modelo *Word2Vec*.

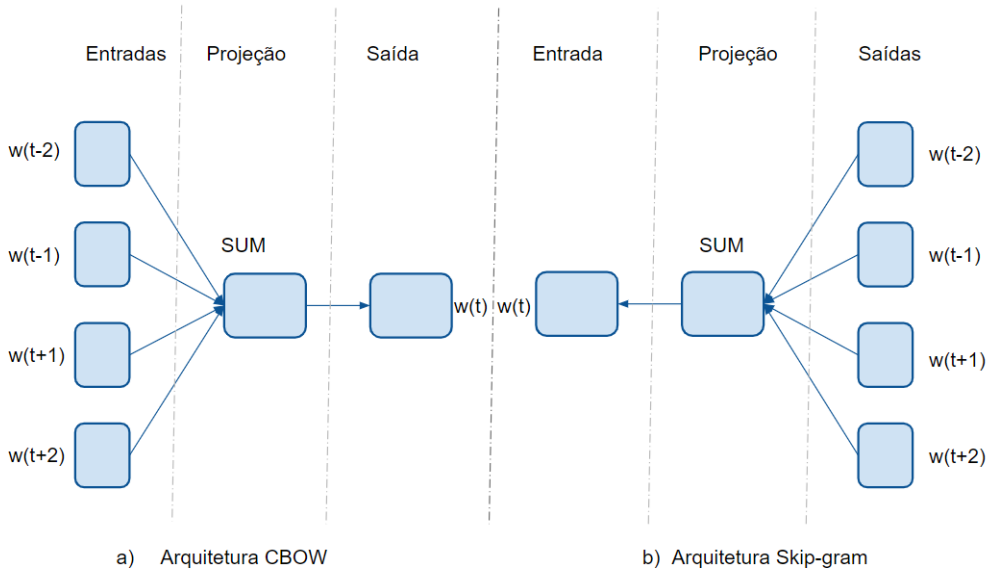

Figura 2 – Arquiteturas utilizadas pelo modelo *Word2Vec*.

Fonte: Adaptado de Mikolov et al. (2013a).

O modelo tem menor complexidade computacional comparada a outras estratégias baseadas em redes neurais presentes na literatura e consegue obter vetores eficientes a

partir de grandes conjuntos de dados, em um tempo relativamente curto (MIKOLOV et al., 2013a; MIKOLOV et al., 2013b).

2.1.2 MODELO *FASTTEXT*

É um modelo para representação de palavras proposto por Bojanowski et al. (2017), tratado como uma extensão do modelo *Word2Vec*, que considera as palavras como um conjunto de n-gramas de caracteres. Dessa forma, o vetor resultante de uma determinada palavra é dado pela soma das sub-representações de seus n-gramas. Essa abordagem de particionamento dos *tokens* possibilita a obtenção de representações para palavra não vistas no conjunto de treinamento, como por exemplo, a ocorrência de sufixos e prefixos. Além disso, palavras raras podem ter representações mais robustas do que aquelas obtidas pelo *Word2Vec* (POLO et al., 2021).

Uma das limitações dos modelos baseados em *Word2Vec* é que produzem vetores únicos para uma determinada palavra. No entanto, é comum haver variações de significados de um termo em uma linguagem, dependendo do contexto em que ele aparece. Nesse sentido, em 2019, Devlin et al. (2019) lançaram o modelo *BERT*, tendo como uma de suas principais características a geração de *embeddings* dinâmicos. Isso significa que, dependendo do contexto em que uma determinada palavra se encontra, um vetor diferente é mapeado. Os modelos baseados em *BERT* utilizam recursos (codificador) da arquitetura *Transformer* para a geração das representações (VASWANI et al., 2017). Esta arquitetura utiliza mecanismos de atenção para aprender as relações contextuais entre palavras (ou sub-palavras) das sentenças de entrada.

2.1.3 *BERT*

O BERT é um modelo para geração de *embeddings* dinâmicas, proposto por Devlin et al. (2019) e disponibilizado pela Google, que utiliza redes neurais *Transformers* para realizar o mapeamento dos textos de entrada (VASWANI et al., 2017) . A arquitetura *Transformer* permite que o *BERT* analise o texto de forma bidirecional em busca das relações contextuais de uma palavra. Uma das principais vantagens do *BERT* é a capacidade de fornecer vetores dinâmicos para as palavras, ou seja, dependendo do contexto em que ela aparece, uma representação diferente é mapeada. Por exemplo, a palavra “juiz” pode se referir a um magistrado em uma determinada sentença e em outra estar relacionada a um árbitro de futebol.

Para que o *BERT* tenha bom desempenho, o modelo precisa ser treinado com um grande volume de dados. Este treinamento é conduzido de forma auto-supervisionada, usando as técnicas *Masked Language Model* (MLM) e *Next Sentence Prediction* (NSP). No MLM, o treinamento visa a previsão de palavras ocultas nas sentenças, também chamadas de palavras mascaradas. Já no NSP, o foco está na previsão binária da próxima sentença,

ou seja, determinar se determinada sentença B é a sentença seguinte de A. Os modelos pré-treinados do *BERT* permitem que ele seja ajustado para tarefas específicas sem a necessidade de grandes quantidades de dados de treinamento do domínio (DEVLIN et al., 2019).

2.2 CONSIDERAÇÕES SOBRE AS REPRESENTAÇÕES *EMBEDDINGS*

Nesta pesquisa, diferentes técnicas para representação textual foram investigadas e a utilização dos modelos mencionados nas subseções anteriores foi adotada nos experimentos computacionais. Essas abordagens foram selecionadas com base em sua robustez científica, maturidade tecnológica e eficácia comprovada. Apesar da existência de outras abordagens notáveis, como o *Global Vectors for Word Representation* (GloVe) (PENNINGTON; SOCHER; MANNING, 2014) e o *Generative Pre-trained Transformer* (GPT) (BROWN et al., 2020), os modelos escolhidos estão consoantes com a literatura e com o contexto de execução do projeto, como recursos computacionais disponíveis e aspectos ligados à privacidade dos dados no momento da implantação do modelo. Ademais, a escolha dos modelos Word2Vec, *FastText* e *BERT* considerou o cenário brasileiro, onde essas metodologias têm se mostrado sólidas e relevantes em vários domínios de aplicação, conforme apresentado no capítulo seguinte.

3 TRABALHOS CORRELATOS

Esta seção discute trabalhos que apresentam técnicas e soluções com foco em representações *embeddings*, destacando aquelas direcionadas à língua portuguesa e também ao domínio jurídico. Inicialmente são apresentadas as principais técnicas de representações, na sequência são destacados trabalhos com modelos *embeddings* treinados com dados da língua portuguesa de modo geral e aplicados a domínios específicos, por fim, são apresentados trabalhos pertencentes ao escopo do trabalho, o âmbito jurídico.

3.1 *EMBEDDINGS* ORIENTADOS A LÍNGUA PORTUGUESA

Um dos principais trabalhos com *embeddings* para língua portuguesa é o de Hartmann et al. (2017), nele os autores treinaram e disponibilizaram modelos *FastText*, *GloVe*, *Wang2Vec* e *Word2Vec* com diferentes dimensões, utilizando dados da língua portuguesa europeia e brasileira. Estes modelos foram analisados de forma intrínseca, por meio de análises sintáticas e semânticas; e extrínseca, utilizando-os em *Part-of-Speech tagging* (PoS tagging) e análise de similaridades semânticas. De acordo com os autores, a utilização de tarefas finais de PLN são preferíveis na avaliação dos modelos em detrimento a análises intrínsecas.

Em Cunha, Almeida e Simões (2022) os autores realizaram treinamentos de modelos *embeddings* utilizando *corpus* da língua portuguesa observando o impacto da parametrização dos modelos (*e.g.*, dimensão do vetor), tamanho do *corpus* de treinamento e do domínio. Também analisaram medidas para avaliação dos modelos treinados. Seus experimentos mostraram que as configurações paramétricas dos modelos têm influências significativas nos resultados. Cunha, Almeida e Simões (2022) concluem que avaliação por meio de analogias de palavras, utilizando *corpus* de teste, não são recomendados para modelos treinados em domínios mais segmentadas, alinhando-se as recomendações de Hartmann et al. (2017).

Visando fornecer modelos dinâmicos e contextuais orientados à língua portuguesa, Souza, Nogueira e Lotufo (2020a) treinaram modelos baseados no *BERT*, chamado de BERTimbau, e avaliaram em experimentos com NER, Similaridade de sentenças textuais e no Reconhecimento de Implicação Textual (RIT). Os resultados mostraram que o BERTimbau obteve desempenho superior ao BERT-*Multilingual*, uma versão do *BERT* treinado com textos da *Wikipedia* de mais de cem idiomas diferentes, incluindo o português, confirmando que este pode ser um grande ponto de partida para treinamento de modelos aplicados a domínios específicos da língua.

Em uma análise mais segmentada, Consoli et al. (2020) realizaram experimentos

com *embeddings* treinadas com dados da área do petróleo e gás e fizeram comparações e combinações com modelos pré-treinados na língua portuguesa de modo geral, também utilizando modelos disponibilizados por Hartmann et al. (2017). Os autores argumentaram que o setor contempla termos técnicos exclusivos, exigindo modelos específicos. Avaliações por meio do NER mostraram que combinações entre modelos gerais e do domínio específico, chamados também de *Stacking embeddings*, podem aumentar o desempenho da tarefa final, nesse caso alcançando F1-score de 84.63% nos experimentos reportados.

Ainda no domínio de petróleo e gás, Gomes et al. (2021) reforçam que a linguagem do setor possui características próprias e que palavras do português podem assumir significados completamente diferentes do comum, dificultando o aprendizado de algoritmos que consomem representações mais generalistas. Neste trabalho, os autores treinaram modelos *Word2Vec* e *FastText* em um *corpus* do domínio, contando com mais de 85 milhões de *tokens*. Os modelos foram submetidos a análises intrínsecas, verificando a relação entre pares de palavras; e extrínsecas, observando o desempenho no NER da área da Geociência. Consoante com os trabalhos anteriormente discutidos, os resultados mostraram que os modelos segmentados obtiveram desempenho superior quando comparados com os generalistas. Também neste domínio de aplicação, e reforçando os desafios terminológicos da área, Rodrigues et al. (2022) introduziram o PetroBert, um modelo baseado em *BERT* treinado com documentos do segmento e dispostos na língua portuguesa. O modelo é treinado a partir dos modelos *BERT* multilíngue e BERTimbau. Análises de similaridade de sentenças e de NER mostraram resultados promissores do modelo segmentado.

3.2 *EMBEDDINGS* ORIENTADO AO SEGMENTO JURÍDICO

Tal como no setor de petróleo e gás, a área jurídica compreende uma linguagem com características próprias e que, por vezes, determinadas palavras possuem significados totalmente da linguagem dita natural. Em Smywiński-Pohl et al. (2019), são treinados modelos *Word2Vec* e *GloVe* e visando a criação de dicionário forneça uma interface entre palavras técnicas da justiça polonesa e palavras *extralegal* que podem ser compreendidas por leigos. Os experimentos apontaram resultados superiores para o *Word2Vec* do tipo CBOW. Também ressaltando essa peculiaridade no meio jurídico, Polo et al. (2021) treinaram e disponibilizaram modelos de representações de palavras, a saber: Phraser, Word2Vec, Doc2Vec, *FastText* e *BERT*, utilizando dados públicos da justiça brasileira. Realizaram experimentos com classificação de *status* (arquivado, ativo ou suspenso) de processos judiciais como demonstração de uso dos modelos treinados.

Em Chalkidis e Kampas (2019), foram treinados e disponibilizados modelos *embeddings* a partir de um grande *corpus* de dados jurídicos disponíveis na língua inglesa. O *corpus* utilizado é formado por 123.066 peças jurídicas com aproximadamente 492.000.000 de *tokens*, envolvendo legislações do Reino Unido, União Europeia, Canadá, Austrália,

decisões da Suprema Corte Americana, além de documentos com legislações japonesas e da União Europeia traduzidas para o inglês. As representações treinadas, nomeadas de law2vec[1], utilizaram o modelo *Word2Vec* com a arquitetura *Skip-gram*. Os autores afirmaram não adotar o *FastText* por ser tendencioso a informações sintáticas e dado a formalidade dos textos utilizados, com pouca incidência de erros, não haveria necessidade de adoção de um algoritmo que visa palavras fora do vocabulário buscando contornar possíveis erros ortográficos.

No estudo de Chalkidis et al. (2020) são discutidos as possibilidades de ganhos reais ao se treinar modelos do zero, com dados do domínio, ou utilizar modelos pré-treinados como ponto de partida. Outro destaque do estudo é quanto a capacidade de aprendizado de modelos com arquiteturas mais compactas. Convém pontuar que Douka et al. (2021) também adota essa abordagem no JuriBERT. Nesse sentido, Chalkidis et al. (2020) apresentaram a família LEGAL-*BERT* contendo várias combinações e utilizando o *BERT-base* (DEVLIN et al., 2019) como ponto de partida em uma de suas variações. Ao compará-los com o *BERT-base*, os autores observaram que os modelos especialistas representam melhor tarefas com maior complexidade (*e.g.*, classificação multirrótulo) e que arquiteturas mais compactas conseguem resultados competitivos no segmento.

Em Douka et al. (2021), os autores treinaram um modelo linguístico adaptado a documentos jurídicos franceses, denominado-o de JuriBERT. O estudo experimentou variações de arquiteturas (*e.g.*, camadas *transformers*, *Attention Heads*), treinamentos partindo do zero e utilizando modelos pré-treinados como *checkpoints*. O JuriBERT demonstrou maior eficiência na representação de documentos do segmento quando comparado com os modelos franceses de caso mais geral. Também demonstraram que modelos com arquiteturas reduzidas podem oferecer resultados competitivos e com menor custo computacional. Tais resultados corroboram achados dos trabalhos para língua portuguesa que utilizam o *BERT* para domínio de petróleo e gás e também para a área médica, discutidos na seção anterior

No cenário italiano, Licari e Comandè (2022) utilizaram o modelo generalista ITALIAN XXL BERT como *checkpoint* para o treinamento de um modelo orientado ao domínio. No estudo, os autores reforçam que a linguagem jurídica do país, além de técnica, também tem influências do Latin do velho italiano, tornando sua prática incomum. Dessa forma, balizam a necessidade da construção de modelos que compreendam tais especificidades. Para o treinamento do modelo, denominado ITALIAN-LEGAL-BERT, foi utilizado um grande volume de dados jurídicos públicos de tribunais italianos (documentos de processos civis). Os resultados mostraram que o ITALIAN-LEGAL-BERT foi superior ao modelo generalista tanto em termos de perplexidade quanto em tarefas finais de PLN, como classificação de sequências, similaridade semântica e o NER. Corroborando, dessa

[1] https://archive.org/details/Law2Vec

forma, as hipóteses levantadas pelos autores.

No trabalho de Wang et al. (2020), é utilizado modelos *BERT* como base para construção de arquiteturas híbridas para rotulagem de sequência (do inglês, *sequence labelling*) orientadas à NER. Foram utilizados dados jurídicos brasileiros no processo de avaliação dos modelos desenvolvidos. Ainda na seara do NER, Batista et al. (2021) avaliaram o impacto de representações *embeddings* no processo de extração de entidades em petições iniciais da justiça brasileira. Os resultados mostraram que a configuração com a *stacking* dos modelos *embeddings* de caracteres, de palavras e *pooled Flair* obteve melhores resultados.

Também observando vetores *embeddings* resultantes, Pont et al. (2020) avaliaram o impacto da especificidade e do tamanho do corpus de texto utilizado no treinamento dos vetores. Aplicados a dados jurídicos brasileiros, em vários níveis de segmentação, os resultados mostraram que *corpus* menores capturam melhor as especificidades textos. Ou seja, para um ramo específico da justiça como no escopo analisado pelos autores (transporte aéreo) representações treinadas em *corpus* menores são preferíveis. De modo geral, quanto maior o *corpus* de treinamento, melhor o desempenho.

3.3 CONSIDERAÇÕES ACERCA DOS TRABALHOS RELACIONADOS

Os trabalhos supracitados e sumarizados nas Tabelas 1 e 2 apresentam técnicas de representações textuais e aplicações envolvendo PLN que consomem tais recursos como entrada, também ressaltam medidas avaliativas adotadas para mensurar o desempenho das soluções propostas. Os trabalhos segmentados realçam também a importância de representações orientadas ao domínio do problema para obtenção de melhores resultados. Nesse sentido, o trabalho proposto foca na construção e treinamentos de modelos orientados ao âmbito jurídico que possa discriminar com maior eficácia as especificidades da linguagem do setor.

Autor(es)	Descrição
(HARTMANN et al., 2017)	• Modelos *word embeddings* para a língua portuguesa: *FastText, Word2Vec, Wang2Vec* e *Glove*; • Avaliação intrínseca e extrínseca;
(CUNHA; ALMEIDA; SIMÕES, 2022)	• Modelos *word embeddings* para a língua portuguesa: *Word2Vec*; • avaliações paramétricas: Tamanho do vetor, épocas de treinamento, entre outras; • Avaliação de forma intrínseca.
(CONSOLI et al., 2020)	• Modelos *word embeddings* orientados ao domínio do petróleo e gás: *Word2Vec, Flair Embeddings* e *Stacked Embeddings*; • Avaliação com NER. • Comparação com Modelos de (HARTMANN et al., 2017);
(GOMES et al., 2021)	• Modelos para o segmento do petróleo e gás: *Word2Vec* e *fastText*; • Avaliação com NER. • Comparação com modelos generalistas;
(SOUZA; NOGUEIRA; LOTUFO, 2020a)	• Desenvolveram um modelo baseado no BERT para a língua portuguesa, denominado BERTimbau; • Superou o BERT multilíngue em várias tarefas PLN: NER, Similaridade de sentenças textuais e RIT.
(RODRIGUES et al., 2022)	• Apresentaram o modelo PetroBert; • Utiliza o BERTimbau como *checkpoint*; • Avaliação com NER.

Tabela 1 – Sumarização dos Trabalhos Correlatos

Autor (es)	Descrição
(POLO et al., 2021)	• Apresentaram embeddings focados em documentos jurídicos em português. • Modelos treinados: Phraser, Word2Vec, Doc2Vec, FastText e BERT • Avaliação na classificação de dados; • Diferenças relativamente em comparação com modelos genéricos; • Reforçam a necessidade de modelos segmentados
(CHALKIDIS; KAMPAS, 2019)	• Treinaram embeddings com jurídicas em inglês; • Utilizaram o modelo Word2Vec para treinar o modelo Law2Vec; • Ressaltam que o formalismo dos textos jurídicos minimizam a necessidade de abordagens como o FastText;
(CHALKIDIS et al., 2020)	• Apresentaram a família LegalBert, treinada com dados em inglês. • Várias estratégias de treinamentos e configuração (e.g., versão small). • Referência para outros modelos: JuriBERT (CHALKIDIS et al., 2020), ITALIAN-LEGAL-BERT (LICARI; COMANDÈ, 2022) e o LegalBertpt (SILVEIRA et al., 2023).

Tabela 2 – (Continuação) Sumarização dos Trabalhos Correlatos.

4 MATERIAIS E MÉTODOS

Este capítulo apresenta os procedimentos metodológicos adotados para o desenvolvimento desta pesquisa, detalhando as principais etapas que envolvem a construção e avaliação dos modelos de representação *embeddings* adotados. A primeira seção, discorre sobre o *framework* utilizado para estruturação do projeto, discutindo suas fases e as atividades realizadas em cada uma delas. A segunda detalha as tecnologias utilizadas na implementação e na execução dos experimentos, visando reprodutividades futuras.

4.1 *CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING*

Considerando a natureza abrangente deste projeto de pesquisa, optou-se pela utilização do modelo *Cross-Industry Standard Process for Data Mining* (CRISP-DM) para o desenvolvimento das atividades. O CRISP-DM é um modelo metodológico que fornece uma abordagem estruturada e cíclica para organizar e realizar projetos de mineração de dados. Sua estrutura contempla as fases e suas respectivas tarefas e saídas, formando um ciclo do projeto (WIRTH, 2000). A Figura 3 apresenta o diagrama do modelo.

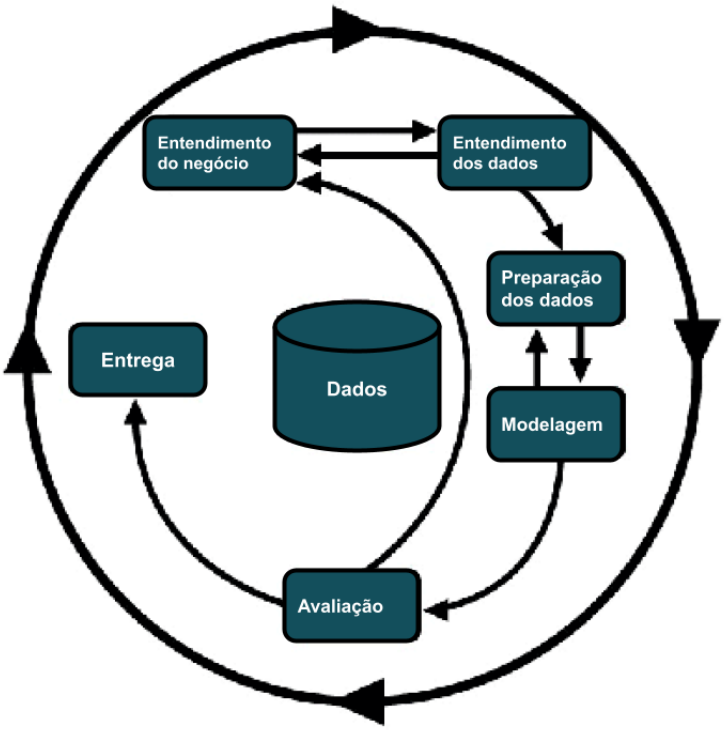

Figura 3 – Diagrama de funcionamento da metodologia CRISP-DM.

Fonte: Adaptado de (WIRTH, 2000).

Wirth (2000) e Schröer, Kruse e Gómez (2021) descrevem seis fases que contemplam um ciclo do CRISP-DM, tal como mostrados na Figura 3, a saber: *i*) Entendimento do negócio, *ii*) Entendimento dos dados, *iii*) Preparação de dados, *iv*) Modelagem, *v*) Avaliação e *vi*) Implantação. Tais fases são detalhadas a seguir:

1. **Entendimento do negócio**: Fase que envolve a compreensão dos objetivos de negócio e dos requisitos para o projeto de mineração de dados, fazendo um planejamento preliminar para a busca dos objetivos;

2. **Entendimento dos dados:** Envolve a coleta e a exploração dos dados para melhor compreender suas características e suas estruturas;

3. **Preparação de dados:** Fase que envolve os procedimentos de limpeza (*e.g.*, remoção de *stopwords*, caracteres especiais, *tags HyperText Markup Language* (HTML) e preparação dos dados para a fase de modelagem;

4. **Modelagem**: Esta fase consiste na construção e teste de diferentes modelos para encontrar o que melhor se adéqua aos dados e ao propósito da aplicação;

5. **Avaliação**: Etapa onde se analisam os resultados dos modelos visando mensurar a sua eficácia em relação aos objetivos de negócio;

6. **Implantação**: Última etapa, que envolve a implementação do modelo e sua utilização em produção.

A construção de modelos de representações *embeddings* para o âmbito jurídico, desenvolvidos nesta pesquisa, possui finalidades práticas, com aplicação prevista em processos de PLN no TJMA. Dessa forma, instanciamos as atividades para compreender todo o processo CRISP-DM. A seguir, são especificadas as atividades desenvolvidas de acordo com cada etapa do modelo.

4.1.1 ENTENDIMENTO DO NEGÓCIO

Considerando os desafios encontrados no sistema jurídico brasileiro em relação ao desenvolvimento de soluções baseadas em IA que promovam a celeridade de suas atividades, esta pesquisa concentra-se que soluções envolvendo PLN, acreditando que podem oferecer resultados eficazes a partir das análises de documentos jurídicos.

Conforme discutido nos capítulos anteriores, um dos principais desafios reside no desenvolvimento de aplicações que compreendam as peculiaridades da linguagem jurídica. Para atender a essa demanda, este trabalho foca no desenvolvimento de modelos de representações (*embeddings*) orientados para a linguagem jurídica do Brasil como forma de fornecer insumos para construção de aplicações com PLN no setor.

Ressalta-se que esta pesquisa é vinculada ao projeto de Cooperação Técnica entra a UEMA e o TJMA, com pretensões de aplicação prática dos modelos desenvolvidos. Dessa forma, em reunião com os *stakeholders* foram discutidos diferentes benefícios e aplicabilidade de vetores *embeddings* treinados com dados jurídicos. A exemplo de aplicação, a análise de petições iniciais para identificação automática de precedentes judiciais do Tribunal a partir conteúdo do documento pode oferecer celeridade aos trâmites processuais. Nesse caso, petições com processos semelhantes mais escritos de forma diferente requerem um fator de compreensão textual que modelos mais simplificados como, como o *Term Frequency - Inverse Document Frequency* (TF-IDF), não podem oferecer. No entanto, dados as características dos modelos *word embeddings*, que já consideram as relações entre as palavras, e dos modelos baseados em *Transformers*, que possuem capacidade de geração dinâmica de representações e conseguem fornecer resultados baseados em análises contextuais dos documentos de entrada.

No entanto, os estudos preliminares mostraram uma lacuna de modelos com tais características no cenário jurídico brasileiro. Outro fator importante está relacionado com os dados de treinamento, percebeu-se a ausência de dados textuais estruturados orientados ao domínio, com isso fez-se necessário um estudo para criação de *corpus* jurídico com dados de diferentes tribunais e regiões dos Brasil, refletindo, dessa forma, tanto a diversidade linguística brasileira quanto a diversidade de peças jurídicas. Em relação aos modelos de representação, foram escolhidos os modelos *Word2Vec* (MIKOLOV et al., 2013a) e *FastText* (BOJANOWSKI et al., 2017) e o *BERT* (DEVLIN et al., 2019), por serem bem recomendados na literatura e já oferecerem resultados práticos em diferentes domínios de aplicações.

Para o processo de avaliação, optou-se pela aplicação em tarefas finais de PLN, tal como recomendado por Hartmann et al. (2017) e Polo et al. (2021). Nesse caso, adotamos classificação de documentos jurídicos. Considerando pretensões práticas dos modelos, utilizou-se a classificação de petições inicias para identificação de precedentes do tribunal. A avaliação é feita de forma indireta de acordo com a desempenho do modelo classificador. Em resumo, o projeto foi mapeado/delimitado em três focos centrais: construção do *corpus* jurídico, treinamento dos modelos *embeddings*, aplicação em tarefas finais de PLN para avaliação.

4.1.2 ENTENDIMENTO DOS DADOS

O treinamento de modelos para a geração de *embeddings* exige um *corpus* significativo para a captura das relações entre as palavras, tanto para modelos livres de contexto (*word embeddings*), quanto para modelos contextuais, como o *BERT*. No entanto, apesar de algumas iniciativas como Sousa e Fabro (2019), Araujo et al. (2020), Luz de Araujo et al. (2018), dados jurídicos públicos e estruturados ainda são insuficientes para suprir

essa demanda. Para isso, a equipe desenvolvedora do projeto realizou um levantamento e extração de documentos legais em domínios públicos, via *web crawlers*.

Os documentos coletados, em sua maioria, são formados por acórdãos dos tribunais superiores de diferentes regiões e esferas do país. Um acórdão é um documento que descreve o resultado do julgamento de determinado processo (SOUSA; FABRO, 2019). Em muitos casos, estes documentos contemplam o entendimento (jurisprudência) daquele tribunal sobre determinado assunto. Outro documento utilizado nos experimentos é a petição inicial, essa peça jurídica é utilizada como primeiro passo para acessar ao Poder Judiciário quando se está representado por um advogado, é nela que está contida a demanda requerida (MARINATO et al., 2022; CARMO et al., 2023). A seguir estão destacadas estas instituições e as características dos dados obtidos para a formação do *corpus* de treinamento:

1. **Supremo Tribunal Federal (STF)**: Foram adotados os dados do *Iudicium Textum Dataset* (ITD), conjunto de dados disponibilizado por Sousa e Fabro (2019) e que contempla textos de acórdãos do tribunal publicados entre os anos de 2010 a 2018. Também foram extraídos dados de jurisprudências do STF disponíveis no portal LexML[1], plataforma especializada na divulgação de informações jurídicas e legislativas. Ainda deste tribunal, também foi realizado precedimentos de raspagem em sua plataforma de jurisprudência[2] buscando por dados de jurisprudenciais de repercussão geral, dispostos no formato de acórdãos. Além disso, foram adotados os dados do Dataset Victor disponibilizado por Araujo et al. (2020), contendo dados de recursos extraordinários da corte.

2. **Tribunal Superior do trabalho (TST):** Foram extraídos dados contendo decisões sobre matérias trabalhistas, disponíveis na forma de acórdãos. Foram coletadas aproximadamente 250 mil amostras dados do portal de jurisprudência do tribunal[3].

3. **Superior Tribunal Militar (STM):** Foram coletados 23.522 documentos com textos de acórdãos publicados pela suprema corte militar em seu portal[4].

4. **Tribunal Superior Eleitoral (TSE):** Foram coletadas amostras de dados contendo jurisprudência do referido tribunal, disponíveis em seu sistema de consulta[5]. A amostragem contempla mais de 84.000 acórdãos na área eleitoral.

5. **Tribunal de Contas da União (TCU):** Foram coletados dados de decisões do Tribunal, em forma de acórdão, compreendendo jurisprudências da instituição. Os

[1] https://www.lexml.gov.br/
[2] https://jurisprudencia.stf.jus.br
[3] https://www.tst.jus.br/jurisprudencia
[4] https://www.stm.jus.br/gestao-da-informacao/pagina-inicial-gest-inform/jurisprudencia
[5] http://www.tse.jus.br/jurisprudencia

dados foram coletados do portal de Pesquisa integrada do TCU[6]. Uma amostra de 15.000 documentos foram extraídos.

6. **Conselho Nacional de Justiça (CNJ):** Foram utilizados dados do Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatórios (BNPR)[7], fornecido pela instituição, contendo dados atualizados sobre processos recorrentes de vários tribunais brasileiros.

7. **Conselho da Justiça Federal (CJF):** Foram utilizados documentos com jurisprudência de vários tribunais regionais e superiores, extraídos do portal de jurisprudência unificada do CJF[8].

8. **TJMA:** foi utilizado um conjunto de dados contendo 11.700 petições iniciais do TJMA. O *dataset* já é anotado de acordo com precedentes relacionados. Esta base será utilizada tanto para treinamentos de modelos quanto na fase de avaliação.

A Tabela 3 apresenta as principais características dos conjuntos dados obtidos e sua representação no *corpus* jurídico final.

Tabela 3 – Distribuição do *corpus* jurídico de treinamento.

Conjunto de Dados/Instituição	Amostras	Incidência (%)
Jurisprudências do CNJ	12.014	0,22
LexML	72.280	1,34
Dataset Victor	2.426.376	44,42
Acórdãos do STF	272	0,01
Temas repetitivos do STJ	1.772	0,03
Jurisprudências do TST	247.084	4,57
Jurisprudências TCU	15.000	0,28
Jurisprudências do STM	23.522	0,44
Jurisprudências do TSE	84.754	1,57
Jurisprudências do TJMA	4.020	0,07
Jurisprudências de Repercussão Geral do STF	1.036	0,02
Temas repetitivos do STJ	1.772	0,03
ITD	41.353	0,77
BNPR	3.255	0,06
Petições iniciais TJMA	11.700	0,22
Modelos de Petições Iniciais	1.814	0,03
Jurisprudências da Justiça Federal	2.453.020	45,42
TOTAL	5.401.044	100

4.1.3 PREPARAÇÃO DOS DADOS

A fase de tratamento e preparação dos dados é uma etapa fundamental em um projeto de mineração de dados, quando não tratados de maneira adequada os resultados finais podem ser seriamente comprometidos (WIRTH, 2000; SILVA et al., 2023). Nesse

[6] https://pesquisa.apps.tcu.gov.br/
[7] https://bnpr.cnj.jus.br/
[8] https://www2.cjf.jus.br/jurisprudencia/unificada/

sentido, Silva et al. (2023) discutem vários procedimentos de limpeza e formatação que podem ser aplicados em textos jurídicos, entre eles: remoção de *stopwords*, remoção de espaços em branco, remoção de números, entre outros.

No presente projeto diferentes abordagens foram utilizadas de acordo com o tipo de modelo de representação adotado. Para os modelos baseados em palavras, utilizou-se como base as técnicas adotadas por (HARTMANN et al., 2017) no treinamento de modelos *word embeddings* para a língua portuguesa. Já para os modelos baseados no *BERT*, adotaram-se técnicas similares as utilizada por Polo et al. (2021) no pré-treinamento do BERTikal. Uma das principais diferenças entre as duas abordagens refere-se as *stopwords*, não recomendadas no treinamento de modelos *word embeddings*, mas comumente preservada em modelos contextuais. A Tabela 4 apresenta os procedimentos adotados em cada uma das abordagens.

Tabela 4 – Filtros de pré-processamento de dados de acordo com o modelos utilizado.

texto original	*Word embeddings*	BumbaBert
Caracteres minúsculos (*lowercase*)	✓	✓
Remoção de *stopwords*	✓	×
Remoção de excesso de espaços em branco	✓	✓
Remoção de quebra de linhas (“/n”)	✓	✓
Remoção de marcadores HTML	✓	✓
Remoção de caracteres duplicados (*e.g.,* aa, aaaa)	✓	✓
Configuração de tokens de email	✓	×
Configuração de token URL	✓	×

Também foram corrigidos de erros de codificação *unicode*, advindas do procedimento do *web crawler*, utilizando a biblioteca *Ftfy* (SPEER, 2019), além da transformação de caracteres para *lowercase*. Os números foram preservados dada a sua importância para o domínio do problema, considerando que tais números são parte importante de um documento jurídico (*e.g.,* número de processos, artigos de lei etc).

4.1.4 MODELAGEM

Com a preparação de dados realizada, Wirth (2000) ressalta que na fase de modelagem são realizados treinamentos com diversos modelos e configurações para a escolha dos mais eficientes. Nesta pesquisa, o foco está na construção e treinamento dos vetores de representação dos textos jurídicos. Dessa forma, foi construído um *framework* contemplando diferentes cenários para a geração de representações de vetores *embeddings* orientados à linguagem jurídica brasileira.

A construção do *framework* experimental desta pesquisa envolve o treinamento de modelos *Word2Vec* e *FastText* considerando vários cenários e configurações para gerações dos vetores. O pré-treinamento de modelos baseados na arquitetura do *BERT*. Ambos os modelos utilizando o *corpus* jurídico destacado nas subseções anteriores. Também foram incorporados no *framework* experimental, modelos pré-treinados e disponibilizados na literatura, tanto orientado para a linguagem portuguesa de modo geral, quanto para

o domínio de aplicação. A Figura 4 apresenta um fluxograma para o treinamento das representações.

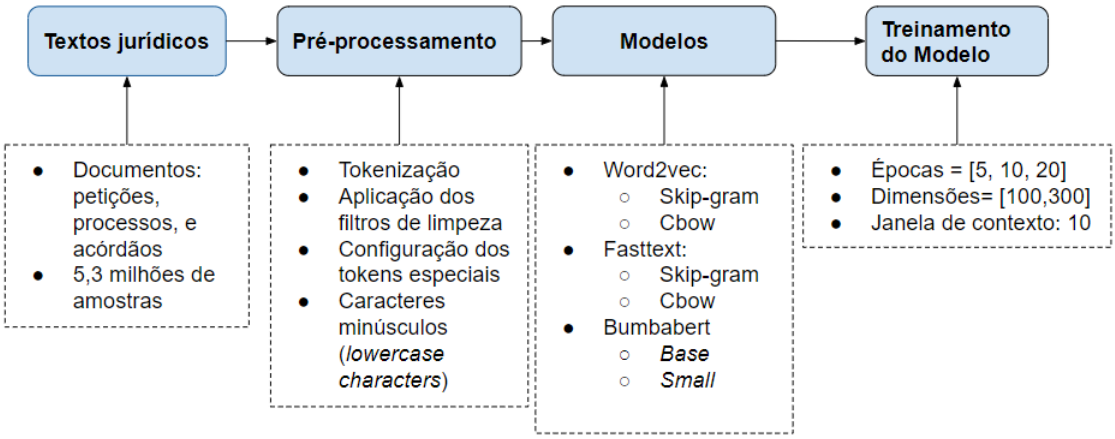

Figura 4 – Fluxograma para o treinamento dos modelos.

Para os modelos *Word2Vec* e *FastText* foram considerados as seguintes configurações: *i)* tipo de arquitetura utilizada: CBOW e Skip-gram; ii) a dimensão do vetor de características: 100 e 300; e iii) épocas de treinamento: 5, 10 e 20. Dessa forma, foram treinados 24 cenários diferentes ($cenários = modelo * arquitetura * númeroDeÉpocasDeTreinamento$). A Tabela 5 apresenta os cenários configurados para os modelos *word embeddings*. Foram utilizadas janelas de contexto de tamanho 10 e taxa de aprendizado de 0,03. Para os demais parâmetros, foram utilizados os valores padrão da biblioteca Python Gensim [9], considerando os resultados de Cunha, Almeida e Simões (2022) sobre o impacto dos parâmetros no modelo resultante. Como modelos pré-treinados, foram incorporados ao *framework*, os modelos *word embeddings* treinados com dados generalistas da língua portuguesa por Hartmann et al. (2017) e disponibilizados pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo [10] e os modelos treinados por Polo et al. (2021) já com dados do segmento jurídico.

Para os modelos baseados na arquitetura do *BERT*, foram construídos modelos, denominados de BumbaBert, com processo de pré-treinamento a partir de modelos pré-treinados (*checkpoints*) na língua portuguesa, nesse caso, utilizando o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020b) e modelos configurados para pré-treinamento do zero. Tais abordagens são utilizadas nos trabalhos de Chalkidis et al. (2020), Silveira et al. (2023) e Douka et al. (2021). Para algumas aplicações, a continuação do treinamento a partir de modelos pré-treinados oferecem resultados satisfatórios, em outros casos, modelos treinados

[9] https://radimrehurek.com/gensim
[10] http://www.nilc.icmc.usp.br/embeddings

Tabela 5 – Modelos *word embeddings* utilizados nos experimentos computacionais.

id	Cenário [11]	dimensão
C1-w2v	word2vec-cbow-5	100
C2-w2v	word2vec-skipgram-5	100
C3-w2v	word2vec-cbow-10	100
C4-w2v	word2vec-skipgram-10	100
C5-w2v	word2vec-cbow-20	100
C6-w2v	word2vec-skipgram-20	100
C7-w2v	word2vec-cbow-5	300
C8-w2v	word2vec-skipgram-5	300
C9-w2v	word2vec-cbow-10	300
C10-w2v	word2vec-skipgram-10	300
C11-w2v	word2vec-cbow-20	300
C12-w2v	word2vec-skipgram-20	300
C1-ft	fasttext-cbow-5	100
C2-ft	fasttext-skipgram-5	100
C3-ft	fasttext-cbow-10	100
C4-ft	fasttext-skipgram-10	100
C5-ft	fasttext-cbow-20	100
C6-ft	fasttext-skipgram-20	100
C7-ft	fasttext-cbow-5	300
C8-ft	fasttext-skipgram-5	300
C9-ft	fasttext-cbow-10	300
C10-ft	fasttext-skipgram-10	300
C11-ft	fasttext-cbow-20	300
C12-ft	fasttext-skipgram-20	300
C1-NILC-w2v	word2vec-Cbow	100
C2-NILC-w2v	word2vec-Skip-gram	100
C3-NILC-w2v	word2vec-Cbow	300
C4-NILC-w2v	word2vec-Skip-gram	300
C5-NILC-ft	fastText-Cbow	100
C6-NILC-ft	fastText-Skip-gram	100
C7-NILC-ft	fastText-Cbow	300
C8-NILC-ft	fastText-Skip-gram	300
C1-LegalNLP-w2v	fastText-Cbow-15	100
C2-LegalNLP-w2v	fastText-Skip-gram-15	100
C3-LegalNLP-ft	fastText-Cbow-15	300
C4-LegalNLP-ft	fastText-Skip-gram-15	300

do zero mostram maior capacidade de aprendizado das características da linguagem do domínio, por isso a consideração de tais cenários.

Outros fatores importantes discutidos por Douka et al. (2021) e Chalkidis et al. (2020) estão relacionados com a configuração da arquitetura utilizada para o treinamento dos modelos baseados em *Transformers*, incluindo aspectos que envolvem experimentos com arquiteturas reduzidas - número de camadas *Transformers*, número de cabeças de atenção, tamanho da camada interna (*hidden size*). Tais experimentos mostraram competitividade no aprendizado e a redução do custo computacional. Dessa forma, também foi incorporado ao *framework* experimental proposto o pré-treinamento de uma versão reduzida do BumbaBert. A Tabela 6 a seguir descreve os modelos e os principais parâmetros adotados. Ressalta-se a adoção da terminologia utilizada por Chalkidis et al. (2020) para os modelos BumbaBert, onde *scratch* (SC) representa os modelos treinados a partir do zero nos dados do domínio; E *further pre-train* (FT) refere-se ao modelo treinado a partir do pré-treinamento do modelo BERTimbau.

Adicionalmente os modelos pré-treinados BERTimbau e o BERTikal foram incorporados ao *framework* experimental proposto. O primeiro representa o estado da arte de modelos baseados em *BERT* para a língua portuguesa e por isso foi considerado como *baseline* de valiação, além da utilização como checkpoint em uma das versões do BumbaBert. Conforme mencionado no Capítulo 3, o BERTikal também utiliza o BERTimbau com ponto de partida.

O processo de tokenização dos modelos BumbaBert foram adotados duas estratégias diferentes, em alinhamento com os trabalhos correlatos. Para a versão FT do BumbaBert optou-se pela utilização da abordagem padrão de tokenização do BERT, conhecido como *WordPiece*. Para isso, aproveitou-se o tokenizador já treinado e disponibilizado com o modelo BERTimbau, o qual foi utilizado como ponto de partida de treinamento. Para os modelos SC, foi treinado um novo tokenizador do zero utilizando a estratégia *Byte-Pair Encoding* (BPE), abordagem adotada pelos autores do Juribert (CHALKIDIS et al., 2020). Esse mapeamento foi feito após testes preliminares com os dados jurídicos trabalhados nesta pesquisa.

Tabela 6 – Modelos para geração de *embeddings* dinâmicas.

Modelo	*checkpoint*	Arquitetura	Total de parâmetros
Família BumbaBert			
BumbaBert (*base*) FT	BERTimbau	(L=12, H=768, A= 12)	110 milhões
BumbaBert (*base*) SC	-	(L=12, H=768, A= 12)	110 milhões
BumbaBert (*small*) SC	-	(L=6, H=512, A=8)	42 milhões
Modelos pré-treinados			
BERTimbau (*base*)	-	(L=12, H=768, A= 12)	110 milhões
BERTikal (*base*)	BERTimbau	(L=12, H=768, A= 12)	110 milhões

Em relação ao processo de construção dos modelos baseados no *BERT*, o pré-

treinamento foi realizado conforme o tamanho da arquitetura do modelo. Para a versão *small* foram consideradas mais de 1 milhão de etapas, tal como recomendado por (DOUKA et al., 2021). Já para as versões *base* foram consideradas cerca 400.000 etapas de treinamentos, ressaltando que as versões *base* possuem maior custo computacional. A Tabela 7 apresenta os principais parâmetros utilizados.

Tabela 7 – Principais configurações de pré-treinamento.

Nome	**Configuração**
Taxa de Aprendizagem	$1e-4$
Otimizador	Adam (parâmetros: $\beta_1 = 0,9$ e $\beta_2 = 0,99$)
Decaimento de Peso	0.1
Tamanho de Lote (*Batch Size*)	8

4.1.5 AVALIAÇÃO

Para a avaliação dos modelos treinados adotaram-se as recomendações apontadas por Hartmann et al. (2017), que indica que aplicações finais de PLN são preferíveis para atestar a capacidade dos modelos *embeddings*. Nesse sentido, tarefas como a classificação de documentos (SMYWIńSKI-POHL et al., 2019) e o reconhecimento de entidades nomeadas (WANG et al., 2020; BATISTA et al., 2021) podem fornecer boas análises. Neste sentido, o desempenho pode ser medido por meio da taxa de acerto do modelo preditor.

A escolha das métricas de avaliação depende dos modelos adotados e das características dos dados utilizados para testes. Neste trabalho é adotado medidas amplamente utilizada na literatura, tais como a acurácia do classificador (ACC), medida que computa a quantidade acertos preditos pelo algoritmo; e o F1-macro, medida que contabiliza o percentual de acertos do modelo de acordo as classes do problema, recomendada para conjunto de dados que possuem alto grau de desbalanceamento. A ACC é definida da seguinte forma:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Onde:

- TP representa os verdadeiros positivos (casos positivos corretamente preditos);
- TN representa os verdadeiros negativos (casos negativos corretamente preditos);
- FP representa os falsos positivos (casos negativos incorretamente preditos como positivos);
- FN representa os falsos negativos (casos positivos incorretamente preditos como negativos).

A métrica F1-*macro* é uma variação da medida *F1-score* para lidar com problemas multicitasses desbalanceadas. F1-*macro* leva em consideração as métricas *recall* e *precision*. A seguir são detalhadas essa composição:

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{4.2}$$

Onde:

- N é o número total de classes;
- $F1_i$ é a medida F1-*score* para a classe i;

A medida F1-*score* para cada classe individual ($F1_i$) é dada por:

$$F1_i = \frac{2 \cdot precision_i \cdot recall_i}{precision_i + recall_i} \tag{4.3}$$

Onde:

- $precision_i$ é a precisão para a classe i;
- $recall_i$ é a revocação para a classe i.

A *precision* para a classe i é definida como:

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{4.4}$$

E o *recall* para a classe i é definida como:

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{4.5}$$

Onde:

- TP_i representa os verdadeiros positivos para a classe i;
- FP_i representa os falsos positivos para a classe i;
- FN_i representa os falsos negativos para a classe i.

Para o processo de avaliação foram adotados vários algoritmos classificação já consolidados na literatura, citam-se o *Support Vector Machine* (SVM), *Random Forest* (RF), *k-Nearest Neighbors* (kNN) e *Logistic Regression* (LR).

Tabela 8 – Conjunto de dados de petições iniciais do TJMA.

Tipo de IRDR	Amostras	Incidência (%)
1	3.581	65,18
2	27	0,49
3	502	9,14
4	54	0,98
5	1.068	19,44
6	4	0,07
7	6	0,11
8	252	4,59
TOTAL	5.494	100

Para a avaliação em uma aplicação final de PLN, os modelos foram aplicados na classificação de Petições iniciais. O objetivo é identificar a qual Incidente de Resolução de Demandas Repetitivas (IRDR) determinada petição inicial pertence. O IRDR é uma técnica utilizada pelos tribunais inferiores da justiça brasileira que visa fornecer uma mesma decisão para casos semelhantes, ou seja, processos semelhantes terão resultados semelhantes, promovendo, dessa forma, isonomia e segurança jurídica. Com isso, a identificação automática de um IRDR baseado no conteúdo do documento de entrada (petição inicial) pode proporcionar celeridade nos trâmites processuais.

Nestes experimentos, foram utilizados os dados das petições iniciais disponibilizadas pelo TJMA como conjunto de testes. Tais petições já estão anotadas de acordo com seu tipo de IRDR. Estas anotações foram feitas por profissionais que atuam no setor, garantindo a eficácia da rotulagem de classes. A Tabela 8 apresenta as características do conjunto de dados.

4.1.6 ENTREGA

A fase de entrega do CRISP-DM são considerados as entregas ou implantação dos artefatos desenvolvidos (WIRTH, 2000). Nesta pesquisa, tal como ressaltado na Seção 1.4 deste documento, contribuições de cunho acadêmico e prático serão entregues. A seguir são apresentados tais artefatos:

1. *Corpus* jurídico: Construção de um *corpus* que reúne informações estruturadas, confiáveis e representativas do âmbito jurídico. Esse recurso visa preencher a lacuna identificada durante a pesquisa, fornecendo subsídios essenciais para o desenvolvimento de novas aplicações de PLN voltadas para o setor. ;

2. Modelos Treinados: Os artefatos produzidos neste trabalho são parte de um projeto de pesquisa que visa soluções com IA para o TJMA, mostrando, dessa forma, um

contributo real para o segmento. Os modelos também serão disponibilizados para comunidade acadêmica visando difundir o conhecimento na área;

3. Cobertura experimental. Os experimentos realizados fornecerão elementos para a reprodução dos treinamentos e avaliação de modelos, bem como aplicação dos mesmos em tarefas finais envolvendo PLN.

4. Publicação de Artigos: Visando reprodutividade e disseminação do conhecimento, os resultados dos experimentos estão sendo publicados em eventos da área:

   - **Artigo 1**: Os resultados dos experimentos preliminares com os modelos *word embeddings* aplicados ao domínio jurídico foram apresentados no XI Workshop de Computação Aplicada em Governo Eletrônico (WCGE 2023). O trabalho, intitulado "*Embeddings* Jurídico: Representações Orientadas à Linguagem Jurídica Brasileira", aborda os procedimentos de treinamento dos modelos, com dados do domínio, e o processo de avaliação do na classificação de dados jurídicos. O artigo correspondente está disponível nos anais do evento, acessível pelo endereço eletrônico: https://sol.sbc.org.br/index.php/wcge/article/view/24876, e no Apêndice A deste documento.
   - **Artigo 2**: O artigo contemplando os resultados dos experimentos com modelos contextuais está em fase de finalização e também será um meio de divulgação desta pesquisa. O trabalho contemplará todo processo de treinamento e avaliação dos modelos denominados BumbaBert, nossos modelos baseados na arquitetura do *BERT* para o segmento jurídico.

4.2 TECNOLOGIAS UTILIZADAS

Esta seção descreve as tecnologias utilizadas para o tratamento dos dados, implementação dos modelos de representação e dos classificadores utilizados para avaliação e para a execução dos experimentos.

A implementação dos procedimentos de pré-processamento dos dados foi realizado utilizando os recursos disponíveis na biblioteca *Spacy*[12], uma biblioteca de código aberto, desenvolvida na linguagem Python, voltado para o PLN. Para a construção dos modelos baseados em *Word embeddings*: *Word2vec* e *FastText*, foi adotado a biblioteca Gensim[13], também amplamente utilizada para aplicações com PLN.

Para a implementação dos algoritmos de classificação foram adotados os recursos da biblioteca Python *Scikit-learn* [14], uma biblioteca amplamente utilizada para construção de modelos de ML.

[12] https://spacy.io/
[13] https://radimrehurek.com/gensim/
[14] https://scikit-learn.org/stable

Para o treinamento dos modelos baseados no *BERT*, foi adotado os recursos da biblioteca *Transformers*[15] e do framework PyTorch [16].

No processo de visualização dos *embeddings* foram adotadas as bibliotecas: *Sentence Transformers*[17], no processo de vetorização e *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP)[18], para a visualização gráfica em baixa dimensão.

A execução da prototipação e condução de experimentos iniciais foi realizada utilizando a plataforma *Google Colaboratory*[19] na versão Pro +, com capacidade de 52GB de Memória RAM e com GPUs do tipo Nvidia Tesla K80, T4 e P100 disponíveis. Os experimentos finais foram conduzidos no Centro Tecnológico de Computação Científica Aplicada da Universidade Federal do Oeste do Pará.

15 https://github.com/huggingface/transformers
16 https://pytorch.org/
17 https://www.sbert.net/
18 https://umap-learn.readthedocs.io/en/latest/index.html
19 https://colab.research.google.com/

5 RESULTADOS E DISCUSSÕES

Este capítulo discorre sobre os resultados obtidos no processo de avaliação dos modelos de acordo com os procedimentos descritos no Capítulo anterior. Mais especificamente, são apresentados e discutidos os desempenhos dos modelos de representações (*Word2Vec*, *FastText* e *BERT*) no processo da classificação de Petições Iniciais do TJMA. Considerando a ampla cobertura experimental, as análises foram realizadas da seguinte forma: avaliação dos modelos *word embeddings*, avaliação dos modelos baseados no *BERT* e Avaliação de desempenho Geral.

5.1 *EMBEDDINGS* PARA CLASSIFICAÇÃO DE PETIÇÕES INICIAIS

O *framework* experimental contempla o treinamento de 12 combinações do modelo *Word2Vec*, 12 combinações do modelo *FastText* e 3 variações do modelo BumbaBert (baseados na arquitetura do *BERT*), totalizando 27 tipos diferentes de representações construídas neste trabalho. Adicionalmente, também foram incorporados modelos pré-treinados com configurações semelhantes como *baseline* comparativos. Nesse caso, modelos *word embeddings* (*Word2Vec* e *FastText*) treinados com dados generalistas da língua portuguesa por Hartmann et al. (2017); e com dados do domínio jurídico por Polo et al. (2021). Como *baselines* para os modelos contextuais, adotou-se o generalista BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020a); e o jurídico BERTikal (POLO et al., 2021). Ambos os cenários estão descritos na seção 4.1.4 deste documento.

5.1.1 RESULTADOS DOS MODELOS *WORD EMBEDDINGS*

Em relação aos modelos baseados em *word embedding*, as Tabelas 9 e 10 apresentam os resultados obtidos na classificação de petições iniciais, destacando em cinza e negrito os maiores resultados. Para a acurácia, apresentada na Tabela 9, o modelo com o maior desempenho geral foi o C11-ft, representando o cenário com: *FastText*, arquitetura CBoW, 20 épocas de treinamento e com dimensão 300. Tal modelo obteve 73% de acurácia média (considerando todos os algoritmos de classificação) e alcançou 77% no melhor cenário, com o classificador k-NN.

Outras variações do modelo *FastText* também atingiram 77% de acurácia (C1-ft, C7-ft e C9-ft), ambas em combinações do *FastText* com o k-NN. Na avaliação com F1-*macro*, detalhada na Tabela 10, as combinações com o modelo C11-ft também obtiveram maior desempenho médio, alcançando 29%, mesmo resultado do cenário utilizando C3-w2v. No entanto, C11-f foi o que obteve o melhor resultado geral, também combinado com o algoritmo k-NN, alcançando 43% para tal métrica.

Considerando o desempenho dos classificadores, os resultados tanto para a acurácia quanto para F1-*macro* mostram superioridade do classificador *k*-NN, atingindo em média até 75% e 40% de acurácia e F1-*macro*, respectivamente. O algoritmo de pior desempenho foi o RF, com acurácia média de 65% para todos os modelos analisados e F1-*macro* atingindo 12%. Ressalta-se que o processo de aprendizado dos algoritmos podem influenciar no resultado da avaliação dos modelos, no entanto, considerando que o *framework* contempla cenários com todos os algoritmos e modelos, é possível observar o poder de influência do modelo vetorizador utilizado no processo de classificação.

Em uma análise comparativa com modelos generalistas pré-treinados por Hartmann et al. (2017), os resultados indicam diferença média relativamente baixa em relação à avaliação via acurácia. Nesse caso, a maior diferença ocorre na via avaliação com o SVM, onde os resultados dos modelos treinados com dados do domínio atingiram média 72% (modelos *Word2Vec* e *FastText* treinados e *FastText* do LegalNLP) enquanto os modelos do NILC obtiveram a média máxima de 69%, utilizando vetores *FastText*. Já na avaliação via F1-*macro*, também para o algoritmo SVM, nota-se uma diferença de até 8 pontos percentuais entre os cenários com modelos segmentados e com modelos generalistas. Para esse algoritmo, quando combinados com os modelos extraídos do NILC, as médias obtiveram média de 17% (*Word2Vec*) e 18% (*FastText*), já os modelos LegalNLP atingiram 24% (*Word2Vec*) e 25% (*FastText*) e os modelos treinados nesta pesquisa alcançaram 25% (*Word2Vec*) e 26% (*FastText*). Tais resultados do SVM são indicativos da influência dos dados do domínio no aprendizado das classes. Os demais classificadores apresentaram resultados alinhados com os experimentos realizados por Polo et al. (2021), mostrando semelhança entre os obtidos por modelos jurídicos e os treinados com dados genéricos no processo de classificação. A exemplo, a variação com acurácia média foi de apenas 1 ponto percentual: 75% e 74% nos cenários com *k*-NN, 73% e 72% com LR e 66% e 65% com RF.

Sob a ótica dos modelos orientados ao domínio, os resultados também mostraram uma competitividade entre os treinados neste estudo e os modelos LegalNLP. Em termos de acurácia, os modelos *FastText* treinados foram superiores, em média, quando combinados com algoritmos *k*-NN e SVM, com 75% e 72% respectivamente, já os modelos LegalNLP, obtiveram 74% e 70%. Para LR e RF, os modelos *FastText* do LegalNLP foram superiores apenas 1 ponto percentual considerando os cenários com maior desempenho, os vetores *FastText* treinados *k*-NN atingiram 77% enquanto os pré-treinados atingiram 76%.

Do ponto de vista do tipo modelo de representação, os resultados mostraram desempenho equilibrado entre modelos *Word2Vec* e *FastText*. Para os cenários treinados, ambos os modelos obtiveram 71% e 26% na média geral, considerando as 12 combinações e os 4 classificadores, para acurácia e F1-*macro*, respectivamente. Esse equilíbrio também é observado nos modelos pré-treinados avaliados.

No entanto, os quatro melhores cenários (C1-ft, C7-ft, C9-ft e C11-ft) que atingiram

Tabela 9 – Resultados da acurácia para classificação de IRDR.

	classificador					
Modelo	k-NN	LR	RF	SVM	Média	Desvio Padrão
C1-w2v	75%	72%	65%	71%	71%	±4,19
C2-w2v	73%	73%	65%	72%	71%	±3,86
C3-w2v	75%	71%	65%	71%	71%	±4,12
C4-w2v	73%	72%	65%	72%	71%	±3,69
C5-w2v	76%	72%	65%	71%	71%	±4,55
C6-w2v	72%	72%	65%	71%	70%	±3,37
C7-w2v	76%	72%	65%	72%	71%	±4,57
C8-w2v	73%	72%	65%	71%	70%	±3,59
C9-w2v	75%	71%	65%	72%	71%	±4,19
C10-w2v	73%	71%	65%	72%	70%	±3,59
C11-w2v	75%	71%	65%	72%	71%	±4,19
C12-w2v	74%	72%	65%	72%	71%	±3,94
Média	74%	72%	65%	72%	-	-
Desvio Padrão	±1,33	±0,62	±0	±0,51	-	-
C1-ft	**77%**	73%	65%	72%	72%	±4,99
C2-ft	73%	72%	65%	72%	71%	±3,69
C3-ft	76%	72%	66%	72%	72%	±4,12
C4-ft	73%	72%	66%	71%	71%	±3,10
C5-ft	76%	73%	65%	72%	72%	±4,65
C6-ft	73%	72%	65%	72%	71%	±3,9
C7-ft	**77%**	73%	66%	72%	72%	±4,55
C8-ft	73%	72%	65%	72%	71%	±3,69
C9-ft	**77%**	73%	66%	73%	72%	±4,57
C10-ft	73%	72%	65%	72%	71%	±3,69
C11-ft	**77%**	73%	68%	72%	73%	±3,69
C12-ft	73%	71%	66%	71%	70%	±2,98
Média	75%	72%	65%	72%	-	-
Desvio Padrão	±1,95	±0,65	±0,89	±0,51	-	-
C1-NILC-w2v	75%	73%	65%	72%	71%	±4,34
C2-NILC-w2v	75%	72%	65%	69%	70%	±4,27
C3-NILC-w2v	75%	73%	65%	65%	70%	±5,26
C4-NILC-w2v	76%	73%	65%	65%	70%	±5,62
Média	75%	73%	65%	68%	-	-
Desvio Padrão	±0,5	±0,5	±0	±3,40	-	-
C5-NILC-ft	75%	72%	65%	72%	71%	±4,24
C6-NILC-ft	73%	73%	66%	72%	71%	±3,37
C7-NILC-ft	75%	73%	67%	65%	70%	±4,76
C8-NILC-ft	74%	73%	65%	65%	69%	±4,92
Média	74%	73%	66%	69%	-	-
Desvio Padrão	±0,96	±0,5	±0,96	±4,04	-	-
C1-LegalNLP-w2v	73%	72%	65%	72%	71%	±3,69
C2-LegalNLP-w2v	75%	73%	65%	68%	70%	±4,57
Média	74%	73%	65%	70%	-	-
Desvio Padrão	±1,41	±0,71	±0	±2,83	-	-
C3-LegalNLP-ft	76%	73%	65%	72%	72%	±4,65
C4-LegalNLP-ft	74%	71%	65%	71%	70%	±3,77
Média	75%	72%	65%	72%	-	-
Desvio Padrão	±1,41	±1,41	±0	±0,71	-	-

77% de acurácia utilizaram vetores *FastText*, mostrando que as características do modelo, de ser tendencioso às informações sintáticas, não é fator limitante para treinamento com dados jurídicos, em desacordo com Chalkidis e Kampas (2019). Apesar disso, enfatiza-se

Tabela 10 – Resultados de F1-*marco* na classificação de IRDR.

	classificador					
Modelo	k-NN	LR	RF	SVM	Média	Desvio Padrão
C1-w2v	40%	30%	10%	25%	26%	$\pm 12,5$
C2-w2v	38%	29%	12%	25%	26%	$\pm 10,8$
C3-w2v	40%	30%	20%	25%	29%	$\pm 8,54$
C4-w2v	38%	29%	10%	25%	26%	$\pm 11,68$
C5-w2v	40%	29%	10%	25%	26%	$\pm 12,50$
C6-w2v	35%	29%	10%	25%	25%	$\pm 10,66$
C7-w2v	38%	29%	10%	26%	26%	$\pm 11,67$
C8-w2v	39%	29%	10%	25%	26%	$\pm 12,03$
C9-w2v	38%	29%	10%	26%	26%	$\pm 11,67$
C10-w2v	35%	29%	10%	25%	25%	$\pm 10,66$
C11-w2v	41%	29%	11%	26%	27%	$\pm 12,34$
C12-w2v	39%	29%	10%	25%	26%	$\pm 12,04$
Média	38%	29%	11%	25 %	-	-
Desvio Padrão	$\pm 1,88$	$\pm 0,39$	$\pm 2,87$	$\pm 0,45$	-	-
C1-ft	38%	30%	10%	26%	26%	$\pm 11,78$
C2-ft	38%	29%	10%	26%	26%	$\pm 11,67$
C3-ft	38%	30%	12%	26%	27%	$\pm 10,88$
C4-ft	35%	29%	11%	24%	25%	$\pm 10,21$
C5-ft	40%	30%	10%	26%	27%	$\pm 12,48$
C6-ft	38%	29%	10%	26%	26%	$\pm 11,67$
C7-ft	42%	30%	13%	27%	28%	$\pm 11,92$
C8-ft	35%	29%	10%	25%	25%	$\pm 10,66$
C9-ft	41%	30%	15%	27%	28%	$\pm 10,69$
C10-ft	35%	29%	10%	25%	25%	$\pm 10,66$
C11-ft	**43%**	31%	17%	25%	29%	$\pm 10,95$
C12-f	35%	29%	13%	25%	26%	$\pm 9,29$
Média	38%	29%	12%	26 %	-	-
Desvio Padrão	$\pm 2,86$	$\pm 0,67$	$\pm 2,34$	$\pm 0,89$	-	-
C1-NILC-w2v	37%	29%	10%	25%	25%	$\pm 11,32$
C2-NILC-w2v	40%	30%	10%	21%	25%	$\pm 12,79$
C3-NILC-w2v	40%	29%	10%	10%	22%	$\pm 14,84$
C4-NILC-w2v	41%	29%	13%	10%	23%	$\pm 14,48$
Média	40%	29%	11%	17%	-	-
Desvio Padrão	$\pm 1,73$	$\pm 0,5$	$\pm 1,5$	$\pm 7,68$	-	-
C5-NILC-ft	40%	29%	10%	25%	26%	$\pm 12,41$
C6-NILC-ft	37%	29%	10%	26%	26%	$\pm 11,32$
C7-NILC-ft	36%	29%	17%	10%	23%	$\pm 11,69$
C8-NILC-ft	38%	29%	10%	10%	22%	$\pm 14,06$
Média	38%	29%	12%	18 %	-	-
Desvio Padrão	$\pm 1,71$	± 0	$\pm 3,5$	$\pm 8,96$	-	-
C1-LegalNLP-w2v	35%	28%	10%	27%	25%	$\pm 10,61$
C2-LegalNLP-w2v	42%	29%	10%	20%	25%	$\pm 13,60$
Média	39%	29%	10%	24 %	-	-
Desvio Padrão	$\pm 4,95$	$\pm 0,71$	± 0	$\pm 4,95$	-	-
C3-LegalNLP-ft	42%	28%	10%	26%	27%	$\pm 13,10$
C4-LegalNLP-ft	35%	29%	10%	24%	25%	$\pm 10,66$
Média	39%	29%	10%	25 %	-	-
Desvio Padrão	$\pm 4,96$	$\pm 0,71$	± 0	$\pm 1,41$	-	-

que o modelo *FastText* tem custo computacional superior aos modelos *Word2Vec*, dado seu procedimento de treinamento partir de sub-representações, o que pode influenciar na escolha final.

Em uma análise sobre a influência das dimensões do vetor, os resultados mostraram que não há diferenças significantes entre os modelos treinados com 100 e 300 dimensões, diferindo dos resultados obtidos no estudo de Cunha, Almeida e Simões (2022).

5.1.2 RESULTADOS DOS MODELOS BASEADOS NO DO BERT

As Tabelas 11 e 12 mostram os resultados adquiridos nos experimentos com os modelos de representações contextuais, todos baseados na arquitetura do *BERT*. Na avaliação com acurácia, vide Tabela 11, os modelos BumbaBert FT (*base*) e BumbaBert SC (*small*) foram os que obtiveram os maiores valores, ambos alcançando o desempenho de 81% nos cenários com o algoritmos k-NN e LR (para o BumbaBert FT (*base*)). Considerando a média de desempenho para os quatro classificadores, o modelo BumbaBert FT (*base*) foi o que obteve melhor resultado com 76%.

Tabela 11 – Resultados da acurácia para classificação de IRDR.

	classificador					
Modelo	k-NN	LR	RF	SVM	Média	Desvio Padrão
Família BumbaBert						
BumbaBert FT (*base*)	**81%**	**81%**	**70%**	73%	76%	$\pm 5,61$
BumbaBert SC(*base*)	64%	65%	65%	65%	65%	± 0.50
BumbaBert SC*(small)*	**81%**	78%	**70%**	72%	75%	$\pm 5,12$
Modelos pré-treinados						
BERTimbau (*base*)	63%	65%	65%	65%	65%	$\pm 1,0$
BERTikal (*base*)	79%	77%	**70%**	**74%**	75%	$\pm 03,92$
Média	74%	73%	68%	70%	-	-
Desvio Padrão	$\pm 9,26$	$\pm 7,63$	$\pm 2,74$	$\pm 4,44$	-	-

No cenário F1-*macro*, disposto na Tabela 12, o modelo BumbaBert FT (*base*) também foi o que obteve maior desempenho geral na avaliação com o classificador k-NN, com 54%, e melhor desempenho médio com 33%.

Tabela 12 – Resultados de F1-*macro* para classificação de IRDR.

	classificador					
Modelo	k-NN	LR	RF	SVM	Média	Desvio Padrão
BumbaBert FT (*base*)	**54%**	**39%**	**15%**	24%	33%	$\pm 17,14$
BumbaBert SC (*base*)	10%	10%	10%	10%	10%	± 0
BumbaBert SC (*small*)	42%	37%	**15%**	22%	29%	$\pm 12,62$
Modelos pré-treinados						
BERTimbau (*base*)	15%	11%	10%	10%	12%	$\pm 02,38$
BERTikal (*base*)	41%	32%	**15%**	**26%**	29%	$\pm 10,90$
Média	32%	26%	13%	18 %	-	-
Desvio Padrão	$\pm 18,95$	$\pm 14,20$	$\pm 2,74$	$\pm 7,79$	-	-

Em relação ao desempenho dos classificadores, o algoritmo k-NN foi o que obteve as melhores para as duas métricas analisadas, com 74% e 32%. Reforça-se que não houve aplicação de qualquer estratégia de ajustes nos dados de testes com foco na otimização do desempenho dos algoritmos. A adoção de mecanismos para o balanceamento do conjunto

de dados, por exemplo, podem favorecer o aprendizado dos classificadores, no entanto, não era foco central deste estudo.

Em uma análise comparativa com os modelos pré-treinados, os resultados do modelos BumbaBert FT (*base*), BumbaBert SC (*small*) e do BERTikal mostram que o treinamento com dados do domínio podem otimizar o desempenho do classificador. Nesse caso, a diferença do modelo genérico BERTimbau para modelos os segmentados foram de até 11 pontos percentuais na avaliação via acurácia e 21 pontos em F1-*macro*.

Tais desempenhos corroboram os resultados reportados por Chalkidis et al. (2020), Douka et al. (2021), Licari e Comandè (2022). Ressalta-se o desempenho relativamente baixo do modelo BumbaBert SC (*base*), com 65 e 10% de acurácia e F1-*macro*, respectivamente. Tal desempenho pode ser justificado pela quantidade de etapas de treinamento adotadas, nesse caso, cerca de 400 mil. Na versão *small*, por exemplo, foram utilizadas cerca de 1 milhão de etapas. Em BumbaBert FT (*base*) também foram adotados cerca de 400 mil, no entanto, o treinado sobre o *checkpoint* de um modelo já pré-treinado.

Outro ponto de destaque é o desempenho do modelo BumbaBert SC (*small*). A versão com arquitetura reduzida demonstrou um desempenho superior ao modelo genérico BERTimbau e competitivo em comparação com o modelo pré-treinado BERTikal e a versão BumbaBert FT (base). Tal fenômeno já havia sido observado por Douka et al. (2021) no contexto do modelo jurídico francês (Juribert) e por Chalkidis et al. (2020) no caso do LegalBert (inglês), indicando a eficácia do aprendizado com arquiteturas mais compactas nesses contextos específicos.

Em comparação com os resultados dos modelos de *word embeddings*, observa-se um notável ganho de desempenho nos modelos derivados do *BERT*, o que era esperado, considerando a capacidade de aprendizado contextual inerente às arquiteturas *Transformers*. Isso contrasta com a abordagem livre de contexto adotada pelos modelos *FastText* e *Word2Vec*. Por exemplo, considerando os melhores cenários, a versão BumbaBert FT (*base*) apresenta resultados superiores para todos os classificadores analisados na avaliação via acurácia, conforme apresentado na Figura 5. Na análise com RL, a diferença chega 8 pontos percentuais.

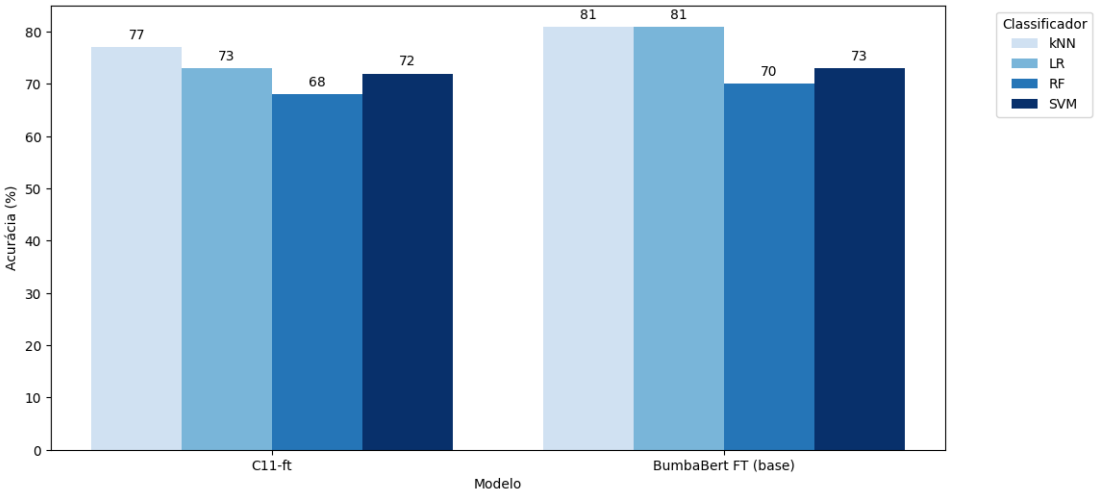

Figura 5 – Comparação dos melhores cenários por tipo de representação *embedding* considerando a acurácia.

Para a métrica F1-macro, mostrado na Figura 6, a diferença alcança 11 pontos percentuais no cenário com o algoritmo k-NN. Essa disparidade entre as duas estratégias de representação de *embeddings* já havia sido indicada por Polo et al. (2021), no entanto, as avaliações realizadas nesta pesquisa revelam uma diferença ainda mais representativa entre as abordagens.

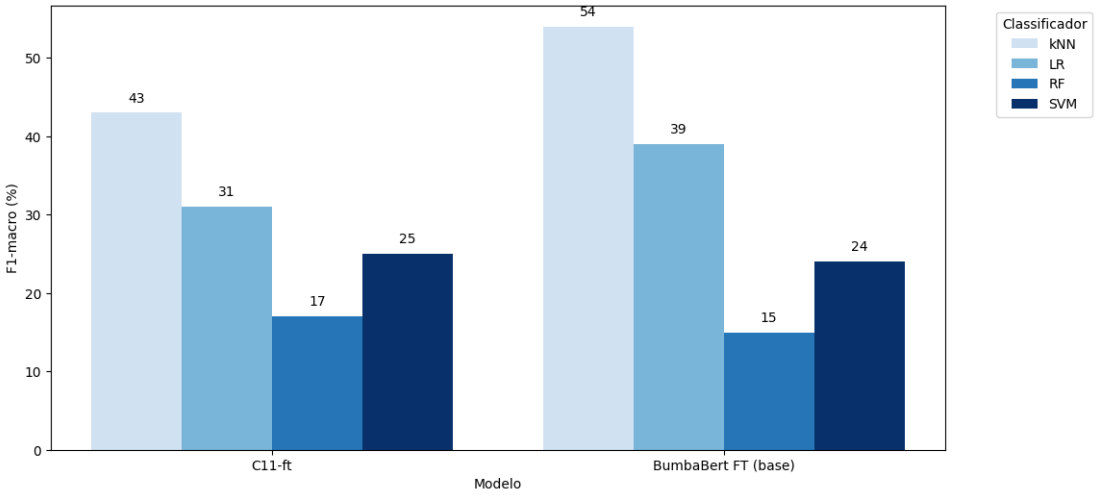

Figura 6 – Comparação dos melhores cenários por tipo de representação *embedding* considerando o F1-*macro*.

5.1.3 PROJEÇÃO DOS VETORES *EMBEDDINGS*

A aprendizagem dos vetores de *embeddings* pode ser visualizada em um espaço de baixa dimensão, revelando assim os agrupamentos contextuais desses *embeddings*. Com esse propósito, empregamos os modelos principais desenvolvidos nesta pesquisa para a projeção em espaço bidimensional.

Para os modelos de *word embeddings*, selecionamos o modelo que apresentou o melhor desempenho nos experimentos, denominado C11-ft. Adotamos dois conjuntos de dados distintos para esse processo: o primeiro abrange dados de petições iniciais, utilizado no processo de classificação nesta pesquisa; o segundo contém jurisprudências do STM, um *dataset* que faz parte do *corpus* jurídico de treinamento, vide Tabela 3. De cada conjunto, selecionamos aleatoriamente 200 documentos. O objetivo é utilizar dados gerais do domínio, no caso do conjunto de petições, e dados de um segmento específico do domínio, no caso, os dados militares. Esse procedimento facilita uma visualização mais clara do aprendizado dos vetores.

A partir da seleção dos conjuntos de dados, utilizou-se o TF-IDF para identificação de palavras mais relevantes em cada um. Em seguida, foram excluídas as palavras em comum, assegurando que cada *dataset* provesse, no mínimo, 50 termos distintos. Essas palavras estão listadas no Apêndice B deste documento e a Figura 7 apresenta a projeção resultante.

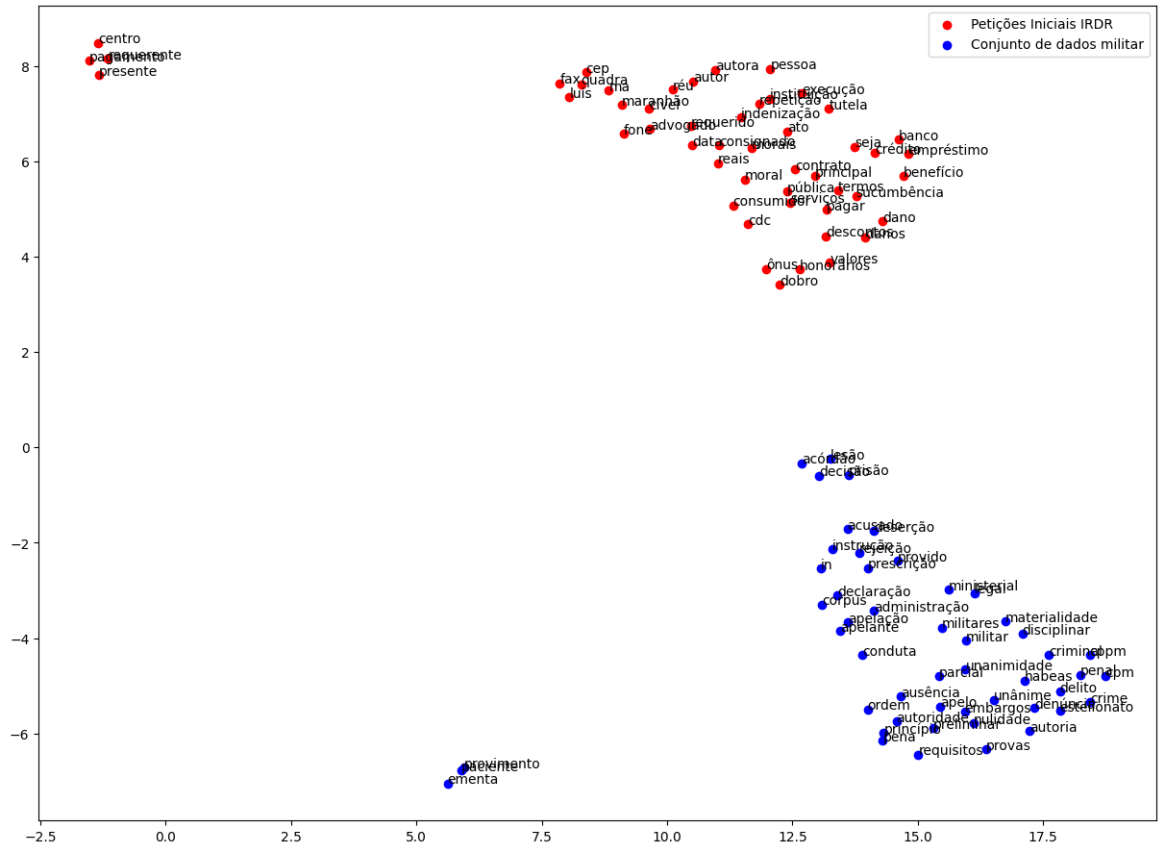

Figura 7 – Projeção dos embeddings utilizando o modelo C11-ft.

A Figura 7 evidencia o aprendizado contextual do modelo C11-ft. Percebe-se a clara divisão entre os dois conjuntos de dados, mostrando que o modelo tem capacidade para capturar aspectos de segmentação dentro do universo jurídico. Também percebe-se que palavras com contextos similares estão próximas no espaço vetorial, evidenciando o

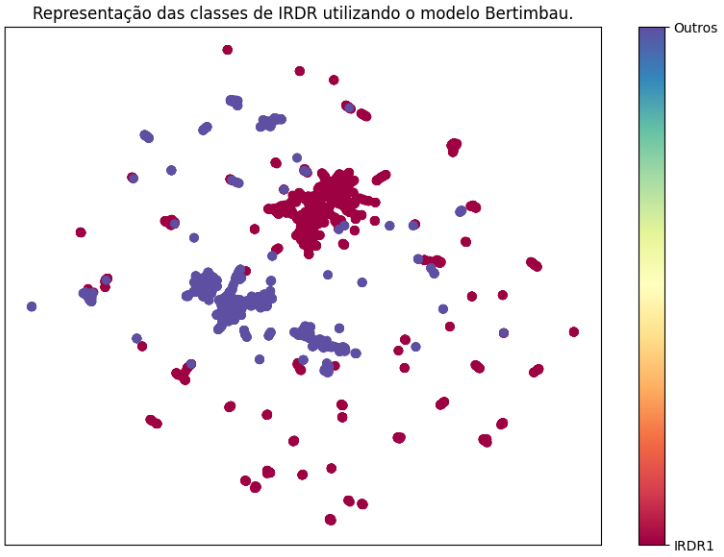

Figura 8 – Projeção dos embeddings utilizando BERTimbau.

aprendizado advindo das relações contextuais das palavras. A exemplos, temos os termos "militar" e "militares", no cenário militar, e "banco", "empréstimo" e "crédito", no cenário geral do domínio.

Para o cenário com modelos contextuais, selecionamos os principais modelos treinados para apresentar a visualização dos vetores de sentenças jurídicas. Para isso adotamos o conjunto de dados de IRDR, vide Tabela 8. Devido o processo de desbalanceamento de banco de dados, optamos por configurar um processo binário de visualização, considerando a classe majoritária IRDR1 como classe 0 e o agrupamento das demais como classe 1. As Figuras 8, 9 e 10 apresentam as projeções.

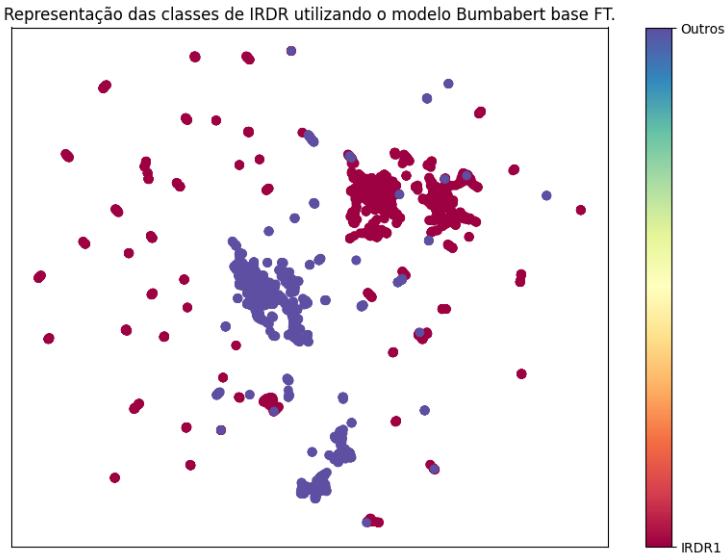

Figura 9 – Projeção dos embeddings utilizando BumbaBert *base* FT.

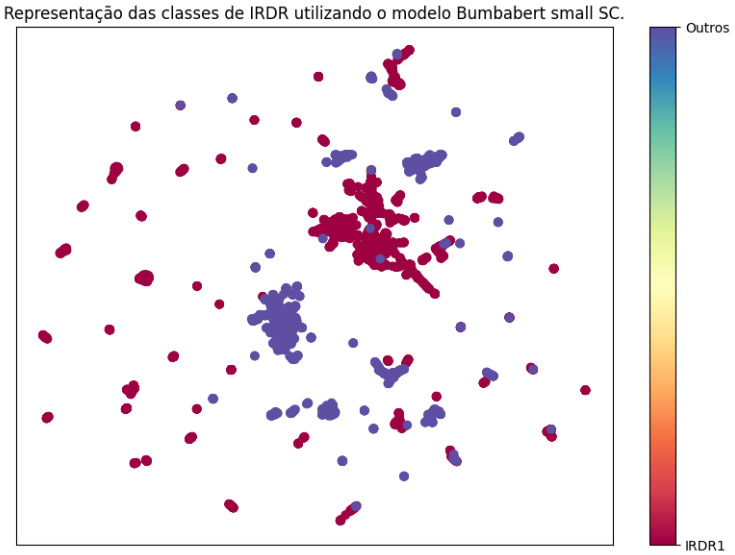

Figura 10 – Projeção dos embeddings utilizando BumbaBert *small* SC.

As Figuras 8, 9 e 10 apresentam os resultados das projeções dos vetores de *embeddings* gerados utilizando a similaridade do cosseno. Para ambos os modelos, evidenciam-se núcleos ou agrupamentos de vetores para cada uma das classes, contudo, não é possível mensurar qual modelo é superior. No entanto, esses núcleos destacam o aprendizado contextual dos textos jurídicos de entrada.

É importante ressaltar alguns fatores de impacto nos resultados das projeções:

i) Os diferentes tipos de IRDRs na classe "Outros", evidenciando uma diversidade contextual dentro da própria classe (subclasses);

ii) O processo de truncamento das sentenças de entrada devido às limitações dos modelos: 768 *tokens* para as versões *base* e 512 *tokens* na versão *small*. Esse processo resulta na consideração apenas da parte inicial do documento jurídico;

iii) Padronização dos documentos jurídicos utilizados, a parte introdutória da petição inicial é geralmente composta por *tokens*/vocábulos comuns a ambas as classes mapeadas (*e.g.,* "excelentíssimo", "senhor", "juiz").

Esses fatores contribuem para aumentar a complexidade do mapeamento contextual dos vetores de *embeddings*.

5.2 CONSIDERAÇÕES SOBRE OS RESULTADOS

Os resultados dos experimentos computacionais apresentados neste estudo mostram que os modelos de representações *embeddings* treinados com dados específicos do domínio jurídico incorporam de maneira mais acentuada as nuances da linguagem inerente ao segmento, em comparação com modelos treinados utilizando dados genéricos. Esta

constatação reforça os resultados de estudos anteriores, como destacado no capítulo 3. Ressalta-se os resultados obtidos pelos modelos contextuais, nos quais a diferença foi melhor evidenciada em ambas as métricas adotadas.

Ainda em relação aos modelos baseados no BERT, destaca-se a notável capacidade de compreensão dos modelos construídos, mesmo utilizando apenas a parte introdutória do documento da petição inicial devido à limitação da arquitetura dos modelos, resultando no truncamento dos textos longos das peças jurídicas, o que não ocorre nos modelos baseados em palavras. Salienta-se que essa limitação, somada à falta de balanceamento no conjunto de dados de teste, pode ter influenciado nos resultados obtidos pelos classificadores. Tais pontos são passíveis de aprimoramento e investigação futura na busca por uma compreensão mais completa do desempenho desses modelos no contexto jurídico.

6 CONSIDERAÇÕES FINAIS

A utilização de aplicações baseadas em PLN representa uma tendência crescente na esfera jurídica atualmente. No cenário brasileiro, já se observa a presença de soluções com resultados práticos e promissores em uma gama de tarefas, como a classificação e agrupamento de documentos e processos, NER, entre outras aplicações. Nesse ensejo, um dos desafios encontrados está relacionado às particularidades da linguagem utilizada, carregada de formalismos e termos técnicos inerentes à área jurídica. Tais peculiaridades adicionam complexidade ao aprendizado de algoritmos que utilizam representações treinados com dados generalistas em soluções para o setor. Dessa forma, buscar soluções orientadas a linguagem jurídica faz-se necessário.

Atuando nesse contexto, o presente trabalho desenvolveu um *framework* experimental envolvendo diferentes modelos de representações *embeddings*, direcionados especificamente para a área jurídica brasileira. Os modelos desenvolvidos foram aplicados no processo de classificação de documentos jurídicos visando uma compreensão mais prática possível do desempenho dos modelos.

Os resultados dos modelos no processo de classificação mostraram-se promissores ao serem comparados com os modelos pré-treinados na língua portuguesa de forma geral. Tais resultados indicam que os modelos segmentados são recomendados frente a modelos generalistas para aplicação em tarefas PLN no setor.

Os artefatos desenvolvidos nesta pesquisa, bem como os procedimentos para utilização dos mesmos estão disponíveis no repositório `https://github.com/ToadaLabTJMA/jus-embedding` visando reprodutividade e disseminação dos conhecimentos gerados.

6.1 CONTRIBUIÇÕES TÉCNICAS

Esta subseção destaca o impacto e a relevância deste trabalho no campo da inteligência artificial aplicada ao setor jurídico. Dada sua ampla cobertura experimental, as contribuições técnicas abrangem vários aspectos, cada um visando estrategicamente atuar em lacunas identificadas durante o processo de pesquisa.

Em primeiro lugar, enfatiza-se a criação de um *corpus* jurídico significativo para o treinamento de modelos de linguagem específicos para o segmento jurídico. Esse *corpus* foi utilizado nos experimentos computacionais desta pesquisa, e sua publicação pode fornecer insumos cruciais para o avanço da inteligência artificial no domínio jurídico.

Outra contribuição de destaque são os próprios modelos desenvolvidos. Esses modelos estão proporcionando contribuições tangíveis para um projeto mais amplo voltado

ao TJMA. Em outras palavras, esses artefatos não apenas oferecem soluções fundamentadas em IA para desafios específicos do setor, mas também representam um compromisso prático com a melhoria do sistema judiciário. Adicionalmente, a disponibilização desses modelos para a comunidade acadêmica visa fomentar a disseminação do conhecimento e estimular futuras pesquisas na área.

Outro ponto importante refere-se à cobertura experimental realizada. Os experimentos realizados fornecerão elementos essenciais para a reprodução dos treinamentos e a avaliação de modelos. Além disso, a aplicação prática desses modelos em tarefas finais envolvendo PLN oferece uma perspectiva valiosa sobre a eficácia e a adaptabilidade das soluções propostas.

Por fim, a publicação dos resultados reforça a contribuição técnica desta dissertação. A divulgação por meio de artigo e da dissertação em si proporciona uma visão abrangente dos experimentos realizados, destacando o processo completo de construção e aplicação dos modelos em tarefas finais de PLN, como no caso específico da classificação de dados. A publicação dos resultados no XI Workshop de Computação Aplicada em Governo Eletrônico, edição 2023, representa um exemplo concreto dessa contribuição, disponibilizando elementos valiosos para a comunidade acadêmica e profissional e para o avanço contínuo do conhecimento na área.

6.2 AMEAÇAS À VALIDADE DO ESTUDO

Considerando a diversidade de documentos e áreas da justiça, o *corpus* utilizado pode não refletir a contento a realidade brasileira. Apesar do melhor esforço na coleta e indexação dos dados, é possível que o *corpus* tenha induzido modelos enviesados.

Outro aspecto pertinente é quanto a avaliação, o presente estudo adotou apenas a classificação de precedentes. Neste ensejo, apenas classificadores-padrão da literatura foram utilizados (SVM, RF, k-NN e LR), cabe incluir outros modelos e também testar diferentes parametrizações nos modelos. Uma ampliação da cobertura experimental se faz necessária para englobar outras tarefas como NER, POS *tagging* e similaridade semântica para se obter conclusões mais robustas.

6.3 TRABALHOS FUTUROS

Como pontos em abertos desta pesquisa, destaca-se a necessidade de ampliar ainda mais a cobertura experimental. Isso implica a incorporação de *(i)* outras técnicas de representação, como o *GloVe*, bem como a exploração de outros modelos de linguagem mais robustos, como: GPT (BROWN et al., 2020), LLaMA (TOUVRON et al., 2023), entre outros. Além disso, *(ii)* é crucial considerar novas estratégias de avaliação, indo além da classificação de dados adotada nos experimentos iniciais. A diversificação dos

métodos de avaliação pode proporcionará uma compreensão mais abrangente e robusta do desempenho dos modelos desenvolvidos.

Outro aspecto relevante a ser abordado é *(iii)* a ampliação do número de amostras de dados jurídicos no treinamento. A inclusão de novas amostras de dados no *corpus* poderá contribuir para a generalização e eficácia dos modelos, permitindo uma representação mais abrangente e precisa da linguagem jurídica. Essa expansão no volume de dados de treinamento também pode impactar positivamente a capacidade dos modelos de lidar com uma variedade maior de contextos jurídicos, enriquecendo, assim, a aplicação prática dessas soluções.

Portanto, a investigação desses pontos em aberto pode fornecer novas possibilidades ou alternativas aos modelos desenvolvidos nesta pesquisa. Essas iniciativas contribuirão significativamente para o avanço contínuo no desenvolvimento de modelos de PLN aplicados ao contexto jurídico.

REFERÊNCIAS

ARAUJO, P. H. Luz de et al. VICTOR: a dataset for Brazilian legal documents classification. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 1449–1458. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.181>. Citado 3 vezes nas páginas 22, 39 e 40.

ARAÚJO, V. S. de; GABRIEL, A. de P.; PORTO, F. R. Justiça 4.0: a transformação tecnológica do poder judiciário deflagrada pelo cnj no biênio 2020-2022. *Revista Eletrônica Direito Exponencial-DIEX*, v. 1, n. 1, p. 1–18, 2022. Citado na página 19.

BAMBROO, P.; AWASTHI, A. Legaldb: Long distilbert for legal document classification. In: *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. [S.l.: s.n.], 2021. p. 1–4. Citado na página 20.

BATISTA, H. et al. A comparative analysis of text embedding approach to extract named entities in portuguese legal documents. In: *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2021. p. 221–232. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view-/18255>. Citado 3 vezes nas páginas 20, 34 e 46.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, MIT Press, v. 5, p. 135–146, 2017. Citado 4 vezes nas páginas 20, 27, 29 e 39.

BROWN, T. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020. Citado 2 vezes nas páginas 30 e 64.

CARMO, F. A. do et al. Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. In: *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*. Porto Alegre, RS, Brasil: SBC, 2023. p. 188–199. ISSN 2763-8723. Disponível em: <https://sol.sbc.org.br/index.php/wcge/article/view/24876>. Citado na página 40.

CHALKIDIS, I. et al. LEGAL-BERT: The muppets straight out of law school. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, 2020. p. 2898–2904. Disponível em: <https://aclanthology.org/2020.findings-emnlp.261>. Citado 5 vezes nas páginas 33, 36, 43, 45 e 56.

CHALKIDIS, I.; KAMPAS, D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, v. 27, n. 2, p. 171–198, Jun 2019. ISSN 1572-8382. Disponível em: <https://doi.org/10.1007/s10506-018-9238-9>. Citado 4 vezes nas páginas 20, 32, 36 e 53.

CONSOLI, B. et al. Embeddings for named entity recognition in geoscience Portuguese literature. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 4625–4630. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.568>. Citado 3 vezes nas páginas 22, 31 e 35.

CUNHA, L. F.; ALMEIDA, J. J.; SIMÕES, A. Reasoning with portuguese word embeddings. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FÜR INFORMATIK. *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. [S.l.], 2022. Citado 4 vezes nas páginas 31, 35, 43 e 55.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: . [S.l.: s.n.], 2019. v. 1. Citado 5 vezes nas páginas 20, 29, 30, 33 e 39.

DOUKA, S. et al. Juribert: A masked-language model adaptation for french legal text. In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. [S.l.: s.n.], 2021. p. 95–101. Citado 5 vezes nas páginas 33, 43, 45, 46 e 56.

GARCIA, A. C. Ética e inteligencia artificial. *Computação Brasil*, n. 43, p. 14–22, 2020. Citado na página 19.

GOMES, D. d. S. M. et al. Portuguese word embeddings for the oil and gas industry: development and evaluation. *Computers in Industry*, Elsevier, v. 124, p. 103347, 2021. Citado 3 vezes nas páginas 22, 32 e 35.

HARIRI, R. H.; FREDERICKS, E. M.; BOWERS, K. M. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, v. 6, n. 1, p. 44, Jun 2019. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-019-0206-3>. Citado na página 19.

HARTMANN, N. S. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2017. p. 122–131. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/4008>. Citado 9 vezes nas páginas 31, 32, 35, 39, 42, 43, 46, 51 e 52.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. *Science*, v. 349, n. 6245, p. 261–266, 2015. Disponível em: <https://www.science.org/doi/abs/10-.1126/science.aaa8685>. Citado na página 19.

LE-KHAC, P. H.; HEALY, G.; SMEATON, A. F. Contrastive representation learning: A framework and review. *IEEE Access*, v. 8, p. 193907–193934, 2020. Citado na página 20.

LICARI, D.; COMANDÈ, G. Italian-legal-bert: A pre-trained transformer language model for italian law. 2022. Citado 3 vezes nas páginas 33, 36 e 56.

Luz de Araujo, P. H. et al. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In: *International Conference on the Computational Processing of Portuguese (PROPOR)*. Canela, RS, Brazil: Springer, 2018. (Lecture Notes on Computer Science (LNCS)), p. 313–323. Disponível em: <https://teodecampos.github.io/LeNER-Br/>. Citado 2 vezes nas páginas 22 e 39.

MACHADO, F. d. V.; COLOMBO, C. Inteligência artificial aplicada à atividade jurisdicional: desafios e perspectivas para sua implementação no judiciário. *Revista da Escola Judicial do TRT4*, v. 3, n. 5, p. 117–141, maio 2021. Disponível em: <https://rejtrt4.emnuvens.com.br/revistaejud4/article/view/113>. Citado na página 19.

MARINATO, M. et al. Classificação automática de petições iniciais usando classificadores combinados. In: *Anais do XVI Brazilian e-Science Workshop*. Porto Alegre, RS, Brasil: SBC, 2022. p. 89–96. ISSN 2763-8774. Disponível em: <https://sol.sbc.org.br/index.php-/bresci/article/view/20479>. Citado na página 40.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: . [S.l.: s.n.], 2013. Citado 5 vezes nas páginas 20, 27, 28, 29 e 39.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Red Hook, NY, USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119. Citado 3 vezes nas páginas 27, 28 e 29.

MOTA, C. et al. Classificação de páginas de petições iniciais utilizando redes neurais convolucionais multimodais. In: *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2020. p. 318–329. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/12139>. Citado na página 20.

PATIL, R. et al. A survey of text representation and embedding techniques in nlp. *IEEE Access*, v. 11, p. 36120–36146, 2023. Citado na página 27.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://aclanthology.org/D14-1162>. Citado na página 30.

PEREIRA, J. C. M.; RODRIGUES, M. V. J. A plataforma sinapses e a continuidade dos modelos de ia no judiciário. In: *ANAIS do Encontro de Administração da Justiça - ENAJUS 2021*. Lisboa: [s.n.], 2021. ISSN 2674-8401. Citado na página 19.

PINTO, H. A. A utilização da inteligência artificial no processo de tomada de decisões: por uma necessária accountability. *Revista de Informação Legislativa: RIL*, v. 57, n. 225, p. 43–60, 2020. Disponível em: <https://www12.senado.leg.br/ril/edicoes/57/225-/ril_v57_n225_p43>. Citado na página 19.

POLO, F. et al. Legalnlp - natural language processing methods for the brazilian legal language. In: *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2021. p. 763–774. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/18301>. Citado 11 vezes nas páginas 20, 22, 29, 32, 36, 39, 42, 43, 51, 52 e 57.

PONT, T. R. D. et al. Impact of text specificity and size on word embeddings performance: An empirical evaluation in brazilian legal domain. In: *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2020. p. 521–535. ISBN 978-3-030-61376-1. Disponível em: <https://doi.org/10.1007/978-3-030-61377-8_36>. Citado na página 34.

QADER, W. A.; AMEEN, M. M.; AHMED, B. I. An overview of bag of words;importance, implementation, applications, and challenges. In: *2019 International Engineering Conference (IEC)*. [S.l.: s.n.], 2019. p. 200–204. Citado na página 28.

RODRIGUES, R. B. M. et al. Petrobert: A domain adaptation language model for oil and gas applications in portuguese. In: *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2022. p. 101–109. ISBN 978-3-030-98304-8. Disponível em: <https://doi.org/10.1007/978-3-030-98305-5_10>. Citado 2 vezes nas páginas 32 e 35.

SCHNEIDER, E. T. R. et al. Biobertpt-a portuguese neural language model for clinical named entity recognition. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. [S.l.: s.n.], 2020. p. 65–72. Citado na página 22.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, Elsevier, v. 181, p. 526–534, 2021. Citado na página 38.

SILVA, M. da et al. Preprocessing applied to legal text mining: analysis and evaluation of the main techniques used. In: *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2023. p. 1010–1021. ISSN 2763-9061. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/25760>. Citado 2 vezes nas páginas 41 e 42.

SILVA, R. A. F.; FILHO, A. I. d. S. Inteligência artificial em tribunais brasileiros: retórica ou realidade? *Encontro de Administração da Justiça: anais do ENAJUS. Curitiba: IBEPES*, 2020. Citado na página 21.

SILVEIRA, R. et al. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In: NALDI, M. C.; BIANCHI, R. A. C. (Ed.). *Intelligent Systems*. Cham: Springer Nature Switzerland, 2023. p. 268–282. ISBN 978-3-031-45392-2. Citado 3 vezes nas páginas 22, 36 e 43.

SMYWIńSKI-POHL, A. et al. Automatic construction of a polish legal dictionary with mappings to extra-legal terms established via word embeddings. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. New York, NY, USA: Association for Computing Machinery, 2019. (ICAIL '19), p. 234–238. ISBN 9781450367547. Disponível em: <https://doi.org/10.1145/3322640.3326727>. Citado 2 vezes nas páginas 32 e 46.

SOUSA, A. W.; FABRO, M. D. D. Iudicium textum dataset uma base de textos jurıdicos para nlp. In: *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD*. [S.l.: s.n.], 2019. Citado 4 vezes nas páginas 19, 22, 39 e 40.

SOUZA, C. M. F. de; SALLES, S. de S. Acesso à justiça em tempos de pandemia: a experiência do núcleo permanente de métodos consensuais de tratamento de conflitos do tjrj. *Conhecimento & Diversidade*, v. 14, n. 33, p. 166–185, 2022. Citado na página 19.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian conference on intelligent systems*. [S.l.], 2020. p. 403–417. Citado 3 vezes nas páginas 31, 35 e 51.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. Citado na página 43.

SPEER, R. *ftfy*. 2019. Zenodo. Version 5.5. Disponível em: <https://doi.org/10.5281-/zenodo.2591652>. Citado na página 42.

TOUVRON, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. Citado na página 64.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado na página 29.

VIANNA, D.; MOURA, E. S. de; SILVA, A. S. da. A topic discovery approach for unsupervised organization of legal document collections. *Artificial Intelligence and Law*, Springer, p. 1–30, 2023. Citado na página 20.

VIANNA, D.; MOURA, E. Silva de. Organizing portuguese legal documents through topic discovery. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. [S.l.: s.n.], 2022. p. 3388–3392. Citado na página 20.

WANG, Z. et al. Named entity recognition method of brazilian legal text based on pre-training model. *Journal of Physics: Conference Series*, IOP Publishing, v. 1550, n. 3, p. 032149, may 2020. Disponível em: <https://dx.doi.org/10.1088/1742-6596/1550/3-/032149>. Citado 3 vezes nas páginas 20, 34 e 46.

WIRTH, R. CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, n. 24959, p. 29–39, 2000. Citado 5 vezes nas páginas 37, 38, 41, 42 e 48.

ZHONG, H. et al. How does NLP benefit legal system: A summary of legal artificial intelligence. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 5218–5230. Disponível em: <https://aclanthology.org/2020.acl-main.466>. Citado na página 20.

APÊNDICE A – ARTIGO PUBLICADO NO *WCGE* 2023

Embeddings Jurídico: Representações Orientadas à Linguagem Jurídica Brasileira

Fabrício A. do Carmo[1], **Ferdinando Serejo**[2], **Antonio F. L. Jacob Junior**[1], **Ewaldo E. C. Santana**[1], **Fábio M. F. Lobato**[1,3]

[1] Programa de Pós-Graduação em Engenharia de Computação e Sistemas (PECS)
Universidade Estadual do Maranhão - (UEMA)

[2]Tribunal de Justiça do Estado do Maranhão (TJMA)

[3]Instituto de Engenharia e Geociências
Universidade Federal do Oeste do Pará (UFOPA)

fabrycio30@gmail.com, fabio.lobato@ufopa.edu.br

Resumo. *O processamento automático de textos jurídicos dispostos em linguagem natural proporciona o desenvolvimento de diversas aplicações para o setor, como a classificação de processos por assunto, sumarização de documentos, tradução para linguagem cidadã etc. Nesse sentido, o judiciário brasileiro lançou o programa Justiça 4.0, buscando soluções que ofereçam celeridade nas atividades processuais. Convém pontuar que a linguagem técnica predomina nesse domínio de aplicação, o que adiciona desafios para modelagem dos dados, exigindo modelos especializados para o segmento. Frente ao exposto, esse trabalho tem como objetivo a construção de modelos embeddings orientados ao âmbito jurídico visando alimentar aplicações na área. Para isso, foram extraídos aproximadamente 500.000 documentos de instituições de justiça do Brasil das mais variadas esferas (civil, criminal, trabalhista etc). Os modelos foram avaliados por meio da classificação de petições iniciais e os resultados mostraram-se competitivos quando comparados a modelos generalistas da língua portuguesa. Tais resultados mostram que modelos treinados com documentos jurídicos compreendem melhor as especificidades da linguagem do segmento e têm o potencial de fomentar novas aplicações para o setor.*

1. Introdução

O uso de aplicações baseadas em Inteligência Artificial (IA) e *Big Data* vem apoiando a tomada de decisões em diversos segmentos da sociedade [Schaulet and Trez 2021, Garcia 2020, Hariri et al. 2019]. No âmbito jurídico, essas soluções podem guiar os profissionais tanto nas atividades administrativas quanto nos trâmites processuais, atuando principalmente sobre o grande volume de dados gerados no dia-a-dia da prestação jurisdicional [Pinto 2020, Parreiras et al. 2022]. No Brasil, onde o sistema judiciário conta com cerca de 77,3 milhões de processos em tramitação, segundo o último relatório "justiça em números"[1] do Conselho Nacional de Justiça (CNJ), já se entende que a celeridade processual passa necessariamente pela adoção de aplicações que utilizam recursos da IA.

[1]https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/

O programa Justiça 4.0[2] é uma das iniciativas do CNJ que incentiva o desenvolvimento de soluções com tecnologias de IA, promovendo aquelas que visem a automatização e otimização dos serviços nos tribunais. Um exemplo dessas iniciativas é a plataforma SINAPSES[3], responsável tanto pelo estabelecimento de parâmetros (legais e técnicos) para o desenvolvimento e implantação de modelos de IA nos tribunais, como para o armazenamento e distribuição dos mesmos [Pereira and Rodrigues 2021].

Dentre as aplicações que utilizam os recursos da IA para análises de dados jurídicos, destacam-se os que consomem os recursos do Processamento de Linguagem Natural, conhecido por *Natural Language Processing* (NLP). NLP é uma área multidisciplinar que envolve a computação e a linguística e contempla estudos e desenvolvimento de métodos e procedimentos que permitam a compreensão e o processamento automático de textos dispostos na forma natural [Sousa and Del Fabro 2019, Hirschberg and Manning 2015]. Na seara jurídica, aplicações como classificação de documentos e processos [Polo et al. 2021, Bambroo and Awasthi 2021], do Reconhecimento de Entidades Nomeadas, ou *Named Entity Recognition* (NER) [Wang et al. 2020, Batista et al. 2021], já oferecem resultados práticos.

Um dos principais desafios no trato com documentos jurídicos está na compreensão das especificidades da linguagem utilizada pelo setor, composta por jargões e termos técnicos [Polo et al. 2021]. O tamanho dos textos jurídicos também é um aspecto significativo, geralmente composto de textos longos e dotado de formalismo [Bambroo and Awasthi 2021]. A exemplos, esses documentos são formados por decisões de julgamentos, contratos, normas legais, petições inicias etc [Zhong et al. 2020, Mota et al. 2020]. Dessa forma, treinar modelos de IA eficientes para esse domínio perpassa pela forma como os dados são tratados e estão representados.

As representações de dados textuais visam modelagens vetoriais representativas de um determinado texto. Elas são elementos fundamentais em NLP, dado a sua capacidade de transformar os documentos de entrada em vetores numéricos preservando informações originais e fornecendo entradas interpretáveis por modelos de *Machine Learning* (ML) e *Deep Learning* (DL) [Le-Khac et al. 2020]. Nesse sentido, os modelos *embeddings* são amplamente utilizados dado sua capacidade de adicionar informações semânticas no processo representacional [Chalkidis and Kampas 2019, Mikolov et al. 2013a]. Dentre os exemplos de modelos para representações para palavras e textos, destacam-se o *Word2vec* [Mikolov et al. 2013a], o *FastText* [Bojanowski et al. 2017] e, mais recentemente, o BERT [Devlin et al. 2019].

Os modelos baseados em *word embeddings* utilizam redes neurais rasas para o aprendizado dos vetores de representação. Tal procedimento permite o mapeamento de palavras e suas respectivas relações no *corpus* de treinamento, produzindo vetores que capturam informações semânticas e sintáticas advindas dessas relações contextuais [Mikolov et al. 2013a]. Dessa forma, por exemplo, as palavras "juiz" e "magistrado" estariam próximas em um espaço vetorial devido ao seu grau de similaridade.

Frente ao exposto, esta pesquisa busca fomentar o desenvolvimento de soluções baseadas no Processamento de Linguagem Natural na área jurídica brasileira, por meio

[2]https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/

[3]https://www.cnj.jus.br/sistemas/plataforma-sinapses/

da construção de um *framework* experimental com diferentes modelos de representações *embeddings*, treinados a partir de documentos do setor. Os modelos treinados foram aplicados na classificação de petições inicias para avaliação de desempenho.

Destaca-se que a pesquisa reportada neste artigo faz parte do acordo de cooperação técnica entre a Universidade Estadual do Maranhão (UEMA) e o Tribunal de Justiça do Estado do Maranhão (TJMA), dessa forma, os modelos produzidos serão utilizados como insumos para o desenvolvimento de aplicações PLN no tribunal, gerando, dessa forma, impacto direto no sistema de justiça.

O restante do artigo encontra-se organizado como segue. Trabalhos relacionados são discutidos na Seção 2. A Seção 3 descreve o trabalho proposto. Na Seção 4 são apresentados os resultados e discussões. Por fim, na Seção 5 são apresentadas as conclusões e sugestões de trabalhos futuros.

2. Trabalhos relacionados

À luz da literatura do PLN, esta seção enfatiza trabalhos que apresentam técnicas e soluções com foco em representações *embeddings*, destacando aqueles direcionados a linguagem portuguesa e ao domínio jurídico. Inicialmente, são apresentadas as principais técnicas utilizadas. Na sequência, são mostrados os trabalhos com *embeddings* para a língua portuguesa, tanto para propósito geral e para a análise em domínios específicos. Por fim, são apresentados trabalhos direcionados ao âmbito jurídico.

2.1. Embeddings orientados a língua portuguesa

Um dos principais trabalhos com *embeddings* para língua portuguesa é o de [Hartmann et al. 2017]. Nele, os autores treinaram e disponibilizaram modelos FastText, GloVe, Wang2Vec e Word2Vec com diferentes dimensões, utilizando dados da língua portuguesa europeia e brasileira. Esses modelos foram analisados de forma intrínseca, por meio de análises sintáticas e semânticas, e extrínseca, utilizando-os em *Part-of-Speech* (PoS) *tagging* e análise de similaridades semânticas. Segundo os mesmos, a utilização de tarefas finais de NLP é preferível na avaliação dos modelos em detrimento a análises intrínsecas.

Em [Cunha et al. 2022] também é realizado treinamentos de modelos *embeddings* utilizando *corpus* da língua portuguesa, observando o impacto da parametrização dos modelos (e.g., dimensão do vetor), o tamanho do *corpus* de treinamento e do domínio. Também foram analisadas medidas para avaliação dos modelos treinados. Seus experimentos mostraram que as configurações paramétricas dos modelos têm influências significativas nos resultados. Também afirmam que avaliação por meio de analogias de palavras, utilizando *corpus* de teste, não é recomendada para modelos treinados em domínios mais segmentados, alinhando-se com as recomendações de [Hartmann et al. 2017].

Em uma análise mais segmentada, [Consoli et al. 2020] realizaram experimentos com *embeddings* treinadas com dados da área do petróleo e gás e fizeram comparações e combinações com modelos já treinados na língua portuguesa de modo geral, utilizando modelos disponibilizados por [Hartmann et al. 2017]. Os autores argumentaram que o setor contempla termos técnicos exclusivos, exigindo modelos mais orientados. Avaliações por meio do do NER, mostraram que combinações (*Stacking embeddings*) entre modelos

gerais e do domínio específico podem aumentar o desempenho da tarefa final, nesse caso alcançando *F1-score* de 84,63%.

Ainda no domínio de petróleo e gás, [Gomes et al. 2021] reforçam que a linguagem do setor possui características próprias e que palavras do português podem assumir significados completamente diferentes do comum, dificultando o aprendizado de algoritmos que consomem representações mais generalistas. Nesse trabalho, foram treinados modelos Word2vec e FastText em um *corpus* do domínio (contando com mais de 85 milhões de *tokens*). Os modelos foram submetidos a análises intrínsecas, analisando a relação entre pares de palavras, e extrínsecas, observando o desempenho no NER da área da Geociência. Os resultados mostraram que os modelos segmentados obtiveram melhores resultados quando comparados com os generalistas.

2.2. *Embeddings* Orientado ao segmento jurídico

Tal como no setor de petróleo e gás, a área jurídica compreende uma linguagem com características próprias na qual, por vezes, determinadas palavras possuem significados totalmente diferente da linguagem dita natural. Em [Smywiński-Pohl et al. 2019], são treinados modelos Word2vec e Glove e visando a criação de dicionário que forneça uma interface entre palavras técnicas da justiça polonesa e palavras que possam ser compreendidas por leigos. Os experimentos apontaram resultados superiores para o Word2vec do tipo CBOW. Também ressaltando essa peculiaridade no meio jurídico, [Polo et al. 2021] treinaram e disponibilizaram modelos de representações de palavras (Phraser, Word2Vec, Doc2Vec, FastText, e BERT), utilizando dados públicos da justiça brasileira. Realizaram experimentos com classificação de *status* (arquivado, ativo ou suspenso) de processos judiciais como demonstração de uso dos modelos treinados.

Em [Chalkidis and Kampas 2019], foram treinados e disponibilizados modelos *embeddings* a partir de um grande *corpus* de dados jurídicos disponíveis na língua inglesa. O *corpus* é formado por 123.066 peças jurídicas com aproximadamente 492.000.000 de *tokens* e envolve legislações do Reino Unido, União europeia, Canadá, Austrália, decisões da Suprema Corte Americana, além de documentos com legislações japonesas e da União europeia traduzidas para o inglês. As representações treinadas, nomeadas de law2vec[4], utilizaram o modelo word2vec com a arquitetura *Skip-gram*. Os autores afirmaram não adotar o *FastText* por ser tendencioso a informações sintáticas e dada a formalidade dos textos utilizados, com pouca incidência de erros, não haveria necessidade de adoção de um algoritmo que visa identificar/tratar palavras fora do vocabulário, buscando também contornar possíveis erros ortográficos.

No trabalho de [Wang et al. 2020], são utilizados modelos BERT como base para construção de arquiteturas híbridas para rotulagem de sequência (*sequence labelling*) orientadas à extração de entidades nomeadas. Foram utilizados dados jurídicos brasileiros no processo de avaliação dos modelos desenvolvidos. Ainda na seara da NER, [Batista et al. 2021] avaliaram o impacto de representações *embeddings* no processo de extração de entidades em petições inicias da justiça brasileira. Os resultados mostraram que a configuração com a *stacking* dos modelos *embeddings* de caracteres, de palavras e *pooled Flair* obteve melhores resultados.

[4]https://archive.org/details/Law2Vec

Também observando vetores *embedings* resultantes, [Dal Pont et al. 2020] avaliaram o impacto da especificidade e do tamanho do *corpus* de texto utilizado no treinamento dos vetores. Aplicados a dados jurídicos brasileiros em vários níveis de segmentação, os resultados mostraram que *corpus* menores capturam melhor as especificidades textos. Ou seja, para um ramo específico da justiça, nesse caso processos relacionados ao transporte aéreo), representações treinadas em *corpus* menores são preferíveis. Entretanto, de modo geral, quanto maior o *corpus* de treinamento, melhores são os resultados obtidos.

Os trabalhados supracitados destacam técnicas de representações textuais e aplicações envolvendo PLN que consomem tais recursos como entrada, também detalham medidas avaliativas para as soluções propostas. Os trabalhos segmentados realçam a importância de representações orientadas ao domínio do problema para obtenção de melhores resultados. Nesse sentido, o trabalho proposto foca na construção e treinamentos de modelos orientados ao âmbito jurídico que possa discriminar com maior eficácia as especificidades da linguagem desse domínio de aplicações.

3. Trabalho Proposto

O presente estudo teve como objetivo a construção e a divulgação de modelos *embeddings* orientados ao segmento jurídico brasileiro, com o intuito de fomentar aplicações NLP no setor. Tal como enfatizado em [Polo et al. 2021], a área jurídica compreende uma linguagem peculiar, requerendo representações que discriminem com maior eficiência o comportamento dos documentos jurídicos. Outro ponto que incentiva o desenvolvimento deste estudo é a falta de um volume significativo de dados disponíveis para o treinamento de modelos *embeddings* nesse domínio de aplicação. Destarte, este trabalho visa preencher essa lacuna, treinando modelos a partir de grande volume de dados públicos e privados da justiça brasileira. As subseções seguintes descrevem os dados, os procedimentos e as técnicas utilizadas para a construção e avaliação dos modelos adotados.

3.1. Dados Jurídicos

Para a criação do *corpus* jurídico de treinamento, tal como apresentado na Tabela 1, foram obtidos dados públicos contendo texto de acórdãos do Tribunal Superior do Trabalho (TST)[5]; do Supremo Tribunal Federal (STF), por meio do *Iudicium Textum Dataset* (ITD) disponibilizado por [Sousa and Del Fabro 2019]; do Superior Tribunal Militar (STM)[6]; do Tribunal Superior Eleitoral (TSE)[7]; do Tribunal de Contas da União (TCU)[8]; de processos recorrentes disponibilizados pelo Conselho Nacional de Justiça (CNJ) por meio do Banco Nacional de Dados de Demandas Repetitivas e Precedentes Obrigatórios (BNPR)[9]; e por meio de dados (e.g., decretos, acórdãos e súmulas) disponíveis na plataforma LexML[10], especializada na divulgação de informações jurídicas e legislativas; foram utilizados também dados internos, com textos de petições iniciais, do TJMA. Com exceção do ITD e do TJMA, os dados foram coletados diretamente dos portais, via *web crawlers*. A Tabela 1 apresenta características de cada *dataset*.

[5]https://www.tst.jus.br/jurisprudencia

[6]https://www.stm.jus.br/gestao-da-informacao/pagina-inicial-gest-inform/jurisprudencia

[7]http://www.tse.jus.br/jurisprudencia

[8]https://pesquisa.apps.tcu.gov.br/

[9]https://bnpr.cnj.jus.br/

[10]https://www.lexml.gov.br/

Tabela 1. Características dos conjuntos de dados jurídicos utlizados.

dataset	tipo do documento	amostras	*tokens*	*tokens* únicos
STF-ITD	Acórdãos STF	41.353	12.134.882	147.172
BNRP	Demandas recorrentes	3.255	179.439	15.126
TSE	Jurisprudência	84.754	8.284.523	109.070
STM	Jurisprudência	23.522	1.968.908	35.498
TCU	Jurisprudência	15.000	584.759	15.746
TST	Jurisprudência	247.084	488.794.965	856.537
STJ	Jurisprudência	1.772	122.338	8.743
LexML	Jurisprudência STF	72.280	2.502.145	54.811
TJMA	petições inicias	11.700	24.200.179	604.995
TOTAL	-	500.720	538.772.138	-

3.2. Modelos embeddings

1. **Word2vec**: É um algoritmo para geração de vetores de palavras, proposto por [Mikolov et al. 2013a], amplamente utilizado em PLN. O *Word2vec* tem a capacidade treinar vetores que consideram as relações entre as palavras, superando modelos mais simplificados como o *Bag-of-words* (BoW), onde a representação é focada em computar co-ocorrências dos termos [Qader et al. 2019]. A redução da dimensionalidade e o tratamento da esparsividade do vetor também são vantagens do modelo *word embeddings* sobre o BoW.
   O *Word2vec* utiliza duas estratégias de treinamento com redes neurais rasas: o *Continuous bag of words* (CBOW) e o modelo Skip-gram. No CBOW, o algoritmo tenta prever a palavra central (alvo), com base no contexto em que ela está inserida. Já no modelo Skip-gram a predição é feita de maneira oposta, as palavras do contexto são previstas com base na palavra central. As predições são realizadas utilizando janelas de contextos local, procedimento que seleciona k palavras vizinhas em torno do alvo.
   O modelo tem um custo computacional bem menor comparada a outras estratégias baseadas em redes neurais presentes na literatura e consegue obter vetores eficientes a partir de grandes conjuntos de dados [Mikolov et al. 2013a, Mikolov et al. 2013b].
2. **FastText**: É um modelo para representação de palavras proposto por [Bojanowski et al. 2017], tratado como uma extensão do modelo *Word2Vec*, que considera as palavras como um conjunto de *n*-gramas de caracteres. Dessa forma, o vetor resultante de uma determinada palavra é dado pela soma das subrepresentações de seus *n*-gramas. Essa abordagem de particionamento dos *tokens* possibilita a obtenção de representações para palavras não vistas no conjunto e treinamento (e.g., sufixos e prefixos). Além disso, palavras raras podem ter representações mais robustas do que aquelas obtidas pelos métodos *Word2Vec* [Polo et al. 2021].

3.3. Classificação de Petições Inicias

A petição inicial é um documento utilizado como primeiro passo para acessar ao Poder Judiciário, quando se está representado por Advogado(a), é nessa peça jurídica que está disposta a demanda requerida [Marinato et al. 2022]. Neste trabalho, utilizaram-se textos de petições iniciais fornecidas pelo TJMA para a avaliação das representações *embeddings* construídas. O objetivo é identificar (classificar) a qual Incidente de Resolução de

Demandas Repetitivas (IRDR), do referido tribunal, determinada petição inicial pertence. O IRDR é uma técnica utilizada pelos tribunais inferiores da justiça brasileira que visa fornecer a decisão para casos semelhantes, promovendo isonomia e segurança jurídica. O *dataset* utilizado já é anotado, ou seja, cada petição pertence a um determinado tipo de IRDR do TJMA. A Tabela 2 apresenta as características do conjunto de dados.

Tabela 2. Conjunto de dados de petições iniciais do TJMA.

Tipo de IRDR	Amostras	Incidência (%)
1	3581	65,18
2	27	0,49
3	502	9,14
4	54	0,98
5	1.068	19,44
6	4	0,07
7	6	0,11
8	252	4,59
TOTAL	5.494	100

3.4. Framework Experimental

Esta seção apresenta as etapas dos experimentos computacionais para o treinamento e avaliação dos modelos *embeddings* treinados a partir dos dados jurídicos. Essas etapas incluem: i) fase de tratamento dos textos de entrada; ii) fase de construção e treinamento dos modelos; iii) fase da análise dos modelos treinados.

Na fase de tratamento dos documentos jurídicos, foram aplicados filtros de pré-processamento de acordo com [Hartmann et al. 2017], realizando remoção de *stopwords*, espaços em branco, marcadores *HyperText Markup Language* (HTML) e quebra de linhas. Também foram configurados o reconhecimento de *tokens* de *e-mail* e *Uniform Resource Locator* (URL), além da transformação de caracteres para *lowercase*. Para essa fase, foram utilizados os recursos disponíveis na biblioteca Python Spacy[11],

Na fase de treinamento das representações, foram considerados diferentes configurações paramétricas para os modelos Word2Vec e FastText, visando ampla cobertura experimental. Para cada um dos modelos, foram analisados: i) tipo de arquitetura utilizada: CBOW e Skip-gram; ii) a dimensão do vetor de características: 300 e 600; e iii) épocas de treinamento: 10. Dessa forma, obtém-se uma representação para cada configuração experimental, como demostrado na Tabela 3. Foram utilizadas janelas de contexto de tamanho 5 e taxa de aprendizado de $0,03$. Para os demais parâmetros, foram utilizados os valores padrão da biblioteca.

Os parâmetros Dimensão e Número de Épocas de treinamento são adotados visando a uma análise de impacto na avaliação final, tal como discutido em [Cunha et al. 2022]. Observando se vetores maiores (com maior quantidade de características para determinada palavra) oferecerem melhores representações e se o impacto da quantidade de rodadas (épocas) de treinamento também é significativo, dado o aumento de custo computacional. Para análises comparativas, modelos já treinados na língua portuguesa também foram incorporados no *framework*, utilizados como *baseline* para os experimentos. Nesse caso, adotou-se os modelos treinados por [Hartmann et al. 2017] e

[11]https://spacy.io/

Tabela 3. Modelos utilizados nos experimentos computacionais.

id	modelo	dimensão	id	modelo-NILC	dimensão
C1	word2vec-Cbow	300	C1-NILC	word2vec-Cbow	300
C2	word2vec-Skip-gram	300	C2-NILC	word2vec-Skip-gram	300
C3	word2vec-Cbow	600	C3-NILC	word2vec-Cbow	600
C4	word2vec-Skip-gram	600	C4-NILC	word2vec-Skip-gram	600
C5	FastText-Cbow	300	C5-NILC	FastText-Cbow	300
C6	FastText-Skip-gram	300	C6-NILC	FastText-Skip-gram	300
C7	FastText-Cbow	600	C7-NILC	FastText-Cbow	600
C8	FastText-Skip-gram	600	C8-NILC	FastText-Skip-gram	600

disponibilizados no repositório do Núcleo Interinstitucional de Linguística Computacional (NILC)[12].

A fase de avaliação dos modelos consistiu na realização de experimentos aplicados em tarefas finais de NLP. Nesse caso, adotou-se a classificação de petições iniciais, tal como descrito na subseção anterior. Como métricas avaliativas, foram utilizadas a acurácia do classificador e, devido ao desbalanceamento das classes do conjunto de dados, a F1-macro. Para o aprendizado do modelo, utilizou-se validação cruzada do tipo *k-fold*, utilizando $k = 5$. No processo de classificação, foram adotados algoritmos de ML já bem conhecidos pela literatura: *Support Vector Machine (SVM)*, *Random Forest (RF)*, *K-Neighbors Neighbor (kNN)* e *Logistic Regression (LR)*, utilizando os recursos da biblioteca python Scikit-learn [13] para a implementação dos mesmos. Como hiperparâmetro do *k*-NN, adotou-se $k = 5$. Para o classificador SVM, foi utilizado a recurso *GridSearchCV* para seleção de parâmetros. Para os demais, foram adotados os paramêtros recomendados pela biblioteca.

4. Resultados e Análises

Esta seção apresenta e discute os resultados da avaliação dos modelos no processo de classificação de petições iniciais. Os resultados apontam que os vetores treinados foram superiores numericamente aos disponibilizados por [Hartmann et al. 2017].

A Tabela 4 mostra os resultados dos experimentos, destacando os melhores casos em negrito. Para a Acurácia (ACC), os modelos C2-NILC, C3-NILC e C4-NILC foram os que apresentaram os melhores resultados para a classificação, alcançando 76%, com o algoritmo *k*-NN. Já para F1-macro, os modelos treinados C6 e C7 foram os que ofereceram melhores resultados, com 42%, para o mesmo classificador. O algoritmo *k*-NN também foi o que obteve a melhor desempenho médio tanto para ACC quanto para F1-macro, com 74 e 40% para os modelos treinados e 75 e 39% para os do NILC, respectivamente. Os resultados médios da F1-macro também mostram um ganho de desempenho dos algoritmos no aprendizado de classes com menor incidência, quando utilizado os vetores treinados.

A Figura 1, apresenta a matriz de confusão com o aprendizado do classificador *k*-NN para o modelo C6, configuração com maior desempenho geral para F1-macro. Do ponto de vista de classificação, percebe-se dificuldades do algoritmo na predição para classes minoritárias. No entanto, ressalta-se que não foram aplicados mecanismos para

[12]http://www.nilc.icmc.usp.br/embeddings

[13]https://scikit-learn.org/stable

Tabela 4. Resultados dos experimentos para a classificação de IRDR.

id	*k*-NN		LR		RF		SVM	
	ACC	F1-macro	ACC	F1-macro	ACC	F1-macro	ACC	F1-macro
C1	0,75	0,40	0,72	**0,32**	0,66	0,12	0,66	0,15
C2	0,74	0,41	0,72	0,29	0,65	0,10	0,65	0,10
C3	0,75	0,40	0,72	0,31	**0,68**	0,16	**0,67**	**0,16**
C4	0,75	0,40	**0,73**	0,30	0,66	0,15	0,65	0,10
C5	0,73	0,36	0,71	0,28	0,67	0,15	0,65	0,10
C6	0,75	**0,42**	0,72	0,29	0,66	0,10	0,65	0,10
C7	0,73	**0,42**	0,71	0,30	0,67	0,17	0,66	0,13
C8	0,75	0,40	**0,73**	0,30	0,66	0,13	0,65	0,10
Média	0,74	**0,40**	0,72	**0,30**	**0,66**	**0,14**	**0,66**	**0,12**
C1-NILC	0,75	0,40	**0,73**	0,29	0,65	0,10	0,65	0,10
C2-NILC	**0,76**	0,41	**0,73**	0,29	0,65	0,13	0,65	0,10
C3-NILC	**0,76**	0,41	0,72	0,29	**0,68**	**0,19**	0,65	0,12
C4-NILC	**0,76**	0,41	**0,73**	0,31	0,66	0,15	0,65	0,10
C5-NILC	0,75	0,36	**0,73**	0,29	0,67	0,17	0,65	0,10
C6-NILC	0,74	0,38	**0,73**	0,29	0,65	0,10	0,65	0,10
C7-NILC	0,75	0,36	0,72	0,29	0,66	0,17	0,65	0,12
C8-NILC	0,74	0,39	0,72	0,28	0,66	0,11	0,65	0,10
Média	**0,75**	0,39	**0,73**	0,29	**0,66**	**0,14**	0,65	0,11

o balanceamento do conjunto de dados (Tabela 2) utilizados para classificação, por não ser o foco principal deste trabalho, deixando como ponto em aberto para experimentos futuros.

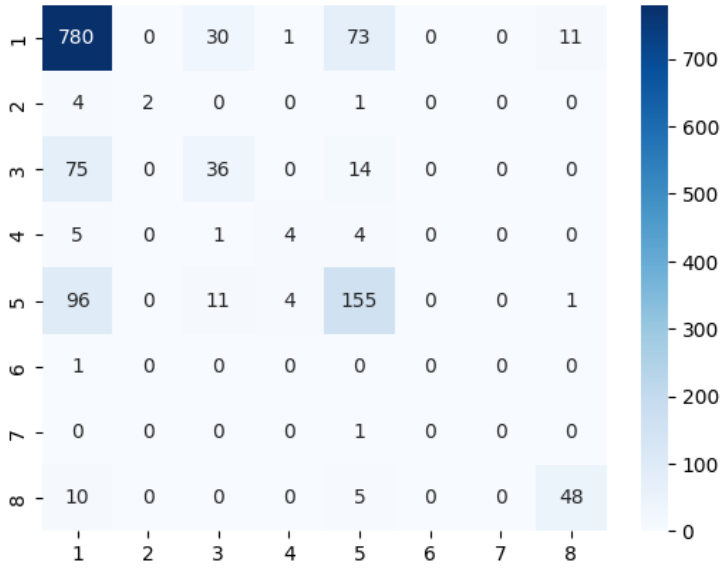

Figura 1. Matriz de Confusão do cenário com melhor desempenho (modelo C6 com algoritmo KNN, vide Tabela 4).

Os resultados obtidos mostram competitividade dos modelos treinados com os dados jurídicos em comparação aos disponibilizados por [Hartmann et al. 2017], treinados com grande volume de dados da língua portuguesa. A diferença no tamanho dos *corpus* indica que a utilização de documentos extraídos do domínio, tal como monstrado em [Gomes et al. 2021], pode fornecer representações que incorporam as especificidades da

área, utilizando quantidade de amostras relativamente menor, reduzindo também o custo computacional.

5. Conclusão

Em vista do Programa Justiça 4.0 do judiciário brasileiro, que busca soluções computacionais que ofereçam celeridade nas atividades processuais, o presente estudo apresentou a construção e avaliação de representações para palavras (*word embeddings*) orientadas à linguagem jurídica brasileira. As particularidades inerentes do domínio de aplicação adicionam complexidade para o aprendizado de algoritmos de *Machine Learning* e *Deep Learning*, principalmente para aqueles que recebem como entrada representações mais generalistas da língua portuguesa.

Como contribuição técnico-científica do estudo, além do *corpus* de mais de 500.000 documentos de instituições de justiça do Brasil das mais variadas esferas, disponibilizam-se os modelos de linguagem treinados com esse corpus e avaliação dos mesmos para a classificação de petições iniciais no repositório `https://github.com/fabiolobato/legal-embeddings-br`. Os resultados obtidos mostraram-se promissores quando comparados com modelos generalistas, indicando que modelos segmentados têm o potencial de melhorar sistemas inteligentes neste domínio. Destaca-se ainda que o presente estudo tem o potencial de fomentar o desenvolvimento de aplicações focadas no processamento de linguagem natural no domínio jurídico.

Ressaltamos que é uma pesquisa em andamento. Dessa forma, como próximos passos, pretendemos ampliar a cobertura experimental, incorporando: (i) outras técnicas de representação (e.g., Glove e o BERT); ii) novas estratégias de avaliação para além da classificação de dados, como agrupamento e mensuração de similaridade semântica; e iii) testes com outros modelos pré-treinados no domínio, como *baselines* comparativos.

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-308334/2020; pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021; e pelo Acordo de Cooperação Técnica N° 02/2021 (Processo N° 38328/2020 -TJ/MA).

Referências

Bambroo, P. and Awasthi, A. (2021). Legaldb: Long distilbert for legal document classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4, Bhilai, India. 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT).

Batista, H., Nascimento, A., Melo, R., Miranda, P., Maldonado, I., and Filho, J. C. (2021). A comparative analysis of text embedding approach to extract named entities in portuguese legal documents. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 221–232, Porto Alegre, RS, Brasil. SBC.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Chalkidis, I. and Kampas, D. (2019). Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2).

Consoli, B., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., and Moreira, V. (2020). Embeddings for named entity recognition in geoscience Portuguese literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4625–4630, Marseille, France. European Language Resources Association.

Cunha, L. F., Almeida, J. J. a., and Simões, A. (2022). Reasoning with Portuguese Word Embeddings. In Cordeiro, J. a., Pereira, M. J. a., Rodrigues, N. F., and Pais, S. a., editors, *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, volume 104 of *Open Access Series in Informatics (OASIcs)*, pages 17:1–17:14, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Dal Pont, T. R., Sabo, I. C., Hübner, J. F., and Rover, A. J. (2020). Impact of text specificity and size on word embeddings performance: An empirical evaluation in brazilian legal domain. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, page 521–535.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Garcia, A. C. (2020). Ética e inteligencia artificial. *Computação Brasil*, 43:14–22.

Gomes, D. d. S. M., Cordeiro, F. C., Consoli, B. S., Santos, N. L., Moreira, V. P., Vieira, R., Moraes, S., and Evsukoff, A. G. (2021). Portuguese word embeddings for the oil and gas industry: development and evaluation. *Computers in Industry*, 124:103347.

Hariri, R. H., Fredericks, E. M., and Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, 6(1):44.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, Porto Alegre, RS, Brasil. SBC.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Marinato, M., Junior, A. J., Lobato, F., and Cortes, O. (2022). Classificação automática de petições iniciais usando classificadores combinados. In *Anais do XVI Brazilian e-Science Workshop*, pages 89–96, Porto Alegre, RS, Brasil. SBC.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Mota, C., Lima, A., Nascimento, A., Miranda, P., and de Mello, R. (2020). Classificação de páginas de petições iniciais utilizando redes neurais convolucionais multimodais. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 318–329, Porto Alegre, RS, Brasil. SBC.

Parreiras, M., Vasconcellos, A., Mangeli, E., Yamamoto, E., Xexéo, G., Metello, I., Costa, L., Marques, P., and Souza, J. (2022). Inteligência artificial aplicada para o aumento da produtividade no atendimento de intimações. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 180–191, Porto Alegre, RS, Brasil. SBC.

Pereira, J. C. M. and Rodrigues, M. V. J. (2021). A plataforma sinapses e a continuidade dos modelos de ia no judiciário. In *ANAIS do Encontro de Administração da Justiça - ENAJUS 2021*, Lisboa.

Pinto, H. A. (2020). A utilização da inteligência artificial no processo de tomada de decisões: por uma necessária accountability. *Revista de Informação Legislativa: RIL*.

Polo, F., Mendonça, G., Parreira, K., Gianvechio, L., Cordeiro, P., Ferreira, J., Lima, L., Maia, A., and Vicente, R. (2021). Legalnlp - natural language processing methods for the brazilian legal language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 763–774, Porto Alegre, RS, Brasil. SBC.

Qader, W. A., Ameen, M. M., and Ahmed, B. I. (2019). An overview of bag of words;importance, implementation, applications, and challenges. In *2019 International Engineering Conference (IEC)*, pages 200–204.

Schaulet, E. and Trez, G. (2021). Big data em organizações de médio e grande porte do setor público brasileiro: Prontidão e situação atual, replicação do estudo holandês de klievink et al. (2017). In *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, pages 13–24, Porto Alegre, RS, Brasil. SBC.

Smywiński-Pohl, A., Lasocki, K., Wróbel, K., and Strzała, M. (2019). Automatic construction of a polish legal dictionary with mappings to extra-legal terms established via word embeddings. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, page 234–238, New York, NY, USA. Association for Computing Machinery.

Sousa, A. W. and Del Fabro, M. D. (2019). Iudicium textum dataset uma base de textos jurıdicos para nlp. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD*, pages 1–11, Fortaleza, Brazil. SBBD.

Wang, Z., Wu, Y., Lei, P., and Peng, C. (2020). Named entity recognition method of brazilian legal text based on pre-training model. *Journal of Physics: Conference Series*, 1550(3):032149.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

APÊNDICE B – TOKENS JURÍDICOS PARA VISUALIZAÇÃO

Tabela 13 – Palavras jurídicas extraídas dos conjuntos de dados usando TF-IDF.

Conjunto de dados militar	Conjunto de dados de petições
“dano”, “danos”, “ma”, “banco”, “honorários”, “contrato”, “autora”, “sucumbência”, “empréstimo”, “moral”, “execução”, “consumidor”, “autor”, “morais”, “requerente”, “luís”, “cep”, “descontos”, “maranhão”, “termos”, “reais”, “seja”, “crédito”, “principal”, “indenização”, “tutela”, “réu”, “presente”, “requerido”, “pagamento”, “fone”, “benefício”, “fax”, “quadra”, “data”, “dobro”, “valores”, “pessoa”, “centro”, “cdc”, “ônus”, “instituição”, “repetição”, “pagar”, “serviços”, “cível”, “advogado”, “pública”, “consignado”, “ato”	“militar”, “decisão”, “ementa”, “cpm”, “penal”, “unânime”, “crime”, “embargos”, “pena”, “preliminar”, “apelo”, “delito”, “acusado”, “apelação”, “deserção”, “criminal”, “ausência”, “corpus”, “habeas”, “denúncia”, “conduta”, “ordem”, “cppm”, “militares”, “declaração”, “princípio”, “provimento”, “unanimidade”, “provido”, “in”, “apelante”, “legal”, “provas”, “prisão”, “nulidade”, “prescrição”, “administração”, “autoria”, “requisitos”, “lesão”, “paciente”, “materialidade”, “rejeição”, “ministerial”, “disciplinar”, “acórdão”, “estelionato”, “parcial”, “autoridade”, “instrução”