CENTRO UNIVERSITÁRIO SENAI CIMATEC

Programa de Pós-Graduação em

Modelagem Computacional e Tecnologia Industrial

RAPHAEL SOUZA DE OLIVEIRA

UMA METODOLOGIA PARA AGRUPAMENTO DE PROCESSOS JUDICIAIS BASEADA EM APRENDIZAGEM PROFUNDA APLICADA À JUSTIÇA TRABALHISTA BRASILEIRA

Salvador

2022

RAPHAEL SOUZA DE OLIVEIRA

UMA METODOLOGIA PARA AGRUPAMENTO DE PROCESSOS JUDICIAIS BASEADA EM APRENDIZAGEM PROFUNDA APLICADA À JUSTIÇA TRABALHISTA BRASILEIRA

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional e Tecnologia Industrial do Centro Universitário SENAI CIMATEC como requisito parcial para a obtenção do título de Mestre em Modelagem Computacional e Tecnologia Industrial. Orientador: Prof. Dr. Erick Giovani Sperandio Nascimento

Salvador

2022

Ficha catalográfica elaborada pela Biblioteca do Centro Universitário SENAI CIMATEC

O48m Oliveira, Raphael Sousa de

Uma metodologia para agrupamento de processos judiciais baseada em aprendizagem profunda aplicada à justiça trabalhista brasileira / Raphael Sousa de Oliveira. – Salvador, 2022.

90 f. : il. color.

Orientador: Prof. Dr. Erick Giovani Sperandio Nascimento.

Dissertação (Mestrado em Modelagem Computacional e Tecnologia Industrial) – Programa de Pós-Graduação, Centro Universitário SENAI CIMATEC, Salvador, 2022.
Inclui referências.

1. Jurídico. 2. Processamento de linguagem natural. 3. Clusterização. 4. Word2vec. 5. Transformers. I. Centro Universitário SENAI CIMATEC. II. Nascimento, Erick Giovani Sperandio. III. Título.

CDD 006.3

NDI - 05

CENTRO UNIVERSITÁRIO SENAI CIMATEC

Mestrado Acadêmico em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, aprova a Defesa de Mestrado, intitulada **"UMA METODOLOGIA PARA AGRUPAMENTO DE PROCESSOS JUDICIAIS BASEADA EM APRENDIZAGEM PROFUNDA APLICADA À JUSTIÇA TRABALHISTA BRASILEIRA**" apresentada no dia 18 de outubro de 2022, como parte dos requisitos necessários para a obtenção do Título de Mestre em Modelagem Computacional e Tecnologia Industrial.

Electronically signed by:
Erick Giovani Sperandio Nascimento
CPF: ***.666.177-**
Date: 10/19/2022 7:26:58 PM -03:00

Orientador: **Prof. Dr. Erick Giovani Sperandio Nascimento**
SENAI CIMATEC

Assinado eletronicamente por:
Hugo Saba Pereira Cardoso
CPF: ***.375.625-**
Data: 20/10/2022 17:43:00 -03:00

Membro Interno: **Prof. Dr. Hugo Saba Pereira Cardoso**
SENAI CIMATEC

Electronically signed by:
Mário de Noronha Neto
CPF: ***.859.519-**
Date: 10/20/2022 8:00:59 AM -03:00

Membro Externo: **Prof. Dr. Mário de Noronha Neto**
IFSC

Av. Orlando Gomes, 1845 – Piatã - CEP: 41650-010 Salvador-Bahia - Tel. (71)3462-9500 Fax: (71)3462-9599

Dedico este trabalho à toda minha família e amigos.

AGRADECIMENTOS

Agradeço a minha família, em especial à minha esposa e filha, Lorena e Gabriela por todo amor, incentivo e dedicação.

Aos meus pais, Genésio e Mariângela, pela confiança e orgulho que depositam em mim, ao meus irmãos Matheus e Nicolly pela parceria de sempre.

Aos meus parentes mais próximos, sogro, sogra, cunhados, cunhadas, sobrinho e sobrinhas pelos momentos em família.

Ao Tribunal Regional do Trabalho da 5ª Região, em especial Dra. Débora Maria Lima Machado, Dra. Dalila Nascimento Andrade, Dr. Firmo Ferreira Leal Neto, Érica Cristina Dórea Rossiter Tavares e Leonardo Rodrigues Barreto, pela confiança e por proporcionar a oportunidade dessa pesquisa e capacitação.

Ao meu orientador, Prof. Dr. Erick Giovani Sperandio Nascimento, um agradecimento especial, por me inspirar na caminhada, acreditando sempre em meu potencial, pelas sugestões, esclarecimentos, paciência, dedicação, disponibilidade e por sempre apresentar oportunidades de crescimento e desenvolvimento acadêmico e científico.

Aos membros da banca pela disponibilidade em avaliar e realizar comentários construtivos, contribuindo com a evolução do trabalho.

Aos professores e colegas dos Programas de Pós-Graduação do SENAI CIMATEC, pela contribuição a cada apresentação e durante as aulas.

Ao SENAI CIMATEC por todo o apoio, pela infraestrutura computacional e disponibilidade de equipe, fundamentais para o desenvolvimento deste trabalho.

A todos aqueles que contribuíram de forma direta ou indireta. Foram muitos que apoiaram esse desenvolvimento.

Meus sinceros agradecimentos.

RESUMO

O ordenamento jurídico brasileiro estabelece medidas para garantir a celeridade da resolução dos processos judiciais, como o princípio da razoável duração das ações judiciais, o princípio da celeridade, da economia processual e do devido fluxo jurídico, com vistas à otimização do processo judicial. No entanto, os Tribunais Judiciais vivem um período de grandes cortes orçamentais e redução do número de magistrados e colaboradores civis. Nesta situação atual, a Tecnologia da Informação, mais especificamente a Inteligência Artificial (IA), tem tido sucesso no enfrentamento de muitos problemas complexos na área de Processamento de Linguagem Natural (do inglês *Natural Language Processing* - NLP), o que pode ajudar a amenizar as dificuldades enfrentadas por essas restrições. Dentre as necessidades que podem ser atendidas com a IA, a automatização da classificação de documentos é de grande interesse, pois, quando corretamente classificados, podem indicar caminhos a serem seguidos pelo judiciário para garantir a resolução célere dos processos judiciais. Nesse contexto, este trabalho busca detectar o grau de similaridade entre documentos judiciais que pode ser alcançado nos grupos inferidos por meio do uso do aprendizado não supervisionado. Em seguida, por meio da aplicação de nove técnicas de Processamento de Linguagem Natural, que são (i) Frequência de Termos - Frequência Inversa de Documentos (TF-IDF); (ii) Word2Vec com CBoW especializado no corpus da língua brasileira; (iii) Word2Vec com Skip-gram especializado no corpus da língua brasileira; (iv) BERT (*Bidirectional Encoder Representations from Transformers*) treinado para fins gerais para o Português Brasileiro; (v) BERT especializado com o corpus do Judiciário trabalhista brasileiro; (vi) GPT-2 treinado para fins gerais para o Português Brasileiro; (vii) GPT-2 especializado no corpus do Judiciário trabalhista brasileiro; (viii) RoBERTa treinado para fins gerais para o Português Brasileiro; e (ix) RoBERTa especializado no corpus do Judiciário trabalhista brasileiro, foi desenvolvido um modelo de agrupamento de ações judiciais, que é calculado com base no cosseno da distância entre os elementos do grupo ao seu centróide. O Recurso Ordinário (RO) foi escolhido como tipo de documento de referência pelo fato de ser o documento que aciona o processo para subir à instância superior e por existir atualmente um grande contingente de ações aguardando julgamento em 2ª instância. Após as etapas de extração de dados, pré-processamento e limpeza, os documentos tiveram seu conteúdo transformado em uma representação vetorial, utilizando as nove técnicas de NLP citadas acima. Para BERT especializado, GPT-2 especializado e RoBERTa especializado foi necessário um passo prévio de geração de vetores densos de representação da informação - *word embeddings*. Preliminarmente, através do estudo apresentado, pode-se perceber que o uso de modelos de *word embedding* especializados é um importante avanço na qualidade dos modelos que utilizam o conteúdo de documentos como recursos, principalmente quando se trata de modelos de NLP baseados na arquitetura *Transformers*.

Palavras-chaves: jurídico; processamento de linguagem natural; clusterização; word2vec; *transformers*;

ABSTRACT

A METHODOLOGY FOR CLUSTERING OF LEGAL PROCEEDINGS BASED ON DEEP LEARNING APPLIED TO THE BRAZILIAN LABOR JUSTICE

The Brazilian legal system establishes measures to ensure the speed of resolution of lawsuits, such as the principle of reasonable duration of lawsuits, the principle of speed, procedural economy and due legal flow, with a view to optimizing the judicial process. However, the Judicial Courts are experiencing a period of major budget cuts and a reduction in the number of magistrates and civil collaborators. In this current situation, Information Technology, more specifically Artificial Intelligence (AI), has been successful in facing many complex problems in the area of Natural Language Processing (NLP), which can help to alleviate the difficulties faced by these restrictions. Among the needs that can be met with AI, the automation of document classification is of great interest, because, when correctly classified, they can indicate paths to be followed by the judiciary to ensure the speedy resolution of legal proceedings. In this context, this work seeks to detect the degree of similarity between court documents that can be achieved in the inferred groups through the use of unsupervised learning. Then, through the application of nine Natural Language Processing techniques, which are (i) Frequency of Terms - Inverse Frequency of Documents (TF-IDF); (ii) Word2Vec with CBoW specialized in the Brazilian language corpus; (iii) Word2Vec with Skip-gram specialized in the Brazilian language corpus; (iv) BERT (Bidirectional Encoder Representations from Transformers) trained for general purposes for Brazilian Portuguese; (v) BERT specialized with the corpus of the Brazilian labor judiciary; (vi) General purpose trained GPT-2 for Brazilian Portuguese; (vii) GPT-2 specialized in the corpus of the Brazilian labor judiciary; (viii) Roberta trained for general purposes in Brazilian Portuguese; and (ix) Roberta, specialized in the corpus of the Brazilian labor judiciary, developed a model for grouping lawsuits, which is calculated based on the cosine of the distance between the elements of the group to its centroid. The Ordinary Appeal (acronym in Portuguese for “Recurso Ordinário” - RO) was chosen as the type of reference document because it is the document that triggers the process to go to the higher court and because there is currently a large contingent of lawsuits awaiting judgment in the 2nd instance. After the steps of data extraction, pre-processing and cleaning, the documents had their content transformed into a vector representation, using the nine NLP techniques mentioned above. For specialized BERT, specialized GPT-2 and specialized RoBERTa, a previous step of generating dense vectors of information representation - word embeddings was necessary. Preliminarily, through the presented study, it can be seen that the use of specialized word embedding models is an important advance in the quality of models that use document content as resources, especially when it comes to NLP models based on Transformers architecture.

Keywords: legal; natural language processing; clustering; word2vec; transformers;

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

BERT	*Bidirectional Encoder Representations from Transformers*
BoW	*Bag of words*
BPE	*Byte-Pair Encoding*
CBoW	*Continuous Bag of Words*
CNJ	Conselho Nacional de Justiça
GPT-2	*Generative Pre-trained Transformer 2*
IAD	Índice de Atendimento à Demanda
NLP	*Natural Language Processing*
PJe	Processo Judicial Eletrônico
TF-IDF	*Term Frequency - Inverse Document Frequency*
RoBERTa	*Robustly optimized BERT approach*TeX

LISTA DE SÍMBOLOS

$\in$	Pertence

SUMÁRIO

1 INTRODUÇÃO

A história recente do Judiciário Brasileiro apresenta importantes transformações no sentido de ter todos os seus atos em formato digital. Em 2012, a Justiça Trabalhista Brasileira implantou o Processo Judicial Eletrônico (PJe) e, desde então, todos os novos processos judiciais são totalmente eletrônicos. De acordo com o Relatório Analítico Anual da Justiça em Números 2021 (ano-base 2020) (CNJ, 2021), produzido pelo Conselho Nacional de Justiça (CNJ), 99,9% dos processos em andamento já se encontram nesta plataforma.

Sabendo-se da limitação dos seres humanos analisarem, em tempo aceitável, uma grande quantidade de dados, sobretudo quando tais dados aparentam não estar correlacionados, é possível auxiliá-los nesse contexto de reconhecimento de padrões, através de métodos estatísticos, computacionais e de análise de dados. Partindo do pressuposto que os dados textuais estão aumentando exponencialmente, a análise de padrões em documentos jurídicos tem se tornado cada vez mais desafiadora.

O ordenamento jurídico brasileiro prevê meios de garantir a tramitação célere dos processos judiciais, tais como: o princípio da duração razoável de um processo, o princípio da celeridade, a economia processual e o devido processo legal com vistas a otimizar o andamento processual (SALUM, 2016). Desta forma, atender com celeridade a crescente demanda judicial é um dos grandes desafios do Judiciário Brasileiro. Portanto, através da utilização de um mecanismo de agrupamento de processos, com uma boa taxa de similaridade entre os documentos a serem analisados, foi possível auxiliar na distribuição de trabalho entre os assessores do gabinete para o qual o processo foi sorteado. Além disso, contribuiu na busca de jurisprudências[1] para julgamento dos processos em questão, resguardando o princípio da segurança jurídica. De acordo com Gomes Canotilho (CANOTILHO, 2003), o princípio geral da segurança jurídica visa garantir ao indivíduo o direito de confiar que as decisões emitidas sobre as suas questões são alicerçadas em normas jurídicas vigentes e válidas.

Desta forma, esta ferramenta de gestão jurídica possibilitou a redução da duração do processo judicial, gerando impactos positivos como a redução dos custos operacionais de um processo a partir da menor alocação dos recursos necessários ao seu julgamento.

Pesquisas recentes têm demonstrado que algoritmos de aprendizado de máquina são ferramentas importantes capazes de resolver problemas de alta complexidade com uso de Processamento de Linguagem Natural (do inglês *Natural Language Processing* -

[1] Termo jurídico que significa um conjunto de decisões judiciais que tiveram uma mesma linha de entendimento.

NLP) (KHAN et al., 2016). Neste sentido, é possível destacar os trabalhos de (WANG; CUI; ZHANG, 2019; MIKOLOV et al., 2013; PENNINGTON; SOCHER; MANNING, 2014; BOJANOWSKI et al., 2016; DEVLIN et al., 2018; RADFORD et al., 2019; LIU et al., 2019), que, levando em consideração o contexto das palavras, aplicam técnicas de geração de *word embeddings*, uma forma de representação vetorial de palavras, e consequentemente de documentos. O uso desses *word embeddings* é imprescindível na análise de um conjunto de dados não estruturados e que se apresentam em forma de documentos de grande volume na Justiça.

Atualmente um especialista faz a triagem dos documentos e distribuição dos processos judiciais a serem julgados entre os membros da equipe, configurando-se um desvio da principal atividade do especialista, que é a produção das minutas de decisão. Isto contribuiu para aumento na taxa de congestionamento (indicador que mede o percentual de casos que permanecem pendentes de solução ao final do ano-base) e para baixa do índice de atendimento à demanda (IAD - indicador que mede o percentual de baixa de processos frente ao quantitativo de casos novos). Esse fato foi evidenciado nos dados consolidados da Justiça do Trabalho contido no Relatório de indicadores da Justiça Trabalhista Brasileira apresentado na Tabela 1, cujos dados foram extraídos do Relatório Analítico Anual da Justiça em Números 2021 (ano-base 2020) (CNJ, 2021) produzido pelo Conselho Nacional de Justiça (CNJ).

Este trabalho objetiva, portanto, através da aplicação de nove técnicas de Processamento de Linguagem Natural, quais sejam: (i) Frequência do Termo - Frequência Inversa do Documento (do inglês *Term Frequency - Inverse Document Frequency* - TF-IDF); (ii) Word2Vec com CBoW (do inglês *Continuous Bag of Words*) treinado para fins gerais para a língua Portuguesa no Brasil (Word2Vec CBoW ptBR); (iii) Word2Vec com Skip-gram treinado para fins gerais para a língua Portuguesa no Brasil (Word2Vec Skip-gram ptBR); (iv) BERT (do inglês *Bidirectional Encoder Representations from Transformers*) treinado para fins gerais para o Português Brasileiro (BERT ptBR); (v) BERT especializado com o corpus do judiciário trabalhista Brasileiro (BERT Jud.); (vi) GPT-2 (do inglês *Generative Pre-trained Transformer 2*) treinado para fins gerais para o Português Brasileiro (GPT-2 ptBR); (vii) GPT-2 especializado com o corpus do judiciário trabalhista Brasileiro (GPT-2 Jud.); (viii) RoBERTa (do inglês *Robustly optimized BERT approach*) treinado para fins gerais para o Português Brasileiro (RoBERTa ptBR); e (ix) RoBERTa especializado com o corpus do judiciário trabalhista Brasileiro (RoBERTa Jud.), apresentar, a partir do uso destas técnicas de NLP, o grau de semelhança entre os documentos judiciais que foi alcançado nos grupos inferidos por meio da aprendizagem não supervisionada.

Tabela 1 – Relatório de Indicadores da Justiça Trabalhista Brasileira

	Descrição	2º Grau	1º Grau	Total
Força de Trabalho				
Magistrados	Autoridade judiciária	561	3.048	3.609
Servidores Judiciário	Funcionário público	6.955	22.647	29.602
Movimentação Processual				
Estoque	Quantidade de processos pendentes	681.888	3.875.625	4.557.513
Casos Novos	Quantidade de processos novos	740.497	2.235.402	2.975.899
Julgados	Quantidade de processos julgados	738.133	2.132.377	2.870.510
Baixados	Quantidade de processos com decisão final	841.727	2.257.139	3.098.866
Indicadores de Produtividade				
IAD	Processos baixados / Casos novos	113,7%	101,0%	104,1%
Taxa de Congestionamento	Processos baixados / (Casos novos + Estoque)	44,8%	63,2%	59,5%
Conhecimento	Fase de conhecimento dos fatos	—	48,4%	48,4%
Execução	Fase de cumprimento da decisão judicial	—	75,6%	75,6%
Indicadores por Magistrado				
Casos Novos	Média de processos novos por Magistrado	1.320	526	659
Carga de Trabalho	Média de processos por Magistrado	3.175	2.403	2.533
Processos Julgados	Média de processos julgados por Magistrado	1.316	765	857
Processos Baixados	Média de processos baixados por Magistrado	1.500	810	925
Indicadores por Servidor				
Casos Novos	Média de processos novos por servidor	110	67	77
Carga de Trabalho	Média de processos por servidor	264	305	295
Processos Baixados	Média de processos baixados por servidor	125	103	108

Fonte: (CNJ, 2021).

Tal grau de similitude indica o desempenho do modelo e resultou da medida da taxa média de similaridade dos documentos dos grupos formados. Essa, por sua vez, foi calculada com base na similaridade cosseno entre os elementos do grupo ao seu centróide e, de forma comparativa, pela média da similaridade cosseno entre todos os documentos do grupo.

Com intuito de delimitar o escopo desta pesquisa, foi extraído um conjunto de dados contendo informações de documentos do tipo Recurso Ordinário Interposto (ROI) de aproximadamente 210 mil processos jurídicos. O Recurso Ordinário Interposto foi usado como referência, pois este é, normalmente, o tipo de documento que dá origem à subida do processo para julgamento em instância superior (2º grau), instituindo assim o Recurso Ordinário (RO). Este é um recurso de fundamentação livre, cabível contra sentenças definitivas e terminativas proclamadas na primeira instância, que busca uma reforma da decisão judicial elaborada por um órgão hierarquicamente superior (OLIVEIRA, 2011).

A partir dos objetivos propostos foram produzidos dois artigos científicos, conforme sintetizado na Tabela 2, dos quais um foi publicado como capítulo do livro digital "*Data Clustering*" da IntechOpen [2] em setembro de 2021 e o segundo foi submetido para a revista científica "*Artificial Intelligence and Law*" [3] da Springer em março de 2022.

Tabela 2 – Síntese dos artigos científicos produzidos

Tipo Publicação	Titulo	Situação
Capítulo de Livro	*Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches (OLIVEIRA; NASCIMENTO, 2021)*	Publicado
Revista Científica	*Brazilian court documents clustered by similarity together using natural language processing approache with Transformers*	Submetido

1.1 Objetivo

O objetivo deste trabalho foi desenvolver uma metodologia que se utiliza de técnicas de processamento de linguagem natural e aprendizagem de máquina profunda para o agrupamento automatizado de processos judiciais da Justiça Trabalhista Brasileira, aplicado a dados de processos de um Tribunal Regional do Trabalho no Brasil.

[2] Endereço de acesso na internet: <https://www.intechopen.com/>
[3] Endereço de acesso na internet: <https://www.springer.com/journal/10506>

1.2 Objetivos Específicos

Para alcançar o objetivo do trabalho, foi proposto como objetivos específicos:

1. Levantar e preparar um banco de dados de processos judiciais da Justiça Trabalhista Brasileira para realização de agrupamento de processos judiciais;
2. Desenvolver e validar um estudo comparativo de técnicas tradicionais de NLP e de aprendizagem de máquina para agrupamento de processos judiciais;
3. Desenvolver e especializar os modelos de aprendizagem de máquina profunda para a língua Portuguesa com corpus jurídico trabalhista;
4. Desenvolver uma metodologia baseada em aprendizagem profunda para agrupamento de processos judiciais, comparando técnicas tradicionais com as baseadas em aprendizagem profunda;
5. Testar e validar as técnicas de processamento de linguagem natural aplicadas, buscando consolidar uma metodologia para a Justiça Trabalhista Brasileira.

1.3 Organização do Documento

Este trabalho está disposto de acordo com os seguintes seções:

- **Seção 1 - Introdução**: apresenta o contexto do assunto que foi tratado, qual o problema que se propôs, os objetivos gerais e específicos e a organização de todo o trabalho;
- **Seção 2 - Revisão da Literatura**: aborda a fundamentação teórica e o estado da arte atual na literatura sobre o tema da pesquisa;
- **Seção 3 - Manuscrito 1**: apresenta um capítulo de livro publicado na *IntechOpen* com o título *Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches* em que se mostra o grau de semelhança entre os documentos judiciais que foi alcançado nos grupos inferidos por meio da aprendizagem não supervisionada através da aplicação de três técnicas de Processamento de Linguagem Natural, quais sejam: (i) TF-IDF; (ii) Word2Vec CBoW ptBR; e (iii) Word2Vec Skip-gram ptBR;
- **Seção 4 - Manuscrito 2**: apresenta um artigo submetido à revista Artificial Intelligence and Law (Editora Springer, Fator de Impacto (FI): 2.723), que utiliza como linha de base os resultados discutidos pela pesquisa *"Clustering by Similarity of*

Brazilian Legal Documents Using Natural Language Processing Approaches" (OLIVEIRA; NASCIMENTO, 2021) comparando-os com o grau de semelhança entre os documentos judiciais alcançado nos grupos inferidos por meio da aprendizagem não supervisionada, através da aplicação de seis técnicas de Processamento de Linguagem Natural, quais sejam: (i) BERT (*Bidirectional Encoder Representations from Transformers*) treinado para fins gerais para o Português Brasileiro (BERT ptBR); (ii) BERT especializado com o corpus do judiciário trabalhista Brasileiro (BERT Jud.); (iii) GPT-2 (*Generative Pre-trained Transformer 2*) treinado para fins gerais para o Português Brasileiro (GPT-2 ptBR); (iv) GPT-2 especializado com o corpus do judiciário trabalhista Brasileiro (GPT-2 Jud.); (v) RoBERTa (*Robustly optimized BERT approach*) treinado para fins gerais para o Português Brasileiro (RoBERTa ptBR); e (vi) RoBERTa especializado com o corpus do judiciário trabalhista Brasileiro (RoBERTa Jud.);

- **Seção 5 - Conclusões**: apresenta as conclusões e propostas de trabalhos futuros desta dissertação.

2 REVISÃO DA LITERATURA

2.1 Fundamentação Teórica

Nesta subseção serão apresentados os principais fundamentos teóricos e conceituais necessários para o entendimento, desenvolvimento e validação desta pesquisa.

2.1.1 Aprendizagem de Máquina (*Machine Learning* - ML)

A aprendizagem de máquina (do inglês *Machine Learning* - ML) é um subdomínio da IA, o ML fornece às máquinas a capacidade de aprender automaticamente sem a ajuda de programação explícita. Os algorítimos de ML são categorizados principalmente em quatro tipos como detalhado a seguir (JAIN; MURTY; FLYNN, 1999; DUTTON; CONROY, 1997):

- Supervisionado: utilizam dados adquiridos com rótulos para prever eventos. Essa abordagem inicia a partir do processo de treinamento do conjunto de dados, o ML desenvolve uma função de inferência para prever os valores de saída. O sistema é capaz de fornecer resultados a um dado de entrada com processo de treinamento adequado. O algoritmo ML compara os resultados obtidos com os resultados reais e esperados para identificar erros para ajustar o modelo. Essa categoria pode ser ainda agrupada em casos de classificação (quando a variável de saída é categórica) ou regressão (quando a variável de saída é um valor contínuo);

- Não supervisionado: são empregadas quando os dados de treinamento não estão classificados ou rotulados previamente. Ele analisa como o sistema pode deduzir uma função para explicar os padrões ocultos dos dados não rotulados. Essa categoria pode ser ainda agrupada em casos de agrupamento (quanto tenta desvendar agrupamentos inerentes aos dados, ou seja, características em comum) ou associação (quando tenta descobrir regras de associação);

- Semi supervisionado: situam-se entre as técnicas de ML supervisionadas e não supervisionadas, onde usa dados rotulados e não rotulados para o processo de treinamento. Geralmente, considera uma quantidade menor de dados rotulados e uma quantidade maior de dados não rotulados. Esses tipos de técnicas podem se ajustar para obter maior precisão;

- Por reforço: interagem com o ambiente por meio de ações e localizam erros ou recompensas. Busca por tentativa e erro, e recompensas atrasadas são algumas das

características comuns do método de reforço. Se o erro for grande, então a penalidade é alta e a recompensa baixa. Se o erro for pequeno, então a penalidade é baixa e a recompensa alta. Está técnica permite a determinação automática do comportamento ideal dentro de um contexto específico para maximizar o desempenho desejado.

Destaca-se portanto, que as categorias de algorítimos de aprendizagem de máquina semi supervisionado e não supervisionado foram utilizados nesta pesquisa respectivamente durante o treinamento dos modelos de processamento de linguagem natural e durante o treinamento para agrupamento de processos similares.

2.1.2 Processamento de Linguagem Natural (*Natural Language Processing* - NLP)

O processamento de linguagem natural (do inglês *Natural Language Processing* - NLP) é um campo de inteligência artificial e ciência de dados em rápido desenvolvimento que lida com tecnologias de processamento de fala e texto. O objetivo desta área de conhecimento é o desenvolvimento de métodos para análise automática e apresentação da linguagem humana (CAMBRIA; WHITE, 2014). A NLP usa uma variedade de metodologias para interpretar as ambiguidades na linguagem humana, incluindo sumarização automática, marcação de parte da fala, desambiguação, extração de entidade e relação, análise de sentimento, compreensão de linguagem natural e reconhecimento de fala.

2.1.3 Bag of words

O *Bag of words* (BoW) é uma técnica primordial de NLP usada para representar os textos a partir da contagem do número de ocorrências de cada palavra do corpus utilizado em cada instância da coleção. Conforme apresentado na Tabela 3, esta técnica não consegue capturar semântica das palavras, a ordem dos termos é ignorada, é bastante esparsa e possui alta dimensionalidade já que o tamanho de cada vetor é igual ao tamanho do vocabulário do corpus.

Tabela 3 – Exemplo da representação do texto com a técnica *Bag of Words*.

	banco	com	da	do	estou	gerente	no	praça	só
Estou no banco da praça	1	0	1	0	1	0	1	1	0
Só com gerente do banco	1	1	0	1	0	1	0	0	1
Estou com a gerente do banco no banco da praça	2	1	1	1	1	1	1	1	0

2.1.4 TF-IDF

Também considerada uma técnica de *Bag of Words*, a técnica Frequência do Termo - Frequência Inversa do Documento (do inglês *Term Frequency - Inverse Document Frequency* - TF-IDF) é uma estatística numérica que pretende indicar o quão importante uma palavra ou *token* é para um documento em um corpus (LESKOVEC; RAJARAMAN; ULLMAN, 2014). O valor é calculado multiplicando-se o valor de $tf_{t,d}$, o número de ocorrências de um termo t em documento d, pelo $idf_{t,D}$, o logaritmo da fração inversa de documentos d que contém aquele termo no corpus D, ou seja,

$$idf_{t,D} = log\frac{N}{|d \in D : t \in d|}, \tag{2.1}$$

onde N é o número total de documentos no corpus. Desta forma, o TF-IDF é dado pela Equação 2.2 (MANNING; RAGHAVAN; SCHüTZE, 2008)

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_{t,D}. \tag{2.2}$$

Assim, ao contrário do *Bag of Words*, o TF-IDF cria uma contagem normalizada em que cada contagem de palavras é dividida pelo número de documentos em que essa palavra aparece. No entanto, conforme apresentado na Tabela 4, esta técnica, assim como BoW, não consegue capturar semântica das palavras, a ordem dos termos é ignorada, é bastante esparsa e possui alta dimensionalidade já que o tamanho de cada vetor é igual ao tamanho do vocabulário do corpus.

Tabela 4 – Exemplo da representação do texto com a técnica TF-IDF.

	banco	com	da	do	estou	gerente	no	praça	só
Estou no banco da praça	0,362	0,000	0,466	0,000	0,466	0,000	0,466	0,466	0,000
Só com gerente do banco	0,336	0,433	0,000	0,433	0,000	0,433	0,000	0,000	0,569
Estou com a gerente do banco no banco da praça	0,506	0,326	0,326	0,326	0,326	0,326	0,326	0,326	0,000

Para possibilitar a redução de dimensionalidade da representação vetorial gerada pelo TF-IDF, algumas técnicas podem ser aplicadas como por exemplo remoção de *stop words* (artigos, preposições, etc - conforme apresentado na Tabela 5), stemização (redução à raiz da palavra) e lematização (retirada das inflexões das palavras).

Tabela 5 – Exemplo da representação do texto sem os *stop words* com a técnica TF-IDF.

	banco	gerente	praça
Estou no banco da praça	0,613	0,000	0,790
Só com gerente do banco	0,613	0,790	0,000
Estou com a gerente do banco no banco da praça	0,739	0,476	0,476

2.1.5 Word embeddings

Collobert et al. (2011) propuseram os *word embeddings*, as primeiras técnicas de criação de representação de palavras em espaço vetorial através de dados não anotados. Para geração de um *word embedding* são consideradas as palavras próximas a palavra alvo, fazendo que o contexto seja capturado sem a necessidade de dados anotados por especialistas. Dessa forma, representações de palavras semelhantes ocupam espaços vetoriais próximos e portanto, torna-se possível calcular o grau de similaridade das palavras e realizar operações matemáticas com a representação vetorial das palavras para encontrar a representação vetorial da palavra destino, por exemplo:

$$vetor(rainha) - vetor(mulher) + vetor(homem) \approx vetor(rei)$$

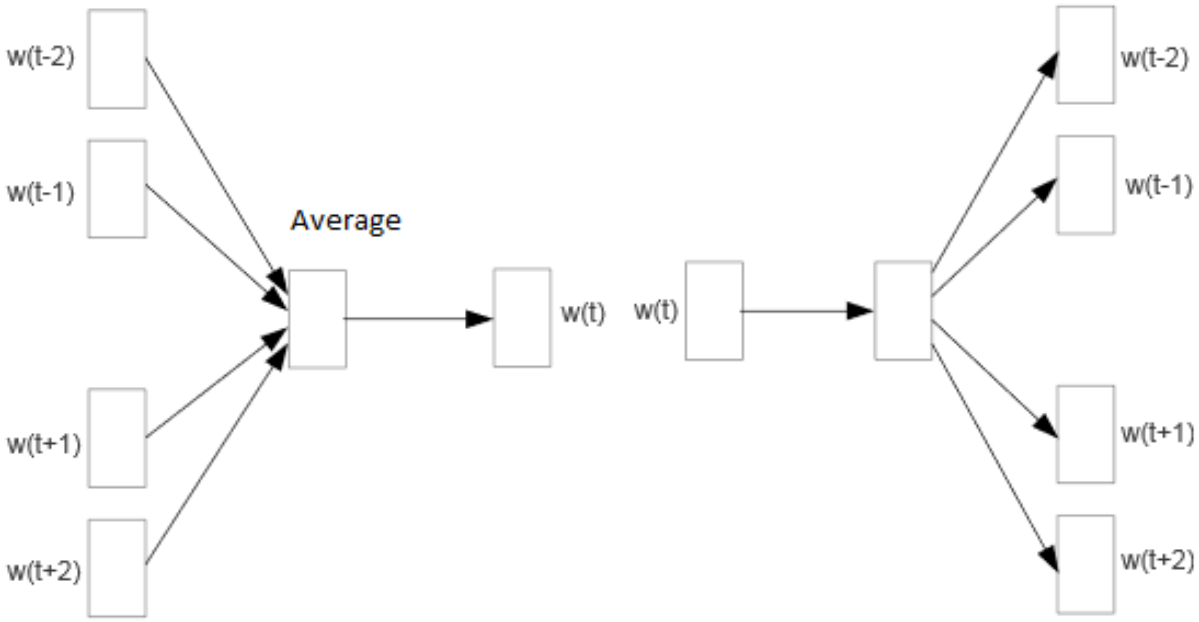
2.1.6 Word2Vec

Mikolov et al. (2013) propuseram uma técnica de aprendizado não-supervisionado, intitulada de Word2Vec, para geração de *word embedding* com duas possíveis abordagens detalhadas a seguir e apresentada na Figura 1.

- *Continuous Bag of Words* (CBoW): objetiva predizer uma palavra alvo a partir de uma quantidade de palavras de contexto ("janela de contexto"), localizadas antes e depois dessa palavra;
- Skip-gram: objetiva prever uma quantidade de palavras de contexto a partir de uma palavra de entrada (alvo).

De acordo com o artigo original, Mikolov et al. (2013), o CBoW possui um tempo de treinamento menor que o Skip-gram e pode representar melhor as palavras mais frequentes. No entanto, o Skip-gram pode representar melhor palavras menos frequentes e funciona bem com pequenos conjuntos de dados. Outra diferença, é que o CBoW captura melhor as relações sintáticas entre as palavras. Por sua vez, o Skip-gram consegue captura melhor as relações semânticas entre palavras. Para melhor ilustrar a diferença, a partir da palavra "filho", o CBoW identifica vetores próximos como "filhos" e "filha" ou "filhas". Enquanto que o Skip-gram obtém vetores próximos como "criança", "bebê" ou "prole".

Figura 1 – Arquitetura Word2Vec.



(a) CBoW (b) Skip-gram

Fonte: (MIKOLOV et al., 2013).

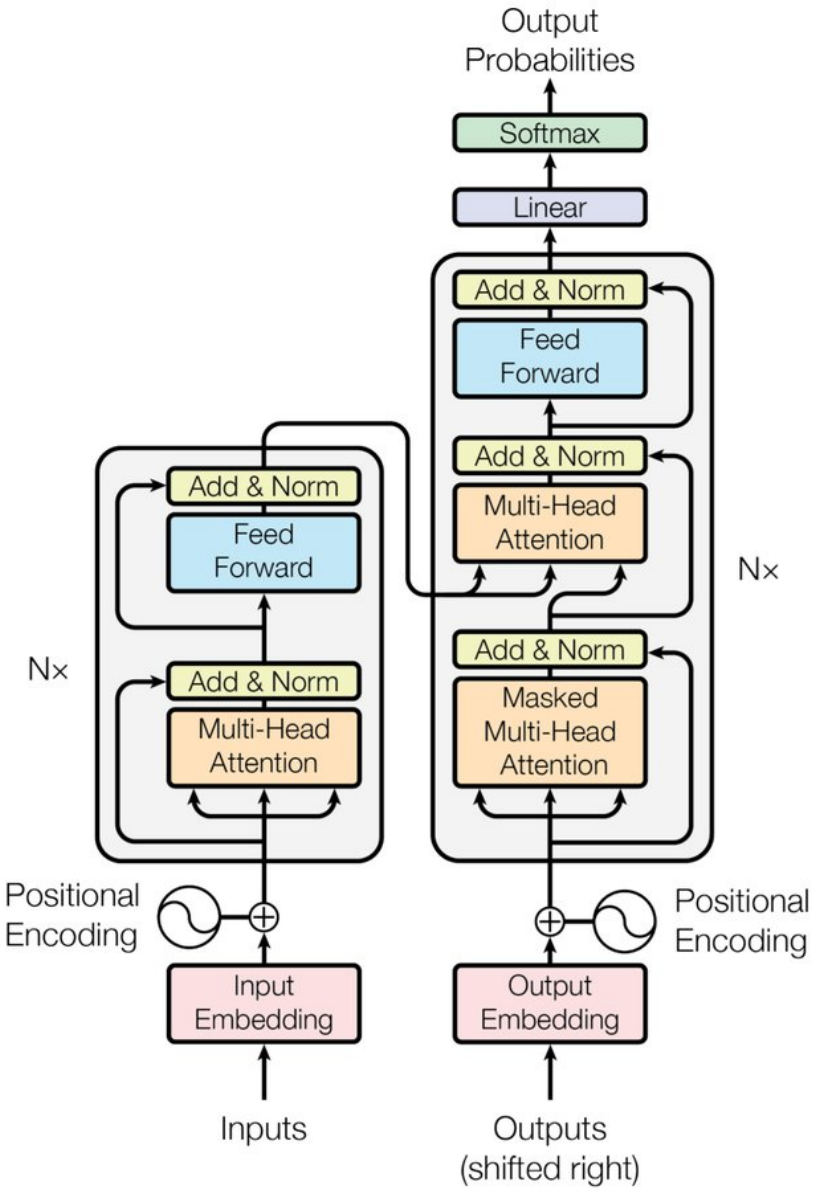
O Word2Vec produz *word embedding* livres de contexto, ou seja, cada palavra do seu vocabulário é representada pelo mesmo vetor numérico independente de qual é o sentido da palavra em uma determinada frase. Então, sua maior desvantagem é não conseguir diferenciar palavras que são homógrafas, por exemplo a palavra "banco" nas frases "Estou no banco da praça" e "Só com gerente do banco" possuem o mesmo *word embedding*. Outro ponto fraco é que o modelo baseado em Word2Vec não consegue criar *word embedding* para palavras que não estão presentes em seu vocabulário.

2.1.7 Arquitetura Transformers

Proposta por Vaswani et al. (2017), a arquitetura *Transformers* originalmente é composta por seis *encoders* e seis *decoders* idênticos que não compartilham dos mesmos pesos, possuindo 110 milhões de parâmetros. Conforme ilustrado na Figura 2, cada *encoder* possui uma camada de Auto-Atenção, *MultiHead Self-Attetion*, e uma de *feed-forward*. O *Transformer* recebe como entrada uma sequência de palavras e na sua entrada está presente o *Positional Encoding*, responsável por atribuir a posição de cada palavra em uma sentença, preservando assim a ordem das palavras. Além disso, o *Positional Encoding* é adicionado no final ao *word embedding* correspondente de cada palavra.

A camada de Auto-Atenção é responsável por gerar um *embedding* para cada palavra. A representação de uma palavra é baseada na soma ponderada de todas as outras palavras da sequência, onde as palavras mais importantes para a palavra alvo,

Figura 2 – Arquitetura do modelo *Transformer*. O *encoder* pode ser visto à esquerda e o *decoder* à direta.



Fonte: (VASWANI et al., 2017)

receberão maior peso. A ideia desse mecanismo é sinalizar que o sentido da palavra alvo pode ser melhor explicado olhando as outras palavras da sequência, guardando em cada representação informação contextual.

A camada de Auto-Atenção vai pegar cada *word embedding* e multiplicar por três matrizes iniciadas com pesos aleatórios: *Query* (Q), *Keys* (K) e *Values* (V). Então, cada *word embedding* vai dar origem a três vetores : Q, K e V. O vetor Q será multiplicado por cada vetor K das demais palavras da sequência e assim será calculado um score para cada palavra em relação a palavra alvo. Em seguida, esse score é dividido por uma constante e é aplicada uma camada de *Softmax* para normalizar esse valor entre 0 e 1. Por fim, é

realizada a multiplicação de cada score da palavra pelo vetor V correspondente. Os vetores com valores mais altos representam as palavras mais relevantes. A função de atenção está representada a seguir na Equação 2.3. A constante $\sqrt{d_k}$ representa a raiz quadrada da dimensão do vetor e é um hiperparâmetro do modelo.

$$Atenção(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.3}$$

A camada de *feed-foward*, por sua vez, recebe como entrada a saída produzida pelo mecanismo de atenção e a envia para o próximo *encoder*. A estrutura do *decoder* é similar ao do *encoder*, com a adição de uma camada intermediária de atenção que sinalizará ao *decoder* para "ficar atento" nas palavras mais relevantes da sentença. Após o último *decoder*, existe uma camada de transformação linear e uma de *Softmax*.

A partir da proposição desta nova arquitetura de geração de *word embedding*, ao contrário de arquiteturas de Redes Neurais Recorrentes, permitiu a realização do treinamento em paralelo para várias sentenças. Desta forma, tal independência, viabilizou o treinamento desses modelos em grandes corpus, tornando-os bem poderosos. Como consequência, passamos a conseguir trabalhar com estes grandes modelos pré-treinados usufruindo da aprendizagem por transferência em qualquer tarefa de NLP.

2.1.8 Aprendizagem por Transferência - *Transfer Learning*

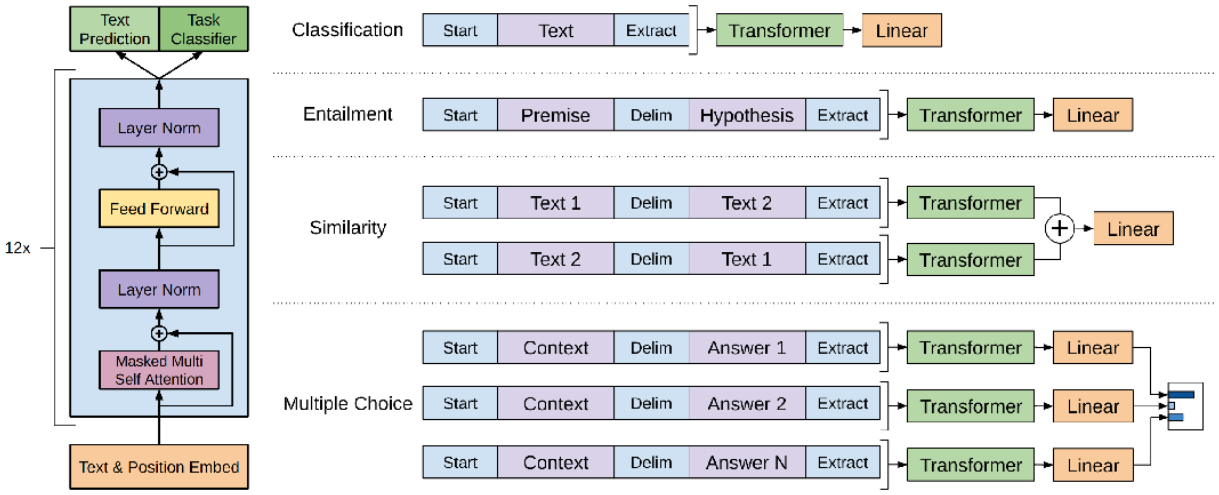
Pesquisas mostram que a aprendizagem de transferência, ou adaptação de domínio, tem sido aplicada em uma variedade de áreas onde existe sinergia entre conjuntos de dados coletados independentemente, possibilitando ganhos significativos em relação a qualidade do modelo, ao tempo de processamento e ao custo computacional (MCCANN et al., 2018). Saenko et al. (2010) adaptaram modelos de reconhecimento de objetos desenvolvidos para um domínio visual para as novas condições de imagem.

Inspirados pela transferência bem-sucedida de Redes Neurais Convolucionais (do inglês *Convolutional Neural Network* - CNN) treinadas em ImageNet para outras tarefas em visão computacional, algumas pesquisas em NLP aprofundaram os estudos neste tema. Collobert et al. (2011) aproveitaram as representações aprendidas com o aprendizado não supervisionado para melhorar o desempenho em tarefas supervisionadas, como reconhecimento de entidade nomeada, marcação de parte da fala e agrupamento. McCann et al. (2018) propuseram um método de transferência de representações de nível superior de vetores de palavras usando um método supervisionado para treinar codificador de frases e mostrar que melhora os modelos para classificação de texto e resposta a perguntas sem ajuste fino.

2.1.9 *Generative Pre-Training* - GPT

Proposto por Radford et al. (2018), pesquisadores da OpenAI (laboratório de pesquisa especializado em Inteligência Artificial fundado em 2015), o modelo *Generative Pre-Training* (GPT) foi baseado na arquitetura *Transformers*. A arquitetura proposta utiliza somente o *decoder* com 12 camadas de atenção mascaradas em apenas uma direção. O procedimento de treinamento, conforme apresentado na Figura 3, consistiu em dois estágios (i) treinamento não supervisionado e (ii) ajuste fino supervisionado, utilizando os dados do BooksCorpus (ZHU et al., 2015) que continha mais de 7.000 livros únicos dos mais gêneros.

Figura 3 – Arquitetura do modelo GPT. A estrutura de pré-treinamento pode ser visto à esquerda e a estrutura de ajuste fino à direta.



Fonte: (RADFORD et al., 2018)

Posteriormente Radford et al. (2019) aprimoram a arquitetura GPT a partir do treinamento com textos de mais de oito milhões de documentos extraídos de 45 milhões de links associados a tokenização com a técnica *Byte Pair Encoding* (SENNRICH; HADDOW; BIRCH, 2016). A nova arquitetura proposta, intitulada de GPT-2, ofereceu quatro modelos de tamanhos diferentes conforme apresentado na Tabela 6.

Tabela 6 – Arquitetura de hiperparâmetros para os quatro modelos GPT-2.

Parâmetros	Camadas	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Mais recentemente a terceira geração do GPT, GPT-3, foi proposta por Brown et

al. (2020) sendo um modelo de de linguagem auto regressivo treinado com 175 bilhões de parâmetros, 10 vezes maior que os modelos de linguagem prévios.

2.1.10 *Bidirectional Encoder Representations from Transformers* - BERT

Devlin et al. (2018) propuseram a primeira arquitetura baseada em *Transformers* que conseguiu capturar o contexto da palavra alvo de maneira bidirecional. Conforme ilustrado na Figura 4 existem na estrutura do *Bidirectional Encoder Representations from Transformers* (BERT) duas etapas, quais sejam, pré-treinamento e ajuste fino. Durante o pré-treinamento, o modelo é treinado em dados não rotulados utilizando a estrategia de mascaramento dos dados (do inglês *Masked Language Model* - MLM). Para o ajuste fino, o modelo BERT é inicializado primeiro com os parâmetros pré-treinados e todos os parâmetros são ajustados usando dados rotulados para tarefas específicas de NLP, como por exemplo (i) Pergunta e Resposta (do inglês *Question Answering* - SQuAD); (ii) Reconhecimento de Entidade Nomeada (do inglês *Named Entity Recognition* - NER); e (iii) Inferência de Linguagem Natural (do inglês *Multi-Genre Natural Language Inference* - MNLI), em que cada uma destas tarefas específicas possuem modelos ajustados separadamente.

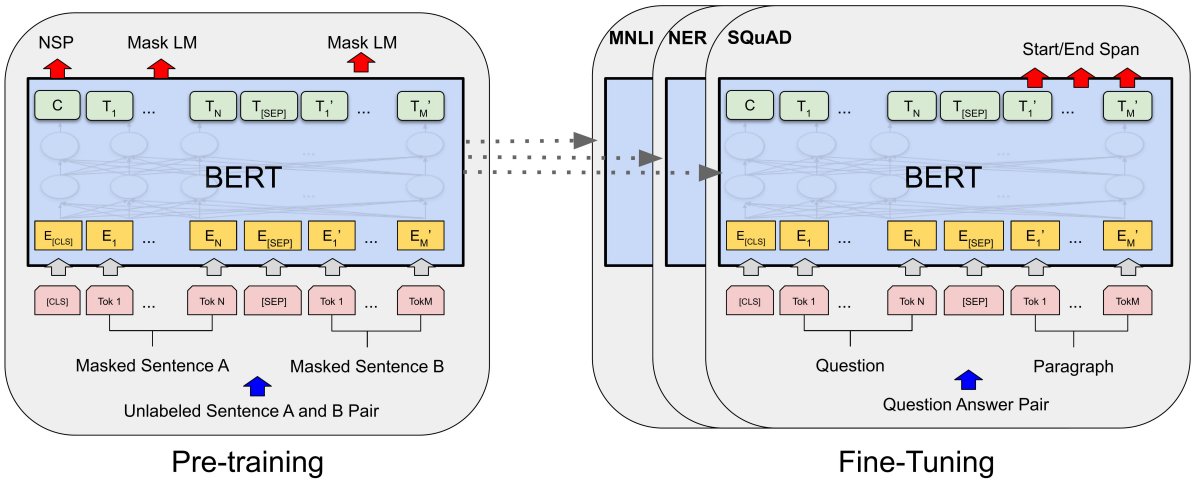
- *Question Answering* (SQuAD): Em tarefas de resposta a perguntas, o software recebe uma pergunta referente a uma sequência de texto e é solicitado a marcar a resposta na sequência. Usando o BERT, um modelo de perguntas e respostas pode ser treinado aprendendo dois vetores extras que marcam o início e o fim da resposta.

- *Named Entity Recognition* (NER): No Reconhecimento de Entidades Nomeadas, o software recebe uma sequência de texto e é obrigado a marcar os diversos tipos de entidades (Pessoa, Organização, Data, etc) que aparecem no texto. Usando BERT, um modelo NER pode ser treinado alimentando o vetor de saída de cada *token* em uma camada de classificação que prevê o rótulo NER.

- *Multi-Genre Natural Language Inference* (MNLI): As tarefas de classificação, como a análise de sentimentos, são feitas de maneira semelhante à classificação da próxima sentença, adicionando uma camada de classificação sobre a saída do *Transformer* para o *token* [CLS].

Ao treinar modelos de linguagem, há um desafio de definir uma meta de previsão, então antes de alimentar sequências de palavras no BERT, 15% das palavras em cada sequência são substituídas por um *token* [MASK]. O modelo então tenta prever o valor original das palavras mascaradas, com base no contexto fornecido pelas outras palavras não mascaradas na sequência.

Figura 4 – Arquitetura do modelo BERT. A estrutura de pré-treinamento pode ser visto à esquerda e a estrutura de ajuste fino à direta.



Fonte: (DEVLIN et al., 2018)

Os pesquisadores da arquitetura BERT, especificaram dois modelos (i) versão Base e (ii) versão *Large* os treinando com textos do *English Wikipedia* com 2.500 milhões de palavras e do BooksCorpus com 800 milhões de palavras (ZHU et al., 2015) associados a tokenização com a técnica WordPiece (WU et al., 2016). A versão Base foi especificada com uma pilha de doze codificadores, doze mecanismos de atenção, possuindo 110 milhões de parâmetros e foi escolhido para ter o mesmo tamanho de modelo do GPT para fins de comparação, conforme sumarizado na Tabela 7. Já a versão *Large* foi especificada com vinte e quatro codificadores, dezesseis mecanismos de atenção e possuindo 335 milhões de parâmetros. As camadas de *feed-foward* são de 768 para a versão Base e 1024 para a versão *Large*.

Tabela 7 – Tabela comparativa entre BERT e GPT.

BERT	GPT
Bidirecional, ou seja, usa todo o contexto circundante	Autorregressivo, ou seja, usa o contexto da(s) palavra(s) prévias
Aplicada em pesquisa aprimorada, assistente de voz, análise de opiniões, etc	Aplicada em geração de código automatizado, escrita de artigos, geração de documentos legais, etc
Usa duas tarefas não supervisionadas em conjunto *Masked language modeling* e *Next Sentence prediction*	Geração de texto direta usando modelagem de linguagem autorregressiva

2.1.11 *Robustly optimized BERT approach* - RoBERTa

Liu et al. (2019) propuseram o *Robustly optimized BERT approach* (RoBERTa), modelo baseado no modelo BERT com as seguintes modificações:

1. Treinar o modelo por mais tempo, com lotes maiores, sobre mais dados;
2. Remover o objetivo de previsão da próxima frase;
3. Treinar utilizando sequências mais longas;
4. Alterar dinamicamente o padrão de mascaramento para os dados de treinamento;
5. Treinar com um grande novo conjunto de dados (CC-NEWS) (NAGEL, 2016);
6. Tokenizar utilizando a técnica *Byte-Pair Encoding* (BPE).

De acordo com os autores, os resultados alcançados a partir das modificações detalhadas apresentadas acima forneceram uma grande melhoria de performance quando comparado com os resultados originais relatados para o modelo BERT *Large*.

2.2 Estado da Arte

Pesquisas mais recentes têm sustentado que algoritmos de aprendizado de máquina possuem um grande potencial para resolver problemas de alta complexidade no contexto Legal (SIL et al., 2019). No contexto desta pesquisa, foi revisada a literatura em busca das produções mais recentes, no período de 2017 a 2022, através das base de dados (i) *Google Scholar*; (ii) *Science Direct*; e (iii) *IEEE Xplorer*, sobre algoritmos de aprendizagem de máquina não supervisionados ou clusterização aplicados à área jurídica utilizando NLP.

As pesquisas revelaram que existem até o momento poucas produções que tratam do tema, o que comprova a sua complexidade. Expandindo a pesquisa para o uso de Processamento de Linguagem Natural aplicado à área judiciária, foi encontrada uma revisão sistemática da literatura dos desafios enfrentados pelo sistema de previsão de julgamento, o qual pode ajudar advogados, juízes e civis a prever a taxa de lucro ou perda, tempo de punição e artigos da lei aplicável a novos casos, usando o modelo de aprendizado profundo. Os pesquisadores descrevem em detalhes a Literatura Empírica sobre Métodos de Previsão de Julgamento Legal, a Literatura Conceitual sobre Métodos de Classificação de Texto e detalhes do modelo *transformers* (.G; JAYARAJU, 2021). Um diferencial relevante deste trabalho de mestrado é a publicação de uma metodologia da aplicação de técnicas NLP primordiais até as mais atuais.

Polo et al. (2021) propuseram um estudo com dados textuais de 6449 processos judiciais de diversos ramos da Justiça Brasileira para realizar classificação para indicar

se o processo está arquivado, ativo ou suspenso. Li et al. (2022) propuseram um modelo de classificação utilizando dados textuais de processos judiciais relacionados à causas que discutiam indenizações de acidentes na construção civil em Hong Kong para indicar se as compensações pedidas tiveram sucesso no julgamento ou não. Hassan e Le (2022) também utilizaram questões judiciais relacionadas à construção civil para propor meios para avaliar contratos e normas com intuito de reduzir disputas jurídicas causadas por contradições contidas nestes tipos de documentos. O presente trabalho de mestrado se diferencia em relação aos achados de pesquisa apresentados anteriormente por utilizar dados textuais da Justiça Trabalhista Brasileira e utilizar nove técnicas de NLP com o propósito de alcançar o agrupamento de processos.

Desta forma, buscou-se então ampliar a pesquisa retirando a restrição para a área jurídica, o que revelou algumas publicações. Em (RENUKA; KIRAN; ROHIT, 2021) foi discutido o uso de sistema de recomendação de conteúdo baseado em agrupamento, com o k-means, de artigos semelhantes através da transformação vetorial do conteúdo dos documentos com o TF-IDF (MACQUEEN, 1967). Em (D'SILVA; SHARMA, 2020), os autores realizaram uma sumarização automática de textos usando o TF-IDF e o k-means para determinar os grupos de sentença dos documentos utilizados na formação do resumo. Verificou-se que essas pesquisas utilizaram o TF-IDF como técnica base para vetorizar o conteúdo textual e o k-means o algoritmo mais utilizado para o aprendizado de máquina não-supervisionado. Destaca-se que nesta presente dissertação utilizou-se, além do TF-IDF, o Word2Vec e a arquitetura *Transformers* para detectar padrões através do agrupamento de processos.

Supõe-se que escolher a melhor técnica de geração de *word embeddings* exige pesquisa, experimentação e comparação dos modelos. Muitas pesquisas recentes demonstraram a viabilidade de utilizar *word embeddings* para melhorar a qualidade dos resultados de algoritmos de IA para detecção de padrão, classificação, entre outros.

Mikolov et al. (2013) propuseram o Word2Vec Skip-gram e CBoW, duas novas arquiteturas para calcular representações vetoriais de palavras consideradas, à época, referência no assunto (MIKOLOV et al., 2013). Em seguida, *Embeddings from Language Models* (Elmo) (PETERS et al., 2018), Flair (AKBIK; BLYTHE; VOLLGRAF, 2018) e context2vec (MELAMUD; GOLDBERGER; DAGAN, 2016), bibliotecas baseadas na Rede *Long Short Term Memory* (LSTM) (HOCHREITER; SCHMIDHUBER, 1997) criaram um *word embeddings* distinto para cada ocorrência da palavra, relacionado ao contexto, que permitiu a captura do significado da palavra. Os modelos LSTM foram amplamente utilizados para reconhecimento de fala, modelagem de linguagem, análise de sentimentos e previsão de texto, e, diferente das Redes Neurais Recorrentes (do inglês *Recurrent Neural Network* - RNN), têm a capacidade de esquecer, lembrar e atualizar as informações dando assim um passo à frente das RNNs (SHERSTINSKY, 2020).

A partir de 2018, novas técnicas de geração de *word embeddings* surgiram, com destaque para: (i) *Bidirectional Encoder Representations from Transformers* (BERT) (DEVLIN et al., 2018), modelo sensível ao contexto com a arquitetura baseada em um modelo de *Transformers* (VASWANI et al., 2017); (ii) *Sentence* BERT (SBERT) (REIMERS; GUREVYCH, 2019), um modelo "siamês" do BERT que foi proposto para melhorar a performance do BERT quando se busca obter a similaridade de sentenças; (iii) *Text-to-Text Transfer Transformer* (T5) (RAFFEL et al., 2019), um *framework* para tratar problemas de NLP como problema de texto para texto, ou seja, entrada ao modelo como texto e a saída do modelo como texto; (iv) *Generative Pre-Training Transformer 2* (GPT-2), modelo baseado em *Transformers* com 1,5 bilhão de parâmetros (RADFORD et al., 2019); e (v) *Robustly optimized BERT approach* (RoBERTa), modelo baseado no modelo BERT, que foi treinado por mais tempo e que utilizou uma maior quantidade de dados (LIU et al., 2019).

A partir dessa análise, é possível verificar que a atual pesquisa avança no atual estado da arte na área de NLP aplicada ao setor jurídico, por desenvolver uma metodologia que se utiliza de técnicas de processamento de linguagem natural e aprendizagem de máquina profunda para o agrupamento automatizado de processos judiciais da Justiça Trabalhista Brasileira, realizando um estudo comparativo e aplicado das técnicas TF-IDF, Word2Vec CBoW, Word2Vec Skip-gram e *Transformers* (BERT, GPT-2 e RoBERTa), utilizando modelos para propósito geral em português brasileiro (ptBR) e modelos *Transformers* especializados no judiciário trabalhista, para a realização do agrupamento de processos jurídicos trabalhistas no Brasil utilizando o algoritmo k-means e a similaridade cosseno para auferir a qualidade dos agrupamentos.

Os dois próximos capítulos contêm publicações que apresentam em detalhes a metodologia desenvolvida e em seguida é apresentada a conclusão da dissertação com perspectivas de trabalhos futuros.

3 MANUSCRITO 1

Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches

Capítulo de Livro publicado em **IntechOpen**. Submetido em 3 de agosto de 2021. Revisado em 10 de agosto de 2021. Publicado em 6 de setembro de 2021.

DOI: 10.5772/intechopen.99875

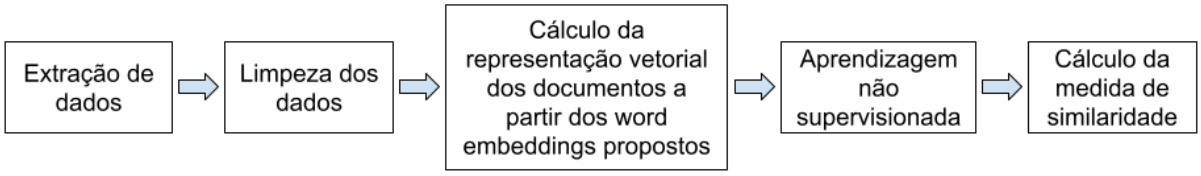
Esta pesquisa apresentou como a utilização de algoritmos de aprendizado de máquina associado a técnicas de Processamento de Linguagem Natural pode auxiliar o Judiciário Brasileiro a atender de forma célere a crescente demanda judicial. Assim, foi levantado e preparado um banco de dados de processos judiciais da Justiça Trabalhista Brasileira para possibilitar desenvolver e validar um estudo comparativo de técnicas tradicionais de aprendizado profundo para agrupamento de processo judiciais.

A partir dessa pesquisa, foi possível avançar no atual estado da arte na área de NLP aplicada ao setor jurídico, a partir do uso de três técnicas tradicionais de Processamento de Linguagem Natural, quais sejam (i) TF-IDF; (ii) Word2Vec CBoW treinado para fins gerais para o Português Brasileiro (HARTMANN et al., 2017); e (iii) Word2Vec Skip-gram treinado para fins gerais para o Português Brasileiro (HARTMANN et al., 2017), associadas à utilização do k-means (MACQUEEN, 1967), um algoritmo de aprendizado de máquina não supervisionado, para realizar o agrupamento de processos jurídicos trabalhistas Brasileiro.

Foi realizada uma revisão do estado da arte em busca de trabalhos científicos relacionados à utilização de algoritmos de aprendizagem de máquina não supervisionados aplicados à área jurídica utilizando Processamento de Linguagem Natural e até momento da publicação desta pesquisa não foi encontrado muitas produções que tratassem do tema em destaque.

Para garantir a reprodutibilidade da pesquisa, a metodologia apresentada foi composta pelas fases (i) extração de dados; (ii) limpeza de dados; (iii) geração de modelos de *word embeddings*; (iv) cálculo da representação vetorial do documento; (v) aprendizagem não supervisionada; e (vi) cálculo da medida de similaridade; conforme ilustrada na Figura 5.

Figura 5 – Diagrama da metodologia aplicada no Capítulo de Livro "*Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches*".



Fonte: Autoria Própria.

Nesta pesquisa de todas as técnicas avaliadas, a técnica Word2Vec Skip-gram ptBR apresentou-se como melhor opção de *word embeddings* para clusterização de documentos judiciais do tipo Recurso Ordinário Interposto.

Chapter

Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches

Raphael Souza de Oliveira
and Erick Giovani Sperandio Nascimento

Abstract

The Brazilian legal system postulates the expeditious resolution of judicial proceedings. However, legal courts are working under budgetary constraints and with reduced staff. As a way to face these restrictions, artificial intelligence (AI) has been tackling many complex problems in natural language processing (NLP). This work aims to detect the degree of similarity between judicial documents that can be achieved in the inference group using unsupervised learning, by applying three NLP techniques, namely term frequency-inverse document frequency (TF-IDF), Word2Vec CBoW, and Word2Vec Skip-gram, the last two being specialized with a Brazilian language corpus. We developed a template for grouping lawsuits, which is calculated based on the cosine distance between the elements of the group to its centroid. The Ordinary Appeal was chosen as a reference file since it triggers legal proceedings to follow to the higher court and because of the existence of a relevant contingent of lawsuits awaiting judgment. After the data-processing steps, documents had their content transformed into a vector representation, using the three NLP techniques. We notice that specialized word-embedding models—like Word2Vec—present better performance, making it possible to advance in the current state of the art in the area of NLP applied to the legal sector.

Keywords: legal, natural language processing, clustering, TF-IDF, Word2Vec

1. Introduction

In recent years, the Brazilian Judiciary has been advancing toward turning all its acts digital. Following this direction, the Brazilian Labour Court implemented in 2012 the Electronic Judicial Process (acronym in Portuguese for "*Processo Judicial Eletrônico*"—PJe), and from this date, all new legal proceedings have already been born electronic. According to the Annual Analytical Report of Justice in Numbers 2020 (base year 2019) [1], produced by the National Council of Justice (acronym in Portuguese for "*Conselho Nacional de Justiça*"—CNJ), more than 99% of the ongoing cases are already on this platform.

Knowing that human beings cannot promptly analyze a large set of data, especially when such data do not appear to correlate, a way to assist in the pattern-recognition process is through statistical, computational, and data analysis methods. From the perspective that an exponential increase in textual data exists, the analysis of patterns in legal documents has become increasingly challenging.

Currently, one of the major challenges in the legal area is to respond quickly to the growing judicial demand. The Brazilian legal system provides for ways to ensure the swift handling of judicial proceedings, such as the principle of the reasonable duration of a case, the principle of speed, the procedural economy, and due process to optimize the procedural progress [2]. Therefore, with the aid of some clustering mechanism, that is, the grouping of processes, with a good rate of similarity between the documents to be analyzed, it was possible to help in the distribution of work among the advisors of the office for which the process was drawn. In addition, it contributed to the search for case law[1] for the judgment of the cases in point, to ensure a speedy trial, upholding the principle of legal certainty. According to Gomes Canotilho [3]:

> *"The general principle of legal certainty in a broad sense (thus encompassing the idea of trust protection) can be formulated as follows: the individual has the right to be able to rely on the law that his acts or public decisions involved in his rights, positions or legal relations based on existing legal norms and valid for those legal acts left by the authorities on the basis of those rules if the legal effects laid down and prescribed in the planning are connected to the legal effects laid down and prescribed in the legal order" (2003, p. 257).*

Thus, this legal management tool created positive impacts such as the decrease of the operational costs of a legal proceeding, as a result of reducing its duration, meaning lower expenses on the allocation of the necessary resources for its judgment.

Recently, machine learning algorithms have demonstrated through research that they are powerful tools capable of solving high-complexity problems using natural language processing (NLP) [4]. In this sense, it is possible to highlight the works of [5–9], which apply the techniques of word-embedding generation, a form of vector representation of terms, and consequently of documents, taking into account their context. The use of these word embeddings is essential when analyzing a set of unstructured data presented in the form of large-volume documents in court.

Nowadays, a specialist screens the documents and distributes among the team members the legal proceedings to be judged, setting up a deviation from the main activity of this specialist, which is the production of draft decisions. This contributed to an increase in the congestion rate (an indicator that measures the percentage of cases that remain pending solution at the end of the base year) and to the decrease in the meeting of demand index (acronym in Portuguese for "*Índice de Atendimento à Demanda*"—IAD—an indicator that measures the percentage of proceedings in downtime, compared to the number of new cases). It becomes evident in the consolidated data of the Labor Justice contained in **Table 1**, with data extracted from the Annual Analytical Report of Justice in Numbers 2020 (base year 2019) [1] produced by the National Council of Justice (CNJ).

This work aims, therefore, to present the degree of similarity between the judicial documents that was achieved in the inferred groups through unsupervised learning *via* the application of three techniques of NLP, namely: (i) term frequency-inverse document frequency (TF-IDF); (ii) Word2Vec with CBoW (continuous

[1] A legal term meaning a set of previous judicial decisions following the same line of understanding.

	Description	2° Degree	1° Degree	Total
Workforce				
Magistrates	Legal authority	559	3077	3636
Legal workers	Public administration employee	6911	22,785	29,696
Legal load handling				
Stockpile	Number of pending cases	792,223	3,741,548	4,533,771
New cases	Number of new cases	898,104	2,632,093	3,530,197
Judged	Number of cases judged	989,324	3,036,686	4,026,010
Closed	Number of cases with final decision	941,356	3,244,652	4,185,708
Productivity indexes				
IAD	Closed cases/new cases	104.8%	123.3%	118.6%
Congestion tax	Closed cases/(new cases + stockpile)	45.7%	53.6%	52.0%
Knowledge	Fact awareness phase	—	35.1%	35.1%
Execution	Judgment enforcement phase	—	72.7%	72.7%
Indexes per magistrate				
New cases	Average number of new cases per magistrate	1607	662	821
Workflow	Average number of cases per magistrate	3583	2794	2,927
Judged cases	Average number of cases judged per magistrate	1770	1103	1216
Closed cases	Average number of cases closed per magistrate	1684	1179	1264
Indexes per legal worker				
New cases	Average number of new cases per worker	135	83	95
Judged cases	Average number of cases judged per worker	300	351	339
Closed cases	Average number of cases closed per worker	141	148	146

Table 1.
Report of indicators of Brazilian labor justice.

bag of words) trained for general purposes for the Portuguese language in Brazil (Word2Vec CBoW pt-BR); and (iii) Word2Vec with Skip-gram trained for general purposes for the Portuguese language in Brazil (Word2Vec Skip-gram pt-BR).

This degree of congruence signals the model's performance and is set from the average similarity measure of the grouped files, based on the similarity cosine between the elements of the group to its centroid and, comparatively, by the average cosine similarity among all the documents of the group.

Aiming to delimit the scope of this research, a dataset containing information from documents of the Ordinary Appeal Interposed (acronym in Portuguese for "*Recurso Ordinário Interposto*"—ROI) type was extracted from approximately 210,000 legal proceedings. The Ordinary Appeal Interposed was used as a reference, as this is usually the type of document that induces the legal proceedings for judgment in the higher instance (2nd degree), thus instituting the Ordinary

Appeal (acronym in Portuguese for "*Recurso Ordinário*"—RO). That is a free plea, an appropriate appeal against definitive and final judgments proclaimed at first instance, seeking a review of the judicial decision drawn up by a hierarchically superior body [10].

For the present work, a literature review on unsupervised machine learning algorithms applied to the legal area was performed, using NLP, and an overview of recent techniques that use artificial intelligence (AI) algorithms in word-embedding generation. Then, we applied some methods until the results were obtained, comparing and discussing them, and finally, conclusions and future challenges were presented.

2. State-of-the-art review

Machine learning algorithms have in the most recent research demonstrated a great potential to solve high-complexity problems, which follow the categories into (i) supervised machine learning algorithms; (ii) unsupervised; (iii) semi-supervised; and (iv) by reinforcement [11]. In the context of this chapter, the literature review focused on the search for the most recent research on unsupervised machine learning or clustering algorithms applied to the legal area using NLP.

The investigation revealed that there are not many works dealing with the highlighted topic, which proves its complexity. Thus, we sought to expand the research by removing the restriction to the legal area bringing light to other publications. In [12], we discussed the content recommendation system approaches based on grouping for similar articles that used TF-IDF to perform vector transformation of the document contents and, through cosine similarity, applied k-means [13] for clustering them. In [14], the authors automatically summarized texts using TF-IDF and k-means to determine the document's textual groups used to create the abstract. Then, TF-IDF is considered the primary technique for vectorizing textual content and k-means the most used algorithm for unsupervised machine learning.

Therefore, we can assume that choosing the best technique of generating word embeddings requires investigation, experimentation, and comparison of models. Several recent pieces of research have demonstrated the feasibility of using word embeddings to improve the quality of AI algorithm results for pattern detection, classification, among other uses.

In 2013, Mikolov et al. [6] proposed two new architectures to calculate vector representations of words calling them Word2Vec, which was considered, at the time, as a reference in the subject. Subsequently, techniques of word embeddings based on the use of the long short-term memory network (LSTM) [15] became widely used for speech recognition, language modeling, sentiment analysis, and text prediction, and that, unlike the recurrent neural network (RNN) they can forget, remember and update the information thus taking a step forward from the RNNs [16]. Therefore, LSTM-based libraries, such as Embeddings from Language Models (Elmo) [17], Flair [18], and context2vec [19] created a different word embedding for each occurrence of the word, related to the context, that allowed to capture the meaning of the word.

In more recent years, new techniques of word embeddings have emerged, with emphasis on (i) Bidirectional Encoder Representations from Transformers (BERT) [9], context-sensitive model with architecture based on a transformer model [20]; (ii) Sentence BERT (SBERT) [21], a "Siamese" BERT model that was proposed to improve BERT's performance when seeking to obtain the similarity of sentences; and (iii) Text-to-Text Transfer Transformer (T5) [22], a framework for treating NLP issues as a text-to-text problem, that is, input to the template as text and template output as text.

From this analysis, it was possible to advance in the current state of the art in the area of NLP applied to the legal sector, by conducting a comparative study and application of the techniques TF-IDF, Word2Vec CBoW, and Word2Vec Skip-gram to perform the grouping of labor legal processes in Brazil using the k-means algorithm and the cosine similarity.

3. Methodology

This section presents each step necessary to achieve the results and to make it possible to analyze them comparatively. To perform all the implementations of the routines necessary for this study, the Python programming language (version 3.6.9) was used and, among other libraries, (i) Numpy (version 1.19.2) was used; (ii) Pandas (version 1.1.3); (iii) Sklearn (version 0.21.3); (iv) Spacy (version 2.3.2); and (v) Nltk (version 3.5).

Every processing flow (pipeline) consists of the phases: (i) data extraction; (ii) data cleansing; (iii) generation of word-embedding templates; (iv) calculation of the vector representation of the document; (v) unsupervised learning; and (vi) calculation of the similarity measure, as detailed in the following subsections.

3.1 Data extraction

The dataset used for these studies belongs to the Regional Labour Court of the 5th Region (acronym in Portuguese for "Tribunal Regional do Trabalho da 5ª Região"—TRT5). There are approximately 210 (two hundred and ten) thousand documents of the Ordinary Appeal Interposed type, incorporated into the Electronic Judicial Process (PJe) system, originally added to the PJe in portable document format (PDF) or hypertext markup language (HTML). As the PJe has a tool for extracting and storing the contents of documents, there was no need for further processing in obtaining the text of such files.

In addition to the content of the documents, the following information was extracted: (i) the name of the parts of the proceedings to which such documents belonged; (ii) the list of labor justice issues from the Unified Procedural Table[2] (acronym in Portuguese for "*Tabela Processual Unificada*"—TPU) of the Labour Justice branch (made available by the National Council of Justice [CNJ] and consolidated by the Superior Labour Court [acronym in Portuguese for "*Tribunal Superior do Trabalho*"—TST]); and (iii) list of abbreviations (acronyms) with their full translation according to tables made available by the Supreme Court (acronym in Portuguese for "Supremo Tribunal Federal"—STF).[3]

3.2 Data cleaning

Preprocessing is a fundamental step for the application of artificial intelligence techniques and involves the following: (i) data standardization (when there is a large discrepancy between the values presented to the technique); (ii) the withdrawal of null values; and (iii) the reorganization and adequacy of the structure of the dataset. In this case, it is usually necessary for experts to conduct an exploratory analysis of the data used in advance to determine the direction of preprocessing.

[2] Labour Justice Unified Procedural Table. Available at: https://www.tst.jus.br/web/corregedoria/tabelas-processuais

[3] Table of abbreviations (and acronyms) made available by the Supreme Court. Available at: https://www.stf.jus.br/arquivo/cms/publicacaoLegislacaoAnotada/anexo/siglas_cf.pdf

For this phase, this study uses two forms of preprocessing: (i) detection of the subjects of the Unified Procedural Table (contained in the extracted documents) and (ii) cleaning the contents of the documents.

For the detection of the subjects of the TPU present in the extracted documents, regular expression matching was used as the search technique to measure the occurrences of these words in the files marking them with "tags" referring to the subject found.

For cleaning the contents of documents, usually using a regular expression, the steps were as follows:

- HTML tags: removed the html tags found in the document, such as <script>, <body>, <style> etc.;
- TPU subjects: replaced the subject text with a subject tag, for example, "*hora extra*" (overtime) changed to *hora_extra*;
- Related Persons: replaced the name of the individuals linked to the legal cases of the documents, such as the name of the author(s) and defendant(s), by the "tag" "*parteprocesso*" (part in the process);
- Judicial process number: replaced the number of the judicial process (according to the standard formatting defined nationally by the CNJ, NNNNNNN-NN. NNNN.N.NN.NNNN where N is a numeral) by the "tag" "*numeroprocesso*" (process number);
- Standardization of abbreviations: replacement of abbreviations (acronyms) by the full translation as drawn STF list as reported in Section 3.1, for example, CLT was transformed into "*Consolidação das Leis do Trabalho*" (Consolidated Labour Law);
- Addresses: replaced the addresses contained in the document with the "tag" "*enderecoprocesso*" (addresses in the process);
- Links: removed Internet links contained in the text;
- Date and Time: replacement of date and time content with "*datahora*" (datetime) tag;
- Time: replacement of the time content with the "*hora*" (hour) tag;
- Days of the week: removed the days of the week found in the document;
- Document ids: replacement of PJe document ids referenced in the document with "tag" "*sequenciadocumento*" (document sequence). These ids are typically composed of alphanumeric characters;
- Unit of measure: replaced the units of measurements and their values by the "tag" "*unidademedida*" (measurement unit);
- Numbers: replaced the numbers in full, ordinal numbers, and numerical sequences by the "tag" "*numeral*" (number);
- Judging bodies: replaced the judging bodies (e.g., "*Tribunal Regional do Traabalho*" [Regional Labour Court]) by the "*orgaojulgador*" (organjudge) tag;

- Months of the year: removed the months of the year found in the document;
- Judicial Stopwords: only when the technique employed is TF-IDF. The common words were removed in all texts of the judiciary, such as (i) "*magistrado*" (magistrate) and (ii) "*processo*" (legal proceeding), among others;
- Stopwords:
  - TF-IDF: removed all stopwords from the Portuguese language, such as "*de*" (from), "*da*" (of), "*a*" (the), "*o*" (the), "*esta*" (this) etc.;
  - Other techniques: removed only the non-adverbs of the Portuguese language, for example, the words "*não*" (no), "*mais*" (more), "*quando*" (when), "*muito*" (very), "*também*" (also), and "*depois*" (after) remain in the document;
- Line breaks: replaced line breaks by space;
- Punctuation marks:
  - TF-IDF: removed all the punctuation marks contained in the documents;
  - Other techniques: removed the punctuation marks except dot (.), comma (,), exclamation (!), and interrogation (?);
- Lemmatization:
  - TF-IDF: applied the technique to replace words with its root, for example, words such as "*tenho*" (have), "*tinha*" (had), and "*tem*" (have) had belong of the same root "*ter*" (have);
  - Other techniques: lemmatization has not been applied;

In addition to the preprocessing detailed above, when the technique used was TF-IDF, the tags inserted in the text during this phase were removed.

3.3 Generation of word-embedding templates

An essential technique in solving machine learning problems, involving NLP, is the use of vector representation of words, in which numerical values indicate some correlation of words in the text. This chapter uses word embeddings generated and shared for the Portuguese language, such as Word2Vec CBoW and Word2Vec template with Skip-gram. These templates were created based on more than 1 billion and 300,000 tokens, with results published in the article "Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks" presented at the Symposium in Information and Human Language Technology - STIL 2017 [23].

3.4 Calculation of the vector representation of the document

Different from the TF-IDF technique, which has the vector representation of the document based on the statistical measurement of each term of the document in relation to all known corpus, and whose vector dimension is equal to the size of

the vocabulary of the corpus, the other techniques (i) Word2Vec CBoW ptBR and (ii) Word2Vec Skip-gram pt-BR need to go through a change to calculate the vector representation of the document (document embeddings). This happens because for these techniques what you can get is the vector representation of the word (word embeddings).

Thus, to calculate the vector representation for the documents some alternatives are suggested, such as (i) average of the word embeddings of the words of the document; (ii) sum of the word embeddings of the words in the document by pondering them with the TF-IDF and then dividing by the sum of the TF-IDF of the words of the document; and (iii) weighted average with the TF-IDF of the word embeddings of the words of the document, the latter being the technique chosen for presenting the best result.

3.5 Unsupervised learning

The use of unsupervised learning techniques is relevant when the intention is to detect patterns among court documents. The k-means algorithm, whose basic concepts were proposed by MacQueen [13], is the technique adopted in this study. In general, this technique seeks to recognize patterns from the random choice of K initial focal points (centroid), where K is the number of groups that one wishes to obtain and, iteratively, position the elements whose Euclidean distance is the minimum possible concerning the centroid of the group.

Since one does not have an ideal K to offer the algorithm, an approach usually used to support such a decision is to calculate the inertia, based on how well the dataset was grouped through k-means.

The inertia calculation is based on the sum of the square of the Euclidean distance from each point to its centroid and seeks to obtain the lowest K with the lowest inertia. However, the higher the K value reaches, the tendency is that inertia will be lower, and then, the elbow method was used to find the point where the reduction in inertia begins to decrease.

Hence, 31 values for K were used within the range from 30 to 61, considering an interval for each unit, selecting the K that generated the best grouping. In addition, the strategy of creating submodels, limited to two, was used for the documents of the groups whose average similarity rate did not reach a value greater than 0.5.

3.6 Similarity measure calculation

The similarity measure is an important tool for the measurement of the quality of inferred groups. In this study, the cosine similarity measure is adopted, which is a measure that calculates the cosine of the angle between two vectors projected in the multidimensional plane, the result of which is between 0 and 1, in which 1 represents that the two vectors are totally similar, and 0 represents that they are totally different. Given two vectors, X and Y, the cosine similarity is presented using a scalar product according to Eq. (1).

$$\text{similarity} = \cos(\theta) = \frac{X \cdot Y}{|X| \cdot |Y|} \tag{1}$$

Consequently, to decide whether, after the clustering of the chief model, it was necessary to generate up to two more submodels, using the average cosine similarity among all elements of the group. Although the computational cost of calculating

the similarity between all files in the group is relevant, we sought to reduce the distance between documents that were part of the same group, although they were located near the centroid. To assess the final efficiency of the technique, another form of calculation was adopted, computing for each group the average cosine similarity between the group elements and its centroid. Thus, as a measure of global similarity of each approach, we calculated the average of the average of the groups, so that the one that reached a value closer to 1 (one) was considered the best technique.

4. Results and discussions

This research shows, as per the methodology presented in the previous sections, how machine learning algorithms associated with NLP techniques are important allies in optimizing the operational costs of the judicial process. It is evidenced from the result, for example, of document screenings and procedural distribution, which allows an expert to devote oneself to their chief activity optimizing working time.

While using the k-means unsupervised learning algorithm, it was necessary to choose the best K for each NLP technique studied. In this scenario, the elbow method was applied based on the calculated inertia of each of the 31 K tested, as shown in **Figure 1**, thus achieving a better result for each technique.

From the attainment of the best K, the k-means model was trained and, from the grouping performed by this technique, we could reach the average similarity between the documents of each group. Those groups that did not make the cutting line of at least 0.5 of average had the group files submitted for creating up to two

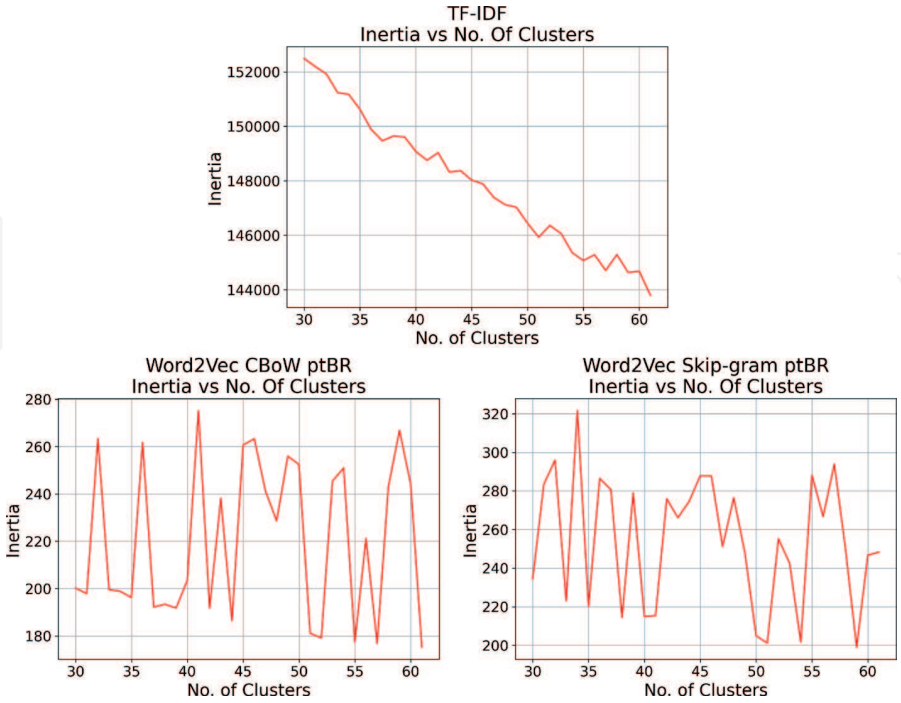


Figure 1.
Inertia charts constructed by using the elbow method for determining the best number of clusters for each approach.

submodels. As expected, only for TF-IDF technique groupings is there a need to generate submodels to improve performance.

Table 2 shows the average similarity of the groups obtained using the TF-IDF technique, as well as the result of the Word2Vec CBoW pt-BR technique. It achieved a little better measure of similarity than the Word2Vec Skip-gram pt-BR technique; however, the latter achieved its result with a smaller number of groups, which places it, in general, as the best technique.

After the groups were formed, the statistical data resulting from each approach were calculated, as shown in **Table 3** and in the comparative graph of distributions between the techniques (**Figure 2**). The cosine similarity of the group elements to

	Model		Submodel 1		Submodel 2		Final	
Type	**Groups**	**Mean**	**Groups**	**Mean**	**Groups**	**Mean**	**Groups**	**Mean**
TF-IDF	37	0.3696	43	0.4001	48	0.4002	48	0.4002
Word2Vec CBoW ptBR	59	0.9060	—		—		59	0.9060
Word2Vec Skip-gram ptBR	34	0.9044	—		—		**34**	**0.9044**

Table 2.
Mean cosine similarity between all elements of the group. The best results are highlighted in bold.

Type	**Groups**	**Mean**	**Std.**	**Min.**	**25%**	**50%**	**75%**	**Max.**
TF-IDF	49	0.6241	0.1718	0.2466	0.5021	0.5864	0.1639	0.9644
Word2Vec CBoW ptBR	59	0.9475	0.0632	0.7640	0.9352	0.9790	0.991	0.9999
Word2Vec Skip-gram ptBR	**34**	**0.9481**	**0.0609**	**0.7960**	**0.9248**	**0.9763**	**0.9924**	**0.9995**

Table 3.
Statistics of the cosine similarity of the group elements to the centroids. The best results are highlighted in bold.

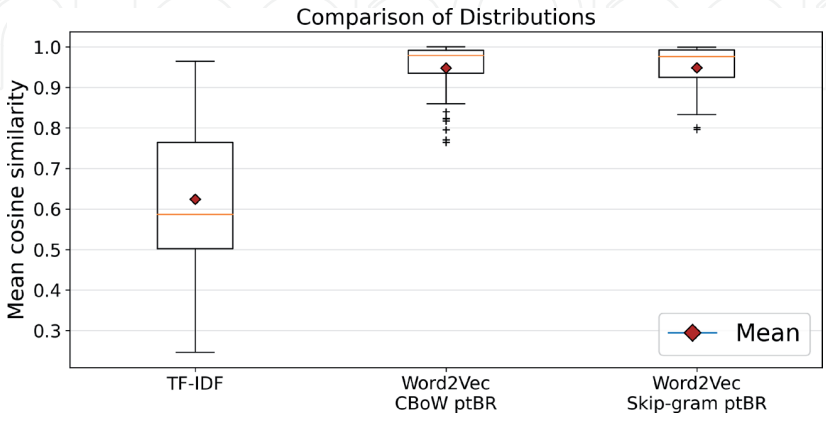


Figure 2.
Boxplots showing the distributions of the clusters calculated by each technique. The more cohesive the boxes and the less number of outliers, the better.

its centroid was used as a metric, showing the proximity of the results between the techniques with Word2Vec and highlighting the technique Word2Vec Skip-gram ptBR for the smaller amount of generated groups.

When comparing the values presented in **Tables 2** and **3**, it is noteworthy that the results presented in **Table 2** are worse in all cases. It is inferable from this observation that the similarity measure calculations shown in **Table 2** can reduce the similarity rates since there may be elements in the group positioned on completely opposite sides. From **Figure 2**, it is also possible to verify that the groupings generated by the Word2Vec technique were more cohesive than those generated by the TF-IDF technique, especially the Word2Vec Skip-gram technique, which created fewer groupings in the range of outliers than Word2Vec CBoW, demonstrating its superiority by allowing fewer groups but maintaining consistent quality and cohesion.

Given the aforesaid, among all the techniques evaluated, the Word2Vec Skip-gram pt-BR technique presented itself as the best option for word embeddings for clustering legal documents of the Ordinary Appeal Interposed type. Although the Word2Vec CBoW pt-BR technique achieves slightly better rates, it stands out from the previous one for reaching a much smaller number of groups.

The result achieved by each approach can be visualized by projecting in two dimensions of the groups formed from the three techniques: (i) TF-IDF; (ii) Word2Vec CBoW pt-BR; and (iii) Word2Vec Skip-gram pt-BR, respectively, presented in **Figures 3-5**. It is evident in the figures that the groups formed from Word2Vec are much better defined, especially skip-gram, which confirms the findings previously explained in this work.

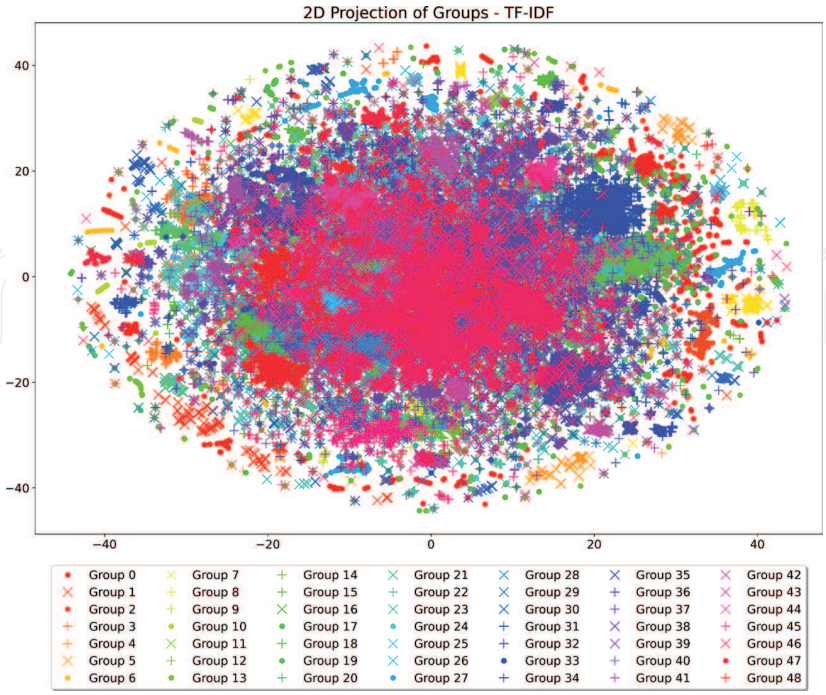


Figure 3.
2D projection of the entire test dataset, showing for each document its corresponding group formed by TF-IDF.

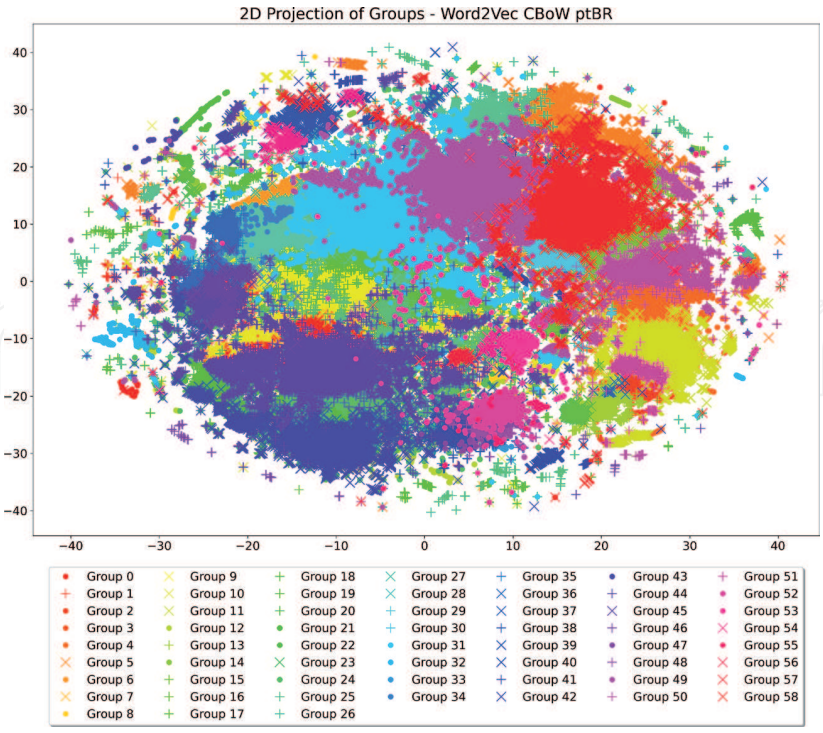


Figure 4.
2D projection of the entire test dataset, showing for each document its corresponding group formed by Word2Vec CBoW ptBR.

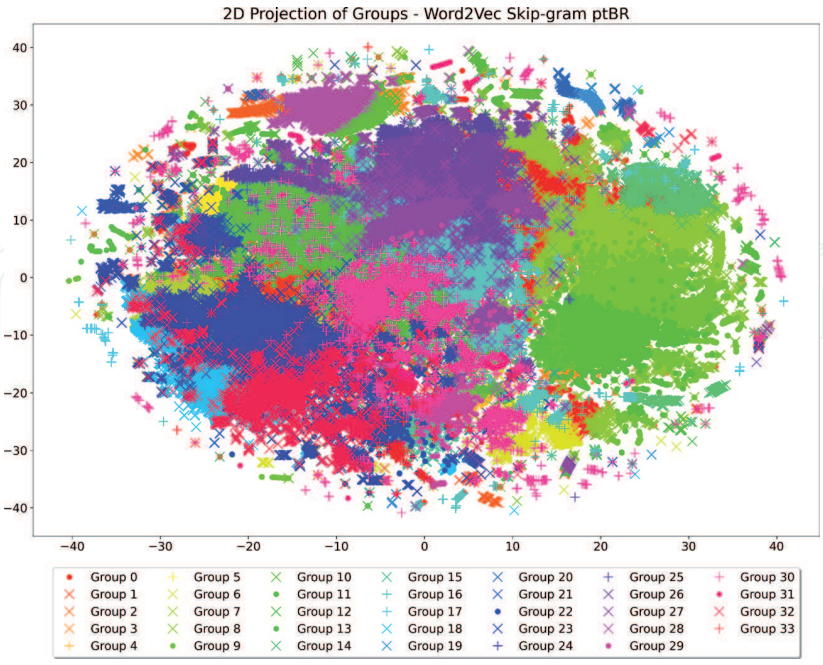


Figure 5.
2D projection of the entire test dataset, showing for each document its corresponding group formed by Word2Vec skip-gram ptBR.

12

5. Conclusion and future work

The use of AI as a standard detection tool based on documents from the judiciary has generally proved to be a viable and helpful solution in the scientific, technological, and practice of legal work. In this chapter, it was possible to present the results considered very promising due to the improvement in the average similarity rate. Thus, we demonstrate the possibility of using word-embedding generation techniques applied on clustering of Ordinary Appeal Interposed using AI algorithms.

Of all the techniques evaluated, the Word2Vec Skip-gram pt-BR technique presented itself as the best option for word embeddings for clustering legal documents of the Ordinary Appeal Interposed type.

We believe that specialized word embeddings have great potential in improving the results. Therefore, comes the suggestion for future study of Word2Vec specialized for the judiciary, in addition to evaluating whether the new embeddings generated provide an opportunity to improve the overall performance of clustering. In addition, using transformer-based techniques, such as BERT, can achieve promising results, using both the Portuguese language word-embedding model and training a specialized BERT model for the judiciary.

Moreover, new possibilities arise for using the techniques discussed in this chapter, such as the draft generation of decisions and classification of documents and processes.

Acknowledgements

The authors thank the Regional Labour Court of the 5th Region for making datasets available to the scientific community and contributing to research and technological development. The authors also thank the Artificial Intelligence Reference Centre and the Supercomputing Centre for Industrial Innovation, both from SENAI CIMATEC.

Author details

Raphael Souza de Oliveira[1] and Erick Giovani Sperandio Nascimento[2*]

1 TRT5—Regional Labor Court of the 5th Region, Salvador, BA, Brazil

2 SENAI CIMATEC—Manufacturing and Technology Integrated Campus, Salvador, BA, Brazil

*Address all correspondence to: ericksperandio@gmail.com

References

[1] CNJ—Conselho Nacional de Justiça. Relatório Analítico Anual da Justiça em Números 2020. 2020. Available from: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/ [Accessed: June 07, 2021]

[2] da Costa Salum G. A duração dos processos no judiciário: aplicação dos princípios inerentes e sua eficácia no processo judicial [Internet], Âmbito Jurídico, Rio Grande. Vol. XIX(145). 2016. Avaliable from: https://ambitojuridico.com.br/cadernos/direito-processual-civil/a-duracao-dos-processos-no-judiciario-aplicacao-dos-principios-inerentes-e-sua-eficacia-no-processo-judicial/ [Accessed: September 01, 2021]

[3] Canotilho JJG. Direito constitucional e teoria da constituição. 7th ed. Coimbra: Almedina; 2003

[4] Khan W, Daud A, Nasir J, Amjad T. A survey on machine learning models for Natural Language Processing (NLP). Computer Science and Engineering. 2016;**43**:95-113

[5] Wang Y, Cui L, Zhang Y. Using Dynamic Embeddings to Improve Static Embeddings. In: arXiv Preprint. arXiv:1911.02929v1. 2019

[6] Mikolov, T, Chen, K, Corrado, G, Dean, J. Efficient Estimation of Word Representations in Vector Space. In: ICLR: Proceeding of the International Conference on Learning Representations Workshop Track, Arizona, USA. 2013.

[7] 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. pp. 1532-1543

[8] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;**5**:135-146

[9] Devlin J, Chang M-W, Lee K, Toutanova K. 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. Minneapolis, Minnesota: Association for Computational Linguistics. 2019; 1:4171-4186. DOI: 10.18653/v1/N19-1423

[10] Oliveira FJV. Os recursos na Justiça do Trabalho [Internet]. Available from: http://www.conteudojuridico.com.br/consulta/Artigos/24853/os-recursos-na-justica-do-trabalho [Accessed: June 10, 2021]

[11] Sil R, Roy A, Bhushan B, Mazumdar AK. Artificial Intelligence and Machine Learning based Legal Application: The State-of-the-Art and Future Research Trends. In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS); 18-19 October 2019; Greater Noida, India: IEEE; 2019. p. 57-62. DOI: 10.1109/ICCCIS48478.2019.8974479

[12] Renuka S, Raj Kiran GSS, Rohit P. An unsupervised content-based article recommendation system using natural language processing. In: Jeena Jacob I, Kolandapalayam Shanmugam S, Piramuthu S, Falkowski-Gilski P, editors. Data Intelligence and Cognitive Informatics (Algorithms for Intelligent Systems). Singapore: Springer; 2021. pp. 165-180. DOI: 10.1007/978-981-15-8530-2_13

[13] MacQueen J. Some methods for classification and analysis of

multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; Berkeley, CA: University of California Press; Vol. 1. 1967. pp. 281-297.

[14] D'Silva J, Sharma U. Unsupervised automatic text summarization of Konkani texts using K-means with Elbow method. International Journal of Engineering Research and Technology. 2020;**13**:2380. DOI: 10.37624/IJERT/13.9.2020.2380-2384

[15] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;**9**:1735-1780. DOI: 10.1162/neco.1997.9.8.1735

[16] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena. 2020;**404**:132306

[17] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. pp. 2227-2237. DOI: 10.18653/v1/N18-1202

[18] Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. pp. 1638-1649

[19] Melamud O, Goldberger J, Dagan I. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning; Berlin, Germany: Association for Computational Linguistics; 2016;. p. 51-61. DOI: 10.18653/v1/K16-1006

[20] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000-10. (NIPS'17).

[21] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019 . p. 3982-92. DOI: 10.18653/v1/D19-1410

[22] Roberts A, Raffel C, Lee K, Matena M, Shazeer N, Liu PJ, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In: arXiv Preprint. arXiv:1910.10683. 2019

[23] Hartmann NS, Fonseca ER, Shulby CD, Treviso MV, Rodrigues JS, Aluísio SM. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: Proceedings of the 11th Brazilian Symposium on Information and Human Language Technology (STIL). Uberlândia, Minas Gerais, Brazil: Brazilian Computing Society - SBC; 2017. p. 122-31.

4 MANUSCRITO 2

Brazilian court documents clustered by similarity together using natural language processing approaches with Transformers

Artigo Submetido para a revista *"Artificial Intelligence and Law"* da Springer em 31 de março de 2022.

Esta pesquisa apresentou um avanço em relação à *"Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches"* (OLIVEIRA; NASCIMENTO, 2021), a partir da utilização de técnicas de geração de *word embeddings* baseadas na arquitetura *Transformers* (VASWANI et al., 2017). Assim, foi desenvolvido e especializado modelos de aprendizagem de máquina profunda para a língua Portuguesa com corpus jurídico trabalhista para junto com os modelos de aprendizagem profunda para a língua Portuguesa desenvolver uma metodologia para agrupamento de processo judiciais comparando técnicas tradicionais com técnicas baseadas em *Transformers*.

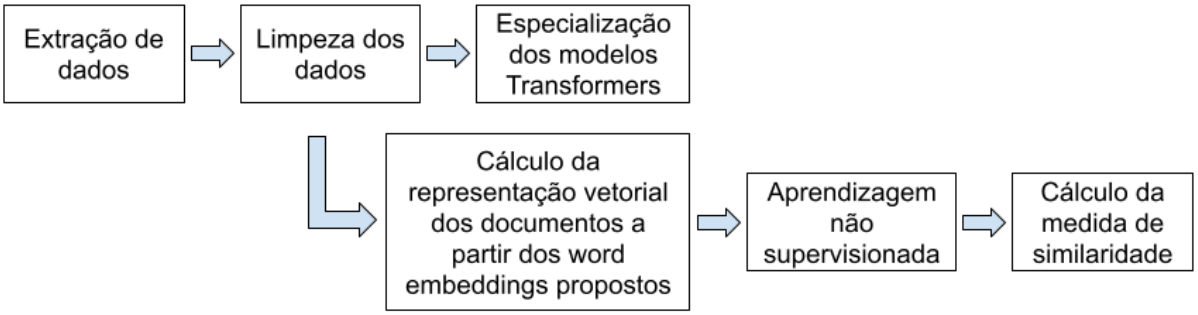
A partir dessa pesquisa, foi possível avançar no atual estado da arte na área de NLP aplicada ao setor jurídico, comparando o grau de semelhança alcançado entre os documentos judiciais nos grupos inferidos por meio da aprendizagem não supervisionada, através da aplicação de seis técnicas de Processamento de Linguagem Natural, quais sejam: (i) BERT (*Bidirectional Encoder Representations from Transformers*) treinado para fins gerais para o Português Brasileiro (SOUZA; NOGUEIRA; LOTUFO, 2020); (ii) BERT especializado com o corpus do judiciário trabalhista Brasileiro; (iii) GPT-2 (*Generative Pre-trained Transformer 2*) treinado para fins gerais para o Português Brasileiro (GUILLOU, 2020); (iv) GPT-2 especializado com o corpus do judiciário trabalhista Brasileiro; (v) RoBERTa (*Robustly optimized BERT approach*) treinado para fins gerais para o Português Brasileiro (SILVA, 2021); e (vi) RoBERTa especializado com o corpus do judiciário trabalhista Brasileiro.

Foi revisada a literatura em busca das produções mais recentes, no período de 2017 a 2021, através das base de dados (i) *Google Scholar*; (ii) *Science Direct*; e (iii) *IEEE Xplorer*, sobre algoritmos de aprendizagem de máquina não supervisionados ou clusterização aplicados à área jurídica utilizando NLP. As pesquisas revelaram que existiam até o momento poucas produções que tratam do tema, o que comprova a sua complexidade. Destaca-se a pesquisa realizada por (OLIVEIRA; NASCIMENTO, 2021) que buscou detectar o grau de semelhança entre os documentos judiciais da Justiça do Trabalho brasileira, por meio de aprendizagem não supervisionada, usando técnicas de NLP como TF-IDF, e Word2Vec com CBoW e Skip-gram treinado para fins gerais para a língua

portuguesa no Brasil.

O protocolo necessário para reproduzir os resultados alcançados e analisá-los comparativamente foi composto pelas fases (i) extração de dados; (ii) limpeza de dados; (iii) geração de modelos de *word embeddings*, com passos adicionais para especialização dos modelos *Transformers*; (iv) cálculo da representação vetorial do documento; (v) aprendizagem não supervisionada; e (vi) cálculo da medida de similaridade; conforme ilustrada na Figura 6.

Figura 6 – Diagrama da metodologia aplicada no Artigo "*Brazilian Court Documents Clustered by Similarity together using Natural Language Processing Approaches with Transformers*".



Fonte: Autoria Própria.

De todas as técnicas avaliadas, a técnica RoBERTa ptBR apresentou-se como melhor opção de *word embeddings* para clusterização de documentos judiciais do tipo Recurso Ordinário Interposto. Também, destaca-se a técnica BERT ptBR, já que apresentou taxas quantitativas ligeiramente melhores que o RoBERTa ptBR, no entanto não alcançou um tempo de execução tão satisfatório quanto o RoBERTa ptBR. Já os modelos especializados com o corpus do judiciário, de forma geral, não alcançaram resultados melhores que os modelos generalistas. A despeito disso, acredita-se que a especialização do BERT, do GPT-2 e do RoBERTa com um corpus jurídico mais robusto possa alcançar resultados ainda melhores.

Brazilian court documents clustered by similarity together using natural language processing approaches with Transformers

Raphael Souza de Oliveira[1†] and Dr Erick Giovani Sperandio Nascimento[2*†]

[1]TRT5 - Regional Labor Court of the 5th Region, Salvador, Bahia, Brazil.
[2*]SENAI CIMATEC - Manufacturing and Technology Integrated Campus, Salvador, Bahia, Brazil.

*Corresponding author(s). E-mail(s): ericksperandio@gmail.com;
Contributing authors: raphael.oliveira@gmail.com;
†These authors contributed equally to this work.

Abstract

Recent advances in Artificial intelligence (AI) have leveraged promising results in solving complex problems in the area of Natural Language Processing (NLP), being an important tool to help in the expeditious resolution of judicial proceedings in the legal area. In this context, this work targets the problem of detecting the degree of similarity between judicial documents that can be achieved in the inference group, by applying six NLP techniques based on transformers, namely BERT, GPT-2 and RoBERTa pre-trained in the Brazilian language and the same specialized using 210,000 legal proceedings. Documents were pre-processed and had their content transformed into a vector representation using these NLP techniques. Unsupervised learning was used to cluster the lawsuits, calculating the quality of the model based on the cosine of the distance between the elements of the group to its centroid. We noticed that models based on transformers present better performance when compared to previous research, highlighting the RoBERTa model specialized in the Brazilian language, making it possible to advance in the current state of the art in the area of NLP applied to the legal sector.

Keywords: legal, natural language processing, clustering; transformers

1 Introduction

The recent history of the Brazilian Justice shows relevant transformations regarding having all its procedural documents in digital format. In 2012, the Brazilian Labor Court implemented the Electronic Judicial Process (acronym in Portuguese for "Processo Judicial Eletrônico" - PJe), and since then, all new lawsuits have become completely digital, reaching 99.9% of cases in progress on this platform in 2020 [1].

Knowing the limitation of human beings analysing, in an acceptable time, a large amount of data, especially when such data appear not to be correlated, it is possible to help them in the patterns' recognition context through data analysis, computational ans statistical methods. Assuming that textual data has been exponentially increasing, patterns' examination in court documents is becoming pronouncedly challenging.

To optimize the procedural progress the Brazilian legal system provides for ways, such as the procedural economy, the principle of speed, due process in order, and the principle of the reasonable duration of a case to ensure the swift handling of judicial proceedings [2]. Hence, one of the major challenges of the Brazilian Justice is swiftly meeting the growing judicial demand. Thus, using a process grouping mechanism, it was possible to assist with the allocation of work among the advisors of the office for which the process was drawn with a good rate of similarity between the documents analysed. Furthermore, it contributed to the search for case-law[1] for the judgment of the cases in point, guarding the principle of legal certainty. According to Gomes Canotilho [3], the general principle of legal certainty aims to ensure the individual the right to trust that the legal rulings made of their issues are based upon current and valid legal norms.

Hence, this legal management tool allowed reducing the length of the judicial process, generating positive impacts such as the decrease of the operational costs of a lawsuit based on the lower allocation of the resources necessary for its judgment.

Recent studies have shown that machine learning algorithms are critical tools capable of solving high-complexity problems using Natural Language Processing (NLP) [4]. To this end, it is possible to highlight the works of [5–11], which, taking into account the context of words, apply techniques of word-embeddings generation, a form of vector representation of names, and consequently of documents. The use of word-embeddings is essential to analyze a large set of unstructured data as presented in court.

At present, a specialist triages the documents and distributes the lawsuits to be judged among the team members, configuring a deviation from the main activity of the specialist, which is the production of the draft decisions. This occurrence reinforced a further increase in the congestion rate (an indicator that measures the percentage of cases that remain pending solution by the end of the base-year) and to the decrease in the supply of demand index (acronym

[1] A legal term meaning a set of previous judicial decisions following the same line of understanding.

in Portuguese for "Índice de Atendimento à Demanda" - IAD - an indicator that measures the percentage of downtime of processes compared to the number of new cases) [1].

This work aims, therefore, to use as a baseline the results discussed by the research "Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches" [12] comparing them with the degree of similarity between the judicial documents achieved in the inferred groups through unsupervised learning, through the application of six techniques of Natural Language Processing, which are: (i) BERT (Bidirectional Encoder Representations from Transformers) trained for general purposes for Portuguese (BERT pt-BR); (ii) BERT specialized with the corpus of the Brazilian labor judiciary (BERT Jud); (iii) GPT-2 (Generative Pre-trained Transformer 2) trained for general purposes for Portuguese (GPT-2 pt-BR); (iv) GPT-2 specialized with the corpus of the Brazilian labor judiciary (GPT-2 Jud); (v) RoBERTa (Robustly optimized BERT approach) trained for general purposes for Portuguese (RoBERTa pt-BR); and (vi) RoBERTa specialized with the corpus of the Brazilian labor judiciary (RoBERTa Jud).

As proposed in [12], the degree of similarity indicates the performance of the model and was a result of the average similarity rate of the documents groups, which was based on the cosine similarity between the elements of the group to its centroid and, comparatively, by the average cosine similarity among all the documents of the group.

To delimit the scope of this research and make a coherent comparison same data as in [12] was applied. Thus, the data set extracted contained information from the Ordinary Appeal Brought (acronym in Portuguese for "Recurso Ordinário Interposto" - ROI) of approximately 210,000 legal proceedings. The Ordinary Appeal Brought was used as a reference, as it is regularly the type of document responsible for sending the case to trial in a higher court (2nd degree), hence creating the Ordinary Appeal (acronym in Portuguese for "Recurso Ordinário" - RO). It serves as a free plea, an appropriate appeal against final and terminative judgments proclaimed at first instance, which seeks a review of the court decision drawn up by a hierarchically superior body [13].

For the present work, a literature review on unsupervised machine learning algorithms applied to the legal area was performed, using NLP, and an overview of recent techniques that use Artificial Intelligence (AI) algorithms in word embeddings generation. Then, applying some methods until obtaining results, comparing them, and finally, proposing future challenges.

2 State-of-the-Art Review

More recent research maintain that machine learning algorithms have great potential for high complexity problem-solving. These machine learning algorithms categories can be: (i) supervised; (ii) unsupervised; (iii) semi-supervised; and (iv) via reinforcement [14]. This research context reviewed the

literature in search of the most recent productions for the period from 2017 to 2021, through the databases (i) Google Scholar; (ii) Science Direct; and (iii) IEEE Xplorer, on unsupervised machine learning algorithms or clustering applied to the legal area using NLP.

The research revealed that, so far, few productions are dealing with the subject, which proves its complexity. We highlight the research conducted by Oliveira and Nascimento [12] that sought to detect the degree of similarity between the judicial documents of the Brazilian Labor Court through unsupervised learning, using NLP techniques such as (i) inverse frequency of the term document frequency (TF-IDF); (ii) Word2Vec with CBoW (Continuous Bag of Words) trained for general purposes for the Portuguese language in Brazil; and (iii) Word2Vec with Skip-gram trained for general purposes for the Portuguese language in Brazil.

Expanding the research for the use of Natural Language Processing applied to the judicial area, a systematic review of the literature of the challenges faced by the system of trial prediction was found, which can assist lawyers, judges and civil servants to predict the rate of profit or loss, time of punishment and articles of law applicable to new cases, using the deep learning model. The researchers describe in detail the Empirical Literature on Methods of Prediction of Legal Judgment, the Conceptual Literature on Text Classification Methods and details of the transformers model [15].

Therefore, we then sought to expand the research by removing the restriction for the legal area, which revealed some publications. [16] Discusses using a content recommendation system based on grouping, with k-means, in similar articles through the vector transformation of the content of documents with the TF-IDF [17]. In [18], the authors performed an automatic summarization of texts using TF-IDF and k-means to determine the sentence groups of the documents used in creating the summary. It concludes that these studies used TF-IDF as the primary technique to vectorize textual content and that k-means is the most commonly used algorithm for unsupervised machine learning.

We assume that choosing the best technique of generating word embeddings requires research, experimentation and comparison of models. Many recent studies prove the feasibility of using word embeddings to improve the quality of the results of AI algorithms for pattern detection and classification, among others.

Mikolov et al. proposed in 2013 Word2Vec Skip-gram and CBoW, two new architectures to calculate vector representations of words considered, at the time, reference in the subject [6]. Then, Embeddings from Language Models (Elmo) [19], Flair [20] and context2vec [21], libraries based on the Long Short Term Memory Network (LSTM) [22] created distinct word embeddings for each occurrence of the word, context-aware, which allowed the capture of the meaning of the word. The LSTM models were used widely for speech recognition, language modelling, sentiment analysis and text prediction, and, unlike

Recurrent Neural Network (RNN), have the ability to forget, remember and update information, thus taking a step ahead of the RNNs [23].

As of 2018, new techniques for generating word embeddings emerged, with emphasis on (i) Bidirectional Encoder Representations from Transformers (BERT) [9], a context-sensitive model with architecture based on a Transformers model [24]; (ii) Sentence BERT (SBERT) [25], a "Siamese" BERT model proposed to improve BERT's performance when seeking to obtain the similarity of sentences; (iii) Text-to-Text Transfer Transformer (T5) [26], a framework for treating NLP issues as a text-to-text problem, i.e. template input as text and template output as text; (iv) Generative Pre-Training Transformer 2 (GPT-2), a Transformers-based model with 1.5 billion parameters [10]; and (v) Robustly optimized BERT approach (RoBERTa), a model based on the BERT model, which was trained longer and used a higher amount of data [11].

With this analysis, it was possible to advance in the current state of the art of NLP applied to the legal sector. By conducting a comparative study and implementation of Transformers techniques (BERT, GPT-2 and RoBERTa), using models for generic purpose in Brazilian Portuguese (pt-BR) and specialized models in the labor judiciary, to carry out the grouping of labor legal processes in Brazil using the k-means algorithm and cosine similarity.

3 Methodology

In this section, the protocol necessary to reproduce the results achieved and to analyze them comparatively is presented. For the implementation of the routines used in this study, we used the Python programming language (version 3.6.9) and the same libraries used in the study by Oliveira and Nascimento [12].

The processing flow (pipeline) was composed of the phases: (i) data extraction; (ii) data cleaning; (iii) generation of word embeddings templates; (iv) calculation of the vector representation of the document; (v) unsupervised learning; and (vi) calculation of the similarity measure, of which phases (i), (ii), (v) and (vi) followed the same steps described by Oliveira and Nascimento [12] and the other phases are detailed in sections to be follow.

3.1 Generation of word embeddings templates

The usage of vector representation of words, whose numerical values indicate some relationship between words in the text, is an essential technique in the machine learning problem-solving process when the data used by the model is textual.

Thus, in this research, word embeddings generated and shared for the Portuguese language were used, such as (i) BERT (large) model generated based on brWaC corpus [27], composed of 2 billion and 700 thousand tokens, and published in the article BERTimbau: Pretrained BERT Models for Brazilian Portuguese [28]; (ii) GPT-2 (Small) model generated based on texts extracted from Wikipedia in Portuguese, and published in article GPorTuguese-2 (Portuguese GPT-2 small): a Language Model for Portuguese text generation (and

more NLP tasks...) [29]; and (iii) RoBERTa (Base) model generated based on texts extracted from Wikipedia in Portuguese, entitled roberta-PT-BR and published in Hugging Face [30].

In addition to these pre-trained models in the Portuguese language, the most recent literature suggests that using embeddings adherent to the context of the problem proposed to be solved may bring a better result. Thus, using the 210,000 documents extracted, two embedding generation techniques were applied, namely, (i) specialization of the BERTimbau model; (ii) specialization of the GPorTuguese-2 model; and (iii) specialization of the roberta-pt-br model, which will be detailed below.

3.1.1 Specialization of Transformers models

Recent studies show the benefits of applying for learning transfer on generalist models, which, in recent years, has significantly improved the results, reaching the state-of-the-art in NLP [31]. For the specialization of Transformers models, in addition to cleaning the data, it is also necessary to adjust the data to make the most of its benefits. Of the adjustments made, two deserve highlights: (i) definition of the sentence slot; and (ii) definition of the strategy of "disguising" or masking (MASK) of the sentences' tokens, which are detailed below.

Defining the sentence slot is a fundamental step to enable the usage of specialized data in the learning transfer from a pre-trained model. Therefore, inspired by the strategy proposed in the article Transformers: State-of-the-Art Natural Language Processing [32] that, for each batch of 1,000 documents, as presented in Figure 1, all content is concatenated and sentences of 128 tokens created, if the last "sentence" of this lot is less than 128 tokens this "sentence" is disregarded, other detailed approaches have been tested later.

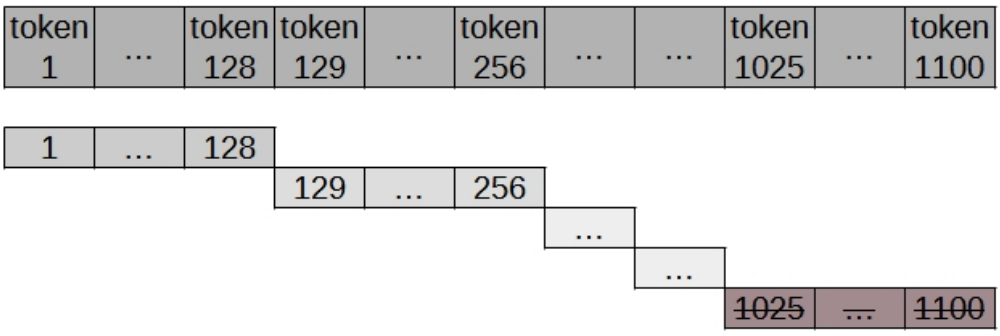


Fig. 1 Slot N - generation of "sentences" with 128 tokens

In order to reduce the loss of context of words at the edges of the sentence, the proposed approach, entitled Slot N/K, generated "sentences" with N tokens from the concatenation of 1,000 documents, as detailed below and illustrated in Figure 2.

- Initial Slot: "sentence" formed by the first N tokens;

- Intermediate slots: "sentence" formed by N tokens counted from the N-K token of the previous "sentence", where K is the number of return tokens;
- Final Slot: "sentence" formed by the last N tokens.

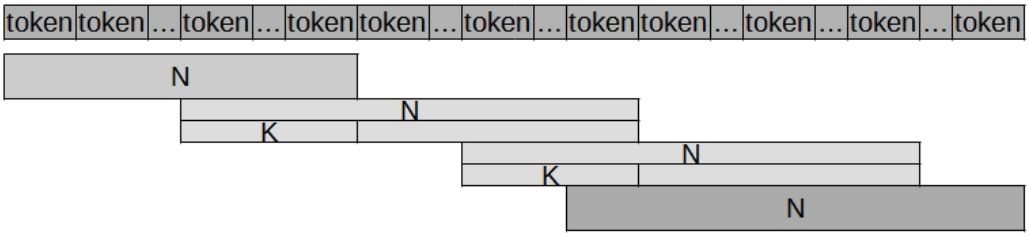


Fig. 2 N/K slot

From the above-detailed approach, simulations performed with the settings (i) Slot 128/16; (ii) Slot 128/32; (iii) Slot 128/64; (iv) Slot 256/64; (v) Slot 512/64; and (vi) Slot 64/16, comparing them with each other and with the approach proposed by [32]. The Slot 128/32 approach was selected for achieving the best performance in the specialization of the Transformers model in Portuguese with the corpus of the judiciary (Figure 3).

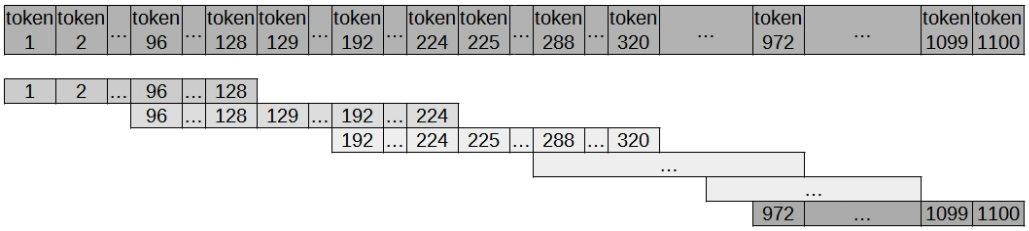


Fig. 3 Slot 128/32

For learning transfer, depending on the Transformers model used, a token masking strategy for each sentence is applied, using Masked Language Models (MLM) for BERT and Causal Language Models (CLM) models for GPT-2 models. While the CLM is trained unidirectionally in order to predict the next word based on the preceding words [33], the MLM has a two-way approach to predict the masked words of the sentence.

Hence, for the transfer of learning of BERT models, inspired by the article Transformers: State-of-the-Art Natural Language Processing [32] that used the masking rate of 15%, simulations were performed using the masking rate of 15% and the masking rate of 25%, reaching, with the rate of 15%, the best result in the specialization of the BERT model in Portuguese with the corpus of the judiciary.

3.2 Calculation of the vector representation of the document

Vector representation techniques of words (word embeddings) such as (i) BERT; (ii) GPT-2; and (iii) RoBERTa need to undergo a transformation in order to, from the word embeddings, calculate the vector representation of the document (document embeddings).

It is initially necessary to detail how to obtain word embeddings for Transformers techniques. One of the advantages of Transformers techniques over previous word embeddings techniques, such as Word2Vec, is the ability to capture the vector representation of the word according to the global context, meaning that the same word can have more than one vector representation. It becomes more evident when highlighting the word "bank" (banco in Pt-BR) in the following two sentences (i) I go to the bank (banco in Pt-BR) to withdraw money; and (ii) I will sit on the bench (banco in Pt-BR) of the square; where, with Word2Vec, the vector representation of the word "bank" is unique regardless of the phrase and with BERT, GPT-2 and RoBERTa word embeddings are different.

Therefore, for Transformers templates, it is necessary to "divide" the entire document into "slots" of sentences. Considering that, unlike the GPT-2 model, the BERT and RoBERTa models have a limitation of up to 512 tokens per sentence and require that the first and last tokens be special, respectively [CLS] and [SEP], the slot size has been set at 510 tokens per sentence.

Thus, we developed strategies to obtain all the word embeddings of the document, whose words of the generated sentences kept the context according to the complete file. These approaches consist, similar to that presented in Figure 2, in bringing about sentences with 510 tokens as detailed below:

- Initial Sentence: "sentence" formed by the first 510 tokens;
- Intermediate sentences: "sentence" consisting of 510 tokens counted from token N - K of the previous "sentence", where K was set empirically to value 64;
- Final Sentence: "sentence" formed by the last 510 tokens;

Therefore, the sentences generated from each document have coincident tokens chosen to ensure greater adherence to the token context in the file. To this end, we tested two different approaches: (i) averages of word embeddings of coincident tokens; and (ii) use of the first 32 coincident tokens of the previous sentence and the last 32 coincident tokens of the current sentence, which showed better results in the simulations performed.

Hence, as shown in Figure 4, the return tokens that are coincident between the current and previous sentences are used as follows: (i) the first 32 coincident tokens of the previous sentence (for example, tokens 446 to 477 from Slot 1 exemplified in Figure 4); and (ii) the last 32 coincident tokens of the sentence in question (for example, Tokens 478 to 510 from Slot 2 exemplified in Figure 4). It is worth noting that the last sentence slot must contain 510 tokens, as well as the others, and coincident tokens tapped as follows: (i) the first half of the

coincident tokens of the previous sentence (for example, tokens 590 to 773 from Slot 2 exemplified in Figure 4); and (ii) the second half of the coincident tokens of the sentence in question (for example, tokens 774 to 956 from Slot 3 exemplified in Figure 4).

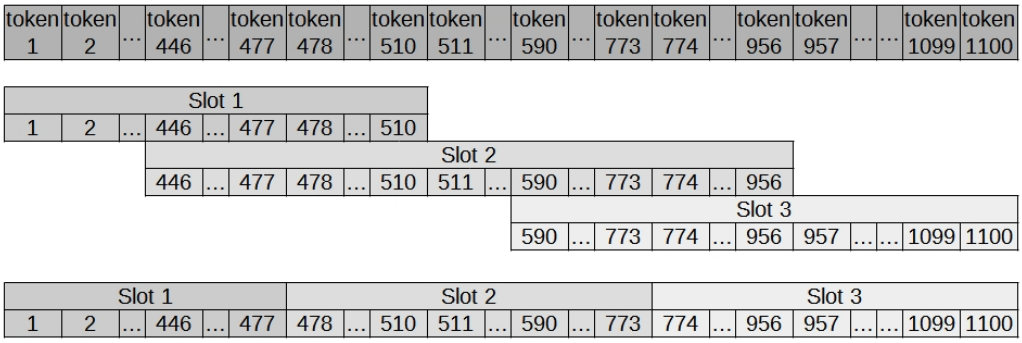


Fig. 4 Word embeddings generation strategy

After obtaining the word embeddings, the same technique used in the research by Oliveira and Nascimento was chosen to generate the document embeddings, that is, the average of the word embeddings of the words in the document, weighting them with the TF-IDF.

Consequently, to enable an overview, Table 1 summarizes the parameters used for training the six models used in this research.

Table 1 Parameters used for training the six models

	BERT imbau	BERT Jud.	GPortu guese-2	GPT-2 Jud.	roberta-pt-br	RoBERTa Jud.
Data for traning	brWac corpus	210,000 ROIs	Wikipedia in Portuguese	210,000 ROIs	Wikipedia in Portuguese	210,000 ROIs
Tokenization type	word-piece		byte-level BPE		byte-level BPE	
Model details	24-layer, 1024-hidden, 16-heads, 340M parameters		12-layer, 768-hidden, 12-heads, 117M parameters		12-layer, 768-hidden, 12-heads, 125M parameters	
Token mask type	Masked		Masked		Causal	

Moreover, in order to allow the graphic representation in two dimensions of the vector representation of the documents, the technique of reduction of T-Distributed Stochastic Neighbor Embedding (t-SNE), which minimizes the divergence between two distributions, measuring the similarities between pairs of input objects and the similarities between pairs of the low dimension points corresponding in the incorporation [34].

4 Results and Discussions

Applying the methodology as previously detailed, this research shows how natural language processing techniques in conjunction with machine learning algorithms are paramount in optimizing the operational costs of the judicial process, such as the aid of document screening and procedural distribution. It grants working time optimization since it allows the experts time to be devoted to their core activity.

In order to use the unsupervised learning algorithm, k-means, it was necessary to select the best K to achieve the best result for each NLP technique studied. This way, the elbow method was applied based on the calculated inertia of each of the 31 K tested, as shown in Figure 5.

From obtaining the best K, the k-means template was trained and, with the grouping performed by the model, we calculated (i) the average similarity between the documents of each group, thus allowing an overview of the distribution of documents in the groups generated by each NLP technique; and (ii) the mean similarity between the group's documents and their centroid, making possible to indicate which technique achieved the best performance.

To demonstrate the progress brought by this research, Table 2 presents the results extracted from the study [12], which established a baseline for research on the use of NLP techniques applied to the legal environment for the same purpose. We highlight the Word2Vec Skip-gram pt-BR technique, which presented itself, in that research, as the best option for generating word embeddings aiming to group judicial documents of the Ordinary Appeal Brought type.

Table 2 Statistical data extracted from the work "Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches" [12]

Type	Groups	Mean	Std.	Min.	25%	50%	75%	Max.
TF-IDF	49	0.624	0.172	0.247	0.502	0.586	0.164	0.964
Word2Vec CBoW ptBR	59	0.947	0.063	0.764	0.935	0.979	0.991	0.999
Word2Vec Skip-gram ptBR	34	0.948	0.061	0.796	0.925	0.976	0.992	0.999

Consequently, the statistical data of the average similarity between the documents of each group and the average similarity of the group documents for their centroid presented respectively in Table 3 and Table 4, highlighted in bold for the best result value of each metric and projected in the comparative distribution chart (Figure 6 and Figure 7), show that the generalist word embeddings in Portuguese (pt-BR) achieved superior results when compared to the specialized legal corpus word embeddings. The proximity of the results among the generalist models is also noteworthy. However, for the expert model, this proximity was observed only between the BERT Jud models and GPT-2 Jud.

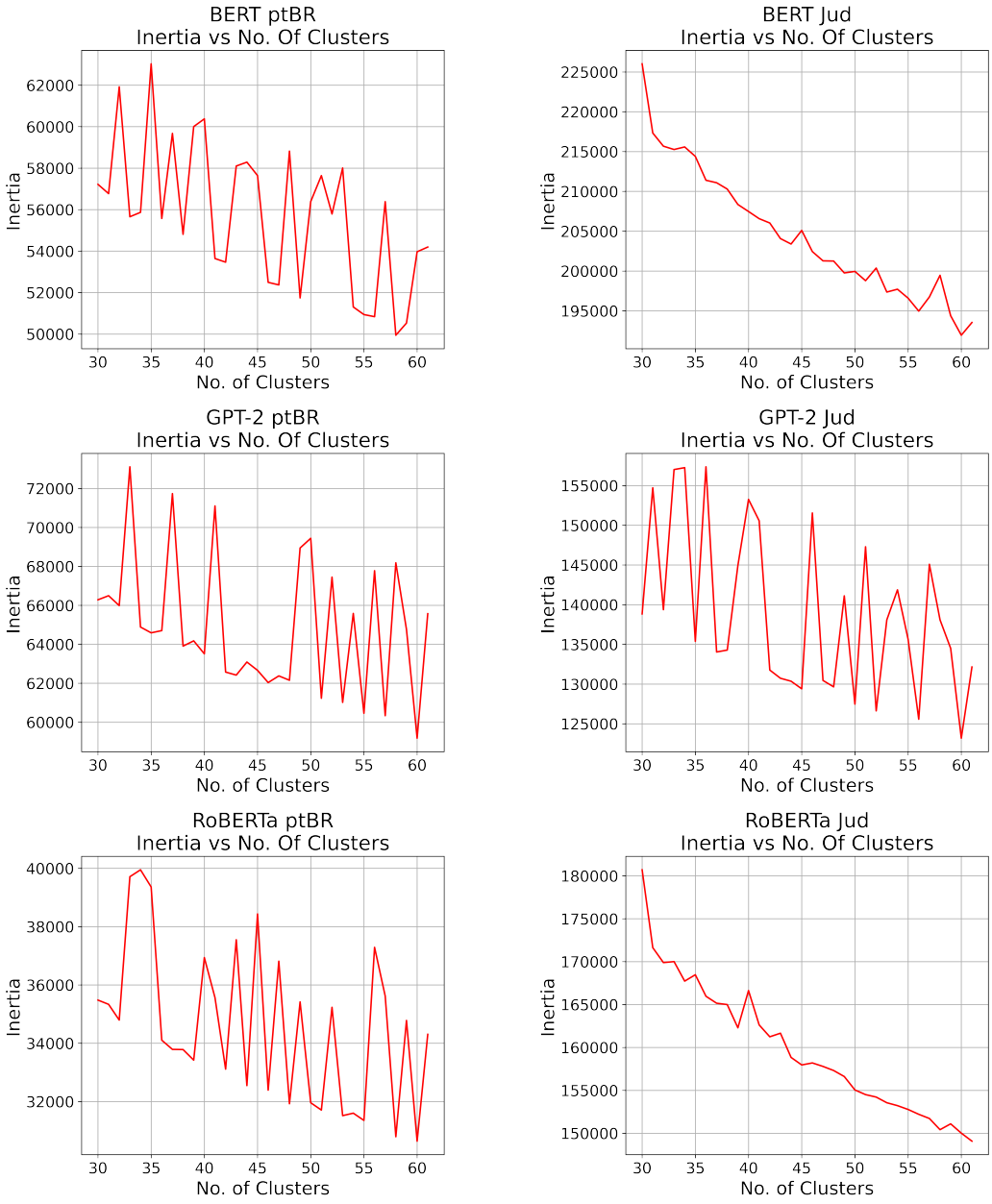


Fig. 5 Inertia charts constructed by using the elbow method for determining the best number of clusters for each approach

When comparing the values presented in Table 3 and Table 4, it is noteworthy that the results in Table 3 are slightly lower in all cases. From this, it is inferable that the measurement of similarity as in Table 3 might reduce the similarity rate since there may be elements positioned altogether opposite in the group. From Figure 6 and Figure 7, it is also possible to verify that the groupings generated by all techniques are very cohesive, especially in the generalist techniques cases, which created fewer groupings in the range of outliers than the expert techniques.

Table 3 Statistics of the cosine similarity between all elements of the group. The best results are highlighted in bold

Transformer Model	Groups	Mean	Std.	Min.	25%	50%	75%	Max.
BERT ptBR	35	0.976	**0.012**	**0.937**	0.967	0.979	0.985	0.991
BERT Jud	36	0.943	0.031	0.853	0.938	0.952	0.960	0.981
GPT-2 ptBR	**33**	0.972	0.020	0.906	0.965	0.979	0.985	**0.996**
GPT-2 Jud	36	0.952	0.034	0.847	0.947	0.964	0.971	0.994
RoBERTa ptBR	34	**0,976**	0,023	0,874	**0,971**	**0,985**	**0,988**	0,993
RoBERTa Jud	39	0,918	0,035	0,835	0,927	0,922	0,941	0,980

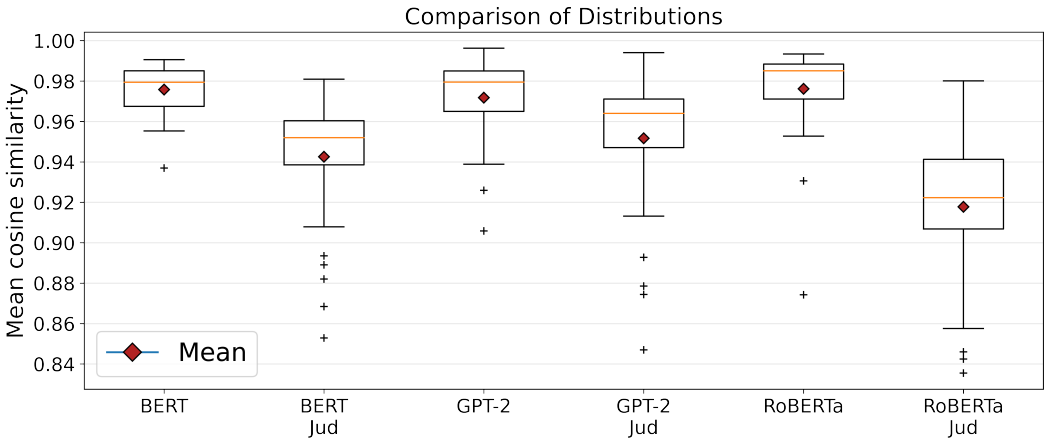


Fig. 6 Comparison chart of the distribution of the average similarity between the group documents. The more cohesive the boxes and the fewer outliers, the better

Table 4 Statistics of the cosine similarity of the group elements to the centroids. The best results are highlighted in bold

Transformer Model	Groups	Mean	Std.	Min.	25%	50%	75%	Max.
BERT ptBR	35	**0.987**	**0.007**	**0.967**	0.983	0.970	0.992	0.995
BERT Jud	36	0.971	0.016	0.923	0.969	0.976	0.980	0.990
GPT-2 ptBR	33	0.985	0.011	0.947	**0.985**	0.990	0.992	**0.998**
GPT-2 Jud	36	0.974	0.021	0.900	0.973	0.980	0.985	0.997
RoBERTa ptBR	34	0.987	0.017	0.905	0.985	**0.992**	**0.994**	0.997
RoBERTa Jud	39	0.958	0.019	0.914	0.952	0.960	0.970	0.990

Since most of the techniques achieved results close to each other, we considered it important to present the time spent for the processing of each Transformer technique, with the use of a computer with 40 physical nuclei and 196 GB of memory, in the generation of numerical representation of approximately 210,000 judicial documents of the Ordinary Appeal Brought type. As presented in Table 5, GPT-2 reached an average vectorization of documents per minute much higher than BERT. However, as expected, RoBERTa further

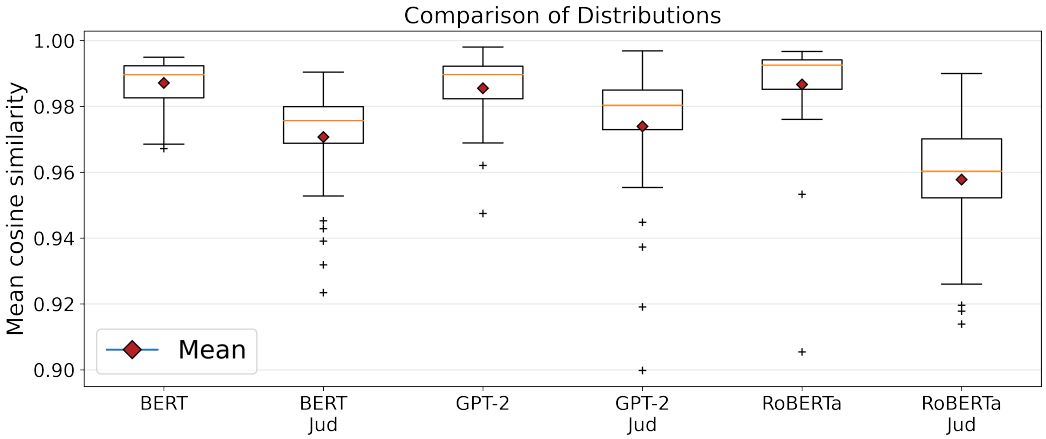


Fig. 7 Comparison chart of the distribution of the average similarity of the group documents to their centroid. The more cohesive the boxes and the fewer outliers, the better

outperformed BERT and GPT-2, as [11] performance can be improved when trained for more extended periods, with larger batches, over more data, without using the prediction of the next sentence strategy, in addition to training longer sequences with dynamically changed standard masking.

Table 5 Average processed documents per minute for each model highlighted in bold for the best result value

Transformer Model	Average number of documents processed per minute
BERT ptBR	6.45
BERT Jud	9.62
GPT-2 ptBR	29.40
GPT-2 Jud	29.03
RoBERTa ptBR	**55.31**
RoBERTa Jud	53.73

Given the above, among all the techniques evaluated, the RoBERTa pt-BR technique was the best option for generating word embeddings for judicial documents clustering of the Ordinary Appeal Brought type. Although the BERT pt-BR technique achieved a slightly better result (a difference minor than 0.01), it was computationally inefficient in document processing gpt-2 pt-BR and RoBERTa pt-BR techniques.

It is important to stress that the results of this research (Table 4) showed relevant advances in contrast to the results presented in the previous research (Table 2), in which the best average cosine similarity of the elements of the group to the centroid was, respectively, 0.98 and 0.94.

A fact to be analysed in the results presented is that specialized word embeddings techniques showed slightly worse results. Its occurrence is due to the general techniques in Portuguese being trained with a much larger corpus than the one used to refine the generalist model. This fact is also reported by

Ruder et al. [31], featuring a behaviour similar to that found in the present study, in which the corpus of the base model is much larger than the specialized corpus used.

The result achieved by each approach can be visualized in a two-dimensional projection of the groups formed in the six techniques (i) BERT pt-BR; (ii) BERT Jud.; (iii) GPT-2 pt-BR; (iv) GPT-2 Jud.; (v) RoBERTa pt-BR; and (vi) RoBERTa Jud., respectively, shown in Figure 8, Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13. After a qualitative analysis, it is evident in the images that the groups formed from the RoBERTa pt-BR are much better defined, which corroborates the findings previously explained in this study.

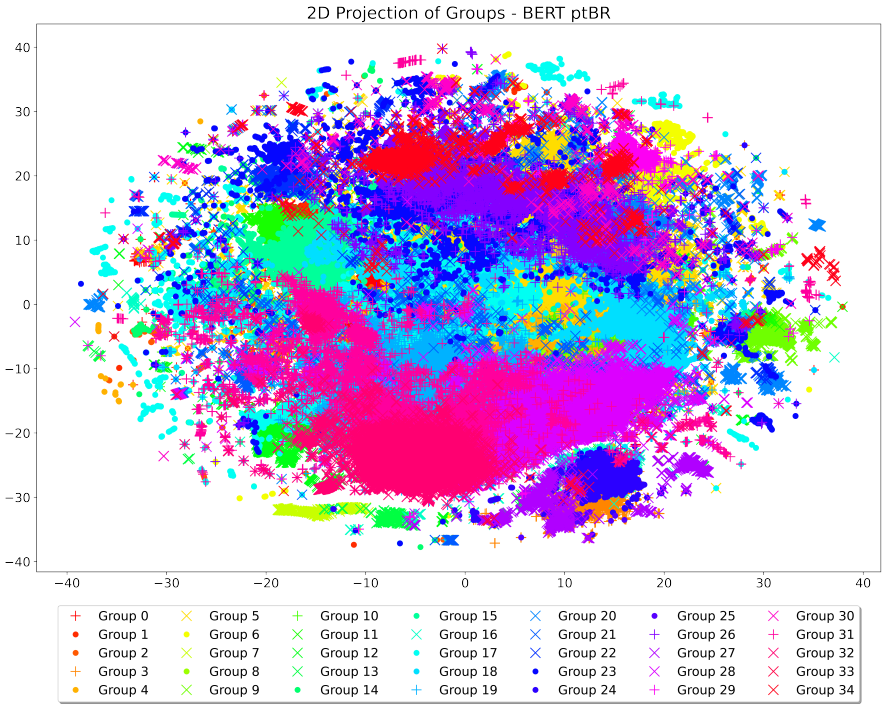


Fig. 8 Groups of documents formed by the BERT pt-BR technique projected in two dimensions using the test dataset.

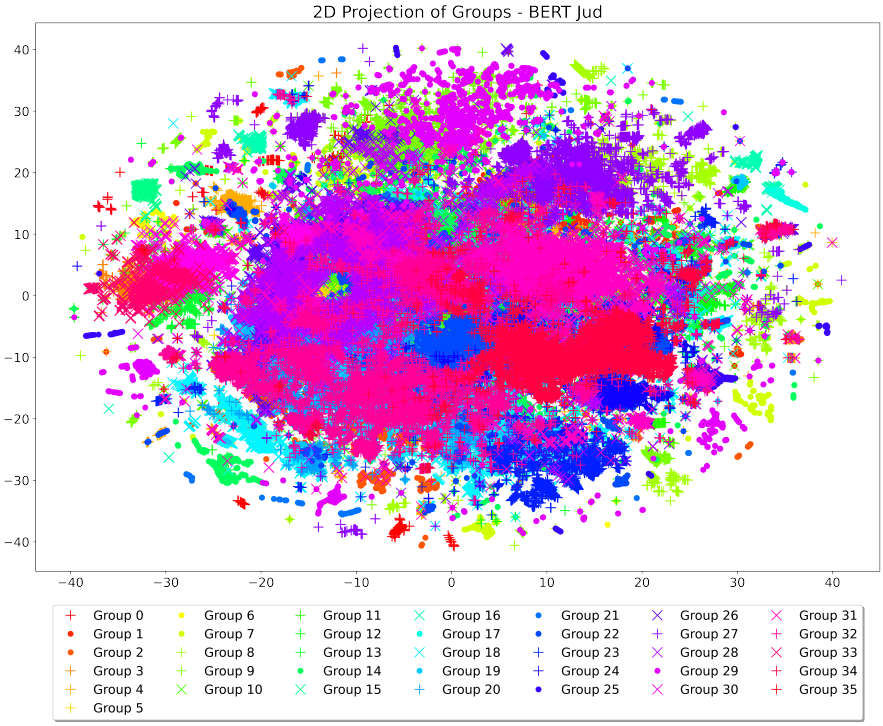


Fig. 9 Groups of documents formed by the BERT Jud technique projected in two dimensions using the test dataset.

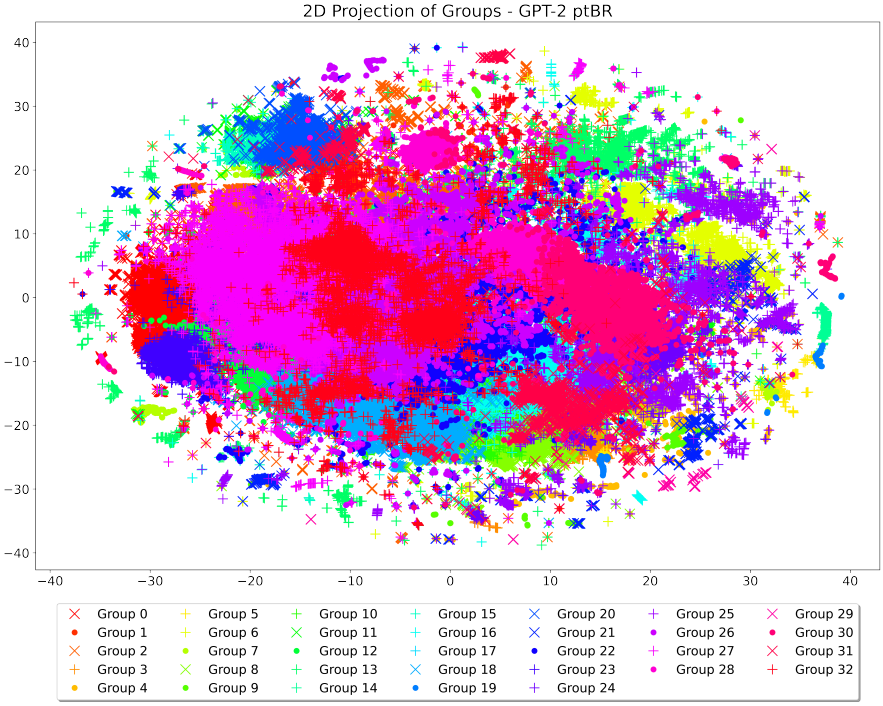


Fig. 10 Groups of documents formed by the GPT-2 pt-BR technique projected in two dimensions using the test dataset.

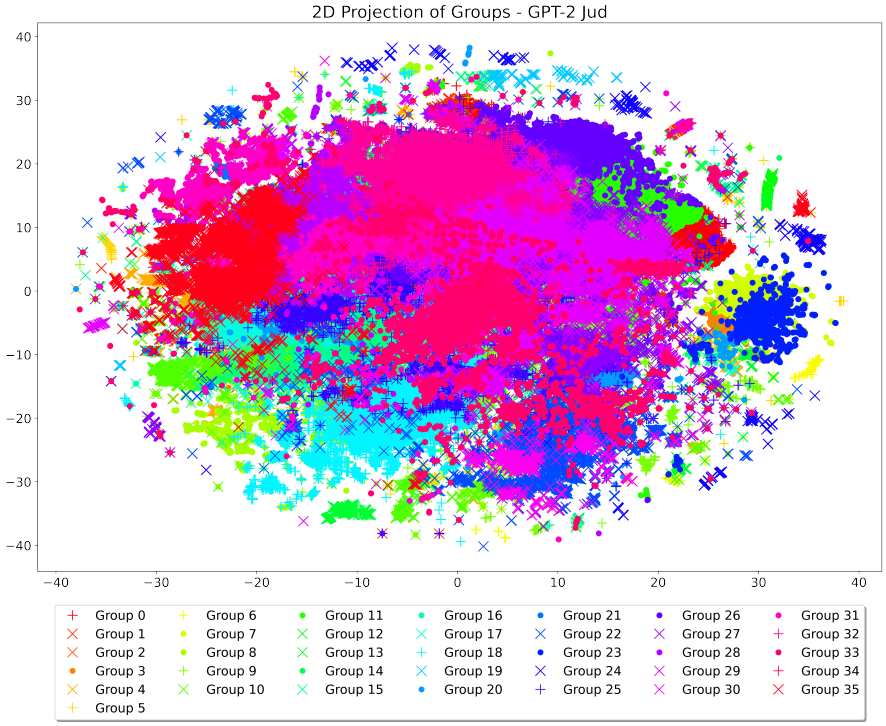


Fig. 11 Groups of documents formed by the GPT-2 Jud technique projected in two dimensions using the test dataset.

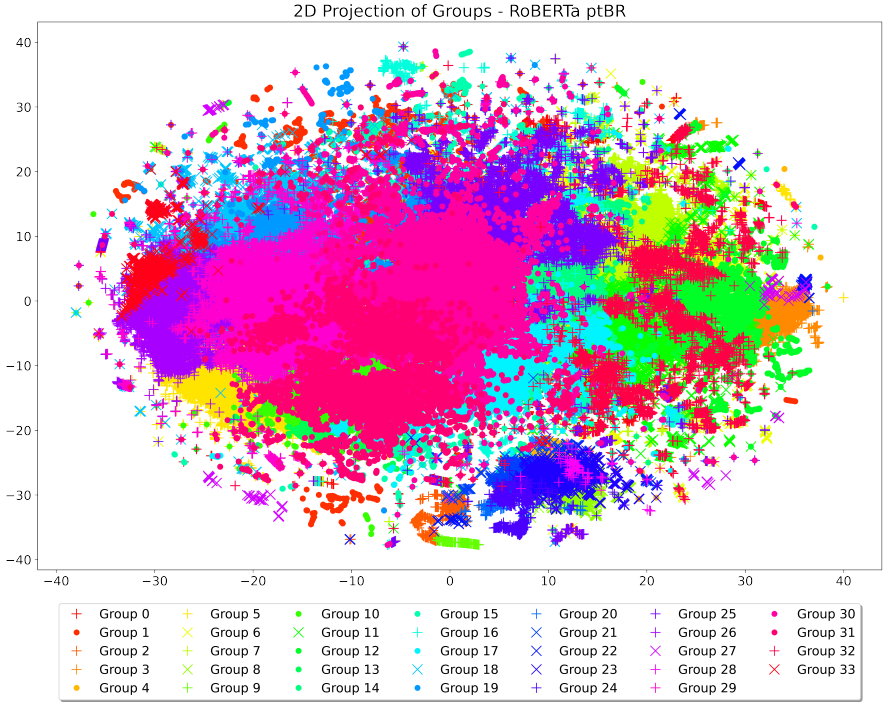


Fig. 12 Groups of documents formed by the RoBERTa pt-BR technique projected in two dimensions using the test dataset.

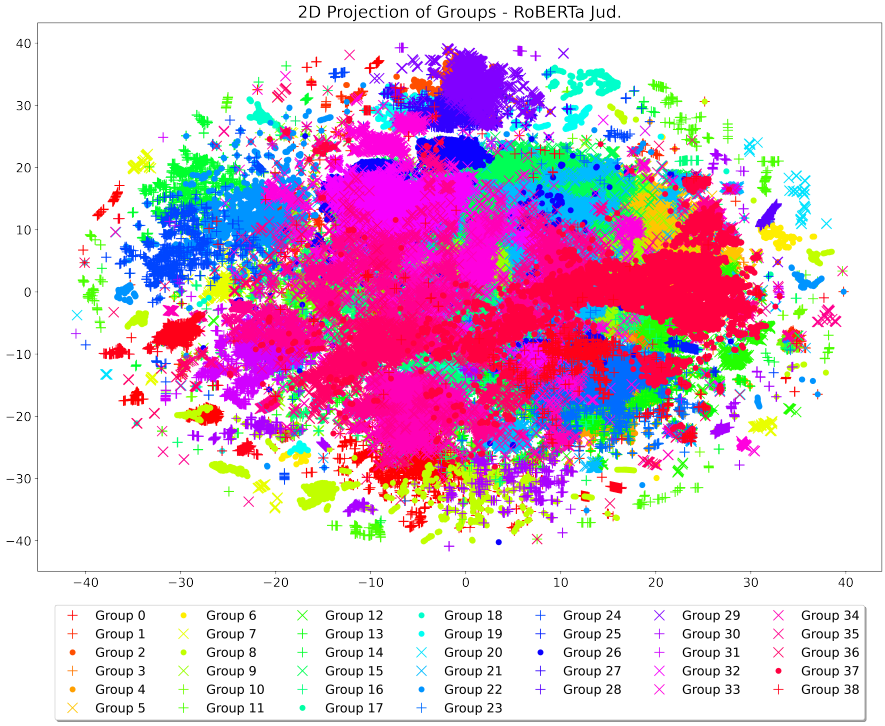


Fig. 13 Groups of documents formed by the RoBERTa Jud technique projected in two dimensions using the test dataset.

5 Conclusions and Future Works

Applying Artificial Intelligence techniques as a tool for pattern detection in legal documents has been proven, in general, as a viable and effective solution in the scientific and technological environment and very satisfactory in the practice of legal work. In this research, the results presented are considered amply promising for improving the Average Similarity Rate compared to previous research conducted by Oliveira and Nascimento [12].

Of all the techniques evaluated, the RoBERTa pt-BR technique stands out as the best option for word embeddings for clustering legal documents of the Ordinary Appeal Brought type. The BERT pt-BR technique is also in evidence since it presented slightly better quantitative rates than RoBERTa pt-BR, even though it did not reach an execution time as satisfactory as RoBERTa pt-BR. On the other hand, the specialized models with the corpus of the judiciary, in general, did not achieve better results than the generalist models. Despite this, we believe that the specialization of BERT, GPT-2 and RoBERTa with a more robust legal corpus can achieve even better results.

Therefore, for future work, there's the suggestion of deepening the specialization of BERT, GPT-2 and RoBERTa for the judiciary and evaluating whether the new embeddings generated will improve the overall performance of clustering. In addition, new possibilities arise, such as validating the word embeddings generated for other types of legal documents and using them

in other applications, such as the generation of decision drafts and classification of documents and processes. It is also worth delving into techniques for texts transformation into their vector representations faster in their word embeddings.

Acknowledgments. The authors thank the Artificial Intelligence Reference Centre and the Supercomputing Centre for Industrial Innovation, both from SENAI CIMATEC. The authors also thank the Regional Labour Court of the 5th Region for making datasets available to the scientific community and contributing to research and technological development.

References

[1] Relatório Analítico Anual da Justiça em Números 2021. CNJ - Conselho Nacional de Justiça (2021). https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/ Accessed 2022-02-19

[2] Salum, G.C.: A duração dos processos no judiciário: aplicação dos princípios inerentes e sua eficácia no processo judicial. In: Direito Processual Civil, vol. 145. Âmbito Jurídico, Rio Grande do Sul/Brazil (2016)

[3] Canotilho, J.J.G.: Direito Constitucional e Teoria da Constituição. Coimbra, Almedina (2003)

[4] Khan, W., Daud, A., Nasir, J., Amjad, T.: A survey on machine learning models for natural language processing (nlp) **43**, 95–113 (2016)

[5] Wang, Y., Cui, L., Zhang, Y.: Using dynamic embeddings to improve static embeddings (2019)

[6] Mikolov, T., Chen, K., Corrado, G.s., Dean, J.: Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR **2013** (2013)

[7] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation, vol. 14, pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162

[8] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5** (2016). https://doi.org/10.1162/tacl_a_00051

[9] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)

[10] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. (2019)

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019)

[12] Oliveira, R., Sperandio Nascimento, E.G.: Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches, (2021). https://doi.org/10.5772/intechopen.99875

[13] Oliveira, F.J.V.: Os recursos na Justiça do Trabalho. Conteudo Juridico (2011). http://www.conteudojuridico.com.br/consulta/Artigos/24853/os-recursos-na-justica-do-trabalho Accessed 2022-02-19

[14] Sil, R., Bhushan, B., Majumdar, A.: Artificial intelligence and machine learning based legal application: The state-of-the-art and future research trends, pp. 57–62 (2019). https://doi.org/10.1109/ICCCIS48478.2019.8974479

[15] Sukanya, G., Priyadarshini, J.: A meta analysis of attention models on legal judgment prediction system. International Journal of Advanced Computer Science and Applications **12**(2) (2021). https://doi.org/10.14569/IJACSA.2021.0120266

[16] Renuka, S., Kiran, G.S.S.R., Rohit, P.: An Unsupervised Content-Based Article Recommendation System Using Natural Language Processing, pp. 165–180 (2021). https://doi.org/10.1007/978-981-15-8530-2_13

[17] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Los Angeles (1967)

[18] D'Silva, J., Sharma, U.: Unsupervised automatic text summarization of konkani texts using k-means with elbow method. International Journal of Engineering Research and Technology **13**, 2380 (2020). https://doi.org/10.37624/IJERT/13.9.2020.2380-2384

[19] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations, pp. 2227–2237 (2018). https://doi.org/10.18653/v1/N18-1202

[20] Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. (2018)

[21] Melamud, O., Goldberger, J., Dagan, I.: context2vec: Learning generic context embedding with bidirectional lstm, pp. 51–61 (2016). https://doi.org/10.18653/v1/K16-1006

[22] Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[23] Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena **404**, 132306 (2020). https://doi.org/10.1016/j.physd.2019.132306

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)

[25] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks, pp. 3973–3983 (2019). https://doi.org/10.18653/v1/D19-1410

[26] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (2019)

[27] Wagner, J., Wilkens, R., Idiart, M., Villavicencio, A.: The brwac corpus: A new open resource for brazilian portuguese (2019)

[28] Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT Models for Brazilian Portuguese, pp. 403–417 (2020). https://doi.org/10.1007/978-3-030-61377-8_28

[29] Guillou, P.: Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...). (2020)

[30] da Silva, J.N.: roberta-pt-br. huggingface https://huggingface.co/josu/roberta-pt-br (2021)

[31] Ruder, S., Peters, M., Swayamdipta, S., Wolf, T.: Transfer learning in natural language processing, pp. 15–18 (2019). https://doi.org/10.18653/v1/N19-5004

[32] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T., Gugger, S., Rush, A.: Transformers: State-of-the-art natural language processing, pp. 38–45 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6

[33] Guo, C., Sablayrolles, A., Jégou, H., Kiela, D.: Gradient-based Adversarial Attacks against Text Transformers (2021)

[34] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)

5 CONCLUSÕES

Aplicar as técnicas de Inteligência Artificial como ferramenta para detecção de padrão em documentos jurídicos tem-se mostrado, de forma geral, uma solução viável e útil no meio científico, tecnológico e bastante satisfatória na prática do trabalho jurídico.

Nesta dissertação foram utilizadas nove técnicas de Processamento de Linguagem Natural aplicadas à aprendizagem de máquina não supervisionada em dados de processos de um Tribunal Regional do Trabalho no Brasil. Para isso levantou-se e foram preparados aproximadamente 210 (duzentos e dez) mil documentos do tipo Recurso Ordinário Interposto incorporados no PJe.

Além disso, em uma primeira fase foi desenvolvido e realizado um estudo comparativo do uso de técnicas tradicionais de NLP na transformação dos dados textuais para agrupamentos de processos judiciais. Posteriormente o estudo foi ampliado para especializar modelos de aprendizagem de máquina profunda para a língua Portuguesa com corpus jurídico trabalhista e então compará-los entre si e com as técnicas tradicionais de NLP. Em ambos os estudos, com o uso do algoritmo de aprendizado não supervisionado k-means, foi necessário escolher o melhor K, dentro da faixa de 30 a 61 definida empiricamente, para cada técnica de NLP estudada. Aplicou-se, assim, o método do cotovelo baseando-se na inércia calculada de cada um dos 31 K testados, alcançando assim melhor resultado para cada técnica.

Desta forma, foi possível estabelecer uma metodologia baseada em aprendizagem profunda para agrupamento de processos judiciais, consolidando-a para a Justiça Trabalhista Brasileira a partir dos testes e validações aplicados. Foi alcançado, então, resultados considerados muito promissores, devido à perceptível melhoria na Taxa de Similaridade Média nos grupos formados a partir do uso de todas técnicas de Processamento de Linguagem Natural objeto deste estudo aplicada à aprendizagem de máquina não supervisionada. Assim, baseada na metodologia estabelecida por esta pesquisa, foi desenvolvida para a Justiça Trabalhista Brasileira uma ferramenta intitulada GEMINI, que permitiu auxiliar a busca por jurisprudência, auxiliar a distribuição do trabalho entre os assessores e auxiliar na detecção de possíveis oportunidades para uniformizar a interpretação do direito no âmbito dos tribunais, ou seja, instaurar Incidentes de Uniformização de Jurisprudência. Tal ferramenta foi disponibilizada para implantação nos vinte e quatro Tribunais Trabalhistas Brasileiro e ajudou, a partir dos agrupamentos sugeridos, agilizar o andamento da resolução dos processos (TRT5, 2020a; CSJT, 2020; TRT5, 2020b; TRT15, 2021).

De todas as técnicas de Processamento de Linguagem Natural avaliadas para clusterização de documentos judiciais do tipo Recurso Ordinário Interposto que alcançaram os

melhores resultados foram Word2Vec Skip-gram ptBR e RoBERTa ptBR, respectivamente, para o primeiro trabalho (Seção 3) e para o segundo trabalho (Seção 4). Destaca-se, portanto, que no primeiro manuscrito não foi necessário utilizar o tempo de processamento como fator decisório da melhor técnica, já que as métricas alcançadas pelas três técnicas abordadas se diferenciavam.

Assim, de forma geral, a técnica RoBERTa ptBR apresentou-se como melhor opção de *word embeddings* para clusterização de documentos judiciais do tipo Recurso Ordinário Interposto. Também, destaca-se a técnica BERT ptBR, já que apresentou taxas quantitativas ligeiramente melhores que o RoBERTa ptBR, no entanto não alcançou um tempo de execução tão satisfatório quanto o RoBERTa ptBR. Já os modelos especializados com o corpus do judiciário, de forma geral, não alcançaram resultados melhores que os modelos generalistas. A despeito disso, acredita-se que a especialização do BERT, do GPT-2 e do RoBERTa com um corpus jurídico mais robusto possa alcançar resultados ainda melhores. Além disso, acredita-se, também, que a criação de modelos generalistas em português para a área jurídica, ou seja, modelos de NLP de fundação de segundo nível para a língua portuguesa com foco no setor jurídico, permitindo com que novos modelos de NLP voltados para a área jurídica na língua portuguesa se especializem a partir deste modelo de fundação de segundo nível, pode alavancar os resultados alcançados já que a linguagem utilizada no meio jurídico tem características próprias.

Desta forma, sugerem-se para trabalhos futuros aprofundar a especialização do BERT, do GPT-2 e RoBERTa para o judiciário além do uso de técnicas como GPT-3 e Megatron, e avaliar se o novo *word embeddings* gerado oportuniza melhora na performance geral da clusterização. Outra oportunidade, é gerar para a Justiça Trabalhista Brasileira *word embeddings* especializado em cada uma das técnicas *transformers* utilizando um corpus mais robusto e diversificado.

Sugere-se também avaliar a possibilidade de segmentar os conteúdos dos documentos para, com foco na abordagem de dados, melhorar a performance. Além disso, oportuniza-se realizar uma análise qualitativa da quantidade de grupos sugeridos pela técnica do cotovelo em relação a quantidade de grupos que especialistas avaliam ser pertinente ao tipo de documento utilizado para o agrupamento.

Ademais, surgem novas possibilidades a serem exploradas como, por exemplo, validar os *word embeddings* gerados para outros tipos de documentos jurídicos e ainda utilizá-los em outras aplicações como, por exemplo, (i) geração de minuta de decisão; (ii) classificação de documentos e processos; (iii) previsão da duração do processo baseado na complexidade dos documentos existentes; (iv) geração de resumo das decisões; (v) tradução da linguagem jurídica para a linguagem geral; e (vi) etc. Vale, também, se aprofundar em técnicas que viabilizem de forma mais célere a transformação dos textos em suas representações vetoriais a partir dos *word embeddings*.

REFERÊNCIAS

AKBIK, A.; BLYTHE, D.; VOLLGRAF, R. Contextual string embeddings for sequence labeling. In: . [S.l.: s.n.], 2018. Citado na página 40.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, v. 5, 07 2016. Citado na página 24.

BROWN, T. et al. Language models are few-shot learners. In: LAROCHELLE, H. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>. Citado na página 37.

CAMBRIA, E.; WHITE, B. Jumping nlp curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine*, v. 9, p. 48–57, 2014. Citado na página 30.

CANOTILHO, J. J. G. *Direito constitucional e teoria da constituição*. Almedina: Coimbra, 2003. Citado na página 23.

CNJ. *Relatório Analítico Anual da Justiça em Números 2021*. CNJ - Conselho Nacional de Justiça, 2021. Disponível em: <https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/>. Citado 3 vezes nas páginas 23, 24 e 25.

COLLOBERT, R. et al. Natural language processing (almost) from scratch. JMLR.org, v. 12, n. null, p. 2493–2537, nov 2011. ISSN 1532-4435. Citado 2 vezes nas páginas 32 e 35.

CSJT. *Gemini*. Conselho Superior da Justiça do Trabalho, 2020. Disponível em: <https://www.csjt.jus.br/web/csjt/justica-4-0/gemini>. Citado na página 83.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 10 2018. Citado 4 vezes nas páginas 24, 37, 38 e 41.

DUTTON, D. M.; CONROY, G. V. A review of machine learning. *The Knowledge Engineering Review*, Cambridge University Press, v. 12, n. 4, p. 341–367, 1997. Citado na página 29.

D'SILVA, J.; SHARMA, U. Unsupervised automatic text summarization of konkani texts using k-means with elbow method. *International Journal of Engineering Research and Technology*, v. 13, p. 2380, 09 2020. Citado na página 40.

.G, S.; JAYARAJU, P. A meta analysis of attention models on legal judgment prediction system. *International Journal of Advanced Computer Science and Applications*, v. 12, p. 532–538, 01 2021. Citado na página 39.

GUILLOU, P. Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...). In: . [S.l.: s.n.], 2020. Citado na página 61.

HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. 08 2017. Citado na página 43.

HASSAN, F. U.; LE, T. State-of-the-art review on the applicability of natural language processing (nlp) methods to address legal issues in construction. In: _____. *Construction Research Congress 2022*. [s.n.], 2022. p. 159–168. Disponível em: <https://ascelibrary.org/doi/abs/10.1061/9780784483978.017>. Citado na página 40.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 11 1997. ISSN 0899-7667. Citado na página 40.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 31, n. 3, p. 264–323, 1999. ISSN 0360-0300. Disponível em: <http://portal.acm.org/citation.cfm?id=331499.331504&coll=Portal&dl=ACM&CFID=26215063&CFTOKEN=18848029>. Citado na página 29.

KHAN, W. et al. A survey on machine learning models for natural language processing (nlp). v. 43, p. 95–113, 10 2016. Citado na página 24.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. *Mining of Massive Datasets*. 2nd. ed. USA: Cambridge University Press, 2014. ISBN 1107077230. Citado na página 31.

LI, R. Y. M. et al. Classification of construction accident court cases via natural language processing in hong kong. In: _____. *Current State of Art in Artificial Intelligence and Ubiquitous Cities*. Singapore: Springer Nature Singapore, 2022. p. 79–89. ISBN 978-981-19-0737-1. Disponível em: <https://doi.org/10.1007/978-981-19-0737-1_5>. Citado na página 40.

LIU, Y. et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. Citado 3 vezes nas páginas 24, 39 e 41.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Los Angeles: University of California Press, 1967. v. 1, p. 281–297. Citado 2 vezes nas páginas 40 e 43.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008. ISBN 978-0-521-86571-5. Disponível em: <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>. Citado na página 31.

MCCANN, B. et al. *Learned in Translation: Contextualized Word Vectors*. 2018. Disponível em: <https://arxiv.org/abs/1708.00107v2>. Citado na página 35.

MELAMUD, O.; GOLDBERGER, J.; DAGAN, I. context2vec: Learning generic context embedding with bidirectional lstm. In: . [S.l.: s.n.], 2016. p. 51–61. Citado na página 40.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, v. 2013, 01 2013. Citado 4 vezes nas páginas 24, 32, 33 e 40.

NAGEL, S. *Cc-news*. 2016. Common Crawl <https://commoncrawl.org/2016/10/>. Citado na página 39.

OLIVEIRA, F. J. V. *Os recursos na Justiça do Trabalho*. Conteúdo Jurídico, 2011. Disponível em: <http://www.conteudojuridico.com.br/consulta/Artigos/24853/os-recursos-na-justica-do-trabalho>. Citado na página 26.

OLIVEIRA, R.; NASCIMENTO, E. G. S. Clustering by similarity of brazilian legal documents using natural language processing approaches. In: _____. [S.l.: s.n.], 2021. Citado 3 vezes nas páginas 26, 28 e 61.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: . [S.l.: s.n.], 2014. v. 14, p. 1532–1543. Citado na página 24.

PETERS, M. et al. Deep contextualized word representations. In: . [S.l.: s.n.], 2018. p. 2227–2237. Citado na página 40.

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 na página 39.

RADFORD, A. et al. Improving language understanding by generative pre-training. In: . [S.l.: s.n.], 2018. Citado na página 36.

RADFORD, A. et al. Language models are unsupervised multitask learners. In: . [S.l.: s.n.], 2019. Citado 3 vezes nas páginas 24, 36 e 41.

RAFFEL, C. et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2019. Citado na página 41.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: . [S.l.: s.n.], 2019. p. 3973–3983. Citado na página 41.

RENUKA, S.; KIRAN, G. S. S. R.; ROHIT, P. An unsupervised content-based article recommendation system using natural language processing. In: _____. [S.l.: s.n.], 2021. p. 165–180. ISBN 978-981-15-8529-6. Citado na página 40.

SAENKO, K. et al. Adapting visual category models to new domains. In: *Proceedings of the 11th European Conference on Computer Vision: Part IV*. Berlin, Heidelberg: Springer-Verlag, 2010. (ECCV'10), p. 213–226. ISBN 364215560X. Citado na página 35.

SALUM, G. C. A duração dos processos no judiciário: aplicação dos princípios inerentes e sua eficácia no processo judicial. In: *Direito Processual Civil*. Rio Grande do Sul/Brazil: Âmbito Jurídico, 2016. v. 145. Citado na página 23.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1715–1725. Disponível em: <https://aclanthology.org/P16-1162>. Citado na página 36.

SHERSTINSKY, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, v. 404, p. 132306, 03 2020. Citado na página 40.

SIL, R. et al. Artificial intelligence and machine learning based legal application: The state-of-the-art and future research trends. In: . [S.l.: s.n.], 2019. p. 57–62. Citado na página 39.

SILVA, J. N. da. *roberta-pt-br*. 2021. Huggingface <https://huggingface.co/josu/roberta-pt-br>. Citado na página 61.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: _____. [S.l.: s.n.], 2020. p. 403–417. ISBN 978-3-030-61376-1. Citado na página 61.

TRT15. *Projetos desenvolvidos no TRT-15 são incorporados ao PJe 2.7*. Tribunal Regional do Trabalho 15ª Região, 2021. Disponível em: <https://trt15.jus.br/noticia/2021/projetos-desenvolvidos-no-trt-15-sao-incorporados-ao-pje-27>. Citado na página 83.

TRT5. *Gemini: Gabinetes do TRT5 participam de projeto-piloto que utiliza inteligência artificial*. Tribunal Regional do Trabalho 5ª Região, 2020. Disponível em: <https://www.trt5.jus.br/noticias/gemini-gabinetes-trt5-participam-projeto-piloto-que-utiliza-inteligencia-artificial>. Citado na página 83.

TRT5. *O PJe trará nova versão com um módulo do projeto Gemini, que tem participação do TRT5*. Tribunal Regional do Trabalho 5ª Região, 2020. Disponível em: <https://www.trt5.jus.br/noticias/pje-trara-nova-versao-com-modulo-projeto-gemini-que-tem-participacao-trt5>. Citado na página 83.

VASWANI, A. et al. Attention is all you need. 06 2017. Citado 4 vezes nas páginas 33, 34, 41 e 61.

WANG, Y.; CUI, L.; ZHANG, Y. Using dynamic embeddings to improve static embeddings. 11 2019. Citado na página 24.

WU, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. Citado na página 38.

ZHU, Y. et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 19–27, 2015. Citado 2 vezes nas páginas 36 e 38.