Preprocessing Applied to Legal Text Mining: analysis and evaluation of the main techniques used

Marcos V. J. da Silva[1], **Ewaldo E. Santana**[1], **Fábio M. F. Lobato**[1,2], **Antonio F. L. Jacob Jr.**[1]

[1]Programa de Pós-graduação em Engenharia da Computação e Sistemas
Universidade Estadual do Maranhão (UEMA)
São Luís – MA – Brasil

[2]Universidade Federal do Oeste do Pará (UFOPA)
Santarém – PA – Brasil

mvjanubis@gmail.com, antoniojunior@professor.uema.br

Abstract. *Text mining in the legal context requires effective preprocessing techniques to prepare the data for analysis. Given the unique legal vocabulary, a meticulous approach is necessary. The choice of preprocessing techniques can significantly influence the relevance of the extracted information. This research investigates the crucial preprocessing tasks involved in Legal Text Mining and systematically evaluates their impact on a classification problem. Through a series of experiments, eight different preprocessing tasks and their combinations were tested. Ultimately, it was found that the tasks with the best combined performance were: removal of numbers/digits; removal of links and emails; conversion of uppercase to lowercase; stemming; lemmatization; and tokenization.*

Resumo. *A mineração de textos no contexto jurídico requer técnicas eficazes de pré-processamento para preparar os dados para análise. Dado o vocabulário jurídico único, é necessária uma abordagem meticulosa. A escolha das técnicas de pré-processamento pode influenciar significativamente a relevância das informações extraídas. Esta pesquisa investiga as tarefas cruciais de pré-processamento envolvidas na Mineração de Textos Jurídicos e avalia sistematicamente seu impacto em um problema de classificação. Por meio de uma série de experimentos, foram testadas oito diferentes tarefas de pré-processamento e suas combinações. Por fim, obteve-se que as tarefas com melhor desempenho combinado foram: remoção de números/dígitos; remoção de links e e-mails; transformação de maiúsculas em minúsculas; stemização; lematização e tokenização.*

1. Introdução

O Poder Judiciário Brasileiro finalizou o ano de 2021 com 77,3 milhões de processos em tramitação, aguardando alguma solução definitiva [CNJ 2022]. Outros milhões de processos entraram nos anos de 2022 e 2023. Dada essa quantidade crescente de processos, é perceptível como o Sistema Judiciário encontra-se sobrecarregado para atender todas as demandas pendentes.

Técnicas de Aprendizado de Máquinas, como classificação, podem auxiliar na categorização de temas em conjuntos de documentos, como petições iniciais [Hagen 2018].

Com esta análise, entidades do setor jurídico podem acelerar a tomada de decisão, obtendo documentos relacionados definidos pelos temas categorizados na classificação.

A etapa de pré-processamento é crucial na mineração de textos para garantir a qualidade dos dados utilizados em tarefas como classificação [Wirth and Hipp 2000]. Ela envolve a limpeza, estruturação e organização dos documentos, proporcionando ganhos significativos nos resultados finais [Grancharova and Jangefalk 2018]. Mesmo com as melhores técnicas e parâmetros, o tratamento adequado dos dados é fundamental para evitar erros e maximizar a eficácia da análise [Ribeiro et al. 2020]. A constante pesquisa na área e a construção de Bases de Dados contribuem para o aperfeiçoamento contínuo dessa etapa [Souza et al. 2021].

Nesse contexto, esse trabalho analisa a utilização de algumas técnicas de pré-processamento de textos voltados para a mineração de dados jurídicos, aplicando as técnicas em documentos de petições iniciais e empregando modelos de Classificação, propondo um fluxo de pré-processamento construído a partir dos estudos realizados. A pesquisa permite identificar as melhores combinações de técnicas a serem aplicadas para os documentos mencionados, contribuindo assim para que futuros estudos na área de mineração de textos possam utilizar deste fluxo de pré-processamento.

2. Trabalhos Correlatos

[Cirqueira et al. 2017] revisaram métodos e técnicas utilizados no pré-processamento de dados em projetos de Análise de Sentimentos em mídias sociais, com foco na língua portuguesa do Brasil. Os autores concluíram que não existe um framework de pré-processamento padronizado na comunidade, com cada estudo seguindo passos específicos.

[Grancharova and Jangefalk 2018] investigaram o impacto de diferentes algoritmos e técnicas de pré-processamento na acurácia dos modelos de classificação. Através de testes, concluíram que a remoção de stopwords e o uso de um corretor ortográfico melhoram a acurácia em todos os classificadores estudados. A pesquisa fornece um fluxo para a realização dos testes e destaca a acurácia como uma métrica eficaz na comparação dos resultados dos classificadores.

[Işik and Dağ 2020] investigaram técnicas de pré-processamento e sua influência na classificação de avaliações de revisões. Concluíram que a remoção de stopwords e palavras comuns, a padronização de letras minúsculas e o uso de n-gramas de 1 a 3 apresentam melhor desempenho na melhoria da acurácia. Também discutiram os efeitos de abreviações, acrônimos, stemização e lematização após a transformação para letras minúsculas.

[Chandrasekar and Qian 2016] investigaram o impacto do pré-processamento de dados na performance do classificador Naive Bayes para identificação de spam em e-mails. Os testes realizados demonstraram que o pré-processamento adequado melhora significativamente os resultados. Os autores contribuem ao estabelecer um ponto de partida para análises comparativas com e sem pré-processamento.

A seguir, apresentam-se pesquisas sobre mineração de textos no contexto judiciário, destacando as principais tarefas de pré-processamento utilizadas. Essas informações permitem identificar as melhores abordagens para impulsionar as métricas de

desempenho.

[Andrade 2015] discute algoritmos de classificação para problemas com várias classes de documentos textuais e propõe um algoritmo de triagem de processos, com uma prova de conceito na triagem de denúncias na Controladoria-Geral da União. O autor empregou várias tarefas de pré-processamento para preparar os elementos textuais.

[de Castro Júnior et al. 2020] propõem uma aplicação de inteligência artificial para identificação de conexões entre fato e tese jurídica em petições iniciais. A ferramenta busca aprimorar a gestão do conhecimento no judiciário, permitindo a mineração de petições iniciais de processos, identificando semelhanças e casos com o mesmo fato e tese jurídica. No pré-processamento, foram utilizadas tarefas como limpeza de caracteres especiais, remoção de stopwords, stemização e tokenização.

[das Neves Junior et al. 2018] propõem uma solução de extração de informações para predição de resultados de sindicâncias, utilizando mineração de dados. A pesquisa utiliza o Diário Oficial de Pernambuco como fonte de dados e aplica tarefas de remoção de stopwords, stemização e tokenização no pré-processamento dos dados.

[Faraco 2020] propõe um modelo de conhecimento para a extração de tópicos em documentos de julgados (acórdãos), com foco na análise de petições iniciais. A Base de Dados consiste em acórdãos em formato JSON, obtidos por meio de um crawler da Justiça Federal da 4ª Região, TRF4. O modelo utiliza a técnica de Modelagem de Tópicos com o algoritmo LDA e aplica tarefas de pré-processamento, como remoção de caracteres especiais, números, acentuação, stopwords, lematização e tokenização.

[Sousa 2019] propõe uma solução de apoio para classificação de processos eletrônicos, utilizando mineração de textos e inteligência artificial. O projeto utiliza um corpus de petição inicial em PDF e aplica classificadores como Naive Bayes, KNN (K-Nearest Neighbors, ou "K-Vizinhos Mais Próximos"), Árvore de Decisão, SVM (Support Vector Machine, ou "Máquina de Vetores de Suporte") e K-Means. No pré-processamento, são realizadas tarefas como remoção de números, pontuações, stopwords, conversão de letras maiúsculas para minúsculas, stemização e tokenização.

[Mastella 2020] aplica mineração de textos para classificação de textos jurídicos, propondo uma metodologia que otimiza a escolha de parâmetros em algoritmos classificadores. O pré-processamento envolve tokenização, conversão para minúsculas, remoção de pontuações e stopwords.

[Castro 2019] emprega aprendizagem profunda para identificar Entidades Nomeadas em documentos jurídicos. O estudo utiliza Redes Neurais Profundas e realiza tarefas de pré-processamento, como remoção de conteúdo XML, números, pontuação e caracteres especiais, e tokenização.

[Ferreira 2018] busca classificar peças processuais do STF. O autor enfrenta problemas de falta de padronização nos documentos, o que prejudica a celeridade dos processos. O pré-processamento envolve remoção de números, espaços em branco, stopwords e tags XML, além de stemização e tokenização.

[Silva et al. 2021] compara abordagens de classificação para documentos jurídicos em português. Cinco metodologias foram avaliadas, com destaque para a vetorização TF-IDF e classificador SVM. O pré-processamento incluiu remoção de acentos, pontu-

ações, espaços em branco, stopwords, padronização para minúsculas, stemização e tokenização.

[Gusmão et al. 2021] investiga técnicas de PLN em Denúncias Criminais do Disque Denúncia RJ. O objetivo é agilizar a análise de mensagens informais e com erros morfológicos. Utilizando SVM, busca-se as melhores técnicas de pré-processamento, incluindo remoção de acentos, dígitos, pontuação, stopwords, padronização em minúsculas, tokenização e stemização.

A plataforma Sinapses/CNJ unifica modelos de inteligência artificial no âmbito jurídico para auxiliar tarefas do poder judiciário. Desenvolvida pelo Tribunal de Justiça de Rondônia, oferece bibliotecas em Python, como "ia utils", que inclui um módulo de normalização de texto. A versão 1.4.1, de 23 de fevereiro de 2018, realiza tarefas como remoção de espaços, generalização de termos, transformação de números para a forma escrita e padronização em minúsculas [Pereira and Rodrigues 2021].

Para a elaboração da Tabela 1 foram verificadas as tarefas exercidas pelos estudos na etapa de Pré-Processamento, seguindo a premissa de marcar apenas as tarefas que o estudo descreveu em sua pesquisa. Logo, para casos em que o estudo não deixa clara a utilização da tarefa, esta não consta na tabela. Para apontar as tarefas que cada pesquisa realizou, foram empregadas seções com tarefas que possuem objetivos similares ou fazem parte de um mesmo escopo, como por exemplo as tarefas de Stemização e Lematização. Assim, as tarefas dividem-se em: T1 - Remoção de Números e/ou Dígitos; T2 - Remoção de Espaços em Branco; T3 - Remoção de Pontuação, Acentuação e/ou caracteres especiais; T4 - Transformação de letras maiúsculas em minúsculas; T5 - Remoção de stopwords; T6 - Stemização e/ou Lematização; T7 - Tokenização; T8 - Remoção de conteúdo XML (tags) e/ou e-mails. Neste cenário, a Tabela 1 apresenta os estudos relacionados com as tarefas encontradas.

Table 1. Técnicas de Pré-Processamento Implementadas por Pesquisa

Tarefa	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**
Andrade (2015)	X	X	X	X	X	X		
Castro Júnior et al. (2020)			X	X	X	X	X	
Junior et al. (2018)					X	X	X	
Faraco (2020)	X		X		X	X	X	
Sousa (2019)	X		X	X	X	X	X	
Mastella (2021)			X	X	X		X	
Castro (2019)	X		X				X	X
Ferreira (2018)	X	X	X		X	X	X	X
Silva et al. (2021)		X	X	X	X	X		
Gusmão (2021)	X		X	X	X	X	X	
Normalizador da Plataforma Sinapses/CNJ	*	X		X		*		
TOTAL	7	4	9	7	9	9	8	2

A marcação (X) representa que o trabalho desenvolvido pelos autores implementa aquele tipo de tarefa de Pré-Processamento. Para o Normalizador da Plataforma Sinapses/CNJ as tarefas T1 e T6 possuem algumas observações (*), pois não empregam a técnica da forma como é descrita. A tarefa T1 (Remoção de Números e/ou Dígitos) não é

implementada pelo Normalizador da Plataforma Sinapses/CNJ, porém existe uma conversão de números para a forma escrita por extenso. Já a tarefa T6 (Stemização/Lematização) também não é incorporada, mas há uma forma de generalização de termos nos textos processados.

3. Framework Experimental

A seguir serão descritas as etapas do framework de experimentação que busca realizar os testes para as técnicas de pré-processamento escolhidas. O framework segue os níveis de incorporação das técnicas: 0 (sem pré-processamento), 1 (técnicas individuais) e 2 (combinação de técnicas). Estes níveis seguem um fluxo definido por [Grancharova and Jangefalk 2018], utilizando a acurácia como métrica de comparação.

3.1. Descrição da Base de Dados e Preparação dos Testes

Os experimentos utilizaram uma base de dados composta por documentos jurídicos de petições iniciais com 3000 elementos fornecida pelo Tribunal de Justiça do Maranhão, como parte da parceria entre o TJMA e a Universidade Estadual do Maranhão. A base de dados é composta por cinco classes de petições: "IRDR1", "RR986", "RR1", "RR1061", "RR1074".

Na fase de preparação, é importante definir quais técnicas de pré-processamento serão implementadas nos testes, incluindo remoção de números, espaços em branco, pontuação, acentuação e caracteres especiais, além de transformação de letras maiúsculas para minúsculas, remoção de stopwords, stemização, lematização, tokenização e remoção de e-mails.

Para realizar a vetorização dos dados textuais dos documentos foi utilizado o modelo TF-IDF (Term Frequency — Inverse Data Frequency), convertendo assim os dados para o formato númerico antes de serem repassados para os classificadores. A classificação foi escolhida como algoritmo alvo para os testes, extraindo a acurácia de três classificadores diferentes: SVM (Support Vector Machine, ou "Máquina de Vetores de Suporte"), KNN (K-Nearest Neighbors, ou "K-Vizinhos Mais Próximos") e LR (Logistic regression, ou "Regressão Logística"). Foi utilizada a técnica de validação cruzada k-fold para avaliação dos modelos com k igual a 10, gerando assim 10 acurácias diferentes para cada classificador, sendo registradas as acurácias para o melhor caso e a média entre as 10 obtidas.

A Figura 1 apresenta sucintamente os níveis que compõem os testes para análise, Performance sem Pré-Processamento, Técnicas Individuais e Combinações de Técnicas, que serão descritos a seguir.

3.2. Performance sem Pré-Processamento

O primeiro passo é obter acurácias sem a aplicação das técnicas de pré-processamento (acurácias de baseline). Este é o nível 0, seguindo os passos: aplicar os classificadores à base de dados e anotar suas métricas.

3.3. Técnicas Individuais

O próximo nível de testes é o Nível 1, onde são estudadas as técnicas de forma de singular, sem mesclar nenhuma delas. Dessa forma, esta fase de testes iniciará o entendimento

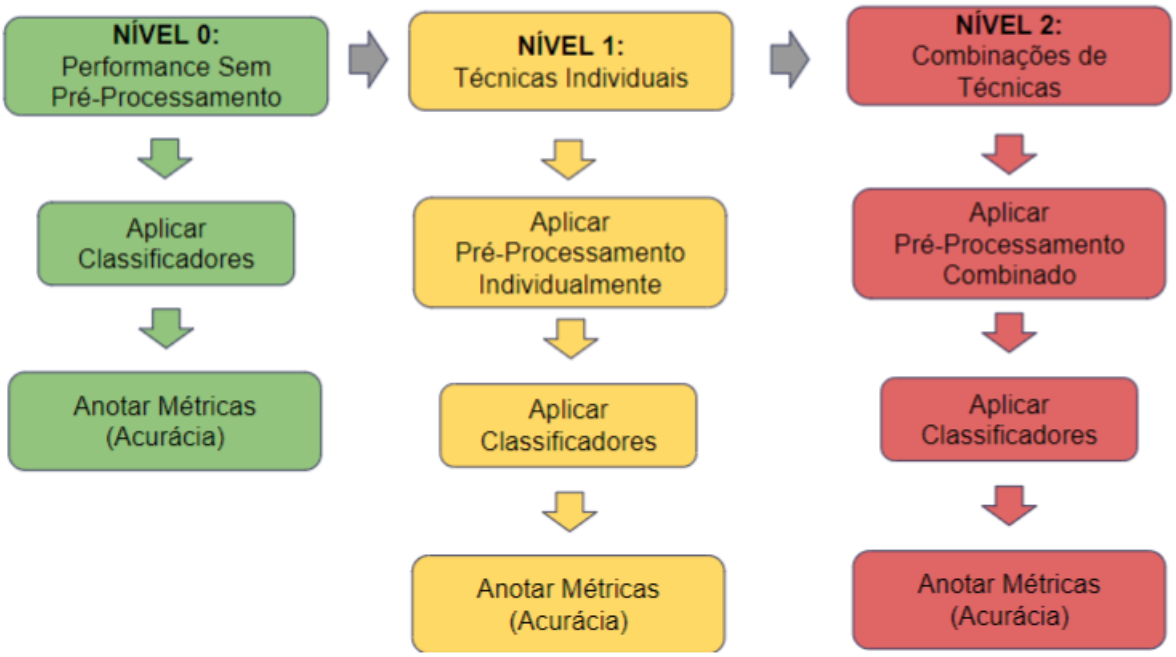

Figure 1. Níveis de Execução do Framework Experimental

de quais são as principais técnicas a serem exploradas na composição de uma estrutura de pré-processamento voltada para textos jurídicos. Os passos de execução do Nível 1 seguem: aplicar o pré-processamento individualmente à base de dados, aplicar os classificadores e anotar as métricas.

3.4. Combinações de Técnicas

Seguindo com a realização dos testes, o Nível 2 de desenvolvimento consiste em realizar combinações das técnicas. Os passos a serem realizados são: aplicar o pré-processamento combinado à base de dados, aplicar os classificadores e anotar as métricas.

Inicialmente, são tomadas tarefas de duas em duas, anotando as acurácias encontradas. Então são tomadas tarefas de três em três para a combinação. Após isso, para avançar nas combinações foram definidos alguns fatores que as combinações precisariam cumprir para avançar para a próxima etapa. As regras que a combinação precisa cumprir estão descritas na tabela 2.

Table 2. Descrição das Tarefas Utilizadas

	Regra	**Exemplo para Combinação de 3**
Regra 1	Cx/.../z > (Tx,...,Tz)	Cx/y/z > (Tx,Ty,Tz)
Regra 2	Cx/.../z > MED(Tx,...,Tz)	Cx/y/z > MED(Tx,Ty,Tz)
Regra 3	Cx/.../z > MAX(Tx,...,Tz)	Cx/y/z > MAX(Tx,Ty,Tz)

A regra 1 é definida como "A acurácia da combinação deve ser maior que a acurácia das tarefas combinadas". A regra 2 representa "A acurácia da combinação deve ser maior que a média das acurácia das tarefas combinadas". Já a regra 3 segue como "A acurácia da combinação deve ser maior que a acurácia máxima entre as tarefas combinadas".

4. Resultados Obtidos

Segundo a arquitetura descrita, foram elaborados e executados os testes para as técnicas de forma individual e as combinações de tarefas. Para melhor apresentação as tabelas seguem a descrição TX (Tarefa X), sendo X o número da tarefa, com a seguinte ordem apresentada na tabela 3.

Table 3. Descrição das Tarefas Utilizadas

Tarefa 1 (T1)	Remoção de Números/Dígitos
Tarefa 2 (T2)	Remoção de links e e-mails
Tarefa 3 (T3)	Remoção de Espaços em Branco especiais
Tarefa 4 (T4)	Remoção de stopwords
Tarefa 5 (T5)	Transformação de maiúsculas em minúsculas
Tarefa 6 (T6)	Remoção de Pontuação, Acentuação e caracteres
Tarefa 7 (T7)	Stemização
Tarefa 8 (T8)	Lematização
Tarefa 9 (T9)	Tokenização

A tabela 4 apresenta as acurácias médias encontradas para os classificadores sem a utilização de técnicas de pré-processamento (Nível 0) e para as Técnicas Individuais (Nível 1).

Table 4. Performances dos Níveis 0 e 1 - Acurácia Média

Tarefa Utilizada	**Regressão Logística %(DP)**	**KNN %(DP)**	**SVM %(DP)**
Sem Pré-Processamento	73,83 (0,03652)	72,66 (0,04353)	68,33 (0,03821)
Tarefa 1	73,80 (0,03773)	73,09 (0.03796)	68,61 (0,03907)
Tarefa 2	73,63 (0,03775)	72,64 (0,04247)	68,04 (0,03604)
Tarefa 3	73,83 (0,03652)	72,66 (0,04353)	68,33 (0,03821)
Tarefa 4	72,24 (0,03529)	72,39 (0,03812)	68,07 (0,03830)
Tarefa 5	73,83 (0,03652)	72,66 (0,04353)	68,33 (0,03821)
Tarefa 6	74,02 (0,03587)	72,61 (0,04217)	68,43 (0,03757)
Tarefa 7	74,04 (0,03945)	73,89 (0,04379)	68,82 (0,03859)
Tarefa 8	73,73 (0,03524)	73,20 (0,03725)	68,59 (0,03889)
Tarefa 9	73,98 (0,03879)	73,85 (0,04118)	68,53 (0,03725)

A tabela 5 apresenta as acurácias de melhor caso encontradas para os classificadores sem a utilização de técnicas de pré-processamento (Nível 0) e para as Técnicas Individuais (Nível 1). Os valores destacados representam as acurácias das Técnicas Individuais (Nível 1) que ultrapassaram os valores base sem pré-processamento para cada classificador.

É possível identificar que, para alguns classificadores, as tarefas de pré-processamento executadas não revelaram nenhum ou quase nenhum ganho. Em alguns casos, os resultados foram até abaixo da acurácia base sem pré-processamento, que podemos chamar de acurácias negativas. As acurácias que foram acima da base podem ser chamadas de positivas.

Table 5. Performances dos Níveis 0 e 1

Tarefa Utilizada	**Regressão Logística (%)**	**KNN (%)**	**SVM (%)**
Sem Pré-Processamento	82,13	79,24	74,95
Tarefa 1	81,79	**80,44**	74,95
Tarefa 2	**82,48**	**79,71**	74,95
Tarefa 3	82,13	79,24	74,95
Tarefa 4	79,24	**79,76**	74,95
Tarefa 5	82,13	79,24	74,95
Tarefa 6	**84,56**	**81,22**	74,95
Tarefa 7	**83,30**	**80,12**	**75,45**
Tarefa 8	79,58	**81,03**	74,60
Tarefa 9	**82,38**	**82,14**	74,24

A partir disso, foi escolhido apenas um classificador para realizar os resultados seguintes. Como observado na tabela 5, o classificador KNN obteve 5 resultados positivos, em comparação com Regressão Logística (4 resultados positivos) e SVM (apenas um resultado positivo). Logo, todos os resultados encontrados para as etapas seguintes são acurácias encontradas com o classificador KNN.

Seguindo para o Nível 2, foi iniciada a etapa de combinação das tarefas. Primeiramente, foram definidas combinações agregando de duas em duas tarefas, como exemplo: Remoção de Números/Dígitos e Remoção de links e e-mails. Ao todo, foram obtidos 72 resultados para esta etapa, combinando as 9 tarefas entre si sem que uma mesma tarefa fosse repetida na combinação. Assim, as tabelas 6 e 7 apresentam as acurácias encontradas para a Combinação de Duas Técnicas, sendo acurácia média e acurácia máxima, respectivamente. As linhas representam qual tarefa foi executada primeiro e as colunas representam a tarefa que foi executada por último.

Table 6. Performances do Nível 2 - Combinação de 2 tarefas - Acurácia Média

	T1 (%)	T2 (%)	T3 (%)	T4 (%)	T5 (%)	T6 (%)	T7 (%)	T8 (%)	T9 (%)
T1	X	72.557	72.563	72.545	**73.090**	72.563	**73.774**	**73.092**	**73.439**
T2	**73.520**	X	72.563	72.557	**73.520**	72.563	**73.774**	**73.092**	**73.439**
T3	**73.090**	72.557	X	72.545	**73.090**	72.563	**73.774**	**73.092**	**73.439**
T4	72.537	72.557	72.554	X	72.545	72.410	**73.774**	**72.459**	**73.500**
T5	**73.090**	72.557	72.563	72.545	X	72.563	**73.774**	**73.092**	**73.439**
T6	72.001	71.442	**72.942**	71.442	72.001	X	72.117	72.942	**73.565**
T7	**74.215**	**73.815**	**72.838**	**73.315**	**73.989**	**73.077**	X	**74.008**	**73.053**
T8	**74.221**	**73.627**	**73.584**	72.582	**73.584**	72.467	**73.188**	X	**73.286**
T9	**73.418**	**73.857**	**73.439**	**73.664**	**73.418**	**73.439**	**73.910**	72.300	X

Com estes resultados, a próxima etapa para o Nível 2 foi realizar a combinação das tarefas tomando de três em três. Para que não fosse necessário realizar várias combinações trocando a ordem de cada tarefa na execução, foi definida uma ordem para as tarefas de acordo com sua relação na combinação de duas em duas. Por exemplo, para as tarefas T1 e T2, quando se toma T1 anteriormente a T2, a acurácia encontrada é **81.910**, já tomando T2 antes de T1 a acurácia se torna **80.444**. Logo, percebe-se que tomando primeiramente

Table 7. Performances do Nível 2 - Combinação de 2 tarefas - Acurácia Máxima

	T1 (%)	T2 (%)	T3 (%)	T4 (%)	T5 (%)	T6 (%)	T7 (%)	T8 (%)	T9 (%)
T1	X	**81.910**	**81.094**	**82.024**	**80.444**	**81.094**	**80.166**	**81.094**	**79.350**
T2	**80.444**	X	**81.094**	**81.910**	**80.444**	**81.094**	**80.166**	**81.094**	**79.350**
T3	**80.444**	**81.910**	X	**82.024**	**80.444**	**81.094**	**80.166**	**81.094**	**79.350**
T4	**82.024**	**81.910**	**78.534**	X	**82.024**	79.002	**80.166**	78.996	78.438
T5	**80.444**	**81.910**	**81.094**	**82.024**	X	**81.094**	**80.166**	**81.094**	**79.350**
T6	**81.612**	78.647	**83.241**	78.647	**81.612**	X	**82.073**	**83.241**	**82.085**
T7	**81.857**	**80.630**	78.994	**80.050**	**81.624**	**80.048**	X	**81.091**	**81.449**
T8	**82.779**	**81.458**	**81.805**	76.912	**81.805**	78.207	**79.654**	X	**83.075**
T9	**81.964**	**80.066**	**79.350**	**80.079**	**81.964**	**79.350**	**81.810**	77.124	X

T1 o resultado é mais relevante.

Com isso, a ordem passada para a combinação de três tarefas se deu acordo com os resultados encontrados na combinação de duas.Dessa forma, tomando as 9 tarefas para combinar foram obtidos 84 resultados de acurácias. A tabela 8 apresenta os resultados para a combinação de três tarefas apenas para aqueles que satisfizeram todas as condições descritas na tabela 2.

Table 8. Combinações de 3 tarefas que satisfizeram as condições

			Ordem de execução	**Acurácia Média%(DP)**	**Acurácia Máxima (%)**
T1	T2	T4	$\mapsto$ T1 $\rightarrow$ T2 $\rightarrow$ T4	73,575 (0,04268)	82.139
T1	T2	T5	$\mapsto$ T1 $\rightarrow$ T5 $\rightarrow$ T2	73,575 (0,04268)	82.139
T1	T7	T9	$\mapsto$ T7 $\rightarrow$ T9 $\rightarrow$ T1	73,727 (0,04196)	81.476
T2	T3	T5	$\mapsto$ T5 $\rightarrow$ T3 $\rightarrow$ T2	73,764 (0,03652)	81.141
T2	T7	T9	$\mapsto$ T7 $\rightarrow$ T9 $\rightarrow$ T2	73,684 (0,04389)	**83.541**
T2	T8	T9	$\mapsto$ T8 $\rightarrow$ T9 $\rightarrow$ T2	73,242 (0,04201)	**83.888**
T3	T4	T5	$\mapsto$ T3 $\rightarrow$ T4 $\rightarrow$ T5	73,683 (0,03848)	82.411
T7	T8	T9	$\mapsto$ T7 $\rightarrow$ T8 $\rightarrow$ T9	73,563 (0,04532)	**85.518**

Caso qualquer uma das regras não fossem cumpridas, a combinação seria descartada. Dos 84 resultados, apenas 8 resultados de acurácia conseguiram satisfazer as condições, eliminando combinações que não produzem ganho significativo.

Também, é possível perceber que algumas combinações alcançaram resultados muito elevados em comparação com a execução sem pré-processamento do KNN, como é o caso para a combinação entre T7, T8 e T9 com acurácia de **85.518**. Os resultados destacados são aqueles que conseguiram ultrapassar as acurácias para a combinação apenas com duas tarefas.

Prosseguindo com os testes do Nível 2, a próxima etapa seria realizar combinações aumentando o número de tarefas combinadas para quatro, cinco e assim sucessivamente. Para encontrar quais seriam as próximas combinações mais prováveis de sucesso, foram escolhidas somente as tarefas que aparecem na tabela 8. Assim, foram mescladas todas as linhas da tabela 8 para encontrar as combinações adequadas. Por exemplo, ao mesclar a linha 1 (T1, T2 e T4) com a linha 2 (T1, T2 e T5) obtemos uma combinação de quatro

tarefas (T1, T2, T4 e T5). Mesclando a linha 1 (T1, T2 e T4) com a linha 3 (T1, T7 e T9) obtemos uma combinação de cinco tarefas (T1, T2, T4, T7 e T9).

Neste contexto, todas as combinações apresentadas na tabela 9 com suas acurácias foram elaboradas seguindo a regra mencionada anteriormente.

Table 9. Combinações com quatro ou mais tarefas

Tarefas Utilizadas	Acurácia Média%(DP)	Acurácia Máxima (%)
T1, T2, T4, T5	73,575 (0,04268)	82.139
T1, T2, T5, T3	73,575 (0,04268)	82.139
T1, T7, T9, T2	72,875 (0,04260)	81.225
T1, T7, T9, T8	72,076 (0,03953)	80.763
T2, T3, T5, T4	71, 994 (0,04013)	78.088
T2, T7, T9, T8	71,872 (0,04356)	81.344
T1, T2, T4, T7, T9	71,896 (0,04349)	81.344
T1, T2, T4, T3, T5	71,877 (0,04138)	78.079
T1, T2, T4, T8, T9	72,082 (0,04594)	83.206
T1, T2, T5, T7, T9	72,031 (0,04474)	81.691
T1, T2, T5, T8, T9	72,155 (0,04428)	81.691
T1, T2, T5, T3, T4	71,928 (0,04148)	78.426
T1, T7, T9, T2, T8	71,965 (0,04417)	81.691
T2, T3, T5, T7, T9	71,961 (0,04415)	81.691
T2, T3, T5, T8, T9	71,996 (0,04400)	81.577
T1, T2, T4, T7, T8, T9	71,946 (0,04413)	81.577
T1, T2, T5, T7, T8, T9	71,946 (0,04413)	81.577
T1, T7, T9, T2, T3, T5	71,946 (0,04413)	81.577
T1, T7, T9, T3, T4, T5	71,946 (0,04413)	81.577
T2, T3, T5, T7, T8, T9	71,946 (0,04413)	81.577
T2, T7, T9, T3, T4, T5	71,946 (0,04413)	81.577
T2, T8, T9, T3, T4, T5	71,969 (0,04418)	81.577
T3, T4, T5, T7, T8, T9	71,946 (0,04413)	81.577

A tabela 10 apresenta o resultado para a combinação com quatro ou mais tarefas que satisfez todas as condições descritas na tabela 2. Como pode ser observado, apenas a combinação de seis (T1, T2, T5, T7, T8 e T9) cumpriu as condições. Assim, como não há mais nenhuma forma de combinação que adicione novas tarefas, pode-se atribuir esta combinação como sendo a que possui maior quantidade de tarefas e que alcançou maior acurácia, dadas as tarefas executadas e a base de dados empregada.

Table 10. Combinações com quatro ou mais tarefas que satisfizeram as condições

Tarefas Utilizadas	Acurácia Média%(DP)	Acurácia Máxima (%)
T1, T2, T5, T7, T8, T9	71,946 (0,04413)	81.577

5. Conclusão

A etapa de Preparação de Dados em mineração de textos, se bem estruturada, consegue agregar diversos ganhos para todo o processo de Descoberta de Conhecimento. Neste contexto, dados jurídicos possuem características próprias que devem ser corretamente tratadas, a fim de serem utilizadas no processo e, consequentemente, gerar informações relevantes.

O presente trabalho teve como objetivos realizar análise e avaliação das principais técnicas empregadas no pré-processamento para bases de texto jurídicas, apresentando os resultados dos testes executados. Ao total, nove tarefas foram testadas individualmente e em conjunto para tratar uma base de dados de petições iniciais.

Apesar dos testes realizarem as combinações de até seis tarefas simultânes, é importante ressaltar que a combinação que conseguiu alcançar maior valor (**85,518%**) foi uma combinação de três tarefas: Stemização, Lematização e Tokenização.

Como trabalhos futuros, pretende-se realizar mais testes com diferentes métodos de classificação e problemas diferentes no contexto jurídico.

Agradecimentos

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

References

Andrade, P. (2015). *Aplicação de Técnicas de Mineração de Textos para Classificação de Documentos: um Estudo da Automatização da Triagem de Denúncias na CGU. Brasília, 2015. 65p*. PhD thesis, Dissertação (Mestrado Profissional em Computação Aplicada). Disponível em ....

Castro, P. V. Q. d. (2019). *Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico*. PhD thesis, Universidade Federal de Goiás.

Chandrasekar, P. and Qian, K. (2016). The impact of data preprocessing on the performance of a naive bayes classifier. In *2016 IEEE 40th annual computer software and applications conference (COMPSAC)*, volume 2, pages 618–619. IEEE.

Cirqueira, D., Jacob, A., Lobato, F., de Santana, A. L., and Pinheiro, M. (2017). Performance evaluation of sentiment analysis methods for brazilian portuguese. In *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19*, pages 245–251. Springer.

CNJ, C. N. D. J. (2022). A justiça em números - relatório analítico 2022.

das Neves Junior, R. B., de Medeiros Melo, W. F., de Araujo Fagundes, R. A., and Maciel, A. M. A. (2018). Extração de informação e mineração de dados no diário oficial de pernambuco. *Revista de Engenharia e Pesquisa Aplicada*, 3(3).

de Castro Júnior, A. P., Calixto, W. P., and de Castro, C. H. A. (2020). Aplicação da inteligência artificial na identificação de conexões pelo fato e tese jurídica nas petições iniciais e integração com o sistema de processo eletrônico. *CNJ*, page 9.

Faraco, F. M. (2020). *Modelo de conhecimento baseado em tópicos de acórdãos para suporte à análise de petições iniciais*. PhD thesis, Universidade Federal de Santa Catarina.

Ferreira, M. H. P. (2018). *Classificação de peças processuais jurídicas: Inteligência Artificial no Direito*. PhD thesis, Universidade de Brasília - UnB, Faculdade UnB Gama - FGA.

Grancharova, M. and Jangefalk, M. (2018). Comparative study of the combined performance of learning algorithms and preprocessing techniques for text classification.

Gusmão, C., Figueiredo, K., and Brito, W. A. (2021). Técnicas de processamento de linguagem natural em denúncias criminais: Automatizaçao e classificaçao de texto em português coloquial. In *Anais do XLVIII Seminário Integrado de Software e Hardware*, pages 172–182. SBC.

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate lda models? *Information Processing & Management*, 54(6):1292–1307.

Işik, M. and Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences*, 28(3):1405–1421.

Mastella, J. O. (2020). Uma metodologia usando ambientes paralelos para otimização da classificação de textos aplicada a documentos jurídicos. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

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

Ribeiro, E. R. et al. (2020). *Impacto de técnicas de pré-processamento de texto na detecção de intenção e extração de parâmetros em sistemas de diálogo orientados a tarefa*. PhD thesis, Universidade Federal do Amazonas.

Silva, J. A., Nogueira Jr, V., Oliveira, H., Barbosa, A., Vieira, T., and Oliveira, K. (2021). Avaliação de abordagens para classificação automática de documentos jurídicos: um estudo comparativo aplicado a petições do tribunal de justiça do estado de alagoas. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, 8(1).

Sousa, R. N. d. (2019). *MINERJUS: solução de apoio à classificação processual com uso de Inteligência Artificial*. PhD thesis, Universidade Federal do Tocantins.

Souza, E., Moriyama, G., Vitório, D., de Carvalho, A. C., Félix, N., Albuquerque, H. O., and Oliveira, A. L. (2021). Assessing the impact of stemming algorithms applied to brazilian legislative documents retrieval. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 227–236. SBC.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.