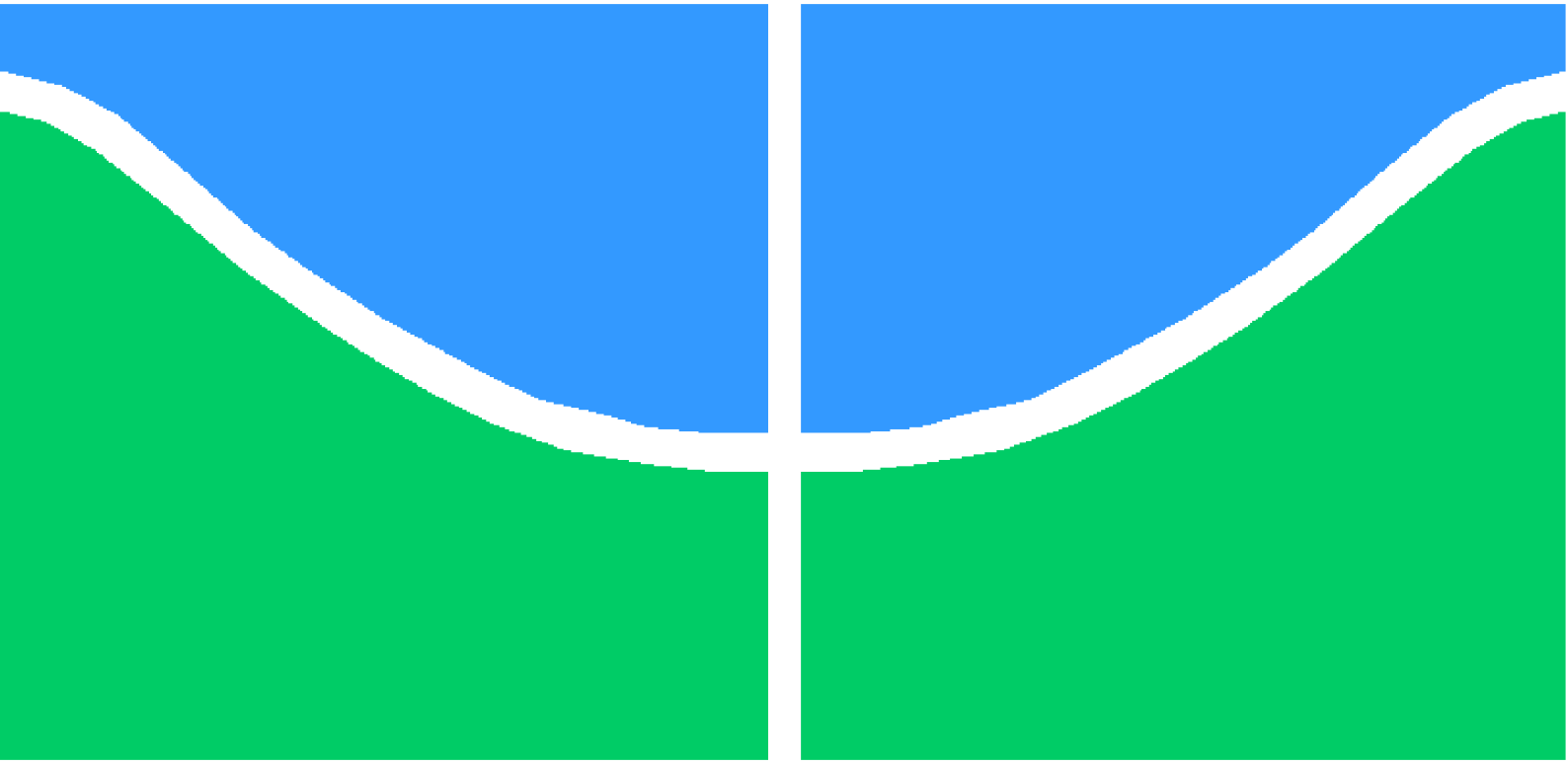


Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Engenharia de Software

Expansão Inteligente de Amostras Rotuladas: Inovações e Desafios na Justiça Brasileira

Autor: Jonathan Jorge Barbosa Oliveira

Orientador: Prof. Dr. Nilton Correia da Silva

Brasília, DF

2023

Jonathan Jorge Barbosa Oliveira

Expansão Inteligente de Amostras Rotuladas: Inovações e Desafios na Justiça Brasileira

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Orientador: Prof. Dr. Nilton Correia da Silva

Coorientadora: Prof[a]. Dra. Debora Bonat

Brasília, DF

2023

Jonathan Jorge Barbosa Oliveira
Expansão Inteligente de Amostras Rotuladas: Inovações e Desafios na Justiça Brasileira/ Jonathan Jorge Barbosa Oliveira. – Brasília, DF, 2023-
101 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Nilton Correia da Silva

Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
Faculdade UnB Gama - FGA , 2023.

1. Rotulagem Automatizada. 2. Precedentes Jurídicos. I. Prof. Dr. Nilton Correia da Silva. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Expansão Inteligente de Amostras Rotuladas: Inovações e Desafios na Justiça Brasileira

CDU

Jonathan Jorge Barbosa Oliveira

Expansão Inteligente de Amostras Rotuladas: Inovações e Desafios na Justiça Brasileira

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 21 de Dezembro de 2023:

Prof. Dr. Nilton Correia da Silva
Orientador

Prof. Dr. Fabricio Ataíde Braz
Convidado 1

Prof. Dr. Fabiano Hartmann Peixoto
Convidado 2

Brasília, DF
2023

Dedico este trabalho à minha mãe, a incrível mulher que, com mãos firmes e coração tenaz, enfrentou o mundo como mãe solo e diarista. Em cada movimento de suas mãos, ensinou-me a importância da resiliência, da dedicação e do amor incondicional. Suas jornadas longas e os sacrifícios diários foram a minha inspiração e motivação para perseguir meus sonhos. Obrigado, mãe, por ser meu farol e por iluminar meu caminho com sua força e sabedoria. Esta conquista é tanto minha quanto sua.

Agradecimentos

Em momentos cruciais como este, onde colhemos os frutos de uma longa e intensa jornada de estudos, é inevitável não olhar para trás e lembrar de todos aqueles que estiveram ao nosso lado, guiando, apoiando e acreditando em nosso potencial.

Primeiramente, minha mais profunda gratidão à minha família. A cada membro que, direta ou indiretamente, ofereceu palavras de conforto nos momentos de dúvida, celebrou pequenas vitórias e esteve sempre presente, minha eterna gratidão. Uma menção especial é reservada à minha mãe, cujo amor incondicional, força e dedicação não apenas me inspiraram, mas me deram as asas necessárias para voar alto. Seu papel em minha vida e neste trabalho é imensurável, e eu sou eternamente grato por tudo o que fez e continua fazendo por mim.

Agradeço, também, aos meus colegas de curso. Nossa caminhada, repleta de desafios, debates e descobertas, foi enriquecida pela diversidade de pensamentos e pelo espírito colaborativo que sempre prevaleceu entre nós. As longas noites de estudo, as dúvidas compartilhadas e os momentos de descontração foram essenciais para formar o profissional e a pessoa que sou hoje.

Reservo um agradecimento especial ao laboratório Ai.Lab, que mais do que um espaço de pesquisa, tornou-se um ambiente de inovação e crescimento. A oportunidade de aprimorar meus conhecimentos, trabalhar com mentes brilhantes e ter acesso a ferramentas e metodologias de ponta foi fundamental para a concretização deste trabalho. A experiência no Ai.Lab não apenas ampliou meus horizontes acadêmicos, mas moldou minha visão sobre a ciência e a pesquisa.

Concluo este agradecimento ciente de que cada pessoa e instituição mencionada aqui foi essencial para que este Trabalho de Conclusão de Curso fosse não apenas um documento, mas a síntese de uma trajetória repleta de aprendizado, superação e gratidão.

"A ciência de hoje é a tecnologia de amanhã."
(EDWARD TELLER, 2000)

Resumo

No campo da aprendizagem de máquina, a rotulação manual de dados, especialmente no setor jurídico, é fundamental, mas desafiadora. Este estudo se concentra no desenvolvimento de um algoritmo de aprendizado semi-supervisionado para expandir uma base de dados jurídica, com o objetivo de propagar precedentes qualificados que não são identificáveis através de expressões regulares (regex) em sentenças de processos. A Análise Exploratória de Dados (AED) é empregada para auxiliar na ampliação dessa rotulação de dados. A validade e precisão do método são asseguradas por avaliações qualitativas conduzidas por profissionais do direito. Ademais, o estudo explora o uso de algoritmos de classificação transdutivos semi-supervisionados para incrementar a eficiência do sistema judiciário. Entre os algoritmos avaliados, o "Mais Próximo"demonstrou ser o mais eficaz, aumentando em 820% as amostras rotuladas na base jurídica, e mostrando grande potencial para melhoria de desempenho com a adição de mais dados.

Palavras-chaves: Aprendizado Semi-Supervisionado, Rotulação de Dados, Processamento de Linguagem Natural, Precedentes Qualificados, Inteligência Artificial Jurídica, Inteligência Artificial e Direito.

Abstract

This study delves into the challenges of manual data labeling in the field of machine learning, particularly within the legal sector. It focuses on developing a semi-supervised learning algorithm aimed at enhancing a legal database. The primary goal is to disseminate qualified precedents that are not identifiable using regular expressions (regex) in legal case texts. To facilitate the expansion of data labeling, Exploratory Data Analysis (EDA) is utilized. The method's validity and accuracy are confirmed through qualitative evaluations conducted by legal professionals. Furthermore, the research investigates the application of transductive semi-supervised classification algorithms to improve the efficiency of the judicial system. Among the algorithms evaluated, the "Nearest Neighbor" algorithm emerged as the most effective, achieving an 820% increase in labeled samples in the legal database. This significant enhancement highlights the algorithm's potential for performance improvement with the integration of additional data.

Key-words: Semi-supervised Learning, Data Labeling, Natural Language Processing, Qualified Precedents, Legal Artificial Intelligence, Artificial Intelligence and Law.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AED	Análise Exploratória de Dados
AI.Lab	Laboratório de inteligência artificial
AUC	Area Under the Curve
BoW	*Bag of Word*
CNJ	Conselho Nacional de Justiça
CPC	Código de Processo Civil
DR.IA	Direito e Inteligência Artificial
HL	Hamming Loss
KNN	K-nearest neighbors
PEDRO	Plataforma de Extração e Descoberta de Precedentes dos Tribunais
PDR	Propagação Dinâmica de Rótulos
PLN	Processamento de Linguagem Natural
RegEx	Regular expression
ROC	*Receiver Operating Characteristic*
Rop	Rotulador de Processos
STF	Supremo Tribunal Federal
STJ	Superior Tribunal de Justiça
TF-IDF	*Term Frequency — Inverse Data Frequency*
UnB	Universidade de Brasília

Lista de símbolos

X	Espaço de Instâncias
λ	Rótulo
L	Conjunto de Rótulos
x	$x = (x_1, x_2, \ldots, x_m) \in X$ - Instância
m	Número de Características
k	$k = \|L\|$ - Número de Rótulos
L	Conjunto de rótulos relevantes (para uma instância)
Y	$Y = (y_1, y_2, \ldots, y_k)$, $y_i \in \{0, 1\}$, $1 \leq i \leq k$ - Vetor de Rótulo Correto
$\mathcal{Y}$	Espaço do Vetor de Rótulo: $\mathcal{Y} = \{0, 1\}^k$
Y_l	$Y_l = (y_{l1}, y_{l2}, \ldots, y_{lp})$, $y_{li} \in L$, $1 \leq i \leq p$
Y_λ	Presença de Rótulo: $Y_\lambda \in \{0, 1\}$
p	Número de rótulos para uma instância
n	Número de Instâncias
Z	$Z = (z_1, z_2, \ldots, z_k)$, $z_i \in \{0, 1\}$, $1 \leq i \leq k$ - Predições do modelo
Z_l	$Z_l = (z_{l1}, z_{l2}, \ldots, z_{lp})$, $z_{li} \in L$, $1 \leq i \leq p$ - Conjunto de rótulos previstos

Sumário

1 Introdução

O laboratório de inteligência artificial da Universidade de Brasília (AI.Lab - UNB) juntamente com o Direto e Inteligência Artificial (DR.IA - UNB) estabeleceram uma parceria com o Conselho Nacional de Justiça (CNJ) para desenvolver o projeto "Plataforma de Extração e Descoberta de Precedentes dos Tribunais" (PEDRO). O objetivo do projeto é fornecer uma solução que auxilie os magistrados na busca por precedentes relevantes ao julgarem casos, facilitando o processo de tomada de decisão.

A colaboração entre AI.Lab - UNB, DR.IA - UNB e CNJ visa modernizar o judiciário ao integrar métodos de inteligência artificial, otimizando a análise de dados e proporcionando julgamentos mais assertivos. Essa iniciativa enfatiza a interseção crucial entre Direito e Tecnologia no atual cenário brasileiro.

A solução proposta consiste em um modelo de agrupamento não supervisionado que identifica precedentes qualificados para uma peça inicial, com base na similaridade com amostras pertencentes a grupos específicos. Além disso, é utilizado um rotulador de processos (RoP) que utiliza expressões regulares (RegEx) para identificar citações de temas do Superior Tribunal de Justiça (STJ) e do Supremo Tribunal Federal (STF) nas sentenças dos processos. Esses temas são então atribuídos à petição inicial dos seus respectivos processo.

Para treinar os modelos, são utilizadas amostras rotuladas e não rotuladas. A qualidade dos *clusters* gerados pelo modelo é avaliada por meio da métrica *multi adjusted rand score*, que leva em consideração os rótulos atribuídos às amostras.

No entanto, devido à natureza do RoP, que utiliza expressões regulares para identificar precedentes, em alguns processos não é possível identificar um tema específico. Isso leva à necessidade de treinar o modelo com amostras não rotuladas. Durante a inferência de uma peça inicial, a solução identifica o grupo ao qual ela pertence e busca por amostras similares para indicar os temas relevantes. Caso encontre uma amostra muito semelhante que não possua rótulo, o sistema indica que o tema é desconhecido, podendo ser um tema novo ou um tema não identificado pelo RoP.

Com o objetivo de aprimorar a qualidade das indicações fornecidas pelo sistema, é necessário aumentar a base de amostras rotuladas. Dessa forma, a ocorrência de sugestões de temas desconhecidos serão reduzidas, contribuindo para uma melhor abrangência da solução.

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo desse trabalho consiste em investigar técnicas de aprendizado semi-supervisionado visando a ampliação automática do volume de amostras rotuladas presentes na base de dados do projeto "PEDRO".

1.1.2 Objetivos específicos

Para alcançar o objetivo geral é preciso alcançar alguns objetivos específicos, sendo eles:

- Acessar e analisar os dados das sentenças utilizados no projeto "PEDRO";
- Implementar técnicas de aprendizado semi-supervisionado para expandir a base de amostras rotuladas;
- Comparar os algoritmos semi-supervisionados implementados em termos de eficácia;
- Avaliar qualitativamente o algoritmo com a melhor eficácia.

2 Referencial Teórico

2.1 Peças processuais

Peças processuais, conforme estabelecido pela Lei nº 13.105, de 16 de março de 2015, que institui o Código de Processo Civil (CPC), são instrumentos jurídicos fundamentais empregados ao longo da tramitação de um procedimento judicial. Estas peças têm a função de formalizar as pretensões, defesas e manifestações das partes, garantindo assim que o processo se desenvolva de forma organizada, previsível e justa (BRASIL, 2015).

A Petição Inicial, sob a égide deste código, representa o marco inaugural de um processo de conhecimento. É por meio dela que o autor, ou quem detém interesse jurídico, expõe ao Poder Judiciário os fatos que, segundo sua interpretação, deram origem ao seu direito. Ademais, nessa peça, o autor traz os fundamentos jurídicos que corroboram sua pretensão e, com base nisso, especifica os pedidos que deseja que sejam atendidos ao término do processo. Esta petição, crucial para a evolução do processo, precisa ser clara e objetiva, facilitando assim a defesa do réu e fornecendo ao juiz os elementos necessários para a condução da ação.

Em contrapartida, a Sentença é a decisão judicial na qual o magistrado resolve a fase de conhecimento do processo, decidindo, ou não, sobre o mérito da causa. Ao proferir a sentença, o juiz analisa as alegações de ambas as partes e as provas produzidas durante o processo. É nesse momento que, frequentemente, o magistrado recorre a precedentes. Estes, decisões anteriores de tribunais superiores ou da mesma instância sobre casos similares, são essenciais para embasar e solidificar a sentença, alinhando-a à interpretação predominante no sistema judiciário sobre determinado tema. Esse uso de precedentes proporciona decisões mais consistentes, contribuindo para a segurança jurídica e a estabilidade das relações sociais.

Portanto, tanto a Petição Inicial quanto a Sentença, bem como inúmeras outras peças processuais, têm suas características e critérios estabelecidos pelo CPC, garantindo a efetiva realização da justiça e a correta aplicação do direito no Brasil.

2.2 Precedentes Qualificados

Os precedentes qualificados emergem como um dos pilares centrais da sistemática processual trazida pelo Código de Processo Civil de 2015, instituído pela Lei nº 13.105. Tal conceito alude àquelas decisões judiciais emanadas de tribunais, principalmente de

instâncias superiores, que, por sua natureza e força vinculante, devem ser observadas e seguidas pelos juízes e tribunais em casos análogos (BRASIL, 2015).

No CPC, o artigo 927 é taxativo ao dispor sobre os precedentes que detêm essa característica vinculante. Conforme o dispositivo legal, são precedentes qualificados: os acórdãos em incidente de assunção de competência ou de julgamento de casos repetitivos, e as súmulas vinculantes. Estas decisões, pela sua magnitude e representatividade, têm a capacidade de orientar e uniformizar a jurisprudência, consolidando a compreensão dos tribunais sobre determinadas matérias (BRASIL, 2015).

A adoção dessa sistemática de precedentes qualificados visa assegurar a coerência, a isonomia e a eficiência do sistema judiciário, proporcionando maior previsibilidade às decisões e robustecendo a segurança jurídica. Este novo enfoque, trazido pelo CPC de 2015, reforça a importância do respeito aos precedentes e alinha o direito processual civil brasileiro a sistemas jurídicos mais maduros no que tange à observância de decisões passadas.

2.3 Repercussão Geral do Supremo Tribunal Federal

A repercussão geral é uma das inovações mais significativas no panorama do direito processual brasileiro contemporâneo, funcionando como um filtro de relevância para a apreciação dos recursos extraordinários pelo Supremo Tribunal Federal. Instituída pela Emenda Constitucional nº 45, de 8 de dezembro de 2004 (BRASIL, 2004), esta ferramenta representa uma reação legislativa à sobrecarga processual enfrentada pelo STF, especialmente considerando o alto volume de recursos extraordinários. A ideia central da repercussão geral é que o STF, enquanto guardião da Constituição, concentre seus esforços nas questões que possuam relevância do ponto de vista econômico, político, social ou jurídico para toda a sociedade, transcendendo, assim, os meros interesses subjacentes às partes litigantes. Desta forma, não basta a existência de uma suposta afronta à Carta Magna para que um recurso seja analisado; é imprescindível que o tema discutido tenha amplo impacto.

O CPC de 2015, instituído pela Lei nº 13.105 (BRASIL, 2015), reforça e detalha os contornos operacionais desse instituto. Além de fornecer diretrizes procedimentais, o CPC coloca a repercussão geral no contexto do sistema de precedentes brasileiro. Quando o STF reconhece a existência de repercussão geral em um tema e decide sobre ele, a decisão torna-se um precedente vinculante, obrigando os demais tribunais e instâncias inferiores a seguirem o mesmo entendimento.

Em suma, a repercussão geral surge como um mecanismo de racionalização do trabalho do STF e de uniformização jurisprudencial, servindo não apenas como critério

de admissibilidade, mas também como meio de consolidar entendimentos sobre questões constitucionais de grande relevância social.

2.4 Repetitivo do Superior Tribunal de Justiça

A figura dos precedentes repetitivos, inserida no contexto do sistema judiciário brasileiro, evidencia uma estratégia jurídica de suma importância para otimizar a prestação jurisdicional. Cabe ressaltar que a ideia central do julgamento por recursos repetitivos não emergiu com o Código de Processo Civil de 2015. De fato, o Código de Processo Civil de 1973, mesmo após suas respectivas alterações, já esboçava a estrutura desse rito. No entanto, o aprofundamento e os aprimoramentos consideráveis desse mecanismo se consolidaram com o CPC de 2015, particularmente nos artigos 1.036 a 1.041 (BRASIL, 2015), fortalecendo sua aplicação pelo Superior Tribunal de Justiça na busca de uniformidade interpretativa de questões de direito recorrentemente apresentadas.

O procedimento se baseia na seleção e julgamento de recursos que versam sobre controvérsias idênticas. A decisão tomada nesses casos torna-se um norte, uma referência para os demais processos que tratam da mesma matéria. A adoção desta metodologia visa, sobretudo, a garantia de coesão nas decisões judiciais, a prevenção de entendimentos conflitantes sobre um mesmo tema e, consequentemente, a oferta de maior segurança jurídica para as partes envolvidas.

Desse modo, os precedentes repetitivos no STJ não apenas agilizam a tramitação de processos que abordam temáticas já decididas, mas também reforçam a função uniformizadora da jurisprudência, assegurando a isonomia e a previsibilidade na aplicação do direito.

2.5 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) refere-se ao conjunto de técnicas e metodologias destinadas a permitir que máquinas compreendam, interpretem e gerem linguagem humana. Fundamentando-se na definição de língua como um composto de regras e símbolos que transmitem informações, o PLN emerge como uma interseção entre Inteligência Artificial e Linguística, com o objetivo primordial de simplificar a interação usuário-computador por meio da linguagem natural. Esta área pode ser dividida em Compreensão de Linguagem Natural, que lida com a interpretação e entendimento do texto, e Geração de Linguagem Natural, responsável por produzir textualmente informações compreensíveis. Ambas as categorias englobam várias subáreas da Linguística, desde a Fonologia, que estuda os sons, até a Pragmática, que investiga o entendimento linguístico em contextos específicos. A contribuição de renomados linguistas, como Noam Chomsky no

século XX, com ênfase em suas teorias sintáticas, tem sido crucial para o avanço do campo. Atualmente, o PLN não se limita ao estudo acadêmico, possuindo aplicações práticas notáveis, como tradução automática e reconhecimento ótico de caracteres, evidenciando sua relevância na tecnologia contemporânea. (KHURANA et al., 2022).

2.6 Pré-processamento de Dados

O pré-processamento de dados é uma etapa crucial em qualquer tarefa de análise de texto. Ele envolve uma série de técnicas que transformam os textos brutos em uma forma mais adequada para a aplicação de algoritmos de aprendizado de máquina. Algumas técnicas comumente utilizadas incluem:

- Tokenização: é o processo de dividir um texto em unidades menores chamadas de tokens, como palavras ou subpalavras. Esta técnica determina a granularidade com a qual o texto é analisado, segmentando um fluxo textual em pedaços menores. Historicamente, essas unidades eram predominantemente palavras, mas abordagens mais recentes exploram fragmentos de texto ainda menores, como n-gramas de caracteres. Isso permite que o texto seja tratado como uma sequência de elementos discretos.
- Remoção de *stop words*: as *stop words* são palavras comuns que geralmente não contribuem significativamente para a análise, como artigos, preposições e pronomes. A remoção dessas palavras pode reduzir a dimensionalidade dos dados e melhorar a eficiência do processamento (SAIF et al., 2014).
- Normalização: envolve a aplicação de transformações para padronizar o texto, como a conversão de todas as letras para minúsculas e a remoção de caracteres especiais e pontuações.
- Stemming e lematização: são técnicas que visam reduzir as palavras à sua forma base ou raiz. O stemming remove os sufixos das palavras (JIVANI, 2011), enquanto a lematização faz uso de um vocabulário e regras gramaticais para encontrar a forma base correta (PLISSON; LAVRAč; MLADENIĆ, 2004). Estes métodos têm o potencial de aprimorar o sinal de um texto como um todo, contudo, também apresentam desafios. Em particular, palavras com significados distintos podem compartilhar uma raiz ou lemma similar, tornando-as indiferenciáveis por meio destas técnicas.

O método escolhido para pré-processamento textual frequentemente depende do objetivo final da análise. Por exemplo, modelos de aprendizado profundo e *transformers* necessitam de abordagens específicas de processamento.

2.7 Projeção em Espaço Vetorial

O tratamento eficaz de dados textuais em aplicações computacionais exige uma transformação dos textos em uma forma que possa ser interpretada e processada por algoritmos. Tipicamente, após etapas iniciais de pré-processamento, onde corpos de texto são convertidos em listas de tokens separados e normalizados, surge a necessidade de uma representação mais sofisticada. Tecnicamente, as palavras do texto são mapeadas para um vocabulário estruturado com base em índices, simplificando assim a representação interna desses tokens. No entanto, para que sejam efetivamente utilizados por máquinas, os tokens ou seus índices correspondentes devem ser transpostos para um formato vetorial. Esta projeção em espaço vetorial tem sido objeto de diversas propostas e metodologias ao longo dos anos, visando otimizar a captura de informações semânticas e estruturais dos textos originais.

2.8 Modelo de Bolsa de Palavras

O modelo de Bolsa de Palavras, conhecido em inglês como *Bag-of-Words* (BoW), é uma abordagem de representação textual que simplifica o conteúdo ao considerar corpos de texto como conjuntos desordenados de palavras, desconsiderando a ordem ou a gramática. Neste modelo, cada documento é representado como um vetor em um espaço de dimensão igual ao tamanho do vocabulário, onde cada palavra única no vocabulário corresponde a uma posição no vetor. O vetor é preenchido com a contagem das ocorrências das palavras correspondentes no documento (PISTELLATO et al., 2019).

No entanto, esta representação tem suas limitações. Ao desconsiderar a ordem e a estrutura das frases, informações contextuais cruciais e relações semânticas entre os elementos da frase são perdidas. Ainda assim, a técnica demonstrou ser eficaz em várias tarefas e tem sido amplamente adotada em diversos domínios do aprendizado de máquina.

2.9 Frequência do Termo - Frequência Inversa do Documento

O TF-IDF, sigla em inglês para *"Term Frequency-Inverse Document Frequency"* e traduzido para o português como Frequência do Termo - Frequência Inversa do Documento, é uma técnica que quantifica a importância de uma palavra em um documento em relação a um corpus. A *Term Frequency* (TF) refere-se à contagem de vezes que uma palavra aparece em um documento. Esta contagem pode ser normalizada para evitar favoritismo em relação a documentos mais longos (JONES, 1988).

A formula para calcular a TF é:

$$\text{TF}(t, d) = \frac{\text{Número de vezes que o termo } t \text{ aparece no documento } d}{\text{Total de termos no documento } d} \tag{2.1}$$

Por outro lado, a *Inverse Document Frequency* (IDF) é uma medida de quão informativa uma palavra é no corpus. Calcula-se tomando o logaritmo do total de documentos dividido pelo número de documentos que contêm a palavra em questão. A fórmula para IDF é:

$$\text{IDF}(t, D) = \log\left(\frac{\text{Total de documentos}}{\text{Número de documentos contendo o termo } t}\right) \tag{2.2}$$

A fórmula do TF-IDF é expressa como:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{2.3}$$

onde t é o termo, d é o documento, e D é o conjunto de todos os documentos.

Ao multiplicar TF por IDF, obtemos o TF-IDF, que é uma ponderação que reflete a importância de um termo em um documento em relação a um corpus. Esta técnica ajuda a superar as limitações do BoW, dando mais relevância a termos que são distintivamente informativos sobre o conteúdo de um documento.

2.10 Aprendizado de Máquina

No cenário contemporâneo da tecnologia da informação, a habilidade de processar e interpretar grandes volumes de dados tornou-se primordial. Neste contexto, o aprendizado de máquina se destaca como uma ferramenta crucial. Segundo Murphy (2012), a aprendizagem automática é caracterizada por um conjunto de métodos capazes de identificar padrões nos dados de forma autônoma. Uma vez reconhecidos, esses padrões podem ser empregados para fazer previsões sobre informações futuras ou auxiliar na tomada de decisões em situações de incerteza, como determinar estratégias ótimas para a coleta de mais dados. Esta competência em discernir padrões e basear decisões neles estabelece o aprendizado de máquina como um pilar central das inovações tecnológicas atuais, alimentando progressos em múltiplos domínios da indústria e pesquisa.

Existem diferentes métodos e abordagens dentro do aprendizado de máquina, cada uma adaptada para diferentes tipos de tarefas e conjuntos de dados. Entre essas abordagens, o aprendizado não-supervisionado e o aprendizado semi-supervisionado serão discutidos em detalhes nos próximos tópicos

2.11 Aprendizado Não-Supervisionado

O aprendizado não supervisionado é uma abordagem fundamental no campo do aprendizado de máquina que se concentra na detecção de padrões intrínsecos em conjuntos de dados, sem a necessidade de rótulos prévios. Ghahramani (2004) sugere que esta forma de aprendizagem vai além da simples identificação de ruídos não estruturados, buscando encontrar estruturas subjacentes e padrões significativos que podem muitas vezes passar despercebidos em uma primeira análise. Esta capacidade de descobrir nuances nos dados torna o aprendizado não supervisionado uma ferramenta poderosa em muitas aplicações práticas, desde a análise de grandes volumes de informação até a extração de conhecimento relevante de conjuntos de dados complexos.

2.12 Aprendizado Semi-Supervisionado

A aprendizagem semi-supervisionada, em sua essência, configura-se como um método intermediário entre os paradigmas de aprendizagem supervisionada e não supervisionada no campo da aprendizagem automática. Enquanto a aprendizagem supervisionada se fundamenta essencialmente no treinamento de modelos com dados previamente etiquetados — isto é, dados que já contam com uma saída ou resposta associada — a aprendizagem não supervisionada, por sua vez, não se vale de saídas predefinidas, focando-se em identificar padrões ou estruturas intrínsecas nos dados por meio de algoritmos (DELALLEAU; BENGIO; ROUX, 2005).

Neste interstício, encontra-se a aprendizagem semi-supervisionada, cuja metodologia combina tanto dados rotulados quanto não rotulados para treinamento. Esta abordagem reconhece a valiosa contribuição que os dados não rotulados podem trazer ao processo, especialmente quando consideramos que, em muitos cenários reais, a obtenção de dados rotulados pode ser onerosa ou limitada em quantidade.

No âmbito da aprendizagem semi-supervisionada, emergem duas abordagens distintas: o aprendizado indutivo e o transdutivo. No aprendizado indutivo, o foco é na construção de um modelo generalizado capaz de fazer previsões para amostras novas que não estavam presentes no conjunto de treinamento original. Por outro lado, o aprendizado transdutivo visa especificamente rotular os dados não rotulados fornecidos durante o treinamento, não buscando uma generalização para pontos fora desse conjunto. Estas duas abordagens oferecem perspectivas complementares, adequando-se a diferentes tipos de tarefas e objetivos dentro do campo da aprendizagem semi-supervisionada.

Assim, ao incorporar também os dados não rotulados, busca-se uma melhoria significativa na precisão e robustez dos modelos gerados. Para fundamentar esse paradigma, diversos pressupostos são considerados. Primeiramente, há o pressuposto da continuidade,

que sugere que pontos próximos no espaço de entrada tendem a pertencer à mesma classe. Em adição, existe o pressuposto do agrupamento, que propõe que dados agrupados em uma região específica do espaço de entrada são prováveis de pertencer à mesma classe. Também é levado em conta o pressuposto da estrutura de manifold, indicando que os dados em alta dimensão podem estar contidos em uma variedade de dimensão mais baixa. Por fim, tem-se o pressuposto da baixa densidade, postulando que a fronteira de decisão entre classes deveria idealmente passar por regiões de baixa densidade de dados. Estes pressupostos coletivamente fornecem uma base sólida para a eficácia e aplicabilidade da aprendizagem semi-supervisionada em uma variedade de contextos. (CHAPELLE; SCHOLKOPF; ZIEN, 2010)

2.12.1 Classificadores Transdutivos Semi-Supervisionados Baseados em Grafos

Dentro do vasto território do aprendizado semi-supervisionado, destacam-se os classificadores transdutivos semi-supervisionados baseados em grafos, uma metodologia que incorpora princípios da técnica de *label propagation*. Esta técnica, inicialmente proposta por Xiaojin Zhu e Zoubin Ghahramani em *Learning From Labeled And Unlabeled Data With Label Propagation* (ZHU; GHAHRAMANI, 2003), fundamenta-se em algoritmos baseados em grafos. Nestes grafos, os nós representam dados, tanto rotulados quanto não rotulados, enquanto as arestas refletem a semelhança entre eles. A propagação de rótulos ocorre a partir dos dados já classificados, permitindo que a informação se dissemine pelo grafo e rotule progressivamente todos os nós (CHAPELLE; SCHOLKOPF; ZIEN, 2010). Esses classificadores são particularmente valiosos em cenários onde os dados estão parcialmente rotulados, capitalizando na estrutura intrínseca dos dados para otimizar a aprendizagem. Eles se diferenciam de abordagens puramente indutivas ou supervisionadas por focarem na inferência direta para amostras específicas, beneficiando-se da combinação de dados rotulados e não rotulados para melhorar a precisão da classificação. Em termos gerais, conforme destacado nas obras de Jebara, Wang e Chang (2009) e de Liu, Wang e Chang (2012), os métodos de aprendizado semi-supervisionado baseados em grafos são estruturados em três etapas fundamentais: a construção do grafo, a atribuição de pesos ao mesmo e o processo de inferência.

A construção de grafos é central para esses classificadores. Os grafos não são apenas uma representação dos dados, mas um meio fundamental para entender as relações entre diferentes instâncias. A escolha de como construir e ponderar estes grafos tem um impacto significativo no desempenho do classificador. As técnicas para construção de grafos incluem a definição de uma matriz de adjacência e a determinação cuidadosa de vértices e arestas, o que requer uma consideração minuciosa da natureza dos dados e do problema em questão.

A ponderação dos grafos é igualmente crucial. Ela determina como a informação é compartilhada entre os nós do grafo, influenciando diretamente a eficácia do classificador.

Os métodos de ponderação variam, mas frequentemente se baseiam em medidas de similaridade ou distância entre as instâncias de dados. Essas ponderações ajudam a definir a estrutura do grafo de uma forma que reflete as relações subjacentes nos dados, facilitando uma classificação mais precisa.

A inferência em grafos é o coração do processo de classificação. Os classificadores utilizam diferentes técnicas de inferência, como cortes mínimos de grafos, campos aleatórios de Markov e campos aleatórios gaussianos (ZHU; GHAHRAMANI; LAFFERTY, 2003), para determinar as etiquetas das instâncias não rotuladas. Esses métodos de inferência contribuem significativamente para a precisão do classificador, abordando desafios como ruído nos dados e irregularidades nos grafos. A inferência em grafos não é apenas um meio para classificar dados; ela reflete uma compreensão profunda das complexas inter-relações presentes no conjunto de dados (SUBRAMANYA; TALUKDAR, 2014).

2.12.1.1 Gráficos mútuos de k-vizinhos mais próximos

O k-vizinhos mútuos (*mutual k-nearest neighbors, mutual k-NN*) é uma variante do tradicional algoritmo *k-nearest neighbors* (k-NN), aplicada principalmente no contexto do aprendizado semi-supervisionado. Enquanto o k-NN padrão baseia-se na ideia de conectar um ponto de dados aos seus k vizinhos mais próximos, independentemente da reciprocidade dessa relação, o k-vizinhos mútuos introduz um critério adicional de mutualidade (SOUSA; REZENDE; BATISTA, 2013). Neste método, uma conexão entre dois pontos é estabelecida somente se ambos os pontos se encontrarem reciprocamente entre seus k vizinhos mais próximos. Esse requisito de reciprocidade implica uma seleção mais criteriosa de vizinhos, levando à formação de conexões que refletem uma relação mais significativa e equilibrada entre os pontos no espaço de dados.

Ozaki et al. (2011) argumentam que os gráficos de k-NN mútuos são particularmente adequados para dados de alta dimensão, como os encontrados no PLN. Eles observam que os gráficos k-NN tradicionais frequentemente produzem “hubs”, ou vértices com um grau extremamente alto, o que é especialmente prevalente em dados de alta dimensão. Esses hubs podem deteriorar a precisão da classificação semi-supervisionada, pois a estrutura de vizinhança no gráfico pode não representar com precisão as relações semânticas ou contextuais no espaço de dados original.

Os gráficos de k-NN mútuos, por outro lado, tendem a ser mais resistentes à formação de hubs devido à sua natureza recíproca. Além disso, todos os vértices em um gráfico de k-NN mútuo têm um grau limitado por k, o que ajuda a prevenir a ocorrência de vértices com grau extremamente alto. Essa característica torna os gráficos de k-NN mútuos mais consistentes com a suposição de cluster subjacente a muitos métodos de classificação semi-supervisionada. Em outras palavras, espera-se que pontos que são semelhantes (ou próximos no espaço de dados) compartilhem a mesma etiqueta ou classificação.

Ozaki et al. (2011) também observam que os gráficos de k-NN mútuos, quando combinados com árvores de abrangência máxima, apresentam desempenho consistentemente superior aos gráficos k-NN em tarefas de classificação de desambiguação de sentido de palavras e classificação de documentos. Além disso, eles mostram que os gráficos de k-NN mútuos podem alcançar uma precisão de classificação igual ou até superior aos gráficos de b-matching, apesar de sua menor complexidade computacional.

Portanto, no contexto do aprendizado semi-supervisionado, os gráficos de k-NN mútuos oferecem uma abordagem mais robusta e eficiente, especialmente para dados de alta dimensão, superando algumas das limitações dos gráficos k-NN tradicionais e proporcionando uma precisão de classificação aprimorada.

2.13 Métricas

O campo do aprendizado de máquina, em sua essência, busca desenvolver modelos que possam aprender padrões a partir de dados. Para entender a eficácia desses modelos, é crucial dispor de ferramentas e métricas de avaliação robustas. No contexto do aprendizado supervisionado tradicional, uma variedade de métricas, incluindo precisão, *F-measure* a área sob a curva ROC (do inglês *Receiver Operating Characteristic*) — frequentemente representada pela sigla AUC, que vem do inglês *Area Under the Curve*, têm sido amplamente adotadas para avaliar o desempenho de generalização dos sistemas de aprendizado. Estas métricas oferecem uma visão clara da capacidade do modelo em prever resultados em conjuntos de dados não vistos anteriormente.

No entanto, à medida que o campo do aprendizado de máquina evolui e abraça problemas mais complexos, como o aprendizado multirrótulo, a necessidade de métricas de avaliação mais sofisticadas torna-se evidente. Diferentemente do ambiente de rótulo único tradicional, no aprendizado multirrótulo, uma única instância pode estar associada a múltiplos rótulos, o que introduz nuances adicionais na avaliação de desempenho. Consequentemente, surgiram várias métricas específicas para abordar as particularidades do aprendizado multirrótulo. Estas métricas podem ser, em geral, categorizadas em dois grupos distintos: métricas baseadas em exemplos (SCHAPIRE; SINGER, 2000) e métricas baseadas em rótulos (TSOUMAKAS; VLAHAVAS, 2007). Enquanto as métricas baseadas em exemplos enfocam a avaliação do desempenho em cada instância individualmente, as métricas baseadas em rótulos voltam-se para a avaliação distinta de cada rótulo de classe. No entanto, é vital reconhecer que além destas, existem outras métricas relevantes que se estendem além do escopo multirrótulo, refletindo a vastidão e diversidade do campo.

2.13.1 Métricas de Avaliação Baseadas em Exemplo

A eficácia de um algoritmo de aprendizado é quantificada pela sua habilidade de fazer previsões próximas dos rótulos reais em conjuntos de dados não vistos anteriormente. Uma forma de abordar acertos parciais é avaliar a diferença média entre as previsões e os rótulos verdadeiros de cada instância e, em seguida, calcular a média para o conjunto de teste completo. Esta técnica é referida como avaliação baseada em exemplos. Enquanto isso, existe uma abordagem alternativa centrada na avaliação por rótulos. No entanto, é essencial entender que métodos focados nos rótulos não capturam as inter-relações entre diferentes classes (ZHANG; ZHANG, 2010).

Neste cenário, considerando o exemplo i:

- Y_i: Representa o conjunto de rótulos verdadeiros para o exemplo i.
- Z_i: Indica o conjunto de rótulos previstos pelo modelo para o exemplo i.

Dentro deste contexto, Godbole e Sarawagi (2004) introduziram as métricas de avaliação em dados parciais: acurácia, precisão, recall e F1-Measure, as quais são detalhadas a seguir.

Acurácia

A acurácia, em termos de exemplos, é a comparação dos rótulos previstos com os rótulos verdadeiros de cada instância. Formalmente, para n instâncias, ela é dada por:

$$\text{Acurácia} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{2.4}$$

Precisão

A precisão avalia a proporção de rótulos previstos que são corretos. Sua formulação é:

$$\text{Precisão} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{2.5}$$

Recall

O *recall* mede a proporção de rótulos verdadeiros que foram corretamente previstos. Sua definição matemática é:

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{2.6}$$

F1-Measure

A *F1-Measure* combina precisão e recall através da média harmônica. Para um conjunto de n instâncias, ela é definida como:

$$F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \tag{2.7}$$

Hamming Loss

A *Hamming Loss* (HL) é uma métrica de desempenho para problemas de classificação multilabel, derivada da distância de Hamming (HAMMING, 1950). Esta métrica quantifica a fração de rótulos previstos incorretamente, refletindo o desalinhamento entre os rótulos reais e previstos.

A *Hamming Loss* considera erros de previsão, quando um rótulo errado é identificado, e erros de omissão, quando um rótulo relevante não é reconhecido. A métrica é normalizada pelo total de classes e pelo número de exemplos. Matematicamente, a *Hamming Loss*, HL, é expressa como:

$$HL = \frac{1}{kn}\sum_{i=1}^{n}\sum_{l=1}^{k}[I(l \in Z_i \wedge l \notin Y_i) + I(l \notin Z_i \wedge l \in Y_i)] \tag{2.8}$$

Onde I representa a função indicadora. Um valor ideal para a *Hamming Loss* é $HL = 0$, indicando ausência de erros. Contudo, valores menores de HL indicam melhor desempenho do algoritmo.

2.13.2 Métricas de Avaliação Baseadas em Rótulos

As métricas que se fundamentam em rótulos realizam avaliações de cada rótulo de forma independente, culminando na média resultante de todas essas avaliações. Tais métricas permitem a integração de critérios típicos de classificadores binários — tais como acurácia, precisão, *recall*, F1 e ROC. Para cada critério adotado, há duas abordagens predominantes: uma que efetua cálculos para cada rótulo em separado, seguido pela obtenção da média geral (denominada média macro), e outra que computa os valores de forma consolidada, levando em consideração todas as instâncias e rótulos simultaneamente (referida como média micro) (YANG, 1999). As seções subsequentes detalharão as metodologias de média macro e micro associadas à precisão, recall e F1.

Média Macro

Para a média macro, calcula-se a métrica para cada rótulo individualmente e, em seguida, tira-se a média desses valores.

$$\lambda\text{-Precisão, } P^{\lambda}_{\text{macro}} = \frac{\sum_{i=1}^{n} Y_i^{\lambda} Z_i^{\lambda}}{\sum_{i=1}^{n} Z_i^{\lambda}}, \qquad \text{Precisão, } P_{\text{macro}} = \frac{1}{k}\sum_{i=1}^{k} P^{\lambda}_{\text{macro}} \tag{2.9}$$

$$\lambda\text{-}\textit{Recall}\text{, } R^{\lambda}_{\text{macro}} = \frac{\sum_{i=1}^{n} Y_i^{\lambda} Z_i^{\lambda}}{\sum_{i=1}^{n} Y_i^{\lambda}}, \qquad \textit{Recall}\text{, } R_{\text{macro}} = \frac{1}{k}\sum_{i=1}^{k} R^{\lambda}_{\text{macro}} \tag{2.10}$$

$$\lambda\text{-}F1^{\lambda}_{\text{macro}} = \frac{2\sum_{i=1}^{n} Y_i^{\lambda} Z_i^{\lambda}}{\sum_{i=1}^{n} Y_i^{\lambda} + \sum_{i=1}^{n} Z_i^{\lambda}}, \qquad F1_{\text{macro}} = \frac{1}{k}\sum_{i=1}^{k} F1^{\lambda}_{\text{macro}} \tag{2.11}$$

Média Micro

Para a média micro, acumula-se a contagem total de verdadeiros positivos, falsos positivos e falsos negativos, e depois calcula-se a métrica.

$$\text{Precisão, } P_{\text{micro}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^{j} Z_i^{j}}{\sum_{j=1}^{k}\sum_{i=1}^{n} Z_i^{j}} \tag{2.12}$$

$$\textit{Recall}\text{, } R_{\text{micro}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^{j} Z_i^{j}}{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^{j}} F1_{\text{micro}} \quad = \frac{2\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^{j} Z_i^{j}}{\sum_{j=1}^{k}\sum_{i=1}^{n} Y_i^{j} + \sum_{j=1}^{k}\sum_{i=1}^{n} Z_i^{j}} \tag{2.13}$$

Onde,

$$Y_i^{\lambda} = \begin{cases} 1 & \text{se } x_i \text{ pertence de fato à classe } \lambda \\ 0 & \text{caso contrário} \end{cases}$$

e,

$$Z_i^{\lambda} = \begin{cases} 1 & \text{se } x_i \text{ é previsto como pertencente à classe } \lambda \\ 0 & \text{caso contrário} \end{cases}$$

Conforme definido, $F1_{\text{macro}}$ é mais sensível ao desempenho das classes com menor número de exemplos, enquanto $F1_{\text{micro}}$ reflete predominantemente o desempenho das classes com uma quantidade maior de exemplos (TANG; RAJAN; NARAYANAN, 2009).

Média Ponderada

A média ponderada, diferentemente da média macro e micro, leva em consideração o desequilíbrio de classes ao ponderar a contribuição de cada classe para a média final. Isso é feito atribuindo pesos proporcionais ao número de instâncias de cada classe no conjunto de dados. As métricas de precisão, recall e F1 são calculadas para cada classe e, em seguida, ponderadas pela frequência da classe correspondente.

$$\text{Precisão, } P_{\text{ponderado}} = \frac{\sum_{i=1}^{k} w_i P_{\text{macro}}^{\lambda}}{\sum_{i=1}^{k} w_i} \tag{2.14}$$

$$\textit{Recall, } R_{\text{ponderado}} = \frac{\sum_{i=1}^{k} w_i R_{\text{macro}}^{\lambda}}{\sum_{i=1}^{k} w_i} \tag{2.15}$$

$$\text{F1, } F1_{\text{ponderado}} = \frac{\sum_{i=1}^{k} w_i F1_{\text{macro}}^{\lambda}}{\sum_{i=1}^{k} w_i} \tag{2.16}$$

Onde, w_i é o peso da classe λ_i, geralmente definido como a proporção de instâncias da classe λ_i no conjunto de dados.

$$w_i = \frac{n_i}{N}$$

Aqui, n_i representa o número de instâncias da classe λ_i, e N é o número total de instâncias em todo o conjunto de dados. Este método assegura que as classes com menos exemplos tenham um impacto proporcional na avaliação do modelo, mitigando o efeito do desequilíbrio de classes.

2.13.3 Métricas de Similaridade

Para avaliar a similaridade entre vetores espaciais, diversas métricas têm sido propostas. Essas métricas podem ser categorizadas em métricas de similaridade e métricas de dissimilaridade. As métricas de similaridade, como a similaridade de cosseno, fornecem uma medida onde valores mais altos indicam maior similaridade entre os elementos comparados. Por outro lado, métricas de dissimilaridade, como a distância euclidiana, operam de maneira inversa: valores mais elevados sinalizam maior distinção ou dissimilaridade entre os elementos analisados.

2.13.4 Similaridade de Cossenos

A similaridade de cosseno, frequentemente empregada em recuperação de informação e áreas correlatas, conceptualiza um texto como um vetor de termos. A semelhança entre dois textos é calculada com base no cosseno do ângulo entre esses vetores em um espaço multidimensional. (RAHUTOMO; KITASUKA; ARITSUGI, 2012) A fórmula matemática para calcular a similaridade de cosseno entre dois vetores A e B é dada por:

$$\text{Similaridade de Cosseno} = \frac{A \cdot B}{\|A\| \times \|B\|} \tag{2.17}$$

Onde:

- $A \cdot B$ é o produto interno dos vetores A e B.

- $\|A\|$ e $\|B\|$ são as magnitudes (ou normas) dos vetores A e B, respectivamente.

Nessa formulação, um valor resultante de 1 indica que os vetores (ou documentos) são perfeitamente alinhados e, portanto, inteiramente similares; já um valor de 0 denota que são ortogonais ou completamente distintos. Esta técnica quantifica a semelhança entre dois documentos ao considerar o ângulo entre seus vetores de termos no espaço multidimensional.

A similaridade de cossenos é uma medida eficaz para comparar a similaridade semântica entre textos e documentos, independentemente do seu tamanho ou dimensionalidade. É amplamente aplicada em algoritmos de recomendação, sistemas de busca por similaridade e detecção de plágio, entre outras aplicações.

2.13.5 Distância Euclidiana

A distância euclidiana é uma das métricas mais amplamente utilizadas para medir a "distância"entre dois pontos em um espaço euclidiano (ou geométrico). Ela é derivada do Teorema de Pitágoras e é útil para identificar a "linha reta"ou a menor distância entre dois pontos em um espaço multidimensional.

A formulação matemática para a distância euclidiana entre dois pontos P e Q com coordenadas (x_1, y_1) e (x_2, y_2) em um espaço bidimensional é:

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.18}$$

Em um espaço multidimensional, onde os pontos P e Q possuem coordenadas $(x_1, x_2, ..., x_n)$ e $(y_1, y_2, ..., y_n)$, respectivamente, a fórmula se expande para:

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.19}$$

Nesta formulação, n representa o número de dimensões ou características dos pontos.

É importante destacar que a distância euclidiana pode não performar adequadamente para vetores que não estão normalizados ou quando trata-se de dados em espaços de alta dimensão. Em espaços de alta dimensão, a distância entre a maioria dos pares de pontos tende a convergir para uma constante, tornando difícil discernir proximidades relativas. Além disso, a falta de normalização pode levar a resultados distorcidos devido à variabilidade na escala dos dados (AGGARWAL; HINNEBURG; KEIM, 2001).

A distância euclidiana é frequentemente utilizada em diversos domínios, como análise de clusters, reconhecimento de padrões e sistemas de recomendação, para mencionar

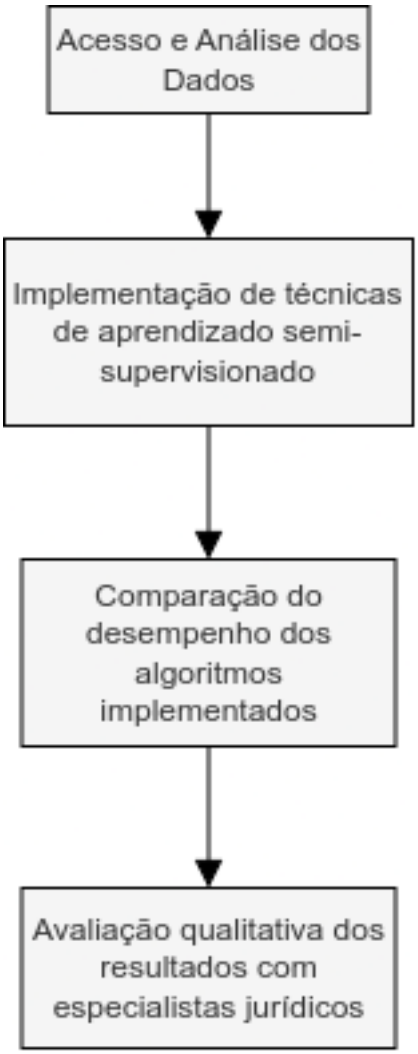
alguns. Ela oferece uma medida intuitiva da proximidade entre dois pontos, permitindo uma interpretação geométrica da similaridade ou dissimilaridade entre eles.

3 Materiais e Métodos

3.1 Plano Metodológico

Neste plano metodológico, detalha-se a abordagem adotada para a realização de uma pesquisa aplicada com design experimental. O objetivo é explorar de maneira prática as soluções para desafios específicos de rotulagem no âmbito do projeto 'PEDRO'. A metodologia empregada é representada de forma sequencial no fluxograma da Figura 1, abrangendo desde o acesso e análise dos dados até a avaliação qualitativa dos resultados. Cada uma das etapas envolvidas será descrita nas seções subsequentes.

Figura 1 – Fluxograma das etapas da pesquisa



3.1.1 Acesso e análise do dados

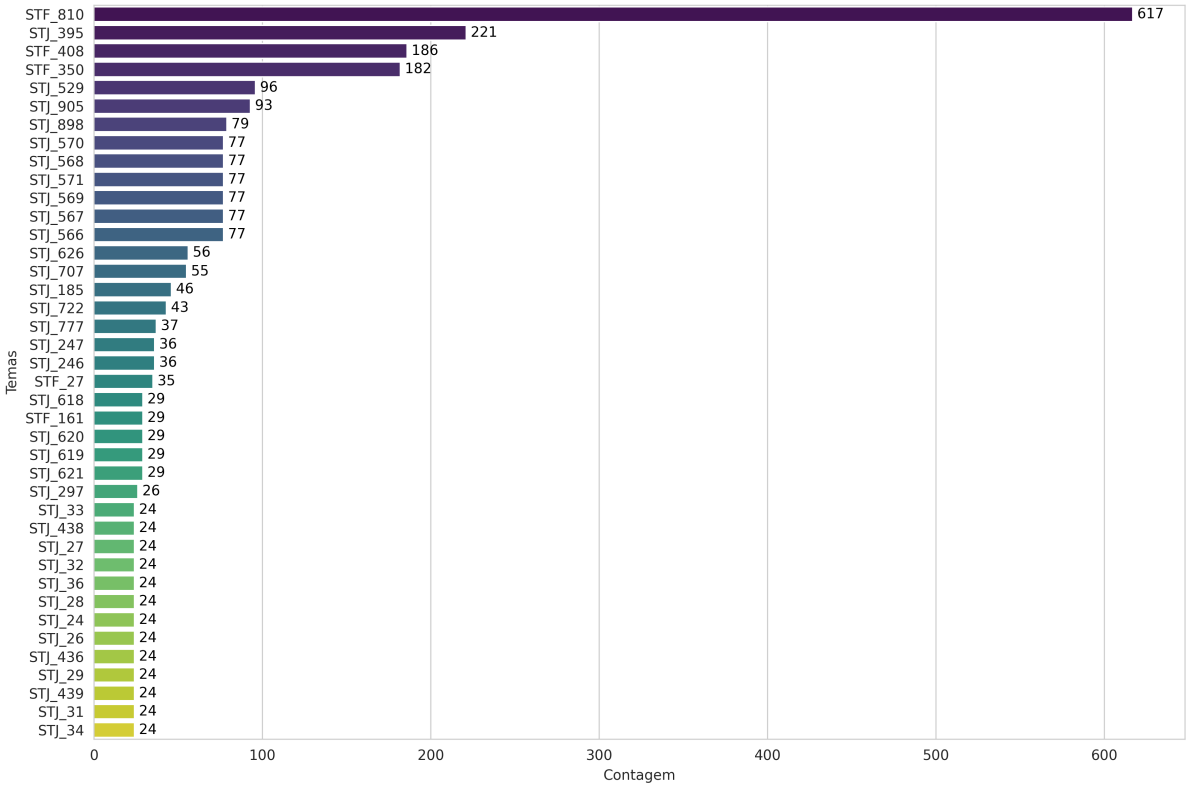
Para entender os dados que será feito o estudo é preciso realizar uma análise dos dados para entender sua estrutura, características e possíveis padrões ou tendências.

3.1.1.1 Análise Exploratória dos Dados

O conjunto de dados analisado compreende 22.503 amostras. Desse total, 2.048 estão associadas a pelo menos um tema identificado, ao passo que 20.455 são categorizadas sob a designação "tema desconhecido", indicando a ausência de rotulação. A Figura 2 representa a distribuição dos 40 temas mais recorrentes no mencionado conjunto. Para uma visualização completa da distribuição dos temas, consulte o Anexo A.

Relativamente à distribuição total sob os temas do STF e STJ, constata-se uma variação considerável. O código 'STF_810' evidencia-se com 617 entradas. Em contrapartida, diversos temas, especificamente os que se estendem de 'STJ_567' a 'STJ_571', registram 77 ocorrências cada.

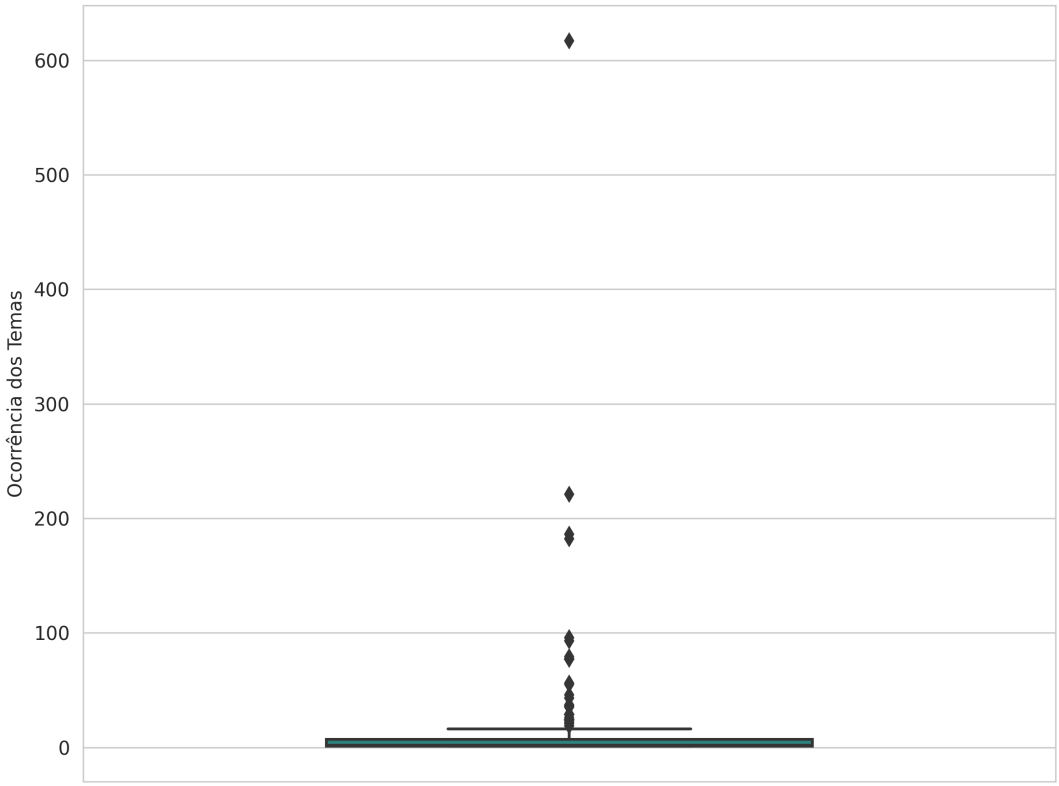
Figura 2 – Distribuição dos 40 temas mais recorrentes



A análise dos quartis, conforme ilustrado na Figura 3, proporciona uma compreensão desta distribuição. Observa-se que 25% dos temas têm uma ocorrência, indicando que um quarto destes é pouco frequente. A mediana, que se situa em 2 ocorrências, revela que metade dos temas é mencionada duas vezes ou menos. Já o 3º quartil, com 7.25 ocorrências,

mostra que três quartos dos temas são citados até 7 vezes. Nota-se também a presença de *outliers*, como o tema mais recorrente com 617 menções, uma quantidade substancialmente distante dos demais valores.

Figura 3 – Box plot da Ocorrência dos temas



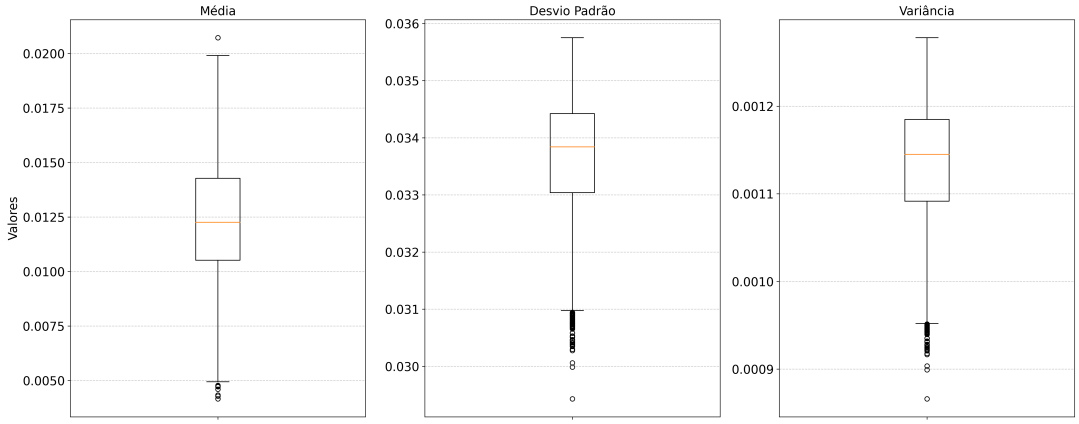
Nota: Q1 (1º Quartil): 1.0; Q2 (Mediana): 2.0; Q3 (3º Quartil): 7.25; Max (Valor Máximo): 617.

Tal constatação sugere que, embora determinados códigos sejam frequentemente adotados ou reportados, vários são aplicados em contextos menos comuns ou específicos. Este padrão de distribuição assimétrica sugere uma variedade nos temas, com uma maioria sendo rara, porém alguns poucos se destacando em frequência. Interessantemente, a baixa recorrência de muitos rótulos pode ser atribuída à não eficiência do *RoP* em capturar certos temas que podem estar sendo mencionados de maneira diversa nas sentenças processadas. Além disso, a baixa recorrência de muitos rótulos pode servir como uma semente eficaz para algoritmos semi-supervisionados, como o de propagação de rótulos, possibilitando a propagação dos rótulos para as amostras que atualmente não os possuem.

Na esfera da análise estatística, a elucidação das propriedades de conjuntos de dados é frequentemente ancorada nas medidas de tendência central e dispersão. Neste contexto, a média emerge como a medida de tendência central mais prevalente, oferecendo um ponto de referência em torno do qual os valores se distribuem. (BLAND, 2000). A

média, contudo, está sujeita à influência de valores extremos, podendo ser deslocada de maneira que não reflete adequadamente a centralidade dos dados.

Figura 4 – Box Plots das Estatísticas do Conjunto de Dados Completo

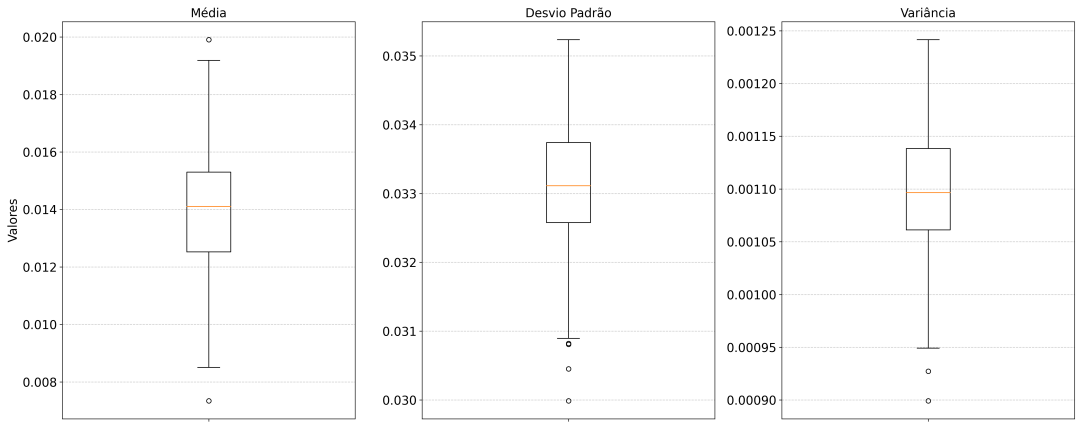


A figura ilustra a média, o desvio padrão e a variância dos pesos TF-IDF para o conjunto de dados completo.

Complementarmente, o desvio padrão emerge como uma medida de dispersão essencial, evidenciando o grau de variação em relação à média (HEALEY, 2014). Valores diminutos de desvio padrão sugerem uma aglomeração de dados próxima à média, enquanto valores mais altos indicam uma dispersão mais abrangente. Esta medida é vital para discernir a variabilidade inerente aos dados e fundamenta a avaliação da robustez das estimativas médias.

A variância, que é o quadrado do desvio padrão, fornece uma indicação da dispersão dos dados, sendo particularmente útil em comparações de variabilidade entre conjuntos de dados (MONTGOMERY; RUNGER, 2010).

Figura 5 – Box Plots das Estatísticas do Conjunto de Dados Rotulados



A figura ilustra a média, o desvio padrão e a variância dos pesos TF-IDF para o conjunto de dados rotulados.

A comparação entre box plots do conjunto total de dados presente na figura 4 e dos dados rotulados presente na figura 5 desvenda diferenças notáveis. Para a média, o conjunto total ostenta uma distribuição ligeiramente mais elevada e a presença de *outliers* superiores, sugerindo que os valores extremos são mais salientes quando considera-se o conjunto em sua totalidade. Em contraste, os dados rotulados apresentam uma distribuição mais compacta da média, indicativa de maior homogeneidade.

Analogamente, o desvio padrão revela uma variabilidade mais pronunciada no conjunto total, evidenciado pela presença de *outliers*. Esta observação contrasta com a distribuição mais restrita do desvio padrão nos dados rotulados, que denota uma maior consistência na dispersão dos dados em torno da média.

Por fim, a variância do conjunto total de dados exibe uma dispersão mais extensa e uma amplitude maior de valores. Isto está em oposição à concentração mais estreita da variância nos dados rotulados, onde os *outliers* são menos frequentes, apontando para uma variação reduzida.

A inferência dessas comparações sugere que os dados rotulados possuem uma uniformidade acentuada e menos variabilidade nas métricas consideradas. Esta constatação pode indicar que o processo de rotulação seleciona dados mais homogêneos. Interessante notar que, do ponto de vista negocial, os *outliers* presentes na base de dados completa podem ser indicativos de processos atípicos, os quais, por sua natureza única, podem ainda não ter sido suficientemente examinados ou classificados por precedentes estabelecidos.

As implicações dessas discrepâncias são profundas para análises subsequentes, particularmente na modelagem preditiva. Se modelos de aprendizado de máquina são treinados com dados rotulados, a sua capacidade de generalizar pode ser comprometida pela não exposição à variabilidade mais vasta presente no conjunto de dados completo.

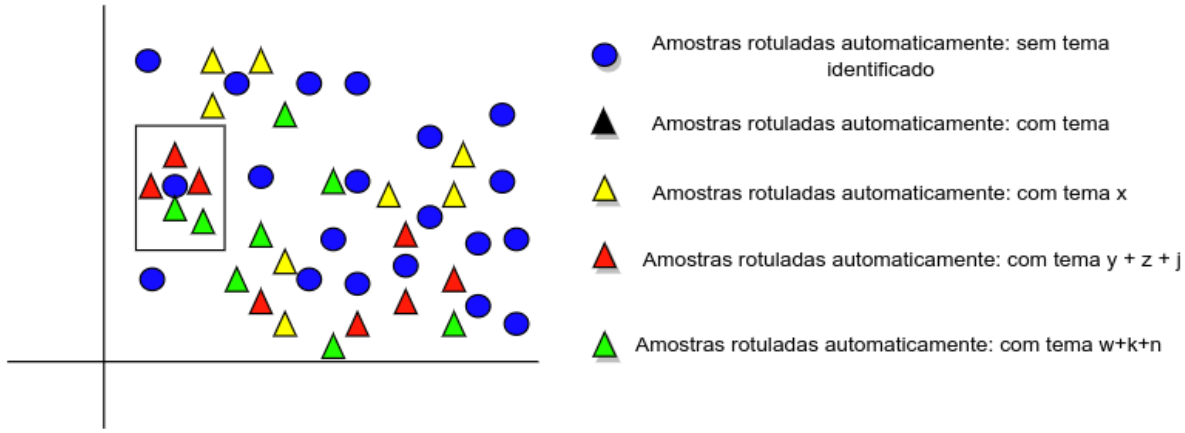
3.1.2 Implementação de técnicas de aprendizado semi-supervisionado

Para garantir a seleção de uma técnica semi-supervisionada que se alinhe às características específicas dos dados multiclasse e multirrótulo do projeto “PEDRO”, proceder-se-á com a avaliação de uma variedade de abordagens. Nos tópicos subsequentes, cada abordagem proposta será detalhada. O espaço bidimensional dos dados, visualizado na Figura 6, será empregado didaticamente para auxiliar na elucidação das abordagens, permitindo uma compreensão mais intuitiva das técnicas e sua implementação.

3.1.2.1 Algoritmo do Mais Próximo

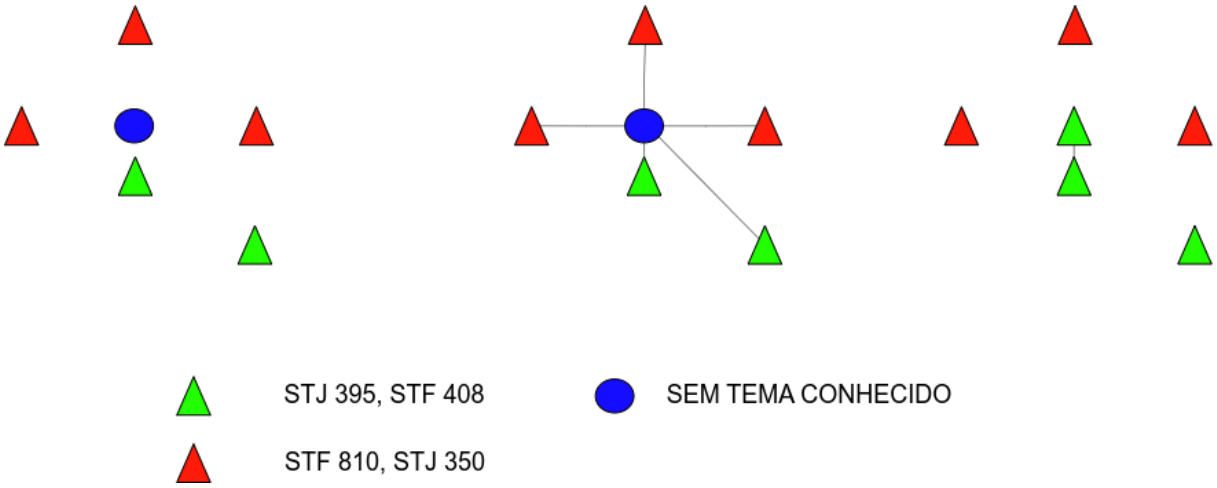
O algoritmo do mais próximo, baseia-se no princípio da continuidade, assume que amostras próximas tendem a compartilhar o mesmo rótulo. Este método envolve a construção de um grafo, onde as arestas representam a similaridade de cosseno entre as

Figura 6 – Espaço Bi-dimensional com amostra rotuladas e não-rotuladas



amostras, e os nós refletem os pesos TF-IDF das amostras. Um aspecto fundamental deste algoritmo é a inclusão de um parâmetro que estabelece um limiar de similaridade, determinando se duas amostras são consideradas vizinhas. Apenas amostras que atingem ou ultrapassam este limiar são levadas em conta na propagação dos rótulos.

Figura 7 – Abordagem 1: Algoritmo do Mais Próximo



A propagação de rótulos ocorre por meio de um processo iterativo, limitado a 10 iterações para evitar ciclos infinitos. Durante a primeira iteração, as instâncias não rotuladas são avaliadas e os rótulos das instâncias vizinhas mais similares, conforme definido pelo limiar de similaridade, são atribuídos a elas. É importante notar que, nesta fase inicial, pode ocorrer uma situação em que uma instância não rotulada esteja próxima de outra também não rotulada. Neste caso, uma delas pode receber um rótulo "virtual". Este rótulo virtual serve como uma espécie de marcador temporário que, nas iterações subsequentes, pode ser transferido para outras instâncias não rotuladas, facilitando assim a propagação progressiva e a estabilização dos rótulos ao longo do processo. Este método assegura que apenas as relações de vizinhança mais pertinentes e confiáveis sejam empregadas

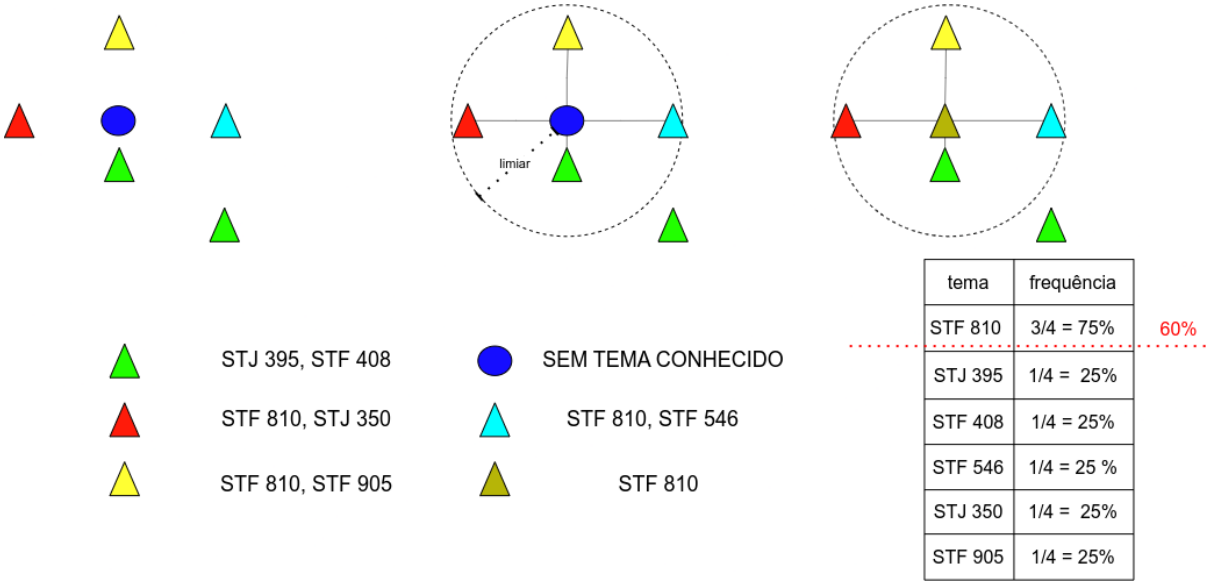
na determinação dos rótulos. Com esse método, pretende-se assegurar que apenas as relações de vizinhança mais pertinentes e confiáveis sejam empregadas na determinação dos rótulos, visando aumentar a precisão e a eficácia da rotulação. A Figura 7 exemplifica essa abordagem: nota-se que o círculo azul, representando uma amostra sem rótulo, está conectado por arestas às amostras mais próximas. Após a identificação das amostras vizinhas com maior similaridade, foi possível atribuir à amostra representada pelo círculo azul o rótulo da amostra verde, que possui os rótulos 'STJ 395' e 'STF 408', demonstrando o processo de inferência de rótulos baseado na similaridade entre as amostras.

3.1.2.2 Algoritmo de Maioria

Este método difere do Algoritmo do Mais Próximo ao considerar para a propagação o rótulo mais frequente entre os vizinhos de uma amostra não rotulada. O fluxo é regido por dois parâmetros principais: um limiar de similaridade que define quais amostras são consideradas próximas e um critério de maioria que determina a frequência necessária para a propagação de um rótulo.

Esse método também segue um protocolo iterativo com um limite de 10 iterações, utilizando um controle que encerra o algoritmo se nenhuma mudança de rótulo ocorrer entre as iterações. Isso garante a convergência da propagação de rótulos e a estabilidade do conjunto de dados.

Figura 8 – Abordagem 2: Algoritmo de Maioria



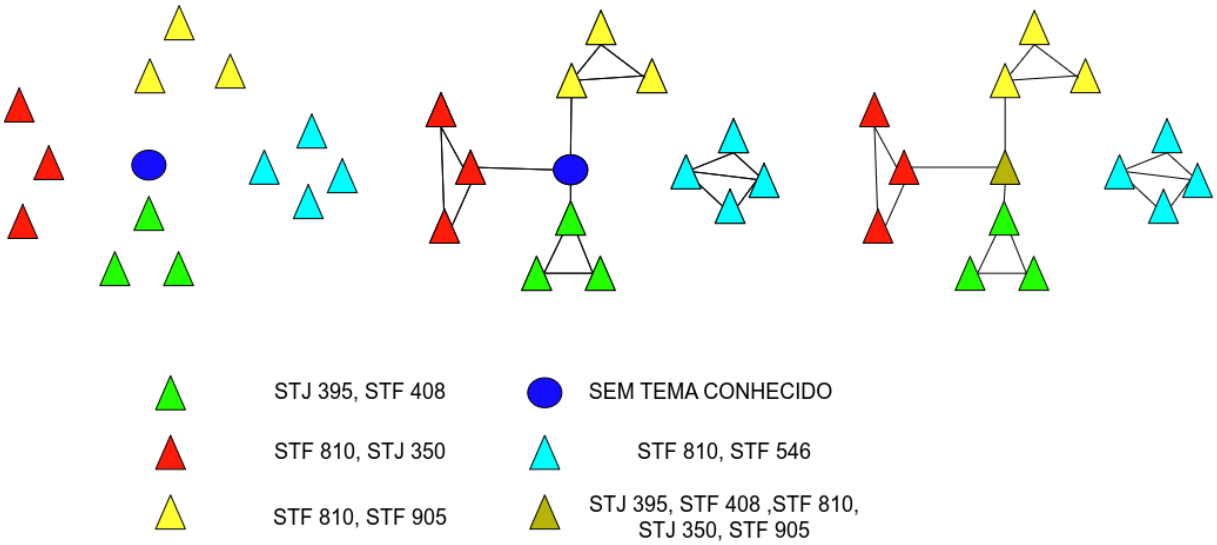
Conforme ilustrado na Figura 8, o algoritmo é visualizado por meio de um grafo onde uma amostra não rotulada, representada pelo círculo azul, é cercada por outras amostras dentro de um raio determinado pelo limiar de similaridade, demarcado pela linha pontilhada. A tabela ao lado direito fornece uma visão geral dos rótulos existentes nas

amostras vizinhas e suas respectivas frequências. Nesta imagem específica, o rótulo 'STF 810' aparece como o mais frequente, estando presente em 75% das amostras vizinhas. No entanto, um critério adicional é aplicado, indicado pela porcentagem de 60% na legenda, sugerindo que um rótulo deve estar presente em pelo menos 60% das amostras para ser atribuído à amostra não rotulada. Portanto, seguindo esta regra, o rótulo 'STF 810', que excede o critério de 60%, é o rótulo escolhido para ser atribuído à amostra central não rotulada. Esta figura destaca a mecânica de decisão do algoritmo da maioria, que se baseia na representatividade dos rótulos na área de influência determinada pelo limiar de similaridade.

3.1.2.3 KNN-Mútuo

Na terceira abordagem, o foco recai sobre a aplicação do algoritmo KNN-Mútuo para a estruturação dos grafos. Através do cálculo de similaridade de cosseno, estabelecem-se as arestas que interligam os nós. O parâmetro fundamental nesta configuração é o K, que define a quantidade de vizinhos mais próximos a serem considerados na construção do grafo. A seleção adequada de K é essencial, visto que afeta a dinâmica de propagação dos rótulos entre as amostras.

Figura 9 – Abordagem 3: KNN-Mútuo



Na etapa de propagação, os rótulos são distribuídos iterativamente. Inicialmente, verifica-se se os K vizinhos de uma amostra sem rótulo possuem rótulos que possam ser agregados a ela. A união dos rótulos dos K vizinhos é então aplicada à amostra, permitindo que, mesmo que uma amostra não possua rótulos inicialmente, ela possa adquiri-los nas iterações subsequentes à medida que os rótulos são propagados através do grafo. Além disso, os rótulos podem ser atualizados ao longo das iterações, já que os vizinhos podem receber novos rótulos em cada ciclo.

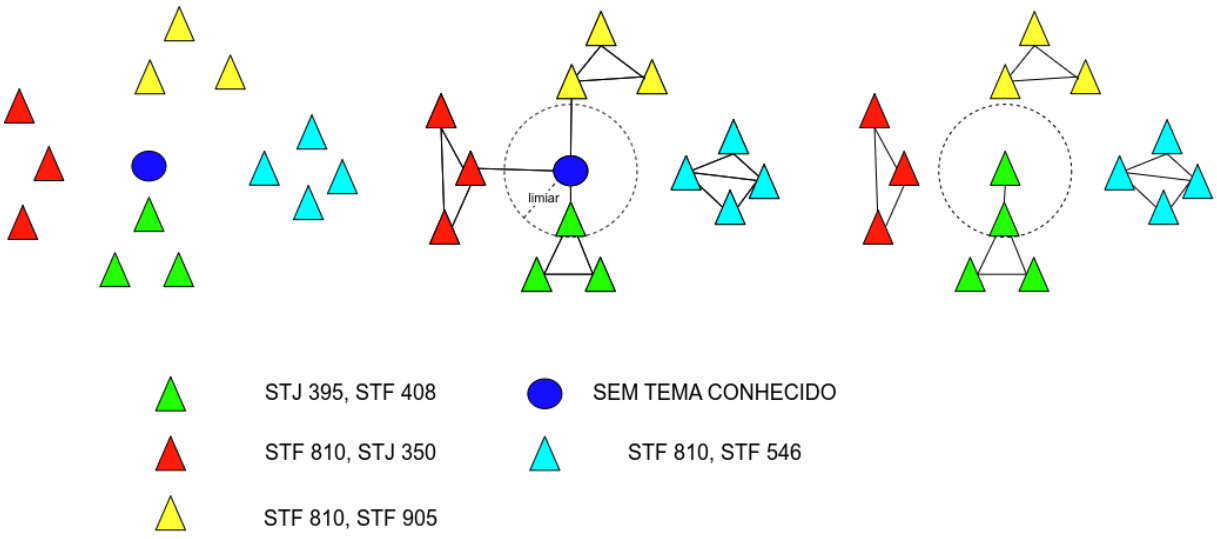
Um mecanismo de controle de iterações é utilizado para monitorar mudanças nos rótulos após cada iteração. Se nenhuma alteração é observada, o algoritmo é finalizado. Caso contrário, o processo segue para a iteração seguinte, com um limite estabelecido em 10 iterações.

Conforme ilustrado na Figura 9, a amostra sem rótulo, representada pelo círculo azul, é inicialmente conectada a três amostras com rótulos distintos. Após um número de iterações suficiente para estabilizar a propagação, esta amostra sem rótulo incorpora os rótulos de todas as amostras vizinhas, refletindo o mecanismo de propagação e atualização de rótulos facilitado pelo algoritmo KNN-Mútuo.

3.1.2.4 KNN-Mútuo com limiar

No contexto da quarta abordagem do estudo, a estratégia KNN-Mútuo é aprimorada com a introdução de um limiar que ajusta a proximidade exigida entre as amostras para estabelecer conexões. Além do parâmetro K, que determina a quantidade de vizinhos mais próximos a serem considerados, a aplicação de um limiar impõe uma condição adicional para a construção do grafo. A escolha do limiar, assim como o K, é ajustável e será variada no intuito de refinar a seleção de vizinhos, impactando diretamente a distribuição de rótulos no conjunto de dados.

Figura 10 – Abordagem 4: KNN-Mútuo com Limiar



Durante a propagação dos rótulos, a abordagem iterativa é mantida. Examina-se inicialmente se os K vizinhos de uma amostra sem rótulo, que também atendem ao critério do limiar, possuem rótulos aplicáveis. A partir daí, a união dos rótulos dos K vizinhos que excedem o limiar é atribuída à amostra. Assim, uma amostra que não possui rótulos pode receber identificações conforme a propagação de informações se ajusta ao limiar definido.

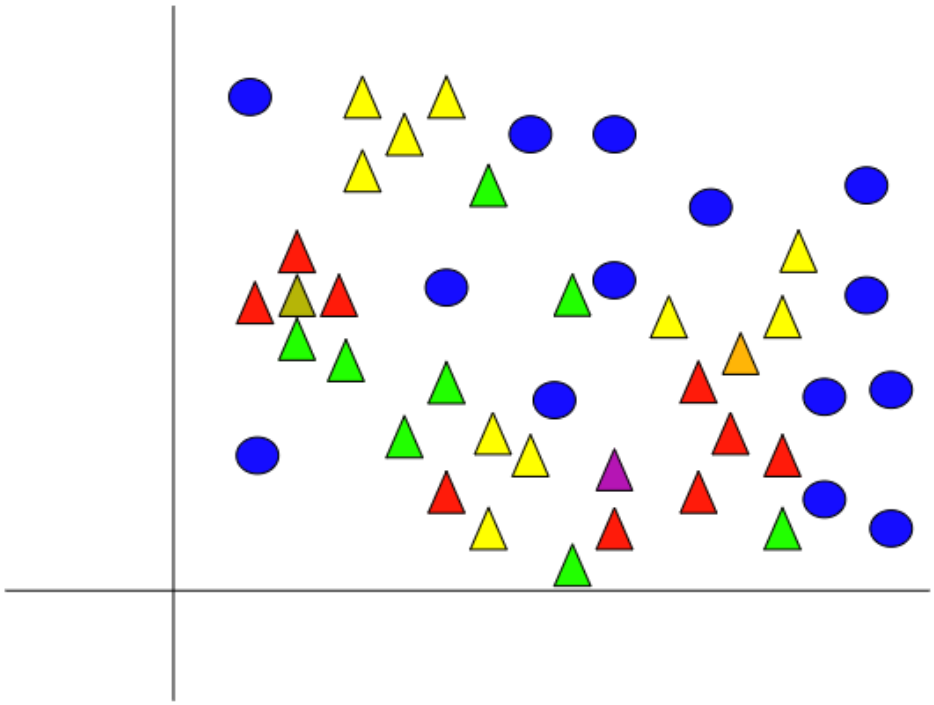
Os rótulos podem também ser revisados durante as iterações, pois os vizinhos podem obter novos rótulos a cada passagem.

Este método também segue o limite de 10 iterações, com um sistema de verificação de mudanças nos rótulos. A estabilidade dos rótulos é alcançada quando não há mais alterações entre as iterações, indicando a conclusão da propagação dos rótulos.

A Figura 10 ilustra a abordagem KNN-Mútuo com limiar. Visualmente, a amostra representada pelo círculo azul, sem rótulo inicial, estabelece conexões com amostras vizinhas que atendem ao critério de limiar. Após o processo de estabilização e propagação dos rótulos, a amostra sem rótulo adota os rótulos das amostras vizinhas que sobrevivem após a aplicação do limiar

Essas abordagens, inspiradas pelos pressupostos da aprendizagem semi-supervisionada, têm o potencial de maximizar a precisão e robustez da propagação de rótulos em no conjunto de dados. Esta implementação considera a dinâmica complexa dos dados multiclasse e multirrótulo, e busca oferecer uma solução para o desafio do dados do projeto "PEDRO".

Figura 11 – Distribuição das Amostras Esperado Pós-Aplicação dos Algoritmos



Após a execução dos algoritmos propostos, espera-se que a base de dados apresente uma combinação de amostras recém-rotuladas e outras que permanecem sem rótulos. Esta disposição ocorre devido à possibilidade de algumas amostras representarem tópicos ainda não abordados no conjunto de dados em questão. Ao observar a figura 11, é evidente a inclusão de rótulos em determinadas amostras, enquanto outras permanecem intencionalmente sem designação.

3.1.3 Comparação do desempenho dos algoritmos implementados

Após a implementação das técnicas de aprendizado semi-supervisionado, torna-se imperativo avaliar e contrastar a performance de cada método. Para essa avaliação, recorre-se a um conjunto de dados previamente rotulados, designados para teste.

O f1-score, que amalgama precisão e recall, será o critério principal de avaliação da eficácia dos algoritmos. Adicionalmente, recorre-se à métrica *Hamming Loss* para discernir a proporção de rótulos incorretamente atribuídos em relação ao conjunto total de predições.

A técnica que evidenciar superioridade em performance será submetida a uma inspeção mais granular. Neste estágio, calcular-se-ão precisão, recall e f1-score para cada tema de forma isolada, visando identificar os temas nos quais a técnica evidencia maior ou menor aptidão na atribuição de rótulos.

3.1.4 Avaliação qualitativa dos resultados com especialistas jurídicos

Subsequentemente à seleção da técnica mais eficaz, preconiza-se uma avaliação qualitativa dos resultados obtidos. Nesta fase, um lote de amostras rotuladas pela técnica escolhida será analisado por um coletivo de especialistas jurídicos.

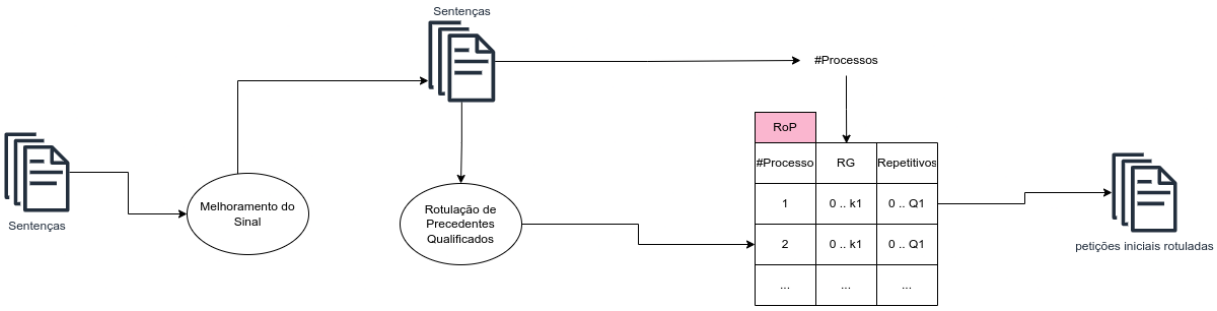
A apreciação desse coletivo tem o propósito de validar ou corrigir as classificações obtidas pela técnica. Este procedimento assegura não somente a eficácia técnica do método, mas também sua aplicabilidade e pertinência no domínio jurídico.

Os apontamentos provenientes dos especialistas serão sistematizados e analisados, culminando na elaboração de gráficos e relatórios que evidenciem a precisão e confiabilidade da técnica no escopo da investigação.

3.1.5 Plataforma de Extração e Descoberta de Precedentes dos Tribunais

O projeto “PEDRO” é estruturado em três etapas fundamentais: construção do *dataset*, com auxilio do RoP, treinamento de um modelo e inferência. Este trabalho foca predominantemente na primeira etapa, examinando a razão pela qual nem todas as amostras são rotuladas.

Figura 12 – Fluxo do projeto “PEDRO”



A fase de construção dos *datasets* é subdividida em três subetapas:

1. Melhoramento do Sinal: Utiliza-se a biblioteca GPAM_Preprocessing (2022) para pré-processar o texto, identificando e transformando citações de leis, artigos e outros termos jurídicos em *tokens*. Essa transformação realça a qualidade do sinal para posterior processamento.

2. Rotulação de Precedentes Qualificados: A partir dos *tokens* gerados, as citações referentes a precedentes são identificadas e atribuídas à petição inicial correspondente a cada sentença.

3. União dos Dados: Nesta fase, os *datasets* são amalgamados. As petições iniciais rotuladas, provenientes da subetapa anterior, são combinadas com as sentenças onde os juízes mencionam os precedentes. O resultado é um *dataset* consolidado que contém o texto da petição inicial e seus temas correlatos, categorizados em dois grupos predominantes: “repercussão geral” e “repetitivos”.

   É importante notar que, devido à variabilidade na maneira como os juízes mencionam os precedentes, nem todas as citações são capturadas pela biblioteca GPAM_Preprocessing (2022). Isso ocorre porque a biblioteca opera com base em expressões regulares. Se um juiz cita integralmente a ementa do precedente, a ferramenta pode não detectar o sinal.

4 Resultados e Discussões

4.1 Conjunto de Dados

O conjunto de dados neste estudo foi inicialmente categorizado em instâncias rotuladas e não rotuladas. Das 2.048 instâncias rotuladas, unificou-se os rótulos de cada uma para considerar as distintas combinações de rótulos como únicas. Após um processo de filtragem, assegurando um mínimo de duas instâncias para cada combinação de rótulos, restaram 1.886 instâncias abrangendo 165 categorias temáticas. Para a divisão entre treino e teste, inicialmente, 80% das amostras rotuladas foram destinadas para treino e 20% para teste. No entanto, verificou-se que algumas combinações de rótulos, especificamente aquelas com apenas duas instâncias, estavam presentes somente no conjunto de treino.

Para corrigir essa disparidade e assegurar a representatividade de todas as combinações de rótulos, adotou-se uma abordagem específica para essas instâncias, dividindo-as igualmente entre treino e teste. Esse ajuste resultou em um total de 1.461 amostras para treino, representando 77,47% do total, e 425 amostras para teste, correspondendo a 22,53%. Este ajuste visou equilibrar adequadamente as instâncias de treino e teste, permitindo uma avaliação mais justa da eficácia do algoritmo na propagação dos rótulos a partir das amostras rotuladas.

Para uma visão detalhada da distribuição das amostras entre treino e teste, é possível consultar o Anexo B, onde essa divisão é explicitada. Esta visualização oferece uma compreensão mais clara do balanceamento das amostras.

4.2 Vetorização

Durante a vetorização, foi adotado os parâmetros já utilizados no projeto PEDRO para a transformação TF-IDF:

- ***max_df*: 0.90** — Este parâmetro exclui termos presentes em mais de 90% dos documentos, pois são considerados comuns e, consequentemente, menos informativos.
- ***min_df*: 0.1** — Inclui termos que aparecem em pelo menos 10% dos documentos, garantindo que sejam estatisticamente significativos.
- ***ngram_range:* (1, 2)** — Habilita a captura de unigramas e bigramas, proporcionando uma análise textual mais robusta que leva em conta termos isolados e a relação entre pares de termos consecutivos.

- ***lowercase*: False** — Mantém a capitalização original dos termos, preservando distinções importantes para a análise, como em nomes próprios.

Esses parâmetros foram escolhidos por refletirem as configurações ótimas de TF-IDF identificadas no projeto "PEDRO", assegurando consistência nos resultados da vetorização em todo o corpus documental.

4.3 Parâmetros dos Algoritmos

No âmbito do estudo presente, procedeu-se à variação sistemática dos parâmetros de cada algoritmo, visando identificar a configuração ótima. A Tabela 1 fornece um resumo detalhado deste processo, incluindo os parâmetros ajustados, os intervalos de variação estabelecidos, os incrementos aplicados em cada etapa, e o número total de combinações testadas para cada algoritmo que totalizaram 513 combinações de parâmetros.

Tabela 1 – Configurações Experimentais Detalhadas para os Algoritmos

Algoritmo: Mais Próximo			
Parâmetro	**Intervalo de Variações**	**Incremento**	**Total de Combinações**
Limiar de Similaridade	60% a 95%	5%	8
Algoritmo: Maioria			
Parâmetro	**Intervalo de Variações**	**Incremento**	**Total de Combinações**
Limiar de Similaridade	60% a 95%	5%	64
Limiar de Maioria	60% a 95%	5%	
Algoritmo: KNN-Mútuo			
Parâmetro	**Intervalo de Variações**	**Incremento**	**Total de Combinações**
Valor de K	2 a 50	1	49
Algoritmo: KNN-Mútuo com Limiar			
Parâmetro	**Intervalo de Variações**	**Incremento**	**Total de Combinações**
Valor de K	2 a 50	1	392
Limiar de Similaridade	60% a 90%	5%	
Total Geral		**513**	

Todos os algoritmos foram parametrizados com um limite máximo de 10 iterações, uma medida adotada para forçar a estabilização dos algoritmos e evitar ciclos infinitos ou extensos de processamento. Tal abordagem permitiu uma análise exaustiva, contribuindo para a identificação das configurações algorítmicas mais eficientes.

4.4 Análise Quantitativa

No estudo realizado, após a execução de todas as combinações experimentais propostas, os resultados obtidos permitiram uma análise da eficácia dos algoritmos em questão. Para a orientação na escolha da melhor combinação de parâmetros para cada algoritmo, a métrica F1-Score Macro foi adotada como principal indicador, devido à sua relevância em contextos de classes desbalanceadas e à sua capacidade de avaliar o desempenho dos modelos de forma equilibrada entre precisão e *recall*. A Tabela 2 detalha as configurações que alcançaram os melhores resultados segundo a métrica F1-Score Macro, que pode ser encontrada detalhadamente na Tabela 3. Além disso, esta tabela inclui a quantidade de iterações que cada algoritmo realizou até estabilizar, e os detalhes das demais métricas serão discutidos a seguir.

Tabela 2 – Configurações Otimizadas por Algoritmo

Algoritmo	**Limiar de Similaridade**	**Limiar de Maioria**	**K**	**Iterações**
Mais Próximo	60%	-	-	2
Maioria	75%	60%	-	1
KNN-Mútuo	-	-	4	2
KNN-Mútuo com Limiar	60%	-	7	2

Conforme apresentado na Tabela 3, que exibe os valores de F1-Score Macro juntamente com precisão e recall, o algoritmo “Mais Próximo” destacou-se com o maior F1-Score Macro, registrando 0.642. Este valor é significativamente importante, pois evidencia um equilíbrio efetivo entre precisão (0.693) e *recall* (0.656) em comparação com os demais algoritmos. Os valores mais baixos observados nas métricas macro para todos os algoritmos podem ser atribuídos à natureza dos dados, que se caracterizam por uma acentuada desproporção entre as classes. Tal desproporção impacta diretamente a métrica Macro, pois esta dá peso igual a todas as classes, independentemente do seu tamanho, o que realça o desafio de classificar corretamente as classes minoritárias.

Conforme apresentado na Tabela 4, o algoritmo “Mais Próximo” destacou-se ao obter o maior F1-Score Micro, com 0.869. Este resultado é acompanhado por uma precisão de 0.860 e um *recall* de 0.879, evidenciando a eficiência do algoritmo na classificação precisa das instâncias em todo o conjunto de dados. Comparativamente, os algoritmos “Maioria”, “KNN-Mútuo” e “KNN-Mútuo com Limiar” também apresentaram desempenhos adequados em precisão e *recall*, mas não alcançaram a mesma efetividade global que o “Mais Próximo”.

Tabela 3 – Métricas Macro dos Algoritmos

Algoritmo	**Precisão**	***Recall***	**F1-score**
Mais Próximo	**0.642**	**0.693**	**0.656**
Maioria	0.550	0.423	0.464
KNN-Mútuo	0.677	0.662	0.641
KNN-Mútuo com Limiar	0.602	0.680	0.621

Por exemplo, o algoritmo "Maioria" registrou a maior precisão (0.962), porém com um *recall* de 0.664 e um F1-Score Micro de 0.786, indicando uma tendência a uma classificação menos abrangente em comparação com o "Mais Próximo".

Tabela 4 – Métricas Micro dos Algoritmos

Algoritmo	**Precisão**	***Recall***	**F1-score**
Mais Próximo	0.860	**0.879**	**0.869**
Maioria	**0.962**	0.664	0.786
KNN-Mútuo	0.838	0.705	0.766
KNN-Mútuo com Limiar	0.762	0.799	0.780

A elevada pontuação do "Mais Próximo" na métrica F1-Score Micro é especialmente relevante no contexto deste estudo, onde os dados apresentam um alto grau de desbalanceamento entre as classes. Em cenários onde algumas classes são significativamente mais prevalentes do que outras, uma alta pontuação na métrica F1-Score Micro indica que o algoritmo conseguiu manter um desempenho eficaz de classificação em todo o conjunto de dados, abrangendo todas as classes.

Conforme indicado na Tabela 5, o algoritmo "Mais Próximo" se destaca nas métricas de avaliação. Esse algoritmo atingiu um F1-Score Ponderado de 0.852, uma precisão de 0.842 e um *recall* de 0.879. Estes resultados demonstram sua eficácia notável na classificação correta das instâncias em todas as classes, ressaltando a importância de levar em consideração o desbalanceamento entre elas.

Tabela 5 – Métricas Ponderadas dos Algoritmos

Algoritmo	**Precisão**	***Recall***	**F1-score**
Mais Próximo	0.842	**0.879**	**0.852**
Maioria	0.841	0.664	0.723
KNN-Mútuo	**0.858**	0.705	0.758
KNN-Mútuo com Limiar	0.785	0.799	0.782

Os algoritmos "KNN-Mútuo" e "KNN-Mútuo com Limiar", por outro lado, apresentaram resultados similares na métrica F1-Score Ponderada, com valores de 0.758 e 0.782, respectivamente. Isso sugere que a adição de um limiar na variante "KNN-Mútuo com

Limiar" não resultou em uma melhoria substancial na performance ponderada, quando comparada ao "KNN-Mútuo". A eficácia desses algoritmos, embora inferior ao "Mais Próximo", ainda é significativa, especialmente considerando as diferentes abordagens e os aspectos específicos de cada algoritmo na gestão das classes desbalanceadas.

Os resultados apresentados nas Tabelas 3, 4 e 5 destacam o desafio inerente ao desbalanceamento de classes e sublinham o desempenho superior do algoritmo "Mais Próximo". Este algoritmo demonstrou ser particularmente eficaz para o contexto dos dados do projeto PEDRO, obtendo as maiores pontuações tanto na métrica Média Macro como nas métricas Média Micro e Média Ponderada. A superioridade nas métricas Macro, Micro e Ponderada indica que o 'Mais Próximo' não só lidou bem com as classes majoritárias, mas também se saiu melhor na classificação das classes minoritárias, em comparação com os outros algoritmos avaliados.

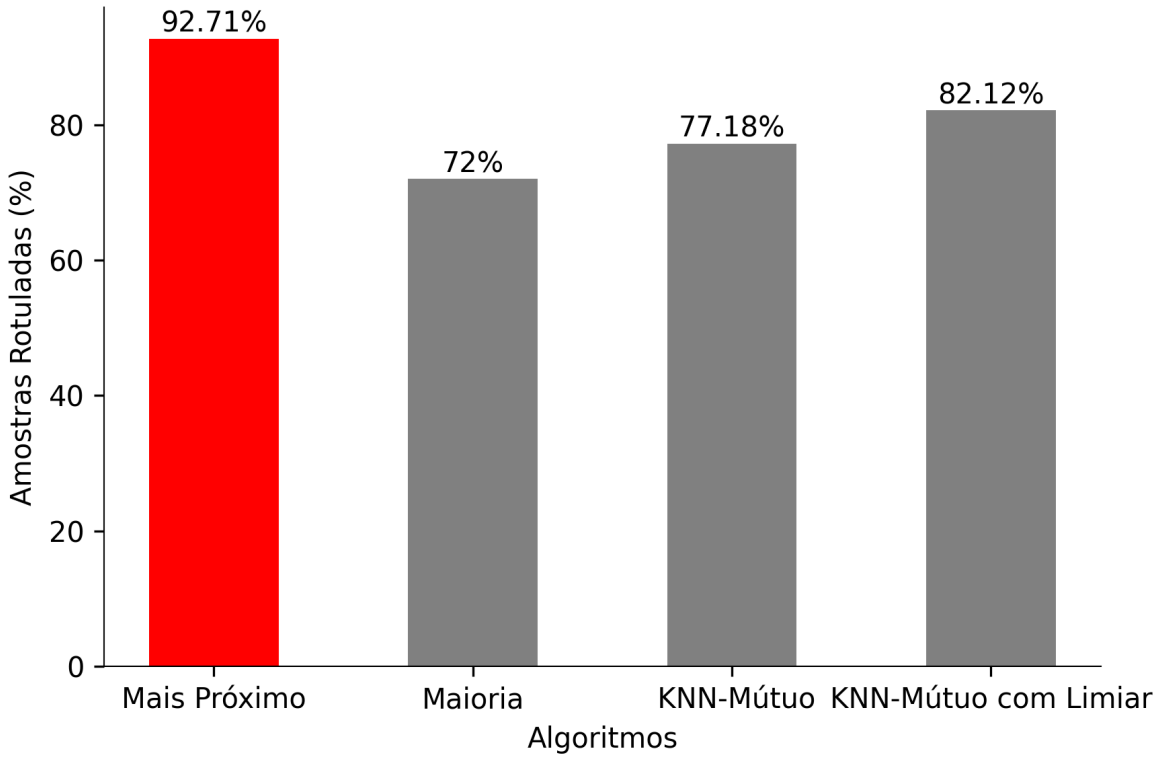
No anexo C, é possível analisar com mais detalhes o resultado de precisão, *recall* e f1-score para cada tema. Nas tabelas 9, 10, 11 e 12, observa-se a performance dos respectivos algoritmos "Mais próximo", "Maioria", "KNN-Mútuo" e "KNN-Mútuo com Limiar" para cada tema. A análise dessas tabelas revela a quantidade de temas com métricas zeradas, resultado da baixa amostragem em determinadas classes. Este fato reforça a relevância do desempenho equilibrado do algoritmo "Mais Próximo".

Tabela 6 – *Hamming Loss* e Acurácia para cada Algoritmo

Nome do Experimento	***Hamming Loss***	**Acurácia**
Mais Próximo	**0.003**	**0.767**
Maioria	0.004	0.609
KNN-Mútuo	0.005	0.598
KNN-Mútuo com Limiar	0.005	0.586

Adicionalmente, a Tabela 6 aborda o *Hamming Loss* e a Acurácia, duas métricas complementares importantes. O *Hamming Loss* quantifica o número médio de rótulos incorretos em relação ao total de rótulos, enquanto a Acurácia reflete a proporção de rótulos corretamente identificados. Novamente, o "Mais Próximo" demonstra uma vantagem, ostentando o menor *Hamming Loss* de 0.003 e a maior Acurácia de 0.767 corroborando a sua superioridade na precisão da classificação em comparação aos outros métodos testados.

Figura 13 – Porcentagem de amostras rotuladas no treino

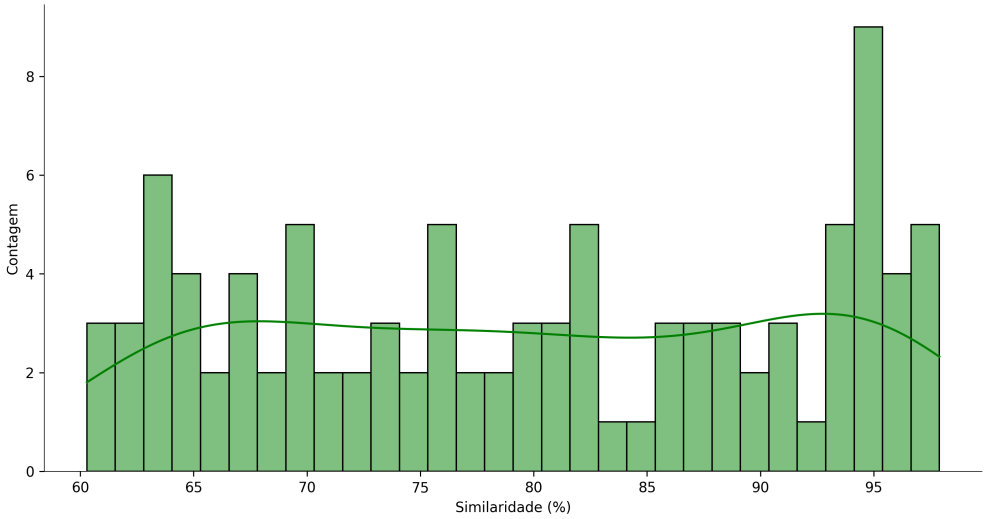


A Figura 13 oferece uma visualização quantitativa da porcentagem de amostras rotuladas no treino por algoritmo. Notavelmente, o algoritmo "Mais Próximo" conseguiu propagar para 92,71% das amostras sem rótulos, uma proporção substancialmente maior do que a alcançada pelos outros métodos. Dado que a base de dados de teste estava integralmente rotulada, o algoritmo "Mais Próximo" demonstrou ser altamente competente, propagando rótulos para a vasta maioria das amostras. Em contraste, as taxas de propagação de rótulos dos algoritmos "Maioria", "KNN-Mútuo" e "KNN-Mútuo com Limiar" foram respectivamente de 72%, 77.18% e 82.12%. Tais resultados indicam uma superioridade do algoritmo 'Mais Próximo' na tarefa de rotulação do conjunto de dados do estudo.

4.5 Análise Qualitativa

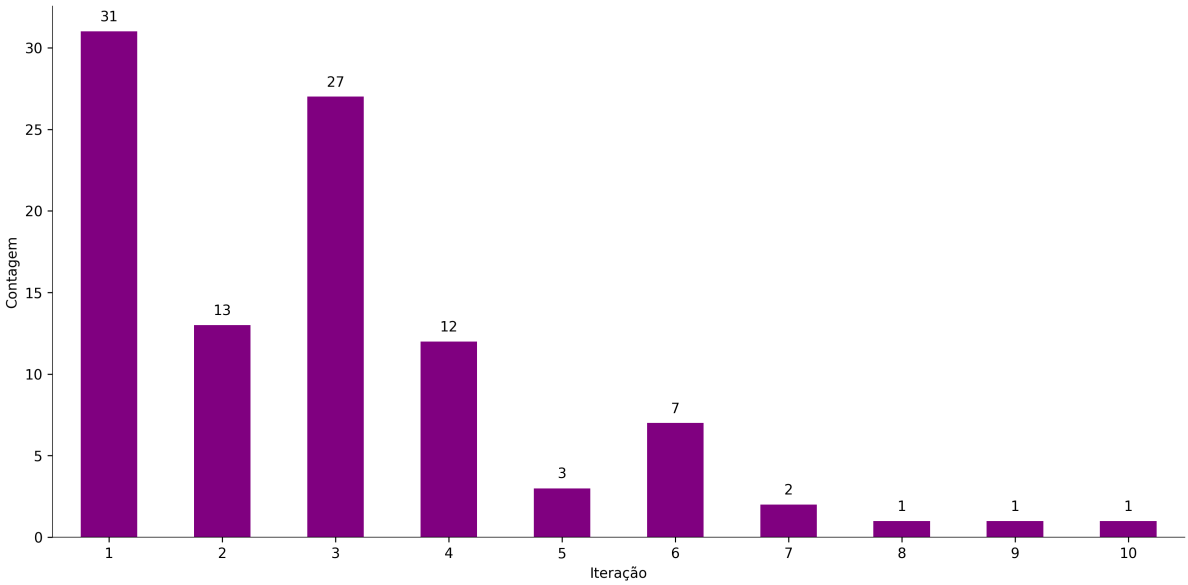
Após análises comparativas, o algoritmo "Mais Próximo" foi identificado como o mais eficaz para a propagação de rótulos na base de teste. O algoritmo foi aplicado no conjunto de 20.455 documentos não rotulados. A propagação dos rótulos foi conduzida com os parâmetros ótimos presente na tabela 2. Foi apresentada para o time de especialistas em Direito do laboratório DR.IA 98 amostras que representa 0.48% do total de amostras não rotuladas.

Figura 14 – Distribuição de Amostras Avaliadas por Similaridade



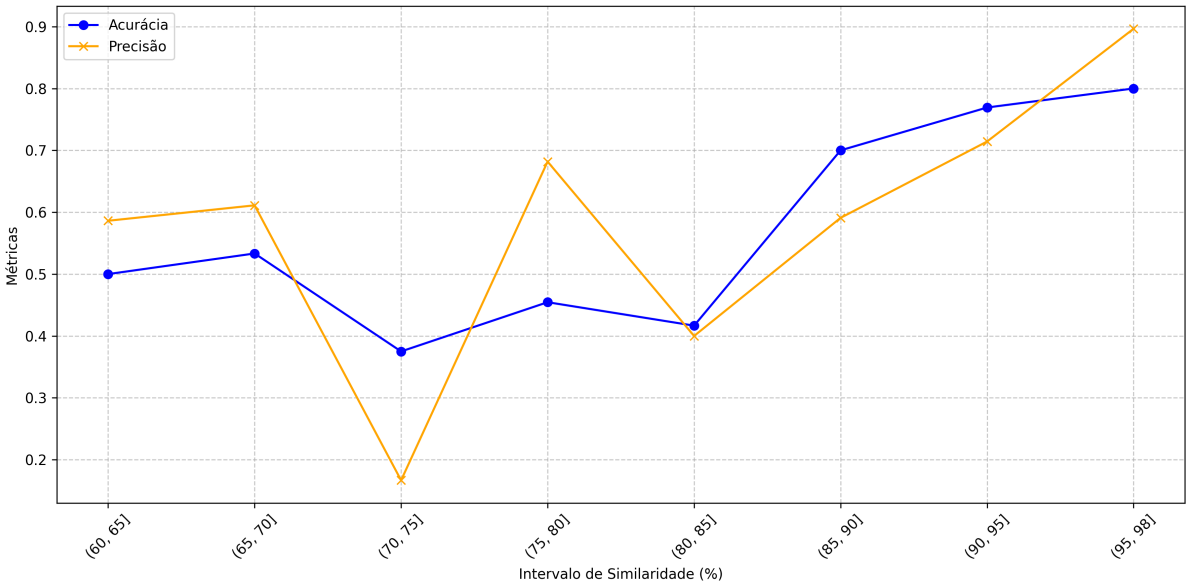
No curso desta análise, foram explorados 25 temas diferentes, com a propagação de rótulos efetuada através de múltiplas iterações do algoritmo. As amostras foram extraídas de maneira aleatória, porém uniforme em relação aos intervalos de similaridade, a fim de garantir uma análise representativa para cada segmento. A Figura 14 demonstra a distribuição das amostras conforme a similaridade, evidenciando a equitativa seleção entre os intervalos. Ademais, a Figura 15 exibe a distribuição das amostras por iteração, revelando a dinâmica de propagação dos rótulos ao longo das iterações.

Figura 15 – Distribuição Amostras Avaliadas por Iteração



Os especialistas do laboratório DR.IA conduziram uma avaliação para determinar a eficácia do algoritmo "Mais Próximo" ao classificar corretamente temas legais em uma série de documentos. Essa análise focou exclusivamente na capacidade do modelo de acertar as classificações, resultando na identificação de verdadeiros positivos — temas corretamente sugeridos pelo algoritmo — e falsos positivos — temas sugeridos incorretamente. Com base nesses dados, foi possível calcular a acurácia e a precisão do modelo, que se mostraram promissoras com 58,16% e 59,69%, respectivamente.

Figura 16 – Acurácia e Precisão por Intervalo de Similaridade

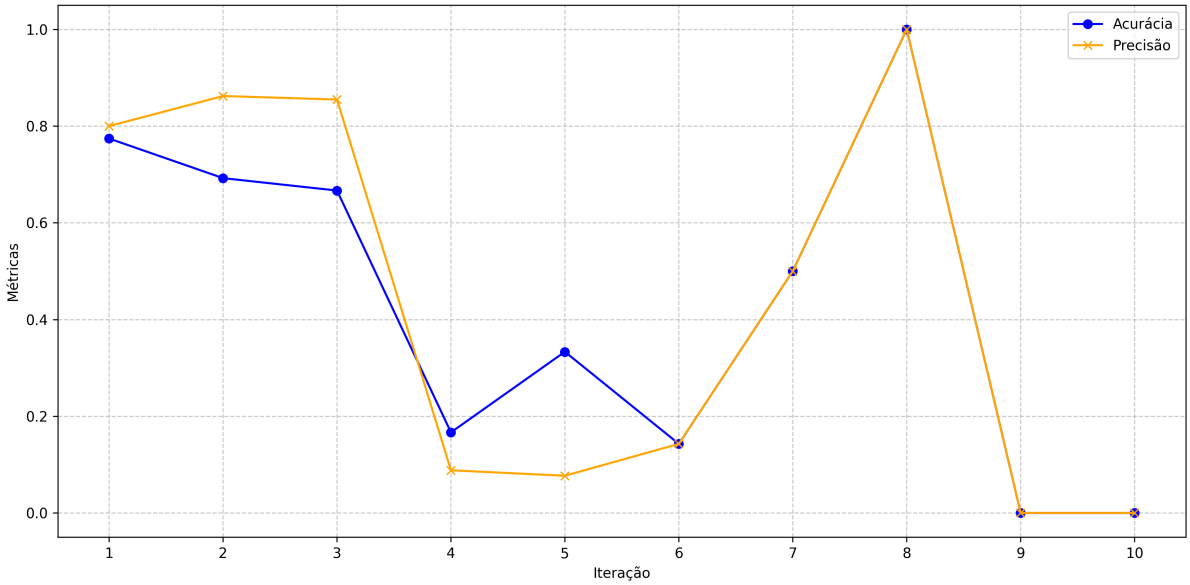


A Figura 16 demonstra como a acurácia e a precisão flutuam em diferentes intervalos de similaridade. Nota-se que o modelo apresenta um desempenho estável, especialmente nas faixas de maior similaridade, o que sinaliza sua competência em realizar classificações acertadas entre documentos com alta similaridade.

A Figura 17 apresenta uma perspectiva sobre a evolução da acurácia e precisão ao longo das iterações do algoritmo. É notável um pico de precisão seguido de um declínio na oitava iteração, o que sugere a possibilidade de um limiar ideal para determinar o número máximo de iterações. A progressão do algoritmo para além deste ponto específico aparenta não contribuir para o aprimoramento das classificações, podendo, de fato, acarretar em sua depreciação.

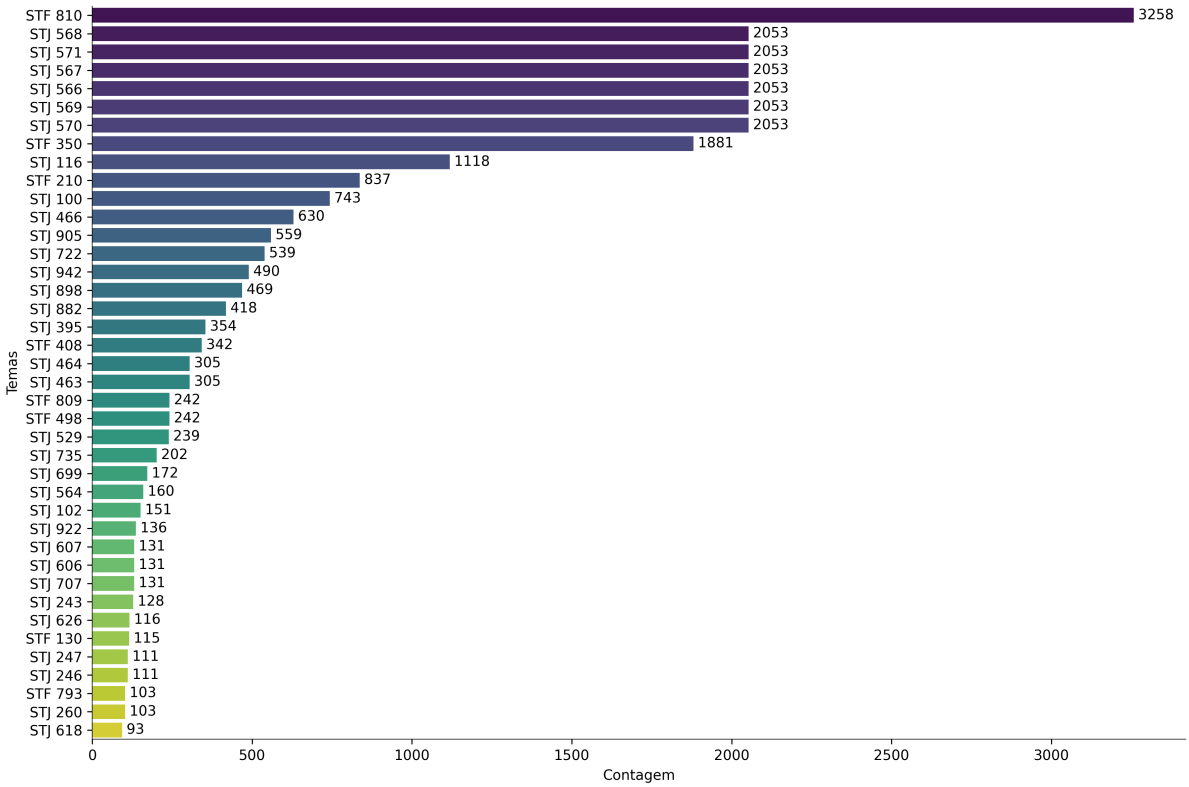
Embora as métricas não sejam perfeitas, elas refletem a capacidade do algoritmo "Mais Próximo" de atuar como uma ferramenta de apoio na classificação de documentos jurídicos, com um desempenho que evidencia o seu valor prático.

Figura 17 – Acurácia e Precisão por Iteração



4.6 Aplicação do Melhor Algoritmo

Figura 18 – Distribuição dos 40 temas mais propagados

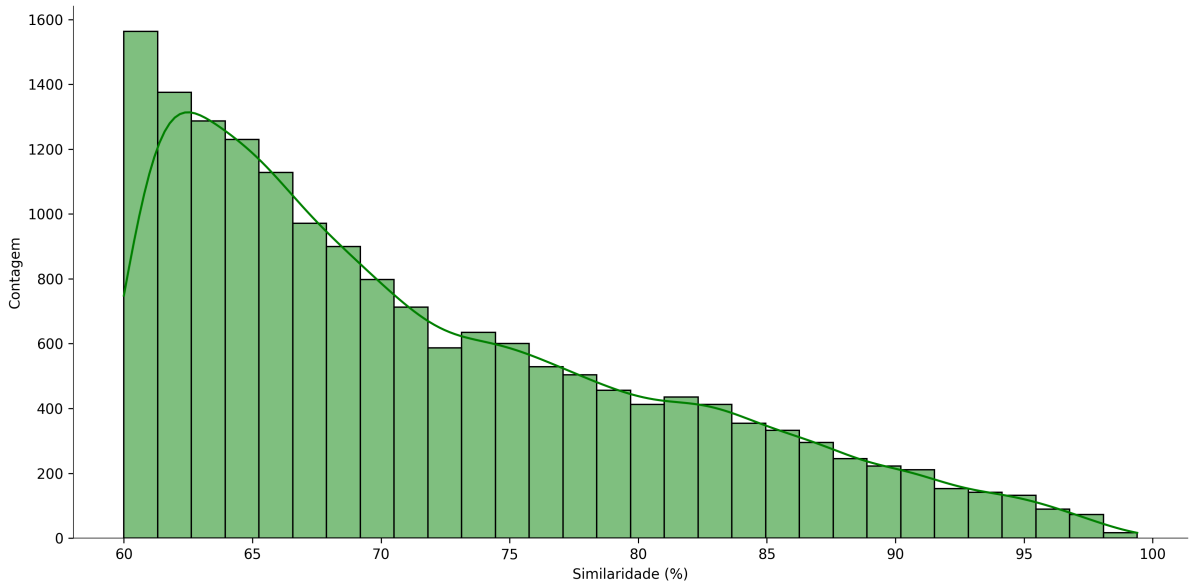


Com a definição do algoritmo “Mais Próximo” como o mais eficaz, procedeu-se à sua aplicação sobre o conjunto de dados não rotulados. Utilizando-se das 2.048 amostras

já rotuladas, o algoritmo foi estimulado para as 20.455 instâncias previamente sem temas identificados. Este processo resultou na rotulação de 16.802 amostras, equivalentes a 82,14% do conjunto originalmente sem rótulos, representando um aumento de 820% na quantidade de amostras rotuladas. As 3.653 amostras restantes, que representam 17,86% do conjunto de dados não rotulado inicialmente, podem ser casos ainda não vinculados a precedentes existentes. Estes casos pendentes de rotulagem oferecem uma oportunidade para análises futuras, que poderão determinar se são passíveis de associação a novos precedentes ou se poderão ser rotulados à medida que novos dados forem integrados à base e processados pelo algoritmo.

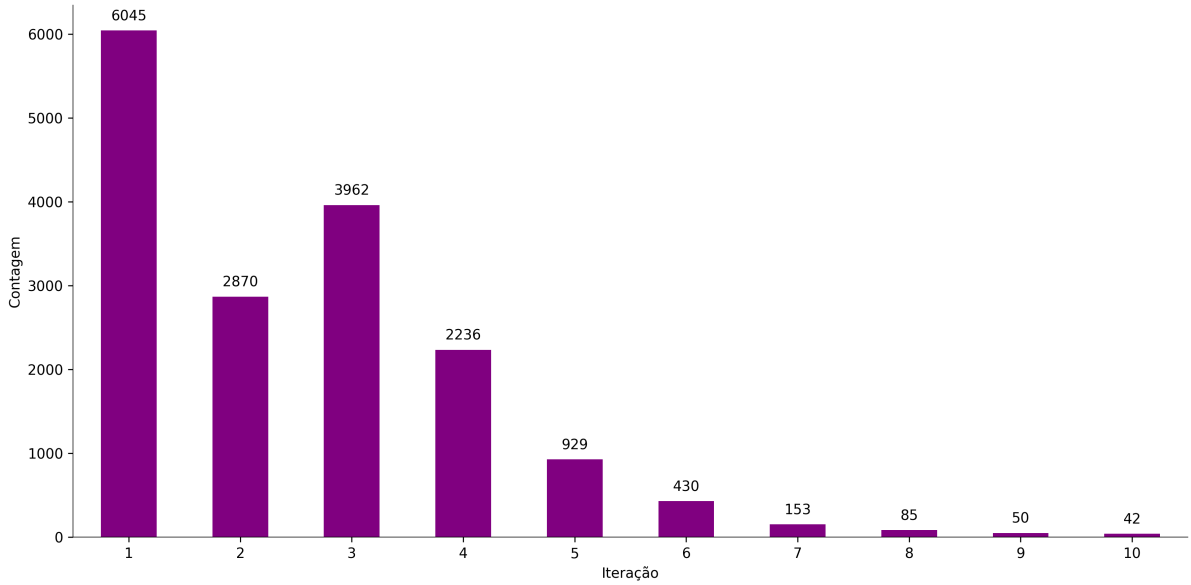
A Figura 18 ilustra a distribuição dos 40 temas mais frequentemente propagados pelo algoritmo. Como pode ser observado, há uma variação considerável na frequência de propagação, com alguns temas alcançando uma disseminação significativamente mais ampla do que outros. Esta distribuição reflete a configuração intrínseca do conjunto de dados rotulados. Detalhes complementares a respeito da propagação temática, englobando temas com menor frequência de ocorrência, podem ser consultados no anexo D.

Figura 19 – Distribuição da amostras propagadas por similaridade



A Figura 19 exibe a distribuição de amostras classificadas conforme o grau de similaridade. Observa-se uma concentração de rotulações em amostras com no mínimo 60% de similaridade, evidenciando uma tendência à diminuição na quantidade de rótulos atribuídos à medida que se ascende na escala de similaridade.

Figura 20 – Distribuição da amostras propagadas por Iteração



Por fim, a Figura 20 apresenta a distribuição de amostras rotuladas ao longo das iterações do algoritmo. Nota-se uma proeminência de classificações nas fases iniciais, com uma declinação notória em iterações subsequentes, o que pode indicar um rápido alcance de um patamar de saturação pelo algoritmo, onde iterações adicionais pouco contribuem para o quantitativo de amostras rotuladas. Tal fenômeno reforça a hipótese de que a implementação de um limiar de corte, conforme discutido na análise qualitativa, teria impacto mínimo no volume final de rótulos propagados, pois a maior parte das rotulações corretas é realizada nas primeiras iterações.

5 Conclusão

O presente trabalho acadêmico conduziu uma investigação sobre a aplicação de algoritmos de classificação transdutivos semi-supervisionados. O objetivo primordial deste estudo foi a ampliação da quantidade de rótulos em base de dados jurídicos, uma abordagem crucial para possibilitar o treinamento eficaz de classificadores em precedentes qualificados no âmbito do sistema judiciário brasileiro. Esta estratégia visa, fundamentalmente, a redução da carga de trabalho manual e repetitiva dos servidores, permitindo que estes se dediquem a tarefas que requerem maior discernimento e análise crítica.

Baseando-se nos princípios do aprendizado semi-supervisionado, foram desenvolvidos quatro algoritmos distintos para a avaliação comparativa de desempenho. Estes incluem o “Mais Próximo” , “Maioria", “KNN-Mútuo” e “KNN-Mútuo com Limiar”. A pesquisa se concentra na análise comparativa desses classificadores, com foco na avaliação de métricas fundamentais como F1-Score, precisão, *recall*, acurácia e *Hamming Loss*, visando identificar o classificador de maior eficiência.

Entre os algoritmos avaliados, o "Mais Próximo"destacou-se por sua notável performance. Enfrentando o desafio do grande desbalanceamento da base de dados, este algoritmo atingiu resultados expressivos nas métricas macro: uma precisão de 0.642, um recall de 0.693 e um F1-Score de 0.656. Embora estes valores sejam relativamente mais baixos em comparação com as métricas obtidas em bases de dados mais equilibradas, eles são de grande importância e demonstram a eficácia do "Mais Próximo"em um contexto desafiador de desbalanceamento de classes.

Um aspecto notável deste estudo é o aumento de 820% na quantidade de amostras rotuladas no projeto “PEDRO” alcançado pelo algoritmo “Mais Próximo”. Além disso, o algoritmo “Mais Próximo” exibe métricas que sugerem um desempenho aprimorado com o aumento do volume de dados. Isso é evidenciado ao analisar as métricas micro e ponderada, que destacam a qualidade da rotulação nos rótulos que possuem maior representação. Considerando a natureza do projeto “PEDRO”, que inclui um sistema de rotulação automática, a integração contínua de novos dados tende a facilitar a construção de um classificador mais robusto e eficiente.

Em síntese, este estudo demonstra que o algoritmo “Mais Próximo”, um classificador transdutivo semi-supervisionado, sobressai não apenas tecnicamente nas métricas de avaliação, mas também oferece um valor prático para a classificação de dados no sistema judiciário brasileiro. A implementação deste algoritmo promete uma gestão de informações mais ágil e eficiente, contribuindo significativamente para o aprimoramento do sistema jurídico do país.

5.1 Trabalhos Futuros

Este trabalho acadêmico, focado na investigação sobre a aplicação de algoritmos de classificação transdutivos semi-supervisionados, delineia várias recomendações para pesquisas futuras com o intuito de aprimorar a eficácia dos classificadores no contexto do sistema judiciário brasileiro.

Uma das sugestões é a realização de testes com os algoritmos em um espectro mais amplo de variação. Isso permitiria avaliar se os algoritmos apresentam melhor desempenho em intervalos de similaridade menores. Outra proposta é a expansão da base de precedentes e de dados já rotulados no fluxo inicial do projeto “PEDRO”, o que pode fornecer percepções adicionais sobre a performance dos algoritmos em um conjunto de dados mais diversificado.

Além disso, recomenda-se testar uma gama mais ampla de algoritmos de propagação de rótulos. O objetivo seria explorar diferentes métodos de propagação, e não apenas focar na variação da construção de grafos em contextos semi-supervisionados. Esta abordagem poderia revelar novas possibilidades de melhoria na eficiência dos classificadores.

Outro aspecto importante é a experimentação com diferentes tipos de vetorizadores. Atualmente, existem vetorizadores capazes de representar de forma mais precisa as nuances dos textos. Assim, variar os parâmetros do TF-IDF e testar outros tipos de vetorizadores pode ser benéfico. Isso poderia resultar em uma representação numérica mais eficaz dos documentos, aumentando a distinção entre documentos com rótulos diferentes e aprimorando a similaridade entre documentos com os mesmos rótulos.

Adicionalmente, uma das propostas de melhoria envolve a variação do limiar de similaridade conforme a iteração aumenta no algoritmo "Mais Próximo". Esta abordagem poderia reduzir a quantidade de erros ao decorrer das iterações, otimizando a precisão do algoritmo na classificação das amostras.

Em resumo, essas recomendações para trabalhos futuros visam não apenas aprimorar os métodos de classificação de dados no sistema judiciário brasileiro, mas também explorar novas fronteiras no campo de classificadores transdutivos semi-supervisionados.

Referências

AGGARWAL, C. C.; HINNEBURG, A.; KEIM, D. A. On the surprising behavior of distance metrics in high dimensional spaces. In: *International Conference on Database Theory*. [s.n.], 2001. Disponível em: <https://api.semanticscholar.org/CorpusID:1648083>. Citado na página 31.

BLAND, M. *An Introduction to Medical Statistics*. 3rd. ed. [S.l.]: Oxford University Press, 2000. Citado na página 35.

BRASIL. Emenda constitucional nº 45, de 8 de dezembro de 2004. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 2004. ISSN 1677-7042. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/Emendas/Emc/emc45.htm>. Citado na página 18.

BRASIL. Lei nº 13.105, de 16 de março de 2015. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 2015. ISSN 1677-7042. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm>. Citado 3 vezes nas páginas 17, 18 e 19.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. (Ed.). *Semi-Supervised Learning*. London, England: MIT Press, 2010. (Adaptive Computation and Machine Learning series). Citado na página 24.

DELALLEAU, O.; BENGIO, Y.; ROUX, N. L. Efficient non-parametric function induction in semi-supervised learning. In: COWELL, R. G.; GHAHRAMANI, Z. (Ed.). *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. PMLR, 2005. (Proceedings of Machine Learning Research, R5), p. 96–103. Reissued by PMLR on 30 March 2021. Disponível em: <https://proceedings.mlr.press/r5/delalleau05a.html>. Citado na página 23.

GHAHRAMANI, Z. Unsupervised learning. In: _____. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 72–112. ISBN 978-3-540-28650-9. Disponível em: <https://doi.org/10.1007/978-3-540-28650-9_5>. Citado na página 23.

GODBOLE, S.; SARAWAGI, S. Discriminative methods for multi-labeled classification. In: DAI, H.; SRIKANT, R.; ZHANG, C. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 22–30. ISBN 978-3-540-24775-3. Citado na página 27.

GPAM_Preprocessing. 2022. Patente: Programa de Computador. Número do registro: BR512022003333-0, data de registro: 06/12/2022, Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial. Citado na página 44.

HAMMING, R. W. Error detecting and error correcting codes. *The Bell System Technical Journal*, v. 29, n. 2, p. 147–160, 1950. Citado na página 28.

HEALEY, J. F. *The Essentials of Statistics: A Tool for Social Research*. 3rd. ed. [S.l.]: Cengage Learning, 2014. Citado na página 36.

JEBARA, T.; WANG, J.; CHANG, S.-F. Graph construction and b-matching for semi-supervised learning. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, 2009. (ICML '09), p. 441–448. ISBN 9781605585161. Disponível em: <https://doi.org/10.1145/1553374.1553432>. Citado na página 24.

JIVANI, A. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl.*, v. 2, p. 1930–1938, 11 2011. Citado na página 20.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. In: _____. *Document Retrieval Systems*. GBR: Taylor Graham Publishing, 1988. p. 132–142. ISBN 0947568212. Citado na página 21.

KHURANA, D. et al. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, Springer Science and Business Media LLC, v. 82, n. 3, p. 3713–3744, jul 2022. Disponível em: <https://doi.org/10.1007%2Fs11042-022-13428-4>. Citado na página 20.

LIU, W.; WANG, J.; CHANG, S.-F. Robust and scalable graph-based semisupervised learning. *Proceedings of the IEEE*, v. 100, p. 2624–2638, 2012. Disponível em: <https://api.semanticscholar.org/CorpusID:554484>. Citado na página 24.

MONTGOMERY, D. C.; RUNGER, G. C. *Applied Statistics and Probability for Engineers*. 5th. ed. [S.l.]: John Wiley Sons, 2010. Citado na página 36.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. [S.l.]: The MIT Press, 2012. ISBN 0262018020. Citado na página 22.

OZAKI, K. et al. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. p. 154–162, 07 2011. Citado 2 vezes nas páginas 25 e 26.

PISTELLATO, M. et al. Robust phase unwrapping by probabilistic consensus. *Optics and Lasers in Engineering*, v. 121, p. 428–440, 2019. ISSN 0143-8166. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0143816618317044>. Citado na página 21.

PLISSON, J.; LAVRAč, N.; MLADENIĆ, D. A rule based approach to word lemmatization. In: . [s.n.], 2004. Disponível em: <https://api.semanticscholar.org/CorpusID:15628229>. Citado na página 20.

RAHUTOMO, F.; KITASUKA, T.; ARITSUGI, M. Semantic cosine similarity. In: . [s.n.], 2012. Disponível em: <https://api.semanticscholar.org/CorpusID:18411090>. Citado na página 30.

SAIF, H. et al. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 810–817. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/292_Paper.pdf>. Citado na página 20.

SCHAPIRE, R. E.; SINGER, Y. *Machine Learning*, Springer Science and Business Media LLC, v. 39, n. 2/3, p. 135–168, 2000. Disponível em: <https://doi.org/10.1023%2Fa%3A1007649029923>. Citado na página 26.

SOUSA, C.; REZENDE, S.; BATISTA, G. Influence of graph construction on semi-supervised learning. In: . [S.l.: s.n.], 2013. v. 8190. ISBN 978-3-642-38708-1. Citado na página 25.

SUBRAMANYA, A.; TALUKDAR, P. *Graph-Based Semi-Supervised Learning*. Morgan & Claypool Publishers, 2014. (Synthesis Lectures on Artificial Intelligence and Machine Learning). ISBN 9781627052023. Disponível em: <https://books.google.com.br/books?id=fzKNBQAAQBAJ>. Citado na página 25.

TANG, L.; RAJAN, S.; NARAYANAN, V. K. Large scale multi-label classification via metalabeler. In: *Proceedings of the 18th International Conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, 2009. (WWW '09), p. 211–220. ISBN 9781605584874. Disponível em: <https://doi.org/10.1145/1526709.1526738>. Citado na página 29.

TSOUMAKAS, G.; VLAHAVAS, I. Random k-labelsets: An ensemble method for multilabel classification. In: KOK, J. N. et al. (Ed.). *Machine Learning: ECML 2007*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 406–417. ISBN 978-3-540-74958-5. Citado na página 26.

YANG, Y. An evaluation of statistical approaches to text categorization. *Information Retrieval*, v. 1, p. 69–90, 1999. Disponível em: <https://api.semanticscholar.org/CorpusID:93891>. Citado na página 28.

ZHANG, M.-L.; ZHANG, K. Multi-label learning by exploiting label dependency. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2010. (KDD '10), p. 999–1008. ISBN 9781450300551. Disponível em: <https://doi.org/10.1145/1835804.1835930>. Citado na página 27.

ZHU, X.; GHAHRAMANI, Z. Learning from labeled and unlabeled data with label propagation. 07 2003. Citado na página 24.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. Semi-supervised learning using gaussian fields and harmonic functions. In: . [S.l.: s.n.], 2003. v. 3, p. 912–919. Citado na página 25.

Anexos

ANEXO A – Distribuição Geral dos Temas

Tabela 7 – Distribuição dos temas nos dados rotulados

Tema	**Quantidade**
STF 810	617
STJ 395	221
STF 408	186
STF 350	182
STJ 529	96
STJ 905	93
STJ 898	79
STJ 568	77
STJ 569	77
STJ 567	77
STJ 570	77
STJ 566	77
STJ 571	77
STJ 626	56
STJ 707	55
STJ 185	46
STJ 722	43
STJ 777	37
STJ 247	36
STJ 246	36
STF 27	35
STJ 618	29
STJ 620	29
STJ 619	29
STJ 621	29
STF 161	29
STJ 297	26
STJ 441	24
STJ 439	24
STJ 438	24

Continua na próxima página

Tema	**Quantidade**
STJ 436	24
STJ 440	24
STJ 383	24
STJ 27	24
STJ 28	24
STJ 24	24
STJ 25	24
STJ 35	24
STJ 33	24
STJ 29	24
STJ 31	24
STJ 26	24
STJ 34	24
STJ 36	24
STJ 32	24
STJ 30	24
STJ 134	22
STF 784	21
STJ 606	19
STJ 607	19
STF 96	16
STJ 957	14
STJ 958	14
STJ 575	13
STJ 638	12
STJ 422	12
STJ 423	12
STF 793	11
STJ 534	11
STF 41	10
STJ 106	10
STJ 554	9
STF 69	9
STJ 466	9
STJ 100	9
STJ 682	9

Continua na próxima página

Tema	**Quantidade**
STJ 683	9
STJ 679	9
STJ 681	9
STJ 680	9
STJ 834	9
STJ 684	9
STF 210	8
STJ 699	8
STJ 942	8
STJ 310	8
STJ 311	8
STF 312	7
STJ 234	7
STJ 233	7
STJ 694	7
STJ 532	6
STJ 533	6
STF 76	6
STF 19	6
STF 624	6
STF 33	6
STF 592	6
STJ 259	6
STJ 492	6
STJ 491	6
STJ 735	5
STJ 692	5
STF 499	5
STJ 537	5
STJ 243	5
STJ 660	5
STJ 648	5
STJ 938	5
STJ 102	5
STJ 166	5
STJ 545	5

Continua na próxima página

Tema	**Quantidade**
STJ 686	5
STJ 546	5
STF 3	5
STJ 1007	4
STJ 179	4
STJ 156	4
STJ 108	4
STF 612	4
STJ 560	4
STJ 662	4
STJ 401	4
STF 88	4
STJ 118	4
STJ 972	4
STJ 922	4
STJ 971	4
STJ 314	4
STJ 210	4
STJ 126	4
STJ 211	4
STF 1093	4
STJ 312	3
STF 897	3
STJ 515	3
STF 313	3
STF 1066	3
STJ 642	3
STJ 464	3
STJ 463	3
STF 498	3
STJ 116	3
STF 4	3
STJ 530	3
STJ 553	3
STJ 52	3
STJ 173	3

Continua na próxima página

Tema	**Quantidade**
STJ 416	3
STJ 886	3
STJ 531	3
STJ 564	3
STJ 577	2
STF 666	2
STJ 544	2
STJ 103	2
STJ 104	2
STF 191	2
STF 118	2
STJ 122	2
STF 485	2
STJ 973	2
STJ 197	2
STF 899	2
STJ 465	2
STJ 265	2
STF 1099	2
STJ 766	2
STF 325	2
STF 227	2
STF 82	2
STF 247	2
STF 888	2
STF 439	2
STJ 260	2
STF 540	2
STF 962	2
STJ 504	2
STJ 505	2
STJ 640	2
STJ 290	2
STJ 896	2
STJ 154	2
STJ 153	2

Continua na próxima página

Tema	**Quantidade**
STJ 155	2
STF 264	2
STJ 38	2
STJ 37	2
STJ 41	2
STJ 40	2
STF 530	2
STJ 84	2
STJ 979	2
STF 884	2
STJ 251	2
STJ 254	2
STJ 253	2
STJ 252	2
STJ 576	2
STF 1075	2
STJ 270	2
STJ 269	2
STF 809	2
STJ 367	2
STJ 725	2
STF 471	2
STJ 82	2
STJ 138	2
STJ 137	2
STF 214	2
STF 201	2
STJ 135	2
STF 930	1
STF 916	1
STJ 673	1
STF 43	1
STF 524	1
STF 362	1
STF 130	1
STF 924	1

Continua na próxima página

Tema	**Quantidade**
STF 26	1
STF 1047	1
STF 835	1
STJ 303	1
STJ 304	1
STF 166	1
STJ 250	1
STJ 953	1
STJ 862	1
STJ 96	1
STJ 97	1
STF 142	1
STJ 174	1
STF 25	1
STJ 146	1
STJ 147	1
STF 117	1
STJ 63	1
STJ 885	1
STF 863	1
STJ 715	1
STJ 825	1
STF 808	1
STF 211	1
STF 60	1
STF 89	1
STF 518	1
STJ 362	1
STJ 83	1
STJ 622	1
STJ 345	1
STJ 970	1
STJ 516	1
STF 265	1
STJ 61	1
STJ 163	1

Continua na próxima página

Tema	**Quantidade**
STF 709	1
STJ 889	1
STJ 98	1
STJ 430	1
STF 456	1
STJ 249	1
STJ 668	1
STJ 875	1
STJ 351	1
STJ 890	1
STJ 947	1
STF 220	1
STJ 209	1
STF 16	1
STJ 999	1
STJ 882	1
STJ 23	1
STJ 998	1
STJ 738	1
STJ 479	1
STJ 478	1
STJ 740	1
STJ 739	1
STJ 737	1
STF 20	1
STJ 57	1
STJ 58	1
STJ 459	1
STJ 539	1
STJ 540	1
STF 832	1
STJ 954	1
STJ 315	1
STJ 481	1
STJ 421	1
STF 1024	1

Continua na próxima página

Tema	**Quantidade**
STJ 261	1
STF 483	1
STF 635	1
STF 906	1
STJ 291	1
STJ 292	1
STF 2	1
STJ 95	1
STJ 613	1
STF 517	1
STJ 375	1
STJ 987	1
STJ 64	1
STJ 70	1
STJ 71	1
STJ 68	1
STJ 67	1
STJ 75	1
STJ 66	1
STJ 72	1
STJ 78	1
STJ 65	1
STJ 74	1
STJ 73	1
STJ 69	1
STJ 743	1

ANEXO B – Distribuição Dos Temas para Treino e Teste

Tema	Treino	Teste
TF 810	473	125
STJ 395	176	44
STF 408	148	37
STF 350	137	38
STJ 529	69	18
STJ 905	66	18
STJ 898	63	16
STJ 566	60	16
STJ 570	60	16
STJ 568	60	16
STJ 571	60	16
STJ 569	60	16
STJ 567	60	16
STJ 707	43	12
STJ 626	43	12
STJ 722	33	9
STJ 185	33	9
STJ 777	29	8
STF 27	24	7
STF 161	23	6
STJ 246	19	9
STJ 247	19	9
STJ 441	18	6
STJ 440	18	6
STJ 436	18	6
STJ 438	18	6
STJ 439	18	6
STJ 297	18	5
STJ 618	17	6
STJ 134	17	5

Continua na próxima página

Tema	Treino	Teste
STJ 619	17	6
STJ 383	17	5
STJ 621	17	6
STJ 620	17	6
STF 784	17	4
STJ 606	15	4
STJ 607	15	4
STJ 25	13	4
STJ 27	13	4
STJ 34	13	4
STJ 24	13	4
STJ 35	13	4
STJ 31	13	4
STJ 32	13	4
STJ 33	13	4
STJ 36	13	4
STJ 28	13	4
STJ 26	13	4
STJ 29	13	4
STJ 30	13	4
STJ 957	11	3
STF 96	11	5
STJ 575	10	3
STJ 638	8	2
STF 41	8	2
STJ 106	7	2
STJ 958	7	3
STJ 554	7	2
STJ 683	7	2
STJ 466	7	2
STJ 834	7	2
STJ 682	7	2
STJ 680	7	2
STJ 684	7	2
STJ 679	7	2
STJ 681	7	2

Continua na próxima página

Tema	Treino	Teste
STJ 942	6	1
STF 210	6	2
STJ 100	6	2
STJ 310	6	2
STJ 311	6	2
STJ 534	6	3
STF 793	6	2
STJ 699	6	2
STF 19	5	1
STF 592	5	1
STF 624	5	1
STJ 423	5	1
STJ 422	5	1
STJ 166	4	1
STJ 735	4	1
STJ 545	4	1
STJ 102	4	1
STJ 211	3	1
STJ 210	3	1
STJ 660	3	1
STJ 243	3	1
STF 88	3	1
STJ 560	3	1
STJ 938	3	1
STJ 1007	3	1
STJ 648	3	1
STJ 179	3	1
STJ 532	3	2
STJ 401	3	1
STJ 533	3	2
STJ 922	3	1
STJ 692	3	1
STJ 314	3	1
STJ 126	3	1
STJ 259	3	3
STF 312	3	2

Continua na próxima página

Tema	Treino	Teste
STJ 662	3	1
STJ 686	3	1
STJ 971	2	1
STF 76	2	1
STF 69	2	1
STJ 515	2	1
STF 499	2	1
STJ 312	2	1
STJ 416	2	1
STF 1066	2	1
STJ 156	2	1
STF 33	2	2
STJ 886	2	1
STJ 531	2	1
STF 612	2	1
STJ 537	2	2
STF 1093	2	2
STF 3	2	2
STJ 465	1	1
STF 888	1	1
STJ 197	1	1
STJ 577	1	1
STJ 270	1	1
STJ 269	1	1
STJ 118	1	1
STJ 104	1	1
STJ 103	1	1
STJ 463	1	1
STJ 108	1	1
STJ 694	1	1
STJ 553	1	1
STJ 766	1	1
STF 201	1	1
STJ 972	1	1
STJ 173	1	1
STF 498	1	1

Continua na próxima página

Tema	Treino	Teste
STF 809	1	1
STJ 464	1	1
STF 439	1	1
STJ 492	1	1
STJ 491	1	1
STJ 564	1	1
STF 897	1	1
STJ 725	1	1
STJ 546	1	1
STJ 116	1	1
STJ 135	1	1
STJ 122	1	1
STJ 233	1	1
STJ 234	1	1
STJ 82	1	1
STJ 642	1	1
STF 485	1	1
STF 471	1	1
STJ 367	1	1
STF 1099	1	1
STJ 530	1	1
STJ 38	1	1
STJ 37	1	1
STJ 41	1	1
STJ 40	1	1
STF 247	1	1
STF 530	1	1

ANEXO C – Relatório de Classificação dos algoritmos

Tabela 9 – Relatório de Classificação do Algoritmo "Mais Próximo"

Tema	Precisão	*Recall*	F1-score	Suporte
STF 810	0.930	0.952	0.941	125
STJ 395	1.000	1.000	1.000	44
STF 350	0.778	0.737	0.757	38
STF 408	1.000	1.000	1.000	37
STJ 905	1.000	0.722	0.839	18
STJ 529	0.667	0.778	0.718	18
STJ 566	1.000	1.000	1.000	16
STJ 567	1.000	1.000	1.000	16
STJ 568	1.000	1.000	1.000	16
STJ 569	1.000	1.000	1.000	16
STJ 570	1.000	1.000	1.000	16
STJ 571	1.000	1.000	1.000	16
STJ 898	1.000	1.000	1.000	16
STJ 707	0.923	1.000	0.960	12
STJ 626	0.846	0.917	0.880	12
STJ 185	0.900	1.000	0.947	9
STJ 722	0.900	1.000	0.947	9
STJ 246	0.818	1.000	0.900	9
STJ 247	0.818	1.000	0.900	9
STJ 777	1.000	1.000	1.000	8
STF 27	1.000	0.857	0.923	7
STJ 436	1.000	0.833	0.909	6
STJ 438	1.000	0.833	0.909	6
STJ 439	1.000	0.833	0.909	6
STJ 440	1.000	0.833	0.909	6
STJ 441	1.000	0.833	0.909	6
STJ 618	0.714	0.833	0.769	6
STJ 619	0.714	0.833	0.769	6

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 620	0.714	0.833	0.769	6
STJ 621	0.714	0.833	0.769	6
STF 161	1.000	0.500	0.667	6
STF 96	1.000	1.000	1.000	5
STJ 134	1.000	1.000	1.000	5
STJ 383	1.000	1.000	1.000	5
STJ 297	0.714	1.000	0.833	5
STJ 24	0.571	1.000	0.727	4
STJ 25	0.571	1.000	0.727	4
STJ 26	0.571	1.000	0.727	4
STJ 27	0.571	1.000	0.727	4
STJ 28	0.571	1.000	0.727	4
STJ 29	0.571	1.000	0.727	4
STJ 30	0.571	1.000	0.727	4
STJ 31	0.571	1.000	0.727	4
STJ 32	0.571	1.000	0.727	4
STJ 33	0.571	1.000	0.727	4
STJ 34	0.571	1.000	0.727	4
STJ 35	0.571	1.000	0.727	4
STJ 36	0.571	1.000	0.727	4
STF 784	0.500	0.750	0.600	4
STJ 606	0.429	0.750	0.545	4
STJ 607	0.429	0.750	0.545	4
STJ 259	1.000	1.000	1.000	3
STJ 575	1.000	1.000	1.000	3
STJ 957	1.000	1.000	1.000	3
STJ 958	0.600	1.000	0.750	3
STJ 534	0.667	0.667	0.667	3
STF 1093	1.000	1.000	1.000	2
STF 210	1.000	1.000	1.000	2
STF 312	1.000	1.000	1.000	2
STF 3	1.000	1.000	1.000	2
STJ 100	1.000	1.000	1.000	2
STJ 106	1.000	1.000	1.000	2
STJ 310	1.000	1.000	1.000	2
STJ 311	1.000	1.000	1.000	2

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 466	1.000	1.000	1.000	2
STJ 532	1.000	1.000	1.000	2
STJ 533	1.000	1.000	1.000	2
STJ 537	1.000	1.000	1.000	2
STJ 554	1.000	1.000	1.000	2
STJ 679	1.000	1.000	1.000	2
STJ 680	1.000	1.000	1.000	2
STJ 681	1.000	1.000	1.000	2
STJ 682	1.000	1.000	1.000	2
STJ 683	1.000	1.000	1.000	2
STJ 684	1.000	1.000	1.000	2
STJ 699	1.000	1.000	1.000	2
STJ 834	1.000	1.000	1.000	2
STF 41	1.000	0.500	0.667	2
STJ 638	1.000	0.500	0.667	2
STF 793	0.500	0.500	0.500	2
STF 33	0.000	0.000	0.000	2
STF 1066	1.000	1.000	1.000	1
STF 19	1.000	1.000	1.000	1
STF 201	1.000	1.000	1.000	1
STF 471	1.000	1.000	1.000	1
STF 592	1.000	1.000	1.000	1
STF 612	1.000	1.000	1.000	1
STF 624	1.000	1.000	1.000	1
STF 76	1.000	1.000	1.000	1
STJ 103	1.000	1.000	1.000	1
STJ 104	1.000	1.000	1.000	1
STJ 108	1.000	1.000	1.000	1
STJ 126	1.000	1.000	1.000	1
STJ 166	1.000	1.000	1.000	1
STJ 173	1.000	1.000	1.000	1
STJ 179	1.000	1.000	1.000	1
STJ 197	1.000	1.000	1.000	1
STJ 210	1.000	1.000	1.000	1
STJ 211	1.000	1.000	1.000	1
STJ 243	1.000	1.000	1.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 269	1.000	1.000	1.000	1
STJ 270	1.000	1.000	1.000	1
STJ 314	1.000	1.000	1.000	1
STJ 367	1.000	1.000	1.000	1
STJ 401	1.000	1.000	1.000	1
STJ 463	1.000	1.000	1.000	1
STJ 464	1.000	1.000	1.000	1
STJ 465	1.000	1.000	1.000	1
STJ 515	1.000	1.000	1.000	1
STJ 545	1.000	1.000	1.000	1
STJ 546	1.000	1.000	1.000	1
STJ 560	1.000	1.000	1.000	1
STJ 662	1.000	1.000	1.000	1
STJ 725	1.000	1.000	1.000	1
STJ 735	1.000	1.000	1.000	1
STJ 766	1.000	1.000	1.000	1
STJ 1007	0.500	1.000	0.667	1
STJ 416	0.500	1.000	0.667	1
STJ 922	0.500	1.000	0.667	1
STJ 971	0.500	1.000	0.667	1
STJ 82	0.250	1.000	0.400	1
STF 1099	0.000	0.000	0.000	1
STF 247	0.000	0.000	0.000	1
STF 439	0.000	0.000	0.000	1
STF 485	0.000	0.000	0.000	1
STF 498	0.000	0.000	0.000	1
STF 499	0.000	0.000	0.000	1
STF 530	0.000	0.000	0.000	1
STF 69	0.000	0.000	0.000	1
STF 809	0.000	0.000	0.000	1
STF 888	0.000	0.000	0.000	1
STF 88	0.000	0.000	0.000	1
STF 897	0.000	0.000	0.000	1
STJ 102	0.000	0.000	0.000	1
STJ 116	0.000	0.000	0.000	1
STJ 118	0.000	0.000	0.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 122	0.000	0.000	0.000	1
STJ 135	0.000	0.000	0.000	1
STJ 156	0.000	0.000	0.000	1
STJ 233	0.000	0.000	0.000	1
STJ 234	0.000	0.000	0.000	1
STJ 312	0.000	0.000	0.000	1
STJ 37	0.000	0.000	0.000	1
STJ 38	0.000	0.000	0.000	1
STJ 40	0.000	0.000	0.000	1
STJ 41	0.000	0.000	0.000	1
STJ 422	0.000	0.000	0.000	1
STJ 423	0.000	0.000	0.000	1
STJ 491	0.000	0.000	0.000	1
STJ 492	0.000	0.000	0.000	1
STJ 530	0.000	0.000	0.000	1
STJ 531	0.000	0.000	0.000	1
STJ 553	0.000	0.000	0.000	1
STJ 564	0.000	0.000	0.000	1
STJ 577	0.000	0.000	0.000	1
STJ 642	0.000	0.000	0.000	1
STJ 648	0.000	0.000	0.000	1
STJ 660	0.000	0.000	0.000	1
STJ 686	0.000	0.000	0.000	1
STJ 692	0.000	0.000	0.000	1
STJ 694	0.000	0.000	0.000	1
STJ 886	0.000	0.000	0.000	1
STJ 938	0.000	0.000	0.000	1
STJ 942	0.000	0.000	0.000	1
STJ 972	0.000	0.000	0.000	1
Média Micro	0.860	0.879	0.869	760
Média Macro	0.642	0.693	0.656	760
Média Ponderada	0.842	0.879	0.852	760

Tabela 10 – Relatório de Classificação do Algoritmo "Maioria"

Tema	Precisão	*Recall*	F1-score	Suporte
STF 810	0.929	0.832	0.878	125
STJ 395	1.000	0.977	0.989	44
STF 350	0.913	0.553	0.689	38
STF 408	0.860	1.000	0.925	37
STJ 905	0.900	0.500	0.643	18
STJ 529	1.000	0.389	0.560	18
STJ 566	1.000	1.000	1.000	16
STJ 567	1.000	1.000	1.000	16
STJ 568	1.000	1.000	1.000	16
STJ 569	1.000	1.000	1.000	16
STJ 570	1.000	1.000	1.000	16
STJ 571	1.000	1.000	1.000	16
STJ 898	1.000	0.938	0.968	16
STJ 707	1.000	1.000	1.000	12
STJ 626	1.000	0.583	0.737	12
STJ 722	1.000	0.778	0.875	9
STJ 185	1.000	0.556	0.714	9
STJ 246	1.000	0.444	0.615	9
STJ 247	1.000	0.444	0.615	9
STJ 777	1.000	1.000	1.000	8
STF 27	0.833	0.714	0.769	7
STJ 618	1.000	0.500	0.667	6
STJ 619	1.000	0.500	0.667	6
STJ 620	1.000	0.500	0.667	6
STJ 621	1.000	0.500	0.667	6
STJ 436	1.000	0.333	0.500	6
STJ 438	1.000	0.333	0.500	6
STJ 439	1.000	0.333	0.500	6
STJ 440	1.000	0.333	0.500	6
STJ 441	1.000	0.333	0.500	6
STF 161	0.000	0.000	0.000	6
STJ 134	1.000	1.000	1.000	5
STJ 297	1.000	1.000	1.000	5
STJ 383	1.000	1.000	1.000	5

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STF 96	0.000	0.000	0.000	5
STJ 24	1.000	0.500	0.667	4
STJ 25	1.000	0.500	0.667	4
STJ 26	1.000	0.500	0.667	4
STJ 27	1.000	0.500	0.667	4
STJ 28	1.000	0.500	0.667	4
STJ 29	1.000	0.500	0.667	4
STJ 30	1.000	0.500	0.667	4
STJ 31	1.000	0.500	0.667	4
STJ 32	1.000	0.500	0.667	4
STJ 33	1.000	0.500	0.667	4
STJ 34	1.000	0.500	0.667	4
STJ 35	1.000	0.500	0.667	4
STJ 36	1.000	0.500	0.667	4
STJ 606	1.000	0.500	0.667	4
STJ 607	1.000	0.500	0.667	4
STF 784	0.000	0.000	0.000	4
STJ 575	1.000	1.000	1.000	3
STJ 958	1.000	0.667	0.800	3
STJ 259	1.000	0.333	0.500	3
STJ 534	0.000	0.000	0.000	3
STJ 957	0.000	0.000	0.000	3
STF 1093	1.000	1.000	1.000	2
STF 3	1.000	1.000	1.000	2
STJ 100	1.000	1.000	1.000	2
STJ 532	1.000	1.000	1.000	2
STJ 533	1.000	1.000	1.000	2
STJ 537	1.000	1.000	1.000	2
STJ 699	1.000	1.000	1.000	2
STF 312	1.000	0.500	0.667	2
STF 793	1.000	0.500	0.667	2
STJ 106	1.000	0.500	0.667	2
STJ 310	1.000	0.500	0.667	2
STJ 311	1.000	0.500	0.667	2
STJ 466	1.000	0.500	0.667	2
STJ 554	1.000	0.500	0.667	2

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 638	0.333	0.500	0.400	2
STF 210	0.000	0.000	0.000	2
STF 33	0.000	0.000	0.000	2
STF 41	0.000	0.000	0.000	2
STJ 679	0.000	0.000	0.000	2
STJ 680	0.000	0.000	0.000	2
STJ 681	0.000	0.000	0.000	2
STJ 682	0.000	0.000	0.000	2
STJ 683	0.000	0.000	0.000	2
STJ 684	0.000	0.000	0.000	2
STJ 834	0.000	0.000	0.000	2
STF 1066	1.000	1.000	1.000	1
STF 19	1.000	1.000	1.000	1
STF 471	1.000	1.000	1.000	1
STF 612	1.000	1.000	1.000	1
STF 624	1.000	1.000	1.000	1
STJ 103	1.000	1.000	1.000	1
STJ 104	1.000	1.000	1.000	1
STJ 108	1.000	1.000	1.000	1
STJ 126	1.000	1.000	1.000	1
STJ 166	1.000	1.000	1.000	1
STJ 173	1.000	1.000	1.000	1
STJ 197	1.000	1.000	1.000	1
STJ 210	1.000	1.000	1.000	1
STJ 211	1.000	1.000	1.000	1
STJ 243	1.000	1.000	1.000	1
STJ 269	1.000	1.000	1.000	1
STJ 270	1.000	1.000	1.000	1
STJ 314	1.000	1.000	1.000	1
STJ 367	1.000	1.000	1.000	1
STJ 401	1.000	1.000	1.000	1
STJ 545	1.000	1.000	1.000	1
STJ 546	1.000	1.000	1.000	1
STJ 560	1.000	1.000	1.000	1
STJ 662	1.000	1.000	1.000	1
STJ 735	1.000	1.000	1.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 766	1.000	1.000	1.000	1
STF 1099	0.000	0.000	0.000	1
STF 201	0.000	0.000	0.000	1
STF 247	0.000	0.000	0.000	1
STF 439	0.000	0.000	0.000	1
STF 485	0.000	0.000	0.000	1
STF 498	0.000	0.000	0.000	1
STF 499	0.000	0.000	0.000	1
STF 530	0.000	0.000	0.000	1
STF 592	0.000	0.000	0.000	1
STF 69	0.000	0.000	0.000	1
STF 76	0.000	0.000	0.000	1
STF 809	0.000	0.000	0.000	1
STF 888	0.000	0.000	0.000	1
STF 88	0.000	0.000	0.000	1
STF 897	0.000	0.000	0.000	1
STJ 1007	0.000	0.000	0.000	1
STJ 102	0.000	0.000	0.000	1
STJ 116	0.000	0.000	0.000	1
STJ 118	0.000	0.000	0.000	1
STJ 122	0.000	0.000	0.000	1
STJ 135	0.000	0.000	0.000	1
STJ 156	0.000	0.000	0.000	1
STJ 179	0.000	0.000	0.000	1
STJ 233	0.000	0.000	0.000	1
STJ 234	0.000	0.000	0.000	1
STJ 312	0.000	0.000	0.000	1
STJ 37	0.000	0.000	0.000	1
STJ 38	0.000	0.000	0.000	1
STJ 40	0.000	0.000	0.000	1
STJ 416	0.000	0.000	0.000	1
STJ 41	0.000	0.000	0.000	1
STJ 422	0.000	0.000	0.000	1
STJ 423	0.000	0.000	0.000	1
STJ 463	0.000	0.000	0.000	1
STJ 464	0.000	0.000	0.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 465	0.000	0.000	0.000	1
STJ 491	0.000	0.000	0.000	1
STJ 492	0.000	0.000	0.000	1
STJ 515	0.000	0.000	0.000	1
STJ 530	0.000	0.000	0.000	1
STJ 531	0.000	0.000	0.000	1
STJ 553	0.000	0.000	0.000	1
STJ 564	0.000	0.000	0.000	1
STJ 577	0.000	0.000	0.000	1
STJ 642	0.000	0.000	0.000	1
STJ 648	0.000	0.000	0.000	1
STJ 660	0.000	0.000	0.000	1
STJ 686	0.000	0.000	0.000	1
STJ 692	0.000	0.000	0.000	1
STJ 694	0.000	0.000	0.000	1
STJ 725	0.000	0.000	0.000	1
STJ 82	0.000	0.000	0.000	1
STJ 886	0.000	0.000	0.000	1
STJ 922	0.000	0.000	0.000	1
STJ 938	0.000	0.000	0.000	1
STJ 942	0.000	0.000	0.000	1
STJ 971	0.000	0.000	0.000	1
STJ 972	0.000	0.000	0.000	1
Média Macro	0.550	0.423	0.464	760
Média Micro	0.962	0.664	0.786	760
Média Ponderada	0.841	0.664	0.723	760

Tabela 11 – Relatório de Classificação do Algoritmo "KNN-Mútuo"

Tema	Precisão	*Recall*	F1-score	Suporte
STF 810	0.875	0.784	0.827	125
STJ 395	1.000	0.750	0.857	44
STF 350	0.683	0.737	0.709	38
STF 408	1.000	0.730	0.844	37
STJ 905	0.722	0.722	0.722	18

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 529	0.667	0.667	0.667	18
STJ 898	1.000	0.812	0.897	16
STJ 566	1.000	0.562	0.720	16
STJ 567	1.000	0.562	0.720	16
STJ 568	1.000	0.562	0.720	16
STJ 569	1.000	0.562	0.720	16
STJ 570	1.000	0.562	0.720	16
STJ 571	1.000	0.562	0.720	16
STJ 707	1.000	0.917	0.957	12
STJ 626	0.909	0.833	0.870	12
STJ 185	1.000	0.889	0.941	9
STJ 722	0.889	0.889	0.889	9
STJ 246	1.000	0.556	0.714	9
STJ 247	1.000	0.556	0.714	9
STJ 777	1.000	0.875	0.933	8
STF 27	1.000	0.714	0.833	7
STF 161	1.000	1.000	1.000	6
STJ 436	1.000	0.833	0.909	6
STJ 438	1.000	0.833	0.909	6
STJ 439	1.000	0.833	0.909	6
STJ 440	1.000	0.833	0.909	6
STJ 441	1.000	0.833	0.909	6
STJ 618	0.750	0.500	0.600	6
STJ 619	0.750	0.500	0.600	6
STJ 620	0.750	0.500	0.600	6
STJ 621	0.750	0.500	0.600	6
STF 96	1.000	1.000	1.000	5
STJ 134	1.000	0.800	0.889	5
STJ 383	1.000	0.800	0.889	5
STJ 297	1.000	0.600	0.750	5
STJ 24	1.000	0.500	0.667	4
STJ 25	1.000	0.500	0.667	4
STJ 26	1.000	0.500	0.667	4
STJ 27	1.000	0.500	0.667	4
STJ 28	1.000	0.500	0.667	4
STJ 29	1.000	0.500	0.667	4

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 30	1.000	0.500	0.667	4
STJ 31	1.000	0.500	0.667	4
STJ 32	1.000	0.500	0.667	4
STJ 33	1.000	0.500	0.667	4
STJ 34	1.000	0.500	0.667	4
STJ 35	1.000	0.500	0.667	4
STJ 36	1.000	0.500	0.667	4
STJ 606	0.429	0.750	0.545	4
STJ 607	0.429	0.750	0.545	4
STF 784	0.200	0.250	0.222	4
STJ 259	1.000	1.000	1.000	3
STJ 575	1.000	1.000	1.000	3
STJ 957	1.000	1.000	1.000	3
STJ 958	1.000	0.667	0.800	3
STJ 534	0.667	0.667	0.667	3
STF 1093	1.000	1.000	1.000	2
STF 210	1.000	1.000	1.000	2
STF 312	1.000	1.000	1.000	2
STF 3	1.000	1.000	1.000	2
STJ 100	1.000	1.000	1.000	2
STJ 466	1.000	1.000	1.000	2
STJ 532	1.000	1.000	1.000	2
STJ 533	1.000	1.000	1.000	2
STJ 537	1.000	1.000	1.000	2
STJ 554	1.000	1.000	1.000	2
STJ 679	1.000	1.000	1.000	2
STJ 680	1.000	1.000	1.000	2
STJ 681	1.000	1.000	1.000	2
STJ 682	1.000	1.000	1.000	2
STJ 683	1.000	1.000	1.000	2
STJ 684	1.000	1.000	1.000	2
STJ 834	1.000	1.000	1.000	2
STJ 106	0.667	1.000	0.800	2
STJ 310	0.667	1.000	0.800	2
STJ 311	0.667	1.000	0.800	2
STJ 638	1.000	0.500	0.667	2

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 699	1.000	0.500	0.667	2
STF 41	0.400	1.000	0.571	2
STF 793	0.333	0.500	0.400	2
STF 33	0.000	0.000	0.000	2
STF 1066	1.000	1.000	1.000	1
STF 592	1.000	1.000	1.000	1
STF 612	1.000	1.000	1.000	1
STF 69	1.000	1.000	1.000	1
STF 76	1.000	1.000	1.000	1
STF 888	1.000	1.000	1.000	1
STJ 102	1.000	1.000	1.000	1
STJ 103	1.000	1.000	1.000	1
STJ 104	1.000	1.000	1.000	1
STJ 108	1.000	1.000	1.000	1
STJ 122	1.000	1.000	1.000	1
STJ 126	1.000	1.000	1.000	1
STJ 135	1.000	1.000	1.000	1
STJ 166	1.000	1.000	1.000	1
STJ 197	1.000	1.000	1.000	1
STJ 210	1.000	1.000	1.000	1
STJ 211	1.000	1.000	1.000	1
STJ 243	1.000	1.000	1.000	1
STJ 269	1.000	1.000	1.000	1
STJ 270	1.000	1.000	1.000	1
STJ 314	1.000	1.000	1.000	1
STJ 401	1.000	1.000	1.000	1
STJ 530	1.000	1.000	1.000	1
STJ 545	1.000	1.000	1.000	1
STJ 546	1.000	1.000	1.000	1
STJ 553	1.000	1.000	1.000	1
STJ 560	1.000	1.000	1.000	1
STJ 642	1.000	1.000	1.000	1
STJ 662	1.000	1.000	1.000	1
STJ 766	1.000	1.000	1.000	1
STJ 886	1.000	1.000	1.000	1
STF 499	0.500	1.000	0.667	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 118	0.500	1.000	0.667	1
STJ 173	0.500	1.000	0.667	1
STJ 416	0.500	1.000	0.667	1
STJ 515	0.500	1.000	0.667	1
STJ 577	0.500	1.000	0.667	1
STJ 735	0.500	1.000	0.667	1
STJ 82	0.500	1.000	0.667	1
STJ 922	0.500	1.000	0.667	1
STF 1099	0.333	1.000	0.500	1
STF 201	0.333	1.000	0.500	1
STJ 367	0.333	1.000	0.500	1
STJ 463	0.333	1.000	0.500	1
STJ 464	0.333	1.000	0.500	1
STJ 465	0.333	1.000	0.500	1
STJ 725	0.333	1.000	0.500	1
STJ 938	0.333	1.000	0.500	1
STJ 971	0.333	1.000	0.500	1
STF 19	0.000	0.000	0.000	1
STF 247	0.000	0.000	0.000	1
STF 439	0.000	0.000	0.000	1
STF 471	0.000	0.000	0.000	1
STF 485	0.000	0.000	0.000	1
STF 498	0.000	0.000	0.000	1
STF 530	0.000	0.000	0.000	1
STF 624	0.000	0.000	0.000	1
STF 809	0.000	0.000	0.000	1
STF 88	0.000	0.000	0.000	1
STF 897	0.000	0.000	0.000	1
STJ 1007	0.000	0.000	0.000	1
STJ 116	0.000	0.000	0.000	1
STJ 156	0.000	0.000	0.000	1
STJ 179	0.000	0.000	0.000	1
STJ 233	0.000	0.000	0.000	1
STJ 234	0.000	0.000	0.000	1
STJ 312	0.000	0.000	0.000	1
STJ 37	0.000	0.000	0.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 38	0.000	0.000	0.000	1
STJ 40	0.000	0.000	0.000	1
STJ 41	0.000	0.000	0.000	1
STJ 422	0.000	0.000	0.000	1
STJ 423	0.000	0.000	0.000	1
STJ 491	0.000	0.000	0.000	1
STJ 492	0.000	0.000	0.000	1
STJ 531	0.000	0.000	0.000	1
STJ 564	0.000	0.000	0.000	1
STJ 648	0.000	0.000	0.000	1
STJ 660	0.000	0.000	0.000	1
STJ 686	0.000	0.000	0.000	1
STJ 692	0.000	0.000	0.000	1
STJ 694	0.000	0.000	0.000	1
STJ 942	0.000	0.000	0.000	1
STJ 972	0.000	0.000	0.000	1
Média Macro	0.677	0.662	0.641	760
Média Micro	0.838	0.705	0.766	760
Média Ponderada	0.858	0.705	0.758	760

Tabela 12 – Relatório de Classificação do Algoritmo "KNN-Mútuo com Limiar"

Tema	Precisão	*Recall*	F1-score	Suporte
STF 810	0.838	0.872	0.855	125
STJ 395	1.000	0.886	0.940	44
STF 350	0.558	0.763	0.644	38
STF 408	1.000	0.892	0.943	37
STJ 905	0.682	0.833	0.750	18
STJ 529	0.591	0.722	0.650	18
STJ 898	1.000	0.938	0.968	16
STJ 566	1.000	0.875	0.933	16
STJ 567	1.000	0.875	0.933	16
STJ 568	1.000	0.875	0.933	16
STJ 569	1.000	0.875	0.933	16
STJ 570	1.000	0.875	0.933	16

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 571	1.000	0.875	0.933	16
STJ 707	1.000	0.917	0.957	12
STJ 626	0.846	0.917	0.880	12
STJ 185	1.000	0.889	0.941	9
STJ 722	1.000	0.889	0.941	9
STJ 246	0.750	0.667	0.706	9
STJ 247	0.750	0.667	0.706	9
STJ 777	1.000	1.000	1.000	8
STF 27	0.545	0.857	0.667	7
STJ 436	1.000	0.833	0.909	6
STJ 438	1.000	0.833	0.909	6
STJ 439	1.000	0.833	0.909	6
STJ 440	1.000	0.833	0.909	6
STJ 441	1.000	0.833	0.909	6
STF 161	0.800	0.667	0.727	6
STJ 618	0.500	0.833	0.625	6
STJ 619	0.500	0.833	0.625	6
STJ 620	0.500	0.833	0.625	6
STJ 621	0.500	0.833	0.625	6
STF 96	1.000	1.000	1.000	5
STJ 134	1.000	0.800	0.889	5
STJ 297	1.000	0.800	0.889	5
STJ 383	1.000	0.800	0.889	5
STJ 606	0.429	0.750	0.545	4
STJ 607	0.429	0.750	0.545	4
STJ 24	0.500	0.500	0.500	4
STJ 25	0.500	0.500	0.500	4
STJ 26	0.500	0.500	0.500	4
STJ 27	0.500	0.500	0.500	4
STJ 28	0.500	0.500	0.500	4
STJ 29	0.500	0.500	0.500	4
STJ 30	0.500	0.500	0.500	4
STJ 31	0.500	0.500	0.500	4
STJ 32	0.500	0.500	0.500	4
STJ 33	0.500	0.500	0.500	4
STJ 34	0.500	0.500	0.500	4

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 35	0.500	0.500	0.500	4
STJ 36	0.500	0.500	0.500	4
STF 784	0.333	0.500	0.400	4
STJ 259	1.000	1.000	1.000	3
STJ 575	1.000	1.000	1.000	3
STJ 957	1.000	1.000	1.000	3
STJ 534	0.750	1.000	0.857	3
STJ 958	0.667	0.667	0.667	3
STF 1093	1.000	1.000	1.000	2
STF 210	1.000	1.000	1.000	2
STF 3	1.000	1.000	1.000	2
STJ 100	1.000	1.000	1.000	2
STJ 466	1.000	1.000	1.000	2
STJ 532	1.000	1.000	1.000	2
STJ 533	1.000	1.000	1.000	2
STJ 537	1.000	1.000	1.000	2
STJ 554	1.000	1.000	1.000	2
STJ 679	1.000	1.000	1.000	2
STJ 680	1.000	1.000	1.000	2
STJ 681	1.000	1.000	1.000	2
STJ 682	1.000	1.000	1.000	2
STJ 683	1.000	1.000	1.000	2
STJ 684	1.000	1.000	1.000	2
STJ 699	1.000	1.000	1.000	2
STJ 834	1.000	1.000	1.000	2
STF 312	0.667	1.000	0.800	2
STJ 310	0.667	1.000	0.800	2
STJ 311	0.667	1.000	0.800	2
STJ 638	1.000	0.500	0.667	2
STJ 106	0.400	1.000	0.571	2
STF 41	0.500	0.500	0.500	2
STF 793	0.333	0.500	0.400	2
STF 33	0.000	0.000	0.000	2
STF 1066	1.000	1.000	1.000	1
STF 19	1.000	1.000	1.000	1
STF 201	1.000	1.000	1.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STF 612	1.000	1.000	1.000	1
STF 624	1.000	1.000	1.000	1
STF 76	1.000	1.000	1.000	1
STJ 1007	1.000	1.000	1.000	1
STJ 103	1.000	1.000	1.000	1
STJ 104	1.000	1.000	1.000	1
STJ 108	1.000	1.000	1.000	1
STJ 126	1.000	1.000	1.000	1
STJ 166	1.000	1.000	1.000	1
STJ 197	1.000	1.000	1.000	1
STJ 210	1.000	1.000	1.000	1
STJ 211	1.000	1.000	1.000	1
STJ 243	1.000	1.000	1.000	1
STJ 269	1.000	1.000	1.000	1
STJ 270	1.000	1.000	1.000	1
STJ 314	1.000	1.000	1.000	1
STJ 367	1.000	1.000	1.000	1
STJ 401	1.000	1.000	1.000	1
STJ 515	1.000	1.000	1.000	1
STJ 545	1.000	1.000	1.000	1
STJ 546	1.000	1.000	1.000	1
STJ 560	1.000	1.000	1.000	1
STJ 642	1.000	1.000	1.000	1
STJ 662	1.000	1.000	1.000	1
STJ 735	1.000	1.000	1.000	1
STJ 118	0.500	1.000	0.667	1
STJ 173	0.500	1.000	0.667	1
STJ 416	0.500	1.000	0.667	1
STJ 463	0.500	1.000	0.667	1
STJ 464	0.500	1.000	0.667	1
STJ 465	0.500	1.000	0.667	1
STJ 694	0.500	1.000	0.667	1
STJ 725	0.500	1.000	0.667	1
STJ 922	0.500	1.000	0.667	1
STF 1099	0.333	1.000	0.500	1
STF 592	0.333	1.000	0.500	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 233	0.333	1.000	0.500	1
STJ 234	0.333	1.000	0.500	1
STJ 577	0.333	1.000	0.500	1
STJ 686	0.333	1.000	0.500	1
STJ 766	0.333	1.000	0.500	1
STJ 938	0.333	1.000	0.500	1
STJ 971	0.333	1.000	0.500	1
STJ 82	0.200	1.000	0.333	1
STF 247	0.000	0.000	0.000	1
STF 439	0.000	0.000	0.000	1
STF 471	0.000	0.000	0.000	1
STF 485	0.000	0.000	0.000	1
STF 498	0.000	0.000	0.000	1
STF 499	0.000	0.000	0.000	1
STF 530	0.000	0.000	0.000	1
STF 69	0.000	0.000	0.000	1
STF 809	0.000	0.000	0.000	1
STF 888	0.000	0.000	0.000	1
STF 88	0.000	0.000	0.000	1
STF 897	0.000	0.000	0.000	1
STJ 102	0.000	0.000	0.000	1
STJ 116	0.000	0.000	0.000	1
STJ 122	0.000	0.000	0.000	1
STJ 135	0.000	0.000	0.000	1
STJ 156	0.000	0.000	0.000	1
STJ 179	0.000	0.000	0.000	1
STJ 312	0.000	0.000	0.000	1
STJ 37	0.000	0.000	0.000	1
STJ 38	0.000	0.000	0.000	1
STJ 40	0.000	0.000	0.000	1
STJ 41	0.000	0.000	0.000	1
STJ 422	0.000	0.000	0.000	1
STJ 423	0.000	0.000	0.000	1
STJ 491	0.000	0.000	0.000	1
STJ 492	0.000	0.000	0.000	1
STJ 530	0.000	0.000	0.000	1

Continua na próxima página

Tema	Precisão	*Recall*	F1-score	Suporte
STJ 531	0.000	0.000	0.000	1
STJ 553	0.000	0.000	0.000	1
STJ 564	0.000	0.000	0.000	1
STJ 648	0.000	0.000	0.000	1
STJ 660	0.000	0.000	0.000	1
STJ 692	0.000	0.000	0.000	1
STJ 886	0.000	0.000	0.000	1
STJ 942	0.000	0.000	0.000	1
STJ 972	0.000	0.000	0.000	1
Média Macro	0.602	0.680	0.621	760
Média Micro	0.762	0.799	0.780	760
Média Ponderada	0.785	0.799	0.782	760

ANEXO D – Distribuição dos temas propagados

Tabela 13 – Distribuição dos Temas Propagados

Tema	**Quantidade**
STF 810	3258
STJ 568	2053
STJ 571	2053
STJ 567	2053
STJ 566	2053
STJ 569	2053
STJ 570	2053
STF 350	1881
STJ 116	1118
STF 210	837
STJ 100	743
STJ 466	630
STJ 905	559
STJ 722	539
STJ 942	490
STJ 898	469
STJ 882	418
STJ 395	354
STF 408	342
STJ 464	305
STJ 463	305
STF 809	242
STF 498	242
STJ 529	239
STJ 735	202
STJ 699	172
STJ 564	160
STJ 102	151

Continua na próxima página

Tema	**Quantidade**
STJ 922	136
STJ 607	131
STJ 606	131
STJ 707	131
STJ 243	128
STJ 626	116
STF 130	115
STJ 247	111
STJ 246	111
STF 793	103
STJ 260	103
STJ 618	93
STJ 619	93
STJ 620	93
STJ 621	93
STF 592	87
STJ 938	74
STJ 297	73
STJ 662	72
STJ 106	69
STJ 37	68
STJ 40	68
STJ 38	68
STJ 41	68
STF 27	67
STJ 638	62
STJ 314	58
STJ 465	57
STJ 312	55
STJ 958	52
STJ 179	51
STJ 533	50
STJ 532	50
STJ 970	49
STJ 439	48
STJ 441	48

Continua na próxima página

Tema	**Quantidade**
STJ 438	48
STJ 436	48
STJ 440	48
STJ 401	46
STJ 166	45
STF 784	45
STJ 530	44
STJ 28	42
STJ 34	42
STJ 27	42
STJ 33	42
STJ 36	42
STJ 26	42
STJ 25	42
STJ 31	42
STJ 32	42
STJ 30	42
STJ 35	42
STJ 24	42
STJ 29	42
STJ 954	40
STJ 95	39
STF 96	38
STJ 577	36
STJ 197	36
STF 897	35
STJ 660	34
STJ 234	32
STJ 233	32
STJ 1007	29
STJ 185	29
STJ 122	28
STF 161	28
STJ 886	27
STJ 971	27
STJ 576	26

Continua na próxima página

Tema	**Quantidade**
STJ 560	26
STJ 575	25
STF 612	24
STJ 554	23
STJ 686	21
STJ 108	20
STJ 210	20
STJ 211	20
STJ 126	20
STJ 766	19
STJ 972	17
STJ 692	17
STJ 540	16
STJ 539	16
STF 26	15
STF 499	12
STJ 642	12
STJ 896	11
STJ 58	9
STJ 57	9
STJ 825	9
STJ 544	9
STF 313	9
STJ 104	9
STJ 103	9
STJ 545	8
STJ 311	8
STJ 310	8
STF 362	7
STJ 156	7
STJ 777	7
STF 312	7
STJ 648	6
STF 247	6
STJ 290	6
STJ 546	6

Continua na próxima página

Tema	**Quantidade**
STF 916	6
STF 43	6
STJ 673	6
STF 191	6
STJ 534	6
STF 220	5
STJ 423	5
STJ 422	5
STF 41	5
STF 517	5
STJ 98	4
STJ 84	4
STF 471	4
STF 60	4
STJ 249	4
STJ 694	4
STJ 147	3
STJ 146	3
STF 88	3
STJ 416	3
STF 214	3
STJ 725	2
STF 863	2
STJ 23	2
STJ 174	2
STF 19	2
STF 624	2
STJ 351	2
STF 25	2
STF 33	1
STJ 82	1
STJ 953	1
STJ 492	1
STJ 491	1
STF 924	1
STF 69	1

Continua na próxima página

Tema	**Quantidade**
STJ 52	1
STF 89	1
STF 483	1
STJ 134	1
STJ 383	1
STJ 291	1
STJ 292	1
STF 325	1
STJ 459	1
STF 82	1
STJ 889	1