Classificação de Documentos Jurídicos Utilizando IA Generativa: Uma Abordagem com RAG e Gemini

Legal Document Classification Using Generative AI: A Retrieval-Augmented Generation (RAG) and Gemini Approach

Ruan Dias Santana[1] e Marcelo Lisboa Rocha[1,2]

[1] *Universidade Federal do Tocantins, Ciência da Computação, Palmas, Tocantins, Brasil*
[2] *Universidade Federal do Tocantins, Inteligência Artificial, Palmas, Tocantins, Brasil*

Data de recebimento do manuscrito: 25/12/2025
Data de aceitação do manuscrito: 26/04/2026
Data de publicação: 02/05/2026

Resumo— O Poder Judiciário brasileiro enfrenta um desafio crítico relacionado ao volume massivo de processos digitais, tornando a triagem manual onerosa e suscetível a erros. Este trabalho investiga a aplicação de Inteligência Artificial Generativa para automatizar a classificação de petições utilizando Modelos de Linguagem de Grande Escala (LLMs). A pesquisa apresenta uma evolução metodológica em três etapas: (i) uma abordagem inicial baseada em *few-shot learning*, que estabeleceu uma linha de base de 56% de acurácia; (ii) o refinamento através de engenharia de *prompt* com N-grams e técnicas de Aumento de Dados para corrigir o desbalanceamento de classes, chegando a 85% de acurácia; e (iii) a implementação de uma arquitetura de Recuperação Aumentada por Geração (RAG), conectando o modelo Gemini 2.5 da Google a uma base de conhecimento vetorial, com uma acurácia de 84%. Os experimentos utilizaram dados reais do Tribunal de Justiça do Tocantins (TJTO). Os resultados finais demonstram que a abordagem RAG alcançou 84% de acurácia em um cenário complexo de 11 classes, mitigando alucinações e ambiguidades semânticas.

Palavras-chave—IA Jurídica, RAG, LLMs, Classificação de Texto, Gemini, Data Augmentation.

Abstract— *The Brazilian Judiciary faces a critical challenge regarding the massive volume of digital lawsuits, making manual screening costly and error-prone. This work investigates the application of Generative Artificial Intelligence to automate the classification of legal petitions using Large Language Models (LLMs). The research presents a methodological evolution in three stages: (i) an initial approach based on few-shot learning, establishing a 56% accuracy baseline; (ii) refinement through prompt engineering with N-grams and Data Augmentation techniques to address class imbalance, which achieved 85% accuracy; and (iii) the implementation of a Retrieval-Augmented Generation (RAG) architecture, connecting Google's Gemini 2.5 model to a vector knowledge base, which achieved 84% accuracy. The experiments utilized real datasets from the Court of Justice of Tocantins (TJTO), covering themes from Superior Courts (STF/STJ). Final results demonstrate that the RAG approach achieved 84% accuracy in a complex scenario of 11 thematic classes, effectively mitigating hallucinations and semantic ambiguities found in previous stages.*

Keywords—*Legal AI, RAG, LLMs, Text Classification, Gemini, Data Augmentation.*

I. Introdução

A sociedade tem passado por um intenso processo de transformação digital nos últimos anos, e o setor jurídico não é exceção. No Brasil, desde 2006, com a promulgação da Lei n° 11.419, que viabilizou a digitalização dos acervos e fluxos judiciais, os tribunais vêm investindo na modernização de seus sistemas. Com a pandemia da Covid-19, a necessidade de distanciamento social acelerou significativamente esse processo, levando o Conselho Nacional de Justiça (CNJ) a determinar que, a partir de março de 2022, o judiciário brasileiro passasse a receber processos exclusivamente em formato digital.

Essa transição resultou em um volume expressivo de documentos digitais, intensificado pelo crescimento contínuo da demanda judicial. De acordo com o relatório Justiça em Números do Conselho Nacional de Justiça (CNJ) [1], o total de novos casos anuais passou de aproximadamente 27 milhões em 2020 para cerca de 39 milhões em 2024.

Diante desse cenário, torna-se necessária a adoção de mecanismos automatizados de organização. A classificação manual de petições iniciais, etapa em que se define o rito processual e a competência do juízo, é uma atividade

Dados de contato: Ruan Dias Santana, ruan.dias@uft.edu.br

repetitiva, onerosa e suscetível a erros humanos. Classificações incorretas podem resultar em tramitações inadequadas, nulidades processuais e atrasos na prestação jurisdicional.

Tradicionalmente, a automação dessa tarefa tem sido abordada por técnicas clássicas de Processamento de Linguagem Natural (PLN), como *Support Vector Machines* (SVM) ou, mais recentemente, por modelos baseados em Transformers como o BERT [2]. No entanto, essas abordagens possuem uma limitação crítica: a dependência de grandes conjuntos de dados rotulados (*labeled datasets*). No domínio jurídico, a criação desses *datasets* é extremamente custosa, exigindo anotação por juristas especializados, e não por rotuladores genéricos [3].

O surgimento da Inteligência Artificial Generativa e dos Modelos de Linguagem de Grande Escala (LLMs), como a família GPT e Gemini, oferece uma mudança de paradigma. Estes modelos, pré-treinados em vastos *corpora* de texto, possuem capacidades de generalização que permitem realizar tarefas complexas com poucos exemplos (*few-shot learning*), reduzindo a necessidade de dados de treinamento específicos [4].

Contudo, a aplicação direta de LLMs no Direito enfrenta desafios como a "alucinação" (geração de informações falsas plausíveis) e a dificuldade em lidar com a intertextualidade jurídica e nuances de teses específicas de tribunais superiores (STF e STJ).

Este artigo apresenta os resultados de uma investigação sobre o uso do modelo Google Gemini para a classificação de documentos do Tribunal de Justiça do Tocantins (TJTO). O trabalho documenta a evolução metodológica de uma abordagem baseada em *few-shot learning* estático para uma arquitetura sofisticada de Recuperação Aumentada por Geração (RAG). O objetivo central é demonstrar como a integração de LLMs com bases de conhecimento vetoriais pode resolver problemas de ambiguidade semântica e desbalanceamento de classes, oferecendo uma solução escalável e de alta precisão.

II. Fundamentação Teórica

Esta seção estabelece os pilares conceituais que sustentam a metodologia proposta, abordando a arquitetura dos LLMs, as estratégias de adaptação e a técnica de RAG.

a. Modelos de Linguagem de Grande Escala

Os LLMs representam o estado da arte em PLN. Fundamentados na arquitetura Transformer, introduzida por Vaswani et al. [5], esses modelos utilizam mecanismos de auto-atenção (*self-attention*) para ponderar a importância de cada palavra em relação às outras em uma sentença, independentemente da distância entre elas. Isso permite capturar dependências de longo prazo e contextos complexos, essenciais para a interpretação de textos jurídicos longos.

O treinamento de um LLM ocorre em duas fases: o pré-treinamento auto-supervisionado em volumes massivos de dados (onde o modelo aprende a estrutura da linguagem e conhecimento de mundo) e o ajuste fino (*fine-tuning*) para tarefas específicas.

Neste trabalho, utilizou-se a família de modelos Gemini (versões 2.0 e 2.5 Flash-Lite) da Google. A escolha destes modelos foi motivada, primordialmente, pela viabilidade econômica, dada a disponibilidade de acesso gratuito à sua API, fator determinante para a execução dos experimentos com recursos limitados. Secundariamente, a escolha justifica-se por sua arquitetura nativamente multimodal e janela de contexto otimizada, permitindo o processamento eficiente de petições extensas sem os custos associados a modelos proprietários concorrentes, como o GPT-4 [6].

b. Fine-Tuning vs. Few-Shot Learning

A adaptação de LLMs para o domínio jurídico geralmente segue dois caminhos distintos:

1. ***Fine-Tuning*:** Consiste no retreinamento dos pesos da rede neural com um *dataset* específico do domínio. Embora resulte em alta performance, exige milhares de exemplos anotados e alto poder computacional, além de apresentar riscos de "esquecimento catastrófico" do conhecimento prévio.

2. ***Few-Shot Learning* (FSL):** O modelo não sofre atualização de pesos. A tarefa é ensinada através de instruções e poucos exemplos fornecidos no contexto do *prompt* (*in-context learning*).

Dada a escassez de dados rotulados e a necessidade de agilidade na prototipagem, este trabalho optou inicialmente pelo FSL, evoluindo posteriormente para o RAG, que pode ser interpretado como uma versão dinâmica e escalável do aprendizado em contexto.

c. Recuperação Aumentada por Geração (RAG)

Uma das principais limitações dos LLMs é o conhecimento estático (limitado à data de corte do treinamento) e a propensão a alucinações quando confrontados com domínios técnicos específicos. O RAG (*Retrieval-Augmented Generation*) mitiga esses problemas conectando o LLM a uma base de conhecimento externa confiável [7].

O fluxo de funcionamento do RAG neste estudo ocorre em três etapas:

- **Indexação Vetorial:** Os documentos jurídicos de referência são convertidos em vetores numéricos densos (*embeddings*) que capturam seu significado semântico.

- **Recuperação (Retrieval):** Quando uma nova petição chega para classificação, ela é vetorizada, e o sistema busca na base os k documentos mais similares matematicamente (geralmente usando similaridade de cosseno).

- **Geração Aumentada:** O LLM recebe a nova petição juntamente com os documentos recuperados como contexto. O *prompt* instrui o modelo a classificar o novo caso baseando-se nos precedentes fornecidos.

Essa abordagem garante que a decisão do modelo seja fundamentada em dados reais e atualizados, aumentando a explicabilidade e a confiabilidade do sistema.

ISSN: 2675-3588

III. METODOLOGIA

A metodologia foi desenvolvida de forma iterativa e incremental, dividida em três fases experimentais desenhadas para superar as limitações de generalização encontradas a cada etapa.

a. Coleta e Preparação dos Dados

Os dados utilizados foram extraídos do sistema E-Proc do Tribunal de Justiça do Tocantins (TJTO). Todos os documentos passaram por um processo de anonimização (remoção de nomes de partes, advogados e CPFs) para garantir conformidade com a LGPD. A classificação "ground truth" foi realizada manualmente por magistrados e assessores jurídicos.

Foram construídas três versões de *datasets* ao longo do projeto:

- ***Dataset* Base (`dataset-temas-novo.csv`):** Versão utilizada como linha de base , composta por 710 documentos distribuídos em 5 classes: TEMA 864, TEMA 986, TEMA 1118, TEMA 1177 e NENHUM. Este conjunto apresentava severo desbalanceamento, com classes majoritárias contendo cerca de 300 exemplos e as minoritárias (como o TEMA 1118) apenas 21 instâncias.
- ***Dataset* Revisado (`arqtemas-ANTIGO01.csv`):** Versão aprimorada com limpeza textual e normalização semântica. Para os experimentos da Fase 2, este conjunto foi submetido a técnicas de *Data Augmentation* (tradução reversa e substituição de sinônimos), expandindo todas as classes para 300 exemplos cada, resultando em um corpus balanceado de 1.500 instâncias.
- ***Dataset* Expandido (`arqtemas-NOVO01.csv`):** Incorporação de temas inéditos de repercussão geral e recursos repetitivos do STF e STJ (TEMAS 793, 1184, 1199, 566, 1132 e 796). Na etapa final (RAG), os conjuntos foram unificados, totalizando 11 classes distintas para validar a robustez do modelo em um cenário de maior complexidade jurídica.

b. Definição das Classes Temáticas

Cada uma das classes temáticas consideradas neste trabalho corresponde a uma jurisprudência específica consolidada pelos tribunais superiores brasileiros, mais precisamente:

- **Supremo Tribunal Federal (STF):** temas de *Repercussão Geral*, que estabelecem entendimentos vinculantes sobre matérias constitucionais.
- **Superior Tribunal de Justiça (STJ):** temas de *Recursos Repetitivos*, que uniformizam a interpretação da legislação infraconstitucional.

Esses temas funcionam, portanto, como categorias jurídicas normativas, nas quais cada texto de pedido judicial pode ou não se enquadrar conforme seu conteúdo e fundamentação. Essa característica reforça a natureza interpretativa e contextual da tarefa de classificação: não se trata apenas de identificar palavras isoladas, mas de compreender a relação semântica entre o pedido judicial e a tese jurídica correspondente.

Já a classe **NENHUM** representa um agrupamento residual crítico, contendo pedidos que não se enquadram em nenhuma das teses previamente fixadas (falsos positivos potenciais).

c. Engenharia de Prompt e Data Augmentation

Na Fase 2, para combater o desbalanceamento, aplicou-se um pipeline de *Data Augmentation* Híbrido: 1. ***Back-translation*:** Tradução dos textos jurídicos para o inglês e re-tradução para o português utilizando APIs de tradução, gerando variações sintáticas naturais mantendo a semântica. 2. **Substituição de Sinônimos:** Troca de termos técnicos por equivalentes (ex: “veículo” por “automóvel”, “demandante” por “autor”), aumentando a variabilidade vocabular.

Adicionalmente, os *prompts* foram enriquecidos com N-grams. Foram extraídas as sequências de palavras (bigramas e trigramas) mais frequentes de cada classe e inseridas no *prompt* como "dicas" explícitas para o modelo focar em termos determinantes.

d. Arquitetura RAG

A solução definitiva (Fase 3) abandonou o augmentation sintético em favor do RAG.

- ***Embeddings*:** Utilizou-se o modelo `text-embedding-004` da Google para vetorizar 80% do *dataset* unificado.
- **Banco Vetorial:** Os vetores foram indexados no *Pinecone*, otimizado para busca de alta dimensionalidade.
- **Processo de Inferência:** Para cada documento de teste (20% restantes), o sistema recuperava os $k = 5$ documentos mais similares da base de treino. O *prompt* final instruía o Gemini a classificar o documento alvo considerando apenas as evidências presentes nos documentos recuperados.

IV. RESULTADOS E DISCUSSÃO

A avaliação experimental foi conduzida de forma sequencial e incremental. Essa abordagem permitiu isolar a contribuição de cada técnica do *prompting* simples à recuperação de contexto para a eficácia da classificação jurídica.

A seguir, discutem-se os desempenhos quantitativos e as implicações qualitativas de cada fase experimental.

a. Fase 1: Limitações do Few-Shot Estático

Na primeira fase, estabeleceu-se a linha de base utilizando o modelo Gemini 2.0 Flash-Lite com *prompts* estáticos (*Few-Shot*), sem acesso a base de conhecimento externa. O objetivo era verificar a capacidade intrínseca do modelo em generalizar padrões jurídicos apenas com o conhecimento pré-treinado.

Como demonstrado na Tabela 1, o desempenho foi insatisfatório, com acurácia global de apenas 56%. O modelo exibiu um forte enviesamento para a classe majoritária (TEMA 864) e para a classe residual (NENHUM), ignorando as especificidades dos temas tributários complexos.

TABELA 1: RELATÓRIO DE CLASSIFICAÇÃO DETALHADO - FASE 1 (LINHA DE BASE)

Classe	Precisão	Recall	F1-Score	Suporte
NENHUM	0.68	0.48	0.57	270
TEMA 1118	0.03	0.10	0.04	10
TEMA 1177	0.27	0.29	0.28	14
TEMA 864	0.56	0.74	0.64	273
TEMA 986	1.00	0.03	0.05	37
Acurácia			0.56	604
Média Macro	0.57	0.34	0.33	604
Média Ponderada	0.63	0.56	0.56	604

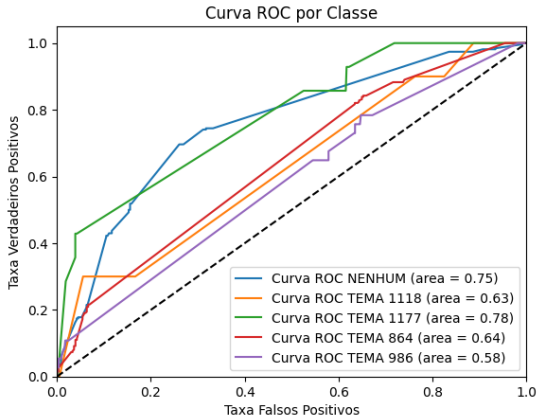

Figura 1: Curva ROC da Fase 1. A baixa convexidade das curvas para as classes minoritárias indica fraca capacidade de separação entre os temas.

A falha crítica ocorre nas classes minoritárias. O TEMA 986, por exemplo, obteve um Recall de apenas 0.03, indicando que o modelo falhou em identificar 97% das petições relativas a este tema.

A Figura 1 corrobora visualmente este cenário. A Curva ROC apresenta uma área sob a curva (AUC) reduzida para as classes desbalanceadas, aproximando-se da linha diagonal (aleatoriedade). Isso evidencia que, sem contexto, o LLM tende a "alucinar" ou recorrer a probabilidades estatísticas genéricas, o que é inaceitável em aplicações jurídicas de alta precisão.

b. Fase 2: Impacto do Data Augmentation

Na segunda fase, buscou-se mitigar o desbalanceamento através de *Data Augmentation* e refinamento de *prompts* com N-grams. Essa estratégia elevou a acurácia para 85% no *Dataset* Revisado. A partir dessa fase foi sempre utilizado o Gemini 2.5 Flash-Lite.

A principal contribuição desta fase foi a redução de falsos positivos na classe "NENHUM". A Matriz de Confusão (Figura 2) revela uma diagonal principal mais definida para os temas originais.

Contudo, a expansão para o ***Dataset* de Novos Temas** (sem a classe residual) revelou o teto desta abordagem baseada apenas em N-grams: a ambiguidade semântica. A Figura 3 ilustra como temas que compartilham vocabulário similar (ex: TEMA 796 e TEMA 1199) geraram confusões que não existiam no cenário anterior.

A Curva ROC para este cenário (Figura 4) confirma que, embora a acurácia geral se mantenha alta, a capacidade de discriminação cai para classes específicas (como o TEMA 796), indicando a necessidade de uma abordagem contextual mais robusta (RAG).

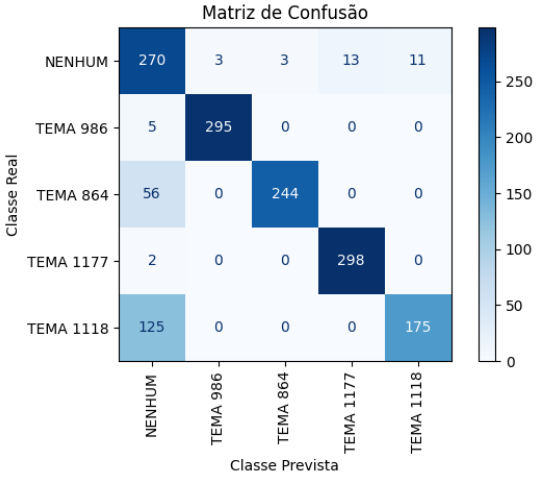

Figura 2: Matriz de Confusão - Fase 2 (*Dataset* Revisado). Nota-se a redução da dispersão na classe NENHUM.

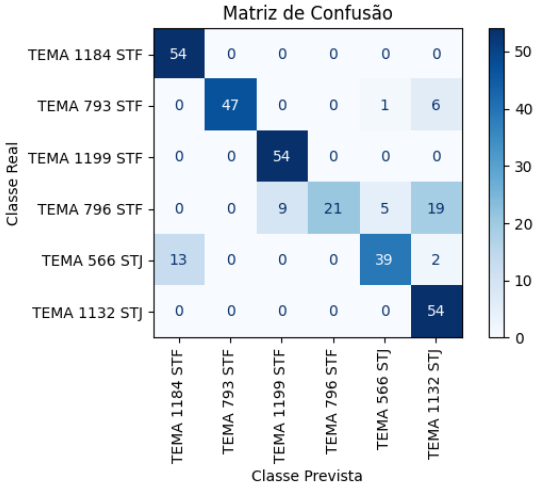

Figura 3: Matriz de Confusão - *Dataset* de Novos Temas. A abordagem de N-grams mostra limitações na distinção de temas semanticamente próximos.

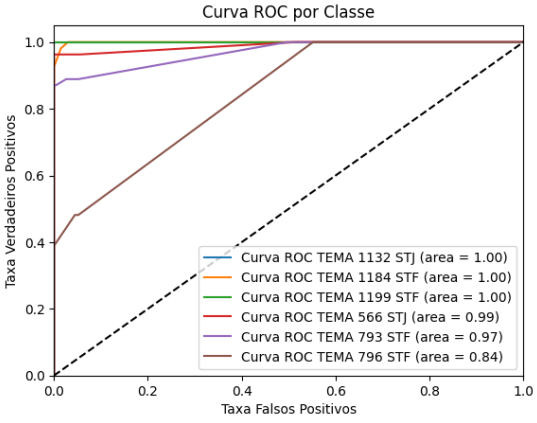

Figura 4: Curva ROC - *Dataset* de Novos Temas.

ISSN: 2675-3588

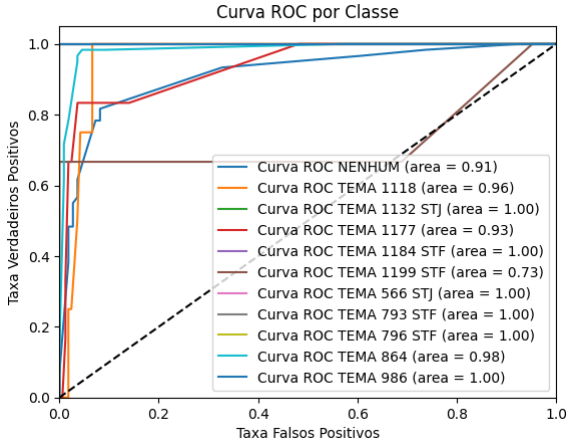

Figura 5: Curva ROC Final (RAG). A alta AUC confirma a capacidade de generalização do modelo frente a múltiplas classes jurídicas.

c. Fase 3: Consolidação com RAG (Resultados Finais)

A implementação da arquitetura RAG (Retrieval-Augmented Generation) representou a evolução definitiva do sistema. Ao ancorar a geração de texto em trechos recuperados de jurisprudência real, foi possível realizar um teste de estresse com 11 classes simultâneas sem a necessidade de balanceamento sintético.

A Tabela 2 apresenta os resultados finais consolidados. A acurácia global estabilizou-se em 84%. Embora numericamente similar à fase anterior, este resultado é qualitativamente superior, pois foi obtido em um cenário de complexidade dobrada (11 classes vs 5 classes).

TABELA 2: RELATÓRIO DE CLASSIFICAÇÃO FINAL COM RAG (11 CLASSES)

Classe	Precisão	Recall	F1-Score	Suporte
NENHUM	0.90	0.63	0.75	60
TEMA 1118	0.21	0.75	0.33	4
TEMA 1132 STJ	1.00	1.00	1.00	2
TEMA 1177	0.44	0.67	0.53	6
TEMA 1184 STF	1.00	1.00	1.00	8
TEMA 1199 STF	1.00	0.67	0.80	3
TEMA 566 STJ	1.00	1.00	1.00	2
TEMA 793 STF	0.92	1.00	0.96	11
TEMA 796 STF	1.00	1.00	1.00	1
TEMA 864	0.91	0.98	0.94	60
TEMA 986	1.00	1.00	1.00	13
Acurácia Global			**0.84**	170
Média Macro	0.85	0.88	0.85	170
Média Ponderada	0.88	0.84	0.85	170

Destaca-se o desempenho perfeito (F1-Score 1.0) em temas complexos como o TEMA 1184 (STF) e TEMA 986. Isso confirma a hipótese de que a recuperação de informação atua como um mecanismo de "ancoragem", transformando a tarefa de alucinação criativa em uma tarefa de verificação semântica.

Por fim, a Figura 5 exibe as Curvas ROC finais. A aproximação das curvas ao canto superior esquerdo para a vasta maioria das classes valida a robustez do classificador RAG, tornando-o apto para auxiliar na triagem de processos em larga escala.

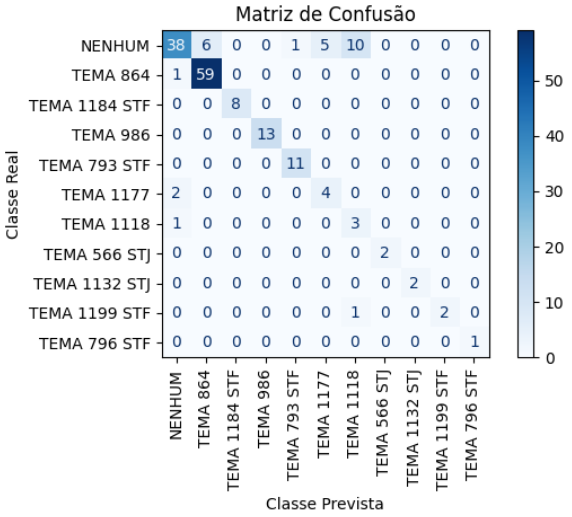

Figura 6: Matriz de Confusão - RAG Final. Note a concentração de falsos positivos na coluna do TEMA 1118 oriundos da classe NENHUM.

d. Análise de Erros e Limitações

Apesar do sucesso global, a análise qualitativa da Matriz de Confusão final (Figura 6) aponta que os erros residuais persistem na classe `NENHUM`. O modelo ainda classifica incorretamente algumas petições genéricas como pertencentes a temas específicos que possuem vocabulário sobreposto, notadamente o TEMA 1118 (questões de IPVA).

Isso ocorre porque, semanticamente, os vetores de uma petição de "Execução de Multa de IPVA" (classe NENHUM) e "Declaratória de Inexistência de Débito de IPVA por Venda" (TEMA 1118) são muito próximos no espaço latente. O mecanismo de recuperação traz ambos os tipos de documentos, e o LLM, na dúvida, tende a optar pela classe específica. Isso sugere que, para aplicações futuras, apenas a busca semântica não basta; pode ser necessário um passo de re-ranking ou uma verificação lógica adicional pós-recuperação.

V. CONCLUSÕES

Este trabalho demonstrou a viabilidade e a eficácia da utilização de Modelos de Linguagem de Grande Escala, especificamente o Gemini, integrados a uma arquitetura de Recuperação Aumentada por Geração (RAG), para a classificação automática de documentos judiciais.

A evolução metodológica, partindo de 56% de acurácia com *few-shot* simples para 84% com RAG, sugere que o uso de contexto dinâmico pode contribuir para a automação jurídica com maior precisão. As técnicas de Data Augmentation e N-grams serviram como etapas intermediárias importantes, mas foi a capacidade do RAG de consultar uma base de conhecimento em tempo real que garantiu a robustez final do classificador frente à complexidade e intertextualidade do Direito brasileiro.

Além da métrica de acurácia, a arquitetura RAG oferece uma vantagem crucial para o setor público: a explicabilidade. Diferente de modelos "caixa-preta", é possível auditar exatamente quais precedentes foram utilizados pelo sistema para fundamentar cada classificação, aumentando a confiança

na tecnologia.

Como trabalhos futuros, sugere-se a exploração de modelos de *embedding* treinados especificamente em corpus jurídico brasileiro (como o LegalBERT-PT) para refinar a etapa de recuperação, e a validação do sistema em ambiente de produção para mensurar ganhos de eficiência processual.

REFERÊNCIAS

[1] Conselho Nacional de Justiça, "Justiça em números 2024," 2024. [Online]. Available: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[3] B. Shukla, S. Gupta, A. K. Yadav, and D. Yadav, "Challenges and issues in legal documents classification," in *AIP Conference Proceedings*, vol. 2754, no. 1. AIP Publishing, 2023.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems 30*, 2017, pp. 5998–6008.

[6] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, G. Nogueira, H. Pua, G. Ross, W.-t. Yih, D. Kiela *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

 ISSN: 2675-3588