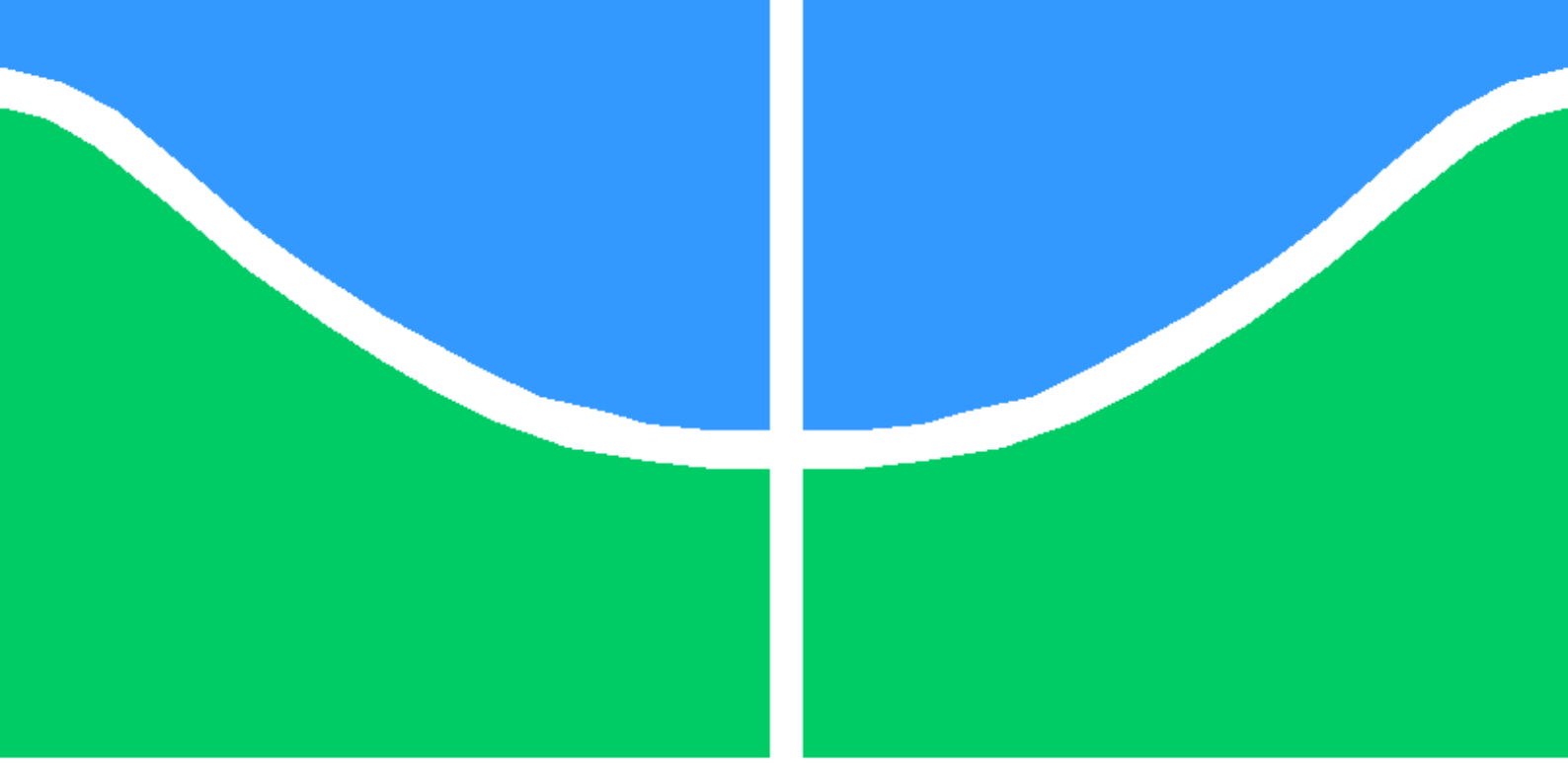


AVALIAÇÃO DE GRANDES MODELOS DE LINGUAGEM (LLMS) PARA A TIPIFICAÇÃO DE DOCUMENTOS

DÁRIO PEREIRA DOS SANTOS

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

AVALIAÇÃO DE GRANDES MODELOS DE LINGUAGEM (LLMS) PARA A TIPIFICAÇÃO DE DOCUMENTOS

DÁRIO PEREIRA DOS SANTOS

Orientador: PROF. DR. DANIEL ALVES, PPEE/UNB

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO PPEE.MP - 093/2025
BRASÍLIA-DF, 15 DE AGOSTO DE 2025.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

AVALIAÇÃO DE GRANDES MODELOS DE LINGUAGEM (LLMS) PARA A TIPIFICAÇÃO DE DOCUMENTOS

DÁRIO PEREIRA DOS SANTOS

DISSERTAÇÃO DE MESTRADO ACADÊMICO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA.

APROVADA POR:

Prof. Dr. Daniel Alves, PPEE/UnB
Orientador

Prof. Dr. Fábio Mendonça, PPEE/UnB
Examinador interno

Prof. Dr. Gilmar dos Santos Marques, FAP/DF
Examinador Externo

BRASÍLIA, 15 DE AGOSTO DE 2025.

FICHA CATALOGRÁFICA
DÁRIO PEREIRA DOS SANTOS
Avaliação de Grandes Modelos de Linguagem (LLMs) para a Tipificação de Documentos: Um Estudo Comparativo
2025xv, 115p., 201x297 mm
(PPEE/FT/UnB, Mestre, Engenharia Elétrica, 2025)
Dissertação de Mestrado - Universidade de Brasília
Faculdade de Tecnologia - DEPARTAMENTO DE ENGENHARIA ELÉTRICA

REFERÊNCIA BIBLIOGRÁFICA

DÁRIO PEREIRA DOS SANTOS (2025) Avaliação de Grandes Modelos de Linguagem (LLMs) para a Tipificação de Documentos: Um Estudo Comparativo. Dissertação de Mestrado em Engenharia Elétrica, Publicação 093/2025, DEPARTAMENTO DE ENGENHARIA ELÉTRICA, Universidade de Brasília, Brasília, DF, 115p.

CESSÃO DE DIREITOS

AUTOR: DÁRIO PEREIRA DOS SANTOS
TÍTULO: Avaliação de Grandes Modelos de Linguagem (LLMs) para a Tipificação de Documentos: Um Estudo Comparativo.
GRAU: Mestre ANO: 2025

_____
DÁRIO PEREIRA DOS SANTOS

Aplicação de Modelos de Linguagem de Grande Escala (LLMs) para Tipificação Automática de Documentos

Autor: Dário Pereira dos Santos
Orientador: Prof. Dr. Daniel Alves da Silva
Programa de Pós-graduação em Engenharia Elétrica - PPEE

O avanço dos Modelos de Linguagem de Grande Escala (LLMs) tem impulsionado significativamente aplicações em Processamento de Linguagem Natural (PLN), especialmente em tarefas como a classificação textual e a organização de grandes volumes de documentos. Este trabalho apresenta um estudo comparativo entre diferentes LLMs aplicados à tipificação automática de documentos digitais.

Foram avaliados oito modelos baseados em arquiteturas transformer da família LLaMA, Mistral, Gemma e DeepSeek, acessados via chamadas assíncronas por API. As análises se basearam em métricas clássicas de desempenho, como acurácia, precisão, revocação, F1-score, perplexidade e log-likelihood, considerando o comportamento dos modelos na classificação multiclasse de textos oriundos de diferentes domínios.

Os resultados mostram que o modelo LLaMA 3 apresentou o melhor desempenho geral, seguido de suas variantes ajustadas. A pesquisa destaca ainda a importância da escolha do modelo de linguagem conforme o contexto e a necessidade da tarefa, contribuindo com evidências empíricas para adoção de LLMs em sistemas de gestão documental automatizada.

Palavras-chave: Grandes Modelos de Linguagem, Tipificação de Documentos, Processamento de Linguagem Natural, LLaMA, Classificação Multiclasse.

Application of Large Language Models (LLMs) for Automatic Document Typification

Author: Dário Pereira dos Santos
Advisor: Prof. Dr. Daniel Alves da Silva
Postgraduate Program in Electrical Engineering - PPEE

The advancement of Large Language Models (LLMs) has significantly driven applications in Natural Language Processing (NLP), especially in tasks such as text classification and the organization of large volumes of documents. This work presents a comparative study of different LLMs applied to the automatic typification of digital documents.

Eight transformer-based models were evaluated, from the LLaMA, Mistral, Gemma, and DeepSeek families, accessed via asynchronous API calls. The analyses were based on standard performance metrics such as accuracy, precision, recall, F1-score, perplexity, and log-likelihood, considering the models' behavior in multiclass classification of texts from different domains.

The results show that the LLaMA 3 model achieved the best overall performance, followed by its fine-tuned variants. The research further emphasizes the importance of choosing the appropriate language model according to the context and task requirements, contributing empirical evidence to support the adoption of LLMs in automated document management systems.

Keywords: Large Language Models, Document Typification, Natural Language Processing, LLaMA, Multiclass Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

Lista de Códigos

LISTA DE TERMOS E SIGLAS

API	Application Programming Interface (Interface de Programação de Aplicações)
AUC	Area Under the Curve (Área Sob a Curva)
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-Words
CSV	Comma-Separated Values (Valores Separados por Vírgula)
FastAPI	Framework web moderno e rápido para APIs em Python
FT	Faculdade de Tecnologia
GPT	Generative Pre-trained Transformer
HTML	HyperText Markup Language (Linguagem de Marcação de Hipertexto)
IA	Inteligência Artificial
JSON	JavaScript Object Notation (Notação de Objetos JavaScript)
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
Matplotlib	Biblioteca Python para visualização de dados
ML	Machine Learning (Aprendizado de Máquina)
MLP	Multilayer Perceptron (Perceptron de Múltiplas Camadas)
NLP	Natural Language Processing (Processamento de Linguagem Natural)
NumPy	Biblioteca Python para computação científica
Pandas	Biblioteca Python para manipulação de dados
PDF	Portable Document Format (Formato Portátil de Documento)
PPEE	Programa de Pós-Graduação em Engenharia Elétrica

ROC	Receiver Operating Characteristic
Scikit-learn	Biblioteca Python para aprendizado de máquina
SEI	Sistema Eletrônico de Informações
SVM	Support Vector Machines (Máquinas de Vetores de Suporte)
TF-IDF	Term Frequency-Inverse Document Frequency
TPU	Tensor Processing Unit (Unidade de Processamento Tensorial)
UnB	Universidade de Brasília

SUMÁRIO

A Publicações 73

Capítulo 1

Introdução

Nos últimos anos, a proeminência dos Grandes Modelos de Linguagem (Large Language Models – LLMs) transformou o campo do Processamento de Linguagem Natural (PLN). Impulsionados por arquiteturas como a Transformer e pelo desenvolvimento de modelos cada vez mais eficientes, os LLMs consolidaram-se como ferramentas essenciais para tarefas avançadas (AKHTAR, 2024; HUPKES et al., 2023; YAO et al., 2024). Dentre suas múltiplas aplicações, destacam-se a análise semântica, a geração de conteúdo e, notadamente, a tipificação textual e a tipificação de documentos (PATIL et al., 2023; SANTOS et al., 2025).

A tipificação de documentos — processo de categorização automática de textos com base em seus temas ou propósitos — constitui um desafio central em setores corporativos, jurídicos, acadêmicos e governamentais. A automação dessa tarefa é vital não apenas para a organização e recuperação eficiente da informação, mas também para a otimização de fluxos de trabalho e a melhoria da gestão documental em larga escala (SANTOS et al., 2023; MUAAD et al., 2022; WEI et al., 2023; ZHANG et al., 2025).

Diante desse cenário, o presente estudo avalia a eficácia de diferentes LLMs de código aberto na tarefa de tipificação de documentos. A pesquisa realiza uma análise comparativa do desempenho de variantes das famílias DeepSeek, Gemma, LLaMA 3 e Mistral, modelos de última geração amplamente reconhecidos pela comunidade científica e técnica (HUANG et al., 2024; KUKREJA et al., 2024). Especificamente, os modelos investigados foram: `deepseek-llm:7b`, `gemma:7b`, `llama3`, `llama3.1:8b`, `llama3.1:latest`, `llama3.2:latest`, `mistral-nemo:latest` e `mistral:7b`.

Cada um desses modelos possui particularidades em sua arquitetura, corpus de treinamento e estratégias de ajuste, o que impacta diretamente sua performance em tarefas específicas. Enquanto alguns são otimizados para tarefas generalistas, outros foram ajustados para maior precisão em domínios técnicos. Tais diferenças tornam a avaliação comparativa fundamental para compreender a adequação de cada modelo à tipificação documental (YIN; NI; WANG, 2024).

A análise experimental conduzida nesta dissertação revelou variações significativas de desempenho entre os modelos. A família LLaMA 3 demonstrou a maior

eficácia na categorização de documentos, com o modelo `llama3` atingindo a maior acurácia (0.8794), seguido por `llama3.1:8b` (0.8726) e `llama3.1:latest` (0.8786). O modelo `gemma:7b` também obteve um desempenho satisfatório (0.8347). Em contrapartida, `deepseek-llm:7b` apresentou um resultado inferior (0.5227), enquanto `llama3.2:latest` obteve a menor acurácia (0.3310), sugerindo dificuldades em lidar com o conjunto de dados utilizado. Os modelos da linha Mistral apresentaram um desempenho intermediário, com `mistral-nemo:latest` alcançando 0.8634 e `mistral:7b`, 0.6881.

Além da acurácia, a análise contempla métricas adicionais, como tempo de processamento e adaptabilidade ao domínio dos documentos. A proposta não é apenas quantificar o desempenho, mas também compreender as limitações e as vantagens de cada arquitetura em contextos distintos.

Este trabalho busca, portanto, oferecer uma visão abrangente sobre o uso de LLMs na tipificação de documentos, contribuindo com recomendações práticas para a escolha de modelos em cenários reais. Por fim, são discutidas possíveis direções para pesquisas futuras, como o uso de modelos híbridos ou adaptações contextuais que aumentem a eficiência na categorização em larga escala.

1.1 Motivação

A onipresença da transformação digital e o consequente crescimento exponencial na produção de dados impõem desafios significativos à gestão da informação nos setores corporativo, jurídico, acadêmico e governamental. Diante de volumes massivos de documentos gerados diariamente, a tarefa de categorizar e estruturar esse conteúdo de forma automática e precisa tornou-se um pilar para garantir agilidade, rastreabilidade e tomadas de decisão assertivas (HUA et al., 2024).

Nesse contexto, a tipificação de documentos, processo que associa textos a categorias predefinidas, emerge como uma solução estratégica. A sua automação otimiza fluxos de trabalho, mitiga erros humanos e democratiza o acesso a informações relevantes, representando um ganho expressivo de produtividade e qualidade organizacional, especialmente em ambientes de alta demanda por processamento textual (KOWSARI et al., 2019; GASPARETTO et al., 2022).

A recente ascensão dos Grandes Modelos de Linguagem (LLMs) como ferramentas de Processamento de Linguagem Natural (PLN) intensificou o interesse em soluções de inteligência artificial para a tipificação documental. Em virtude de suas capacidades robustas de compreensão semântica, generalização e adaptação a diferentes domínios, esses modelos posicionam-se como candidatos promissores para substituir ou complementar as abordagens de tipificação tradicionais (NASUTION; ONAN, 2024).

Contudo, a aplicação eficaz dos LLMs nesta tarefa não é trivial e enfrenta limitações importantes. A performance dos modelos pode variar drasticamente a depender

de fatores como sua arquitetura, o volume e a diversidade dos dados de treinamento, a capacidade de contexto e o alinhamento com a tarefa-alvo. Modelos compactos podem ter dificuldades com linguagens especializadas, enquanto modelos de maior escala podem exigir um custo computacional elevado e um complexo processo de ajuste fino (*fine-tuning*) para alcançar a performance desejada (FIELDS; CHOVANEC; MADIRAJU, 2024).

Essa variabilidade de desempenho e a ausência de um consenso sobre qual modelo é mais adequado para diferentes cenários de tipificação constituem a motivação central deste estudo. A presente pesquisa propõe-se a investigar, de forma sistemática e comparativa, a capacidade de diferentes LLMs para executar essa tarefa com alta acurácia e eficiência. Através de uma análise rigorosa, busca-se não apenas compreender as forças e limitações de cada arquitetura, mas também gerar subsídios práticos que orientem sua adoção em aplicações reais e contribuam para o avanço da pesquisa em PLN, por meio de estratégias de adaptação baseadas em domínio (WEI et al., 2023).

Dessa forma, este trabalho visa oferecer uma visão aprofundada e atualizada do papel dos LLMs na organização automatizada de conteúdos textuais, promovendo sua adoção de forma crítica, estratégica e tecnicamente fundamentada.

1.2 Objetivo

Este estudo tem como objetivo central avaliar e comparar a eficácia de diferentes Grandes Modelos de Linguagem (LLMs) na tarefa de tipificação automatizada de documentos.

Para tanto, a avaliação transcende a análise da acurácia geral, empregando um conjunto de métricas multidimensional, projetado para capturar uma visão completa do desempenho de cada modelo. As métricas investigadas incluem:

- **Acurácia (Accuracy):** a proporção geral de classificações corretas.
- **Precisão (Precision):** a proporção de classificações corretas entre as respostas positivas fornecidas por classe.
- **Revocação (Recall):** a capacidade do modelo de identificar corretamente todos os exemplos relevantes de uma classe.
- **F1-Score:** a média harmônica entre precisão e revocação, oferecendo um balanço entre ambas.
- **Perplexidade (Perplexity):** uma medida da incerteza do modelo na previsão de sequências textuais, indicando sua confiança.
- **Log Likelihood:** a probabilidade atribuída pelo modelo aos dados observados, refletindo sua adequação geral ao corpus.

- **Exact Match:** a taxa de correspondência exata entre a categoria prevista e a real, idêntica à acurácia em cenários de tipificação single-label.

A utilização conjunta dessas métricas permite não apenas quantificar a correção das previsões, mas também avaliar a robustez, a confiança e o comportamento dos modelos em cenários com diferentes distribuições de classes. Com base nessa análise detalhada, busca-se oferecer um panorama empírico sobre as capacidades e limitações de cada LLM, fornecendo subsídios para a escolha informada e estratégica dessas tecnologias em aplicações de gestão documental.

1.3 Objetivos Específicos

Para alcançar o objetivo geral deste estudo, foram definidos os seguintes objetivos específicos:

1. **Realizar uma avaliação quantitativa da performance dos LLMs,** comparando o desempenho dos modelos selecionados na tarefa de tipificação de documentos por meio de um conjunto abrangente de métricas: acurácia, precisão, revocação, F1-Score, perplexidade, log likelihood e *exact match*.
2. **Analisar a variação de desempenho dos modelos entre as diferentes categorias temáticas** do conjunto de dados utilizado, investigando como a performance de cada LLM se altera em função do domínio textual (ex: negócios, esporte, tecnologia).
3. **Identificar as forças e limitações de cada arquitetura,** interpretando os padrões de acerto e erro revelados pelas métricas para determinar em quais contextos cada modelo se mostra mais robusto ou suscetível a falhas.
4. **Prover recomendações práticas baseadas nos resultados empíricos,** indicando os modelos de linguagem que se mostraram mais eficazes e eficientes, de modo a auxiliar profissionais e pesquisadores na seleção de ferramentas para suas aplicações.
5. **Sugerir direções para investigações futuras** que possam aprimorar a tipificação automatizada de documentos, abordando temas como o uso de conjuntos de dados mais complexos, o ajuste fino de modelos e a implementação de sistemas híbridos.

1.4 Metodologia Científica

Esta seção detalha o delineamento metodológico empregado para conduzir a avaliação de Grandes Modelos de Linguagem (LLMs) na tarefa de tipificação de docu-

mentos. A pesquisa foi estruturada com uma abordagem aplicada e natureza quantitativa, focada na investigação de técnicas de categorização textual baseadas em modelos de linguagem (KOWSARI et al., 2019; GASPARETTO et al., 2022).

1.4.1 Caracterização da Pesquisa

Quanto aos fins, a pesquisa é classificada como **aplicada**, pois seu objetivo é gerar conhecimento para a solução de problemas práticos relacionados à organização e tipificação de documentos digitais. Quanto à abordagem do problema, o estudo é **quantitativo**, uma vez que se baseia na mensuração de métricas de desempenho a partir de dados objetivos coletados em um ambiente experimental controlado.

Do ponto de vista dos objetivos, a pesquisa tem um caráter **exploratório**, pois busca examinar a viabilidade e a eficácia do uso de diferentes LLMs para a tipificação automática de documentos, um campo em rápida evolução. A metodologia envolve a formulação de um *pipeline* experimental, a análise estatística dos resultados e a interpretação crítica das evidências observadas.

1.4.2 Procedimentos Metodológicos

A investigação empírica foi conduzida por meio da construção de um *pipeline* computacional modular, dividido em seis etapas principais que refletem o processo completo de tipificação automática:

1. **Aquisição e Definição dos Dados:** Utilização do conjunto de dados público *BBC News* (`bbc-text.csv`), composto por textos jornalísticos previamente rotulados em cinco categorias.

2. **Pré-processamento e Codificação:** Padronização dos textos, limpeza de ruídos e codificação numérica das categorias para viabilizar a análise quantitativa.

3. **tipificação via LLMs:** Submissão assíncrona dos textos a uma API que se comunica com os diferentes modelos de linguagem para obter a predição da categoria.

4. **Normalização das Predições:** Tratamento e padronização dos rótulos preditos para garantir a consistência com o formato das categorias originais.

5. **Avaliação de Desempenho:** Análise das predições com base em métricas quantitativas consagradas, como acurácia, precisão, revocação e F1-Score.

6. **Geração de Relatórios e Visualizações:** Produção automatizada de gráficos e relatórios que permitem a interpretação e a comunicação dos resultados.

1.4.3 Justificativa do Método

A escolha por utilizar LLMs em detrimento de abordagens tradicionais de aprendizado de máquina fundamenta-se na capacidade desses modelos de capturar contextos semânticos complexos e realizar inferência com mínima engenharia manual de atributos (*features*) (BROWN et al., 2020). A implementação de um fluxo de tipificação assíncrona, por sua vez, foi projetada para garantir escalabilidade e eficiência em tempo de resposta, viabilizando a análise de um grande volume de dados.

A avaliação do desempenho foi pautada em métricas consolidadas na literatura de tipificação, assegurando a validade e a comparabilidade dos resultados (SOKOLOVA; LAPALME, 2009).

1.4.4 Limitações da Metodologia

Reconhece-se que a metodologia adotada possui limitações inerentes. O desempenho dos modelos é dependente da qualidade e do domínio do conjunto de dados de entrada. Adicionalmente, a sensibilidade dos LLMs à formulação dos *prompts* é um fator que pode influenciar os resultados (JIANG et al., 2023). Por se tratar de modelos de inferência acessados via API, não há controle direto sobre os parâmetros internos das arquiteturas, o que restringe certas análises mais profundas.

Apesar dessas limitações, a abordagem adotada é robusta para testar as hipóteses da pesquisa e construir uma base sólida de evidências empíricas, abrindo caminhos para trabalhos futuros que explorem técnicas híbridas, ajuste fino de modelos ou a aplicação em outros domínios documentais.

1.4.5 Organização do Trabalho

Este trabalho está estruturado em seis capítulos, além das referências e apêndices, de modo a conduzir o leitor de forma lógica através da pesquisa desenvolvida.

No **Capítulo 1**, realiza-se a introdução, contextualizando a relevância dos Grandes Modelos de Linguagem (LLMs) para a tipificação de documentos. Apresentam-se também a motivação, os objetivos geral e específicos, e a metodologia que norteou o estudo.

O **Capítulo 2** é dedicado à fundamentação teórica, onde são abordados os conceitos centrais sobre LLMs, a arquitetura Transformer, a tarefa de tipificação documental, as métricas de avaliação e os trabalhos relacionados na literatura.

O **Capítulo 3** descreve em detalhes a arquitetura computacional projetada e implementada para os experimentos. São detalhadas as etapas do *pipeline*, desde a aquisição e o pré-processamento dos dados até a tipificação assíncrona e a avaliação de desempenho.

No **Capítulo 4**, são apresentados e discutidos os resultados empíricos obtidos. A

análise comparativa do desempenho dos oito modelos avaliados é detalhada por meio de tabelas e gráficos que ilustram as métricas de performance.

O **Capítulo 5** sugere direções para investigações futuras, apontando caminhos para o aprimoramento e a expansão da pesquisa, como a exploração de novos datasets, técnicas de ajuste fino e o desenvolvimento de uma aplicação prática.

Finalmente, o **Capítulo 6** encerra a dissertação com a conclusão, que sintetiza os principais achados, reitera as contribuições do trabalho e discute suas implicações práticas e teóricas. Após os capítulos, são listadas as referências bibliográficas e um apêndice com as publicações acadêmicas resultantes desta pesquisa.

LLMLarge Language Model (Modelo de Linguagem de Grande Escala)

Capítulo 2

Conceitos e Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que embasam a presente dissertação. A discussão abrange desde a definição da tarefa de tipificação documental até a ascensão dos Grandes Modelos de Linguagem (LLMs) como tecnologia central para sua execução. Adicionalmente, são abordados os desafios inerentes a esses modelos e as métricas utilizadas para sua avaliação, construindo a base conceitual para os capítulos subsequentes.

2.1 Tipificação de Documentos como Desafio Central

A categorização automática de textos é uma das aplicações mais consolidadas em Processamento de Linguagem Natural (PLN), com relevância crescente em domínios como gestão de informações empresariais, sistemas arquivísticos, recuperação da informação e organização de acervos digitais institucionais (NAZI; PENG, 2024; CHANG et al., 2024). Nesse contexto, a **tipificação documental** refere-se ao processo de atribuição automática de uma ou mais categorias a um documento, com base em seu conteúdo semântico.

Essa técnica é amplamente empregada para otimizar sistemas de tipificação de notícias, realizar a triagem de e-mails, analisar sentimentos em larga escala e automatizar *workflows* jurídicos ou administrativos (PŁONKA et al., 2025). Para tal, é imperativo o uso de modelos capazes de extrair representações semânticas precisas dos textos e mapeá-las adequadamente para as classes predefinidas.

2.2 Grandes Modelos de Linguagem e a Arquitetura Transformer

Com os avanços recentes na área de PLN, os Grandes Modelos de Linguagem (LLMs) emergiram como a principal alternativa para tarefas de tipificação textual. Modelos como GPT, BERT, RoBERTa, LLaMA, Mistral e Gemma têm demonstrado

elevada performance em *benchmarks* de categorização textual e compreensão semântica (QIN et al., 2025; XU, 2025; ZHANG; TSUDA, 2025).

A base tecnológica desses modelos é a arquitetura **Transformer**, proposta em um trabalho seminal por Vaswani et al. (2017). Essa arquitetura revolucionou o PLN ao introduzir mecanismos de atenção (*attention mechanisms*), que permitem aos modelos capturar dependências contextuais em longa escala de forma paralela e eficiente, superando limitações de arquiteturas sequenciais anteriores.

2.3 Desafios e Métricas de Avaliação

Apesar do seu potencial, o desempenho dos LLMs pode variar significativamente conforme o tipo de tarefa, o domínio do corpus e os parâmetros do modelo. Fatores como custo computacional, tempo de inferência, adaptabilidade a domínios específicos e tolerância a ruído textual são considerados na literatura como pontos críticos para sua aplicação em ambientes de produção (CHANG et al., 2024; SCHMIDT et al., 2025).

A avaliação de modelos de tipificação textual, por sua vez, exige o uso de métricas consolidadas que permitam uma análise robusta do desempenho. Embora a acurácia seja um indicador comum, estudos comparativos demonstram que a escolha da métrica deve ser sensível ao contexto, sendo essencial o uso de precisão, revocação e F1-Score, principalmente em cenários com distribuição desequilibrada de classes ou em tarefas com múltiplos rótulos (SOKOLOVA; LAPALME, 2009). O sucesso da tipificação depende também de estratégias de pré-processamento, codificação de rótulos e normalização de dados.

Finalmente, do ponto de vista prático, revisões sistemáticas indicam que a integração de LLMs com APIs REST tem se tornado uma abordagem eficiente para a condução de experimentos aplicados, viabilizando testes rápidos, replicáveis e escaláveis com diferentes modelos de linguagem (LIU et al., 2024a).

2.4 Grandes Modelos de Linguagem (LLMs)

Os Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*) representam um marco no avanço do Processamento de Linguagem Natural (PLN), impulsionados por sua capacidade de compreender, gerar e manipular a linguagem humana em múltiplos contextos (IQBAL; QURESHI, 2022; LE et al., 2023; MIROŃCZUK; MÜLLER; PEDRYCZ, 2024). Essencialmente, são redes neurais de larga escala, treinadas sobre vastos corpora textuais, que utilizam arquiteturas de aprendizado profundo — notadamente a arquitetura **Transformer** — para modelar as complexas nuances da comunicação humana (VASWANI et al., 2017). Por meio dessa exposição massiva a dados, os LLMs aprendem representações distribuídas que codificam significados, relações semânticas e estruturas sintáticas, habilitando-os a executar

tarefas como tradução, sumarização e, de especial interesse para este trabalho, a tipificação de documentos.

Uma das principais inovações trazidas pelos LLMs é a sua capacidade de generalização, que lhes permite executar tarefas com poucas ou nenhuma amostra de treinamento, explorando os paradigmas de *few-shot* e *zero-shot learning* (CHAE; DAVIDSON, 2023; BOYINA et al., 2024; BROWN et al., 2020). Essa flexibilidade é uma consequência direta da arquitetura Transformer, que, através de seus mecanismos de atenção, processa sequências de texto de forma paralela e captura dependências de longa distância entre os *tokens*.

A aplicação prática dessa arquitetura foi popularizada por modelos fundacionais com diferentes filosofias de pré-treinamento. O **BERT** (*Bidirectional Encoder Representations from Transformers*) introduziu uma abordagem baseada em codificadores (*encoders*) e na modelagem de linguagem mascarada (*masked language modeling*), na qual o modelo aprende a prever *tokens* omitidos a partir de um contexto bidirecional (esquerdo e direito), gerando ricas representações semânticas (DEVLIN et al., 2019). Em contraste, o **GPT** (*Generative Pre-trained Transformer*) e seus sucessores adotaram uma abordagem baseada em decodificadores (*decoders*), treinados de forma autorregressiva para prever o próximo *token* em uma sequência, o que os torna especialmente aptos para tarefas de geração (RADFORD et al., 2018). Variações posteriores, como RoBERTa, T5, LLaMA e Mistral, refinaram esses esquemas com foco em eficiência, desempenho e generalização.

A evolução contínua dos LLMs é marcada pelo crescimento exponencial no número de parâmetros e pela expansão de suas capacidades. Modelos como GPT-3 e LLaMA 3 demonstram habilidades emergentes em tarefas de PLN de alta complexidade, incluindo a tipificação textual (PREUSS; ALSHEHRI; YOU, 2024; JARADAT et al., 2025). Contudo, esse avanço impõe desafios significativos: o uso desses modelos demanda uma infraestrutura computacional avançada, e sua eficácia pode ser limitada por fatores como o domínio de aplicação, a qualidade dos dados e o alinhamento com a tarefa-alvo (SUMANATHILAKA; MICALLEF; HOUGH, 2024; DONG; WANG, 2024). Apesar dessas limitações, a robustez, a capacidade de generalização e a adaptabilidade dos LLMs consolidam seu potencial para transformar a categorização automática de documentos.

2.5 Tipificação de Documentos

A tipificação de documentos — também referida na literatura como categorização de documentos ou tipificação textual — consiste na tarefa de atribuir rótulos a documentos de forma automática, com base em seu conteúdo semântico. Trata-se de uma técnica fundamental para a gestão da informação, sendo amplamente empregada em ambientes corporativos, acadêmicos, jurídicos e governamentais. Diante da crescente produção de dados textuais em larga escala, métodos automatizados de tipificação tornaram-se essenciais para viabilizar a organização eficiente, a recupe-

ração de informações e a tomada de decisão baseada em dados (KO; NECHES, 2003; CHEN; WARREN, 2013).

Historicamente, o processo pode ser conduzido manualmente, por meio da análise de especialistas, ou de forma automatizada. A abordagem automatizada, especialmente com o advento de modelos de linguagem baseados em redes neurais profundas, permite um aumento significativo na escalabilidade, consistência e eficiência da tarefa, ao mesmo tempo que mitiga vieses subjetivos e reduz custos operacionais (HAQUE et al., 2025).

Do ponto de vista técnico, a tipificação de documentos é comumente formulada como um problema de tipificação supervisionada. Nesse paradigma, um algoritmo é treinado a partir de um conjunto de dados previamente rotulado para que aprenda a prever as categorias de documentos novos, ainda não classificados. Abordagens tradicionais para essa tarefa incluem o uso de algoritmos como Naive Bayes, Support Vector Machines (SVM) e redes neurais multicamadas, que operam sobre representações vetoriais dos textos (GASPARETTO et al., 2022). Contudo, a evolução dos Grandes Modelos de Linguagem (LLMs) expandiu as fronteiras dessa tarefa ao introduzir o uso de representações contextuais densas e de alto nível semântico (ARSLAN; MUNAWAR; CRUZ, 2024).

Os LLMs destacam-se por sua capacidade de capturar o contexto e as estruturas linguísticas complexas, o que os torna particularmente eficazes na tipificação documental. Em contraste com métodos clássicos que dependem de uma extração manual de características (e.g., *bag-of-words* ou TF-IDF), os LLMs fornecem *embeddings* contextualizados que representam o significado do texto em sua totalidade, uma característica com impacto direto na melhoria da precisão e da generalização para diferentes domínios (WANG et al., 2022). Essa capacidade de generalização é amplificada pelo fato de serem pré-treinados em corpora massivos e diversos, o que lhes confere adaptabilidade a múltiplos tipos de documentos — de relatórios técnicos a publicações jornalísticas — mesmo em cenários com baixa disponibilidade de exemplos rotulados (HUA, 2020).

Adicionalmente, a tipificação automatizada pode ser estendida para cenários mais complexos, como a tipificação hierárquica ou multi-rótulo, onde um documento pode pertencer a múltiplas categorias ou subcategorias simultaneamente. Essa abordagem é especialmente útil em bibliotecas digitais e repositórios jurídicos, que exigem alta granularidade na organização da informação (ZHANG et al., 2025).

Apesar desses avanços, persistem desafios relevantes. A escassez de dados rotulados de alta qualidade, a necessidade de adaptação dos modelos a domínios de nicho e as limitações na interpretação de terminologias técnicas continuam sendo barreiras para uma tipificação precisa. Adicionalmente, o alto custo computacional associado à inferência com LLMs pode restringir sua adoção em sistemas com infraestrutura limitada (ZHANG et al., 2025; KHOBOKO; MARIVATE; SEFARA, 2025).

Em síntese, embora os LLMs representem um avanço paradigmático na tarefa de tipificação de documentos, sua adoção em ambientes de produção requer uma

análise criteriosa de escalabilidade, custo e adequação ao domínio. A combinação desses modelos com estratégias de engenharia de atributos, aprendizado ativo ou pré-processamento especializado pode representar uma alternativa promissora para superar tais limitações, consolidando a tipificação como um componente estratégico para a gestão da informação na era digital.

2.5.1 Arquiteturas de Modelos de Linguagem

A vasta maioria dos Grandes Modelos de Linguagem (LLMs) contemporâneos é fundamentada na arquitetura **Transformer**, que representou um ponto de inflexão no campo do Processamento de Linguagem Natural (PLN). Proposta por Vaswani et al. (2017), sua principal inovação reside no uso de mecanismos de atenção (*attention*), que possibilitam ao modelo ponderar a importância de diferentes *tokens* em uma sequência, capturando dependências contextuais de forma paralela e eficiente, em contraste com as abordagens sequenciais que a precederam (ALAKTIF et al., 2024).

Os modelos específicos avaliados neste trabalho — como LLaMA, Mistral, Gemma e DeepSeek — pertencem a uma família de arquiteturas Transformer conhecida como *decoder-only*. Essa estrutura é otimizada para tarefas de compreensão e geração textual, sendo composta por um empilhamento de blocos idênticos. Cada bloco, por sua vez, contém os componentes estruturais essenciais definidos na arquitetura Transformer original (VASWANI et al., 2017):

- **Camadas de atenção multi-cabeça (*multi-head attention*):** Permitem que o modelo foque simultaneamente em diferentes partes e relações da sequência de entrada, capturando uma gama mais rica de informações contextuais.
- **Codificação posicional (*positional embeddings*):** Vetores que são adicionados aos *embeddings* dos *tokens* para injetar informação sobre a posição ou a ordem na sequência, uma vez que a arquitetura em si não processa os dados sequencialmente.
- **Redes *feed-forward*:** Camadas densamente conectadas, aplicadas após cada bloco de atenção, que realizam transformações não lineares sobre as representações, enriquecendo a capacidade do modelo.
- **Normalização e regularização:** Mecanismos como *Layer Normalization* e *Dropout*, aplicados dentro dos blocos para estabilizar o treinamento, acelerar a convergência e prevenir o sobreajuste (*overfitting*).

Embora compartilhem essa base arquitetural, as principais diferenças entre os modelos residem no número de parâmetros, nas estratégias de pré-treinamento e nas técnicas de ajuste fino. Esses fatores influenciam diretamente a capacidade de generalização, o desempenho e a eficiência de cada modelo em tarefas específicas de tipificação documental.

2.5.2 Desafios Técnicos na Aplicação de LLMs à Tipificação

Apesar dos avanços proporcionados pelos LLMs, sua aplicação prática em cenários de tipificação documental apresenta uma série de desafios técnicos e operacionais que demandam consideração cuidadosa:

- **Ambiguidade textual:** Textos curtos, mal estruturados ou com linguagem polissêmica dificultam a correta inferência de categorias, um desafio intrínseco à compreensão de linguagem natural (SUMANATHILAKA; MICALLEF; HOUGH, 2024).
- **Desbalanceamento de classes:** Em conjuntos de dados do mundo real, a distribuição desigual de documentos entre as categorias pode introduzir vieses no modelo, comprometendo sua performance em classes minoritárias (SANTOS et al., 2023).
- **Sensibilidade ao *prompt*:** A performance dos LLMs é altamente sensível à formulação das instruções textuais (*prompts*), onde pequenas variações podem resultar em predições significativamente diferentes (JIANG et al., 2023).
- **Custo computacional elevado:** Modelos de grande porte demandam recursos computacionais intensivos (memória e poder de processamento), tanto para inferência quanto para ajuste, o que pode limitar sua viabilidade e escalabilidade em aplicações práticas (FIELDS; CHOVANEC; MADIRAJU, 2024).
- **Limitações na generalização de domínio:** Modelos pré-treinados com dados genéricos da internet frequentemente apresentam dificuldade em lidar com a terminologia e o contexto de domínios altamente especializados, como documentos técnicos, jurídicos ou administrativos (WEI et al., 2023).

O reconhecimento e a mitigação de tais desafios são essenciais para a implantação bem-sucedida de LLMs em sistemas de produção. Adicionalmente, esses obstáculos servem como um motor para a pesquisa, impulsionando o desenvolvimento de estratégias híbridas e abordagens adaptativas que visam tornar os modelos mais robustos, interpretáveis e eficientes em ambientes diversos.

2.6 Estratégias de Pré-processamento para LLMs

O pré-processamento textual para arquiteturas baseadas em Grandes Modelos de Linguagem (LLMs) difere substancialmente das abordagens tradicionais. Enquanto métodos clássicos exigiam etapas intensivas como *stemming*, lematização e a remoção de *stopwords* para a construção de representações vetoriais (GASPARETTO et al., 2022), os LLMs modernos, com seus tokenizadores e mecanismos de atenção, internalizam grande parte dessa complexidade. No entanto, um conjunto distinto

de práticas de preparação de dados torna-se essencial para garantir a eficácia da inferência, a compatibilidade com os *prompts* e a coerência dos resultados.

As principais estratégias de pré-processamento no contexto de LLMs incluem:

- **Curadoria do Texto de Entrada:** Assegurar que os textos submetidos ao modelo estejam livres de ruídos que possam confundir a análise semântica, como artefatos de formatação (e.g., tags HTML), quebras de linha excessivas ou conteúdo repetitivo. Em contextos experimentais, é comum a truncagem de textos que excedem o limite de *tokens* da janela de contexto do modelo, uma vez que a capacidade de LLMs de processar informações em contextos muito longos pode degradar (LIU et al., 2024b).
- **Normalização de Rótulos:** Em tarefas de tipificação, é fundamental padronizar os rótulos das categorias (e.g., convertendo para minúsculas e removendo caracteres especiais). Essa etapa garante a correspondência exata entre os rótulos de treinamento/validação e as predições geradas pelo modelo, evitando erros de avaliação por inconsistências triviais.
- **Engenharia de *Prompt*:** A formulação das instruções fornecidas ao LLM é um dos fatores mais críticos para o seu desempenho. O *prompt* deve ser construído com instruções claras e inequívocas, contendo a descrição da tarefa e, quando aplicável, a lista explícita das categorias de saída desejadas. A literatura demonstra que a performance de um modelo pode variar drasticamente com pequenas alterações no formato do *prompt* (JIANG et al., 2023; LESTER; AL-RFOU; CONSTANT, 2021).

Embora os LLMs eliminem a necessidade de tokenização e extração de características manuais, a consistência dos dados de entrada permanece crucial. O alinhamento cuidadoso entre o conteúdo textual, a normalização dos rótulos e a engenharia de *prompt* constitui uma etapa crítica para maximizar a acurácia e a eficiência na tarefa de tipificação documental.

2.7 Avaliação de Modelos de tipificação de Texto

A avaliação de modelos de tipificação de texto é uma etapa indispensável no ciclo de desenvolvimento de sistemas de PLN, pois permite quantificar objetivamente a eficácia de diferentes abordagens. Em tarefas de tipificação supervisionada, a prática padrão consiste em comparar as categorias previstas pelo modelo com os rótulos verdadeiros, previamente atribuídos aos documentos, por meio de um conjunto de métricas de desempenho.

As métricas mais tradicionais, amplamente discutidas na literatura, incluem a **acurácia**, a **precisão**, a **revocação** e o **F1-Score**. Enquanto a acurácia oferece uma visão geral da proporção de acertos, a precisão mede a correção das predições positivas, a revocação avalia a capacidade do modelo de encontrar todos os exemplos

relevantes, e o F1-Score representa a média harmônica entre as duas últimas. Este conjunto de métricas é particularmente útil para diagnosticar o comportamento do modelo de forma mais granular do que a acurácia isoladamente (SOKOLOVA; LAPALME, 2009).

Para problemas de tipificação com múltiplas classes (*multiclass*) ou múltiplos rótulos (*multi-label*), como é o caso da tipificação documental, a análise se estende. Utilizam-se variações das métricas padrões, como as médias *macro* e *micro*, que agregam o desempenho por categoria de diferentes maneiras. Ferramentas visuais como a **matriz de confusão**, que detalha os tipos de erro entre as classes, e a **curva ROC** (com sua métrica correspondente, a **AUC**), que avalia a capacidade discriminativa de classificadores probabilísticos, são também de uso corrente e recomendadas para uma análise aprofundada (SOKOLOVA; LAPALME, 2009).

Com a ascensão dos LLMs, o escopo da avaliação foi expandido para incluir métricas que refletem as propriedades desses modelos. Medidas como a **perplexidade** e o **log-likelihood**, embora originárias de tarefas de modelagem de linguagem, são frequentemente empregadas para aferir a confiança ou a incerteza do modelo em suas predições. Adicionalmente, o **exact match score** é comumente utilizado em *benchmarks* contemporâneos para mensurar a taxa de acerto exato em tarefas de geração restrita, como a tipificação (CHANG et al., 2024).

A escolha de quais métricas utilizar depende diretamente dos objetivos da tarefa, da natureza do conjunto de dados — especialmente do equilíbrio entre as classes — e do custo associado a diferentes tipos de erro de tipificação. A literatura recente e a prática da área convergem para a recomendação de uma avaliação combinada e multifacetada. A utilização de um conjunto diverso de métricas, em vez de um único indicador, oferece uma visão mais holística e confiável sobre a eficácia, a robustez e a aplicabilidade de um modelo em cenários do mundo real.

2.8 Abordagens Tradicionais na Tipificação de Documentos

Antes do advento dos modelos de linguagem de grande porte, a tipificação de documentos era majoritariamente realizada por meio de um paradigma de duas etapas: extração de características (*feature extraction*) e aplicação de algoritmos clássicos de aprendizado de máquina supervisionado. Essas abordagens dependiam de uma representação numérica dos textos, que por sua vez era utilizada para treinar um modelo classificador.

A etapa de extração de características era central e geralmente envolvia técnicas como **Bag-of-Words (BoW)**, **n-gramas** ou **TF-IDF (Term Frequency-Inverse Document Frequency)**. Esses métodos transformavam os textos em representações vetoriais de alta dimensionalidade e esparsas, permitindo que os modelos aprendessem padrões estatísticos a partir da frequência e distribuição de termos no corpus

(GASPARETTO et al., 2022).

Sobre essas representações vetoriais, eram aplicados algoritmos de tipificação tradicionais. Entre os mais comuns para essa tarefa estavam o **Naive Bayes**, notável por sua simplicidade e eficiência computacional; o **Support Vector Machines (SVM)**, reconhecido por sua robustez em espaços de alta dimensionalidade; e algoritmos baseados em árvores de decisão, como o **Random Forest**, que oferecem bom desempenho em conjuntos de dados com ruído (KOWSARI et al., 2019).

Outras técnicas também foram exploradas, como redes neurais rasas, a exemplo dos **Multilayer Perceptrons (MLP)**, que capturavam relações não lineares nos dados, e modelos de tópicos, como **Latent Semantic Analysis (LSA)** e **Latent Dirichlet Allocation (LDA)**, empregados para identificar estruturas semânticas latentes nos textos (GASPARETTO et al., 2022; KOWSARI et al., 2019).

A principal limitação dessas abordagens tradicionais residia na sua limitada capacidade de compreensão semântica profunda. O desempenho dos modelos era altamente dependente da qualidade da engenharia de atributos e sensível a variações lexicais (sinonímia, polissemia) e sintáticas, o que restringia sua capacidade de generalização para domínios complexos ou textos com linguagem ambígua.

Ainda assim, essas técnicas não são obsoletas. Elas pavimentaram o caminho para os avanços contemporâneos em PLN e continuam a ser utilizadas como *baselines* importantes em estudos que avaliam novos modelos. Sua simplicidade, interpretabilidade e baixo custo computacional ainda as tornam atrativas em contextos com restrições de recursos ou quando a explicabilidade do modelo é um requisito fundamental.

2.9 Considerações Éticas e de Viés em Modelos de Linguagem

A aplicação de Grandes Modelos de Linguagem (LLMs) em tarefas de tipificação documental, embora tecnicamente eficaz, exige uma análise criteriosa de suas implicações éticas. É amplamente reconhecido na literatura que, apesar de seu desempenho, esses modelos podem refletir e até mesmo amplificar vieses sociais presentes nos dados massivos sobre os quais foram pré-treinados, o que arrisca comprometer a equidade e a imparcialidade das classificações geradas (YAO et al., 2024).

A origem desse desafio reside no próprio processo de treinamento dos LLMs. Eles são desenvolvidos a partir de vastos volumes de texto extraídos de fontes heterogêneas da internet, como artigos, fóruns e redes sociais. Como resultado, o conteúdo dessas bases de dados pode conter estereótipos de gênero e raça, desinformação, ou visões de mundo tendenciosas, que são inevitavelmente aprendidas e codificadas pelo modelo. Tal viés pode impactar diretamente tarefas de categorização documental, gerando resultados problemáticos em domínios sensíveis como o jurídico, o governamental ou o biomédico.

Além do viés algorítmico, destaca-se o desafio da interpretabilidade (ou o problema da "caixa-preta"). Em muitos casos, os LLMs operam de uma maneira que torna extremamente difícil explicar a lógica por trás de uma decisão específica. Essa falta de transparência limita a capacidade de auditar os processos de tipificação, levantando preocupações sobre a confiabilidade e a responsabilização (*accountability*) na adoção dessas ferramentas em ambientes institucionais (YAO et al., 2024).

Outro aspecto relevante envolve a privacidade e a segurança da informação. A utilização de LLMs, especialmente aqueles hospedados em APIs remotas, para processar documentos pode expor dados sensíveis ou pessoais. Essa prática representa riscos de conformidade com legislações de proteção de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil ou o General Data Protection Regulation (GDPR) na Europa. Portanto, a adoção de estratégias que garantam a anonimização e a proteção dos dados é um requisito essencial.

A mitigação desses problemas exige uma abordagem multifacetada, combinando soluções técnicas e organizacionais. Isso inclui o uso de filtros de conteúdo, a realização de auditorias de viés, a aplicação de técnicas de reponderação de amostras durante o ajuste fino e, de forma mais ampla, a incorporação de princípios de IA Ética (*Ethical AI*) desde a concepção até a implantação dos modelos. Avaliações sistemáticas de equidade e robustez são, portanto, recomendadas como parte integral do ciclo de vida de qualquer aplicação baseada em LLMs.

Em suma, o uso ético de modelos de linguagem na tipificação documental transcende a otimização de métricas de desempenho. É mandatório que se considerem os fatores sociais, legais e organizacionais para garantir que os sistemas desenvolvidos operem com justiça, segurança e responsabilidade.

2.10 Trabalhos Relacionados

A tipificação de documentos, uma tarefa fundamental em Processamento de Linguagem Natural (PLN), foi drasticamente transformada pelo advento dos Modelos de Linguagem de Grande Porte (LLMs). Esta seção revisa a trajetória das metodologias de avaliação de LLMs, partindo dos *benchmarks* de ajuste fino até as análises contemporâneas focadas nas capacidades e desafios do aprendizado no contexto (*in-context learning*).

A avaliação rigorosa de modelos de linguagem foi consolidada por *benchmarks* como o GLUE (WANG et al., 2018) e sua versão mais desafiadora, o SuperGLUE (WANG et al., 2019). Desenvolvidos na era de modelos como o BERT, que exigiam um ajuste fino (*fine-tuning*) para cada tarefa, esses conjuntos de dados estabeleceram um padrão para a avaliação de modelos em tarefas de tipificação. Sua principal limitação, contudo, era a dependência de grandes volumes de dados rotulados e o custo de treinar um novo modelo para cada tarefa específica.

Um desafio adicional que emergiu com a escala dos LLMs é o da **contaminação**

de dados nos *benchmarks*. Como os modelos são pré-treinados em trilhões de *tokens* extraídos da internet, há um risco crescente de que os dados de teste dos próprios *benchmarks* (incluindo GLUE e SuperGLUE) tenham sido inadvertidamente incluídos no corpus de treinamento. Esse "vazamento"de informação pode inflar artificialmente as métricas de desempenho, levando a uma avaliação otimista e pouco confiável da verdadeira capacidade de generalização do modelo. Pesquisas recentes buscam desenvolver métodos para detectar e mitigar essa contaminação, garantindo a integridade da avaliação científica (SAINZ et al., 2023).

A introdução de LLMs em grande escala, como o GPT-3, marcou uma mudança de paradigma, consolidada no trabalho seminal de Brown et al. (2020). Os autores demonstraram que LLMs massivos podem executar tarefas com poucas ou nenhuma amostra de treinamento (*few-shot* e *zero-shot learning*), utilizando apenas exemplos fornecidos no *prompt* de entrada. Este método, conhecido como aprendizado no contexto, eliminou a necessidade de atualização de gradientes, mas seu desempenho mostrou-se, em geral, inferior ao de modelos totalmente ajustados.

Investigações posteriores, no entanto, revelaram a **fragilidade do aprendizado no contexto**. O desempenho do modelo mostrou-se altamente sensível não apenas à formulação do *prompt*, mas também à seleção, ao formato e até mesmo à ordem dos exemplos fornecidos no cenário *few-shot*. Zhao et al. (2021) demonstraram que vieses na distribuição dos exemplos (e.g., se a maioria dos exemplos pertence a uma única classe) ou a ordem em que são apresentados podem enviesar drasticamente a predição do modelo, independentemente do texto a ser classificado. Isso indica que o aprendizado no contexto é menos um processo de aprendizado dinâmico e mais um mecanismo sofisticado de reconhecimento de padrões, cujos resultados podem ser instáveis.

Estudos subsequentes, como o de Jiang et al. (2023), focaram em avaliações sistemáticas dos LLMs modernos. Ao comparar o desempenho em cenários *zero-shot*, *few-shot* e de ajuste fino, os autores concluíram que modelos ajustados para seguir instruções (*instruction-tuned models*) superam consistentemente os LLMs padrão no modo *zero-shot* e que nenhum modelo se destaca universalmente em todas as tarefas, evidenciando uma forte variabilidade de desempenho.

Para mitigar os altos custos do ajuste fino completo, surgiram técnicas de ajuste eficientes em parâmetros (*parameter-efficient tuning*). Lester, Al-Rfou e Constant (2021) introduziram o *prompt tuning*, que congela os parâmetros do LLM e ajusta apenas um pequeno vetor contínuo. Simultaneamente, pesquisas sobre *instruction tuning* demonstraram que o ajuste fino de um LLM em uma coleção diversificada de tarefas melhora drasticamente sua capacidade de generalização para novas tarefas (WEI et al., 2021).

Finalmente, investigações sobre as limitações intrínsecas dos LLMs ganharam destaque. Um exemplo notável é o fenômeno "lost in the middle", identificado por Liu et al. (2024b), onde o desempenho de LLMs de ponta degrada significativamente quando a informação relevante está localizada no meio de um longo contexto de entrada.

Essa crescente conscientização sobre as limitações dos *benchmarks* e das métricas tradicionais impulsionou o desenvolvimento de **plataformas de avaliação mais holísticas**. Um exemplo proeminente é o *Holistic Evaluation of Language Models* (HELM) da Universidade de Stanford. Em vez de focar em algumas poucas métricas, o HELM avalia os modelos em um espectro muito mais amplo de cenários e métricas, cobrindo não apenas acurácia, mas também robustez, equidade (*fairness*), viés, toxicidade e eficiência. Essa abordagem multifacetada visa fornecer uma visão mais completa e transparente das verdadeiras capacidades e riscos de cada modelo, movendo o campo para além de uma simples "corrida"por pontuações em *leaderboards* (LIANG et al., 2022).

A literatura revisada demonstra, portanto, uma clara trajetória evolutiva, que vai além da simples medição de desempenho e passa a incorporar questões críticas sobre a confiabilidade dos *benchmarks*, a estabilidade dos métodos de inferência e a necessidade de uma avaliação mais completa e responsável dos modelos de linguagem.

É neste panorama consolidado e complexo que a presente dissertação se insere, com o objetivo de preencher uma lacuna entre as avaliações teóricas de larga escala e a necessidade de *benchmarks* pragmáticos. Diferentemente de estudos que avaliam dezenas de modelos ou focam em aspectos teóricos, este trabalho oferece três contribuições principais:

1. Conduz um estudo comparativo focado em um conjunto de oito modelos de código aberto, acessíveis via API e de alta relevância para implementações no mundo real.

2. Detalha e valida uma arquitetura de software reprodutível, servindo como um guia metodológico para a tipificação de documentos de forma assíncrona.

3. Gera evidência empírica detalhada e multimétrica, que vai além da acurácia e do F1-Score ao incorporar sistematicamente a perplexidade e o log-likelihood como medidas de confiança e incerteza do modelo.

Os resultados aqui apresentados fornecem um guia prático para a tomada de decisão, ao demonstrar a superioridade da família `llama3` no *dataset* avaliado, com acurácias de até 87.94%. Igualmente importante, o trabalho reforça empiricamente a conclusão de Jiang et al. (2023) sobre a variabilidade de desempenho, ao evidenciar a performance drasticamente inferior de modelos como o `llama3-2-latest` (33.10%) e o `deepseek-llm-7b` (52.27%) na mesma tarefa. Dessa forma, este estudo contribui com uma análise quantitativa rica e diretamente aplicável sobre a adequação de LLMs específicos para a tipificação documental.

Capítulo 3

Desenvolvimento e Implementação da Arquitetura de Tipificação Documental com LLMs

Este capítulo apresenta, de forma detalhada, o desenvolvimento e a implementação da arquitetura computacional projetada para a tipificação automatizada de documentos por meio de Grandes Modelos de Linguagem (LLMs). O objetivo central foi construir um *pipeline* de processamento robusto, escalável e reprodutível, integrando técnicas de Processamento de Linguagem Natural, inferência assíncrona via APIs e avaliação quantitativa de desempenho.

A seção a seguir descreve o modelo conceitual da arquitetura, detalhando as seis etapas que compõem o fluxo de trabalho, desde a ingestão dos dados até a geração dos resultados finais. Subsequentemente, cada uma dessas etapas será explorada em maior profundidade nas seções subsequentes deste capítulo.

3.1 Modelo da Arquitetura

A arquitetura implementada nesta pesquisa foi estruturada como um *pipeline* sequencial e modular, composto por seis etapas principais, conforme ilustrado na Figura 3.1. Essa estrutura foi concebida para garantir um fluxo de processamento coeso e preciso, facilitando a automação e a análise comparativa dos modelos. As etapas são:

1. **Aquisição dos Dados:** O processo inicia-se com a utilização de um conjunto de dados público e rotulado, o `bbc-text.csv`, que serve como base para o treinamento implícito (via *prompt*) e a validação do classificador.

2. **Pré-processamento e Codificação:** O conteúdo textual e os rótulos são submetidos a um tratamento para garantir a padronização e a integridade das entradas, incluindo a normalização e a codificação numérica das categorias.

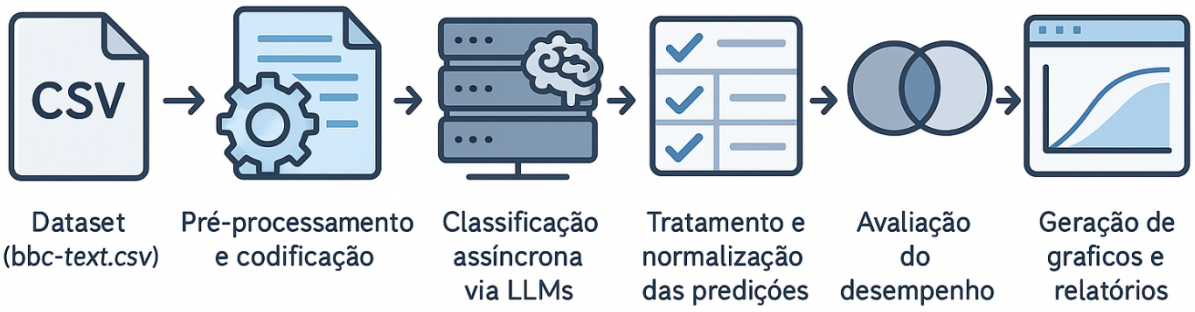
3. **tipificação Assíncrona via LLMs:** Os textos pré-processados são enviados de forma assíncrona para as APIs dos modelos de linguagem, que realizam a tarefa de tipificação. A natureza assíncrona do processo maximiza a eficiência e a escalabilidade da solução.

4. **Tratamento e Normalização das Predições:** As respostas (predições) dos LLMs são tratadas e normalizadas para garantir a compatibilidade com os rótulos originais do *dataset*, viabilizando a comparação.

5. **Avaliação de Desempenho:** O desempenho de cada modelo é avaliado quantitativamente por meio de métricas consolidadas, como acurácia, precisão, revocação e F1-Score, proporcionando uma análise objetiva da eficácia do *pipeline*.

6. **Geração de Relatórios e Visualizações:** Por fim, são gerados relatórios sintéticos e visualizações gráficas (e.g., matrizes de confusão, curvas ROC) que permitem uma interpretação clara dos resultados e a análise exploratória dos erros.

Figura 3.1: Arquitetura de tipificação de documentos em seis etapas.



Fonte: Elaborado pelo autor.

3.2 Dataset

A validação e a experimentação da arquitetura proposta foram realizadas com base no **BBC News Dataset**, um conjunto de dados público e amplamente referenciado em estudos de tipificação textual (GREENE; CUNNINGHAM, 2006). O *dataset*, disponibilizado no formato `bbc-text.csv`, é composto por 2.225 artigos jornalísticos extraídos do portal da BBC entre 2004 e 2005, onde cada instância contém o texto completo do artigo e um rótulo de categoria correspondente.

As notícias estão distribuídas em cinco categorias temáticas bem definidas: *business*, *entertainment*, *politics*, *sport* e *tech*. Conforme detalhado na Tabela 3.1, a distribuição das amostras é razoavelmente equilibrada, sem a presença de classes com representação drasticamente minoritária, o que favorece uma análise de desempenho mais justa e direta. Essa clareza na estrutura de classes, aliada à diversidade

semântica dos textos, proporciona um cenário robusto para a avaliação de modelos de tipificação.

A escolha deste *dataset* é justificada por múltiplas razões metodológicas. Primeiramente, sua ampla utilização na literatura facilita a replicação de experimentos e a comparação de resultados com outras abordagens. Em segundo lugar, a simplicidade estrutural do arquivo — contendo apenas duas colunas (texto e rótulo) — otimiza a etapa de pré-processamento em *pipelines* automatizados. Por fim, a disponibilidade de rótulos de alta qualidade permite a aplicação direta de métricas de avaliação supervisionada, fundamentais para a análise quantitativa do desempenho dos modelos.

Outro fator relevante é a abrangência temática do corpus, que inclui desde notícias factuais sobre negócios e tecnologia até textos mais opinativos sobre política e entretenimento. Essa heterogeneidade textual permite analisar a capacidade de generalização dos modelos de linguagem frente a diferentes estilos, vocabulários e contextos comunicativos, ampliando a validade dos resultados obtidos.

Tabela 3.1: Distribuição das categorias no BBC News Dataset

Categoria	Quantidade de documentos
business	510
entertainment	386
politics	417
sport	511
tech	401
Total	2.225

3.3 Pré-processamento e Codificação

A etapa de pré-processamento e codificação é uma fase fundamental do *pipeline* proposto, sendo responsável por preparar os dados brutos para a tipificação pelos LLMs e para a subsequente avaliação quantitativa. O processo inicia-se com a extração dos textos e dos rótulos categóricos do arquivo `bbc-text.csv`. Para garantir a consistência, a primeira ação de normalização é a conversão de todos os rótulos para letras minúsculas.

Após a normalização, os rótulos textuais são transformados em representações numéricas, um requisito para o cálculo da maioria das métricas de desempenho. Para isso, empregam-se duas técnicas de codificação padrão, implementadas com o auxílio da biblioteca **Scikit-learn** (PEDREGOSA et al., 2011):

- **Codificação de Rótulos (*Label Encoding*):** Técnica que associa um identificador inteiro único a cada categoria textual (e.g., 'business' $\rightarrow$ 0, 'tech' $\rightarrow$ 4). Essa representação é utilizada para a maior parte dos cálculos de métricas, como acurácia e F1-Score.

- **Codificação *One-Hot* (*One-Hot Encoding*):** Adicionalmente, os rótulos inteiros são convertidos para um formato vetorial binário. Nesta representação, cada categoria é um vetor onde apenas o índice correspondente à classe é 1 e os demais são 0. Este formato é necessário para análises mais específicas, como o cálculo da curva ROC por classe.

3.4 tipificação Assíncrona via LLMs

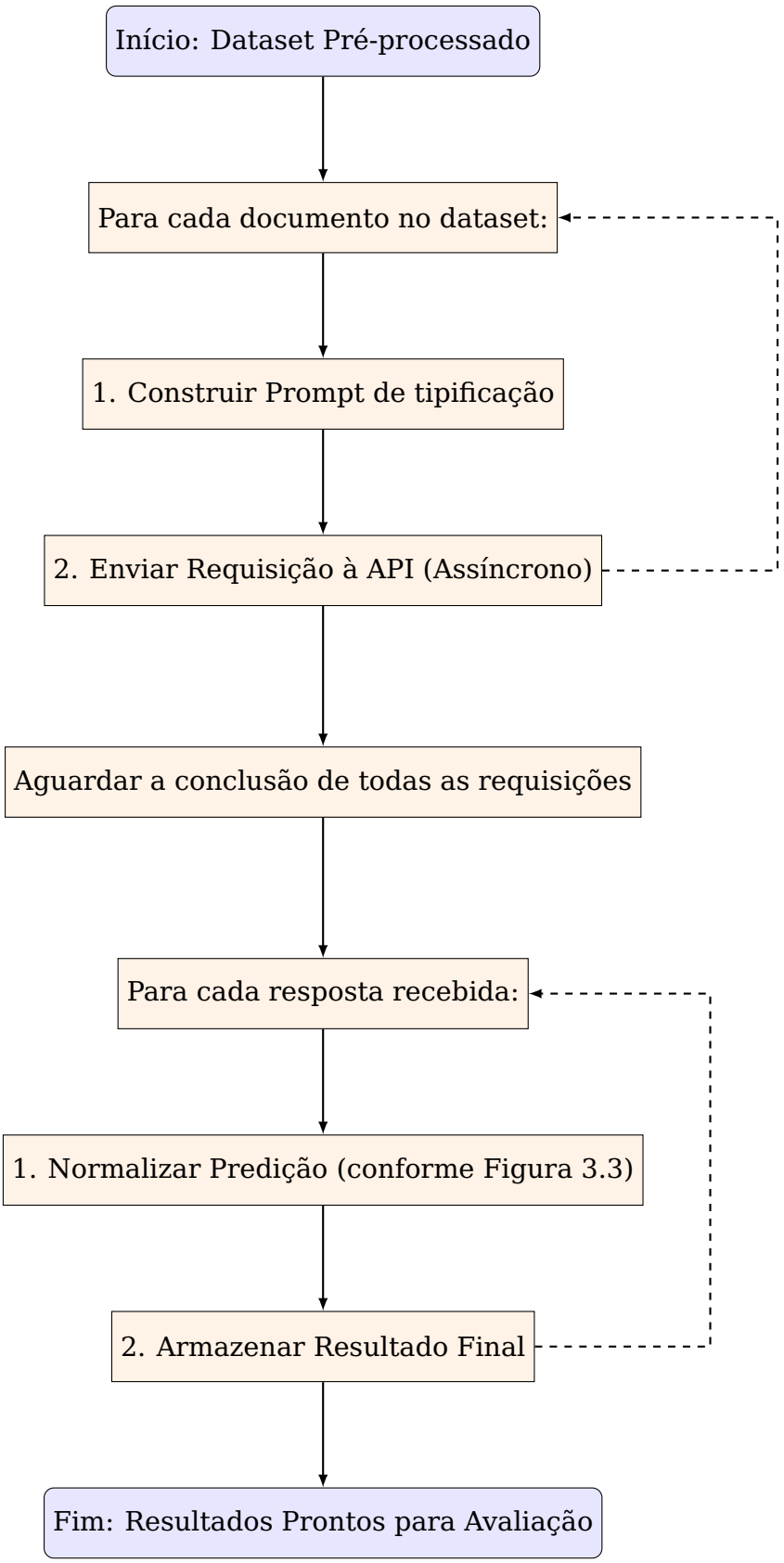
A etapa de tipificação constitui o núcleo operacional da arquitetura de tipificação documental. Nesta fase, os textos pré-processados são submetidos aos Grandes Modelos de Linguagem (LLMs) por meio de chamadas a uma Interface de Programação de Aplicações (API), que retorna a categoria mais apropriada para cada instância textual.

Uma decisão de projeto fundamental nesta arquitetura foi a implementação de chamadas de API de forma **assíncrona**. Diferentemente de um modelo síncrono, que processa uma requisição de cada vez e bloqueia a execução até obter a resposta, a abordagem assíncrona permite que múltiplas requisições sejam enviadas em paralelo. Enquanto o sistema aguarda a resposta de uma chamada (uma operação inerentemente limitada por I/O, ou *Input/Output-bound*), ele pode iniciar o processamento de outras. Essa capacidade de concorrência otimiza drasticamente o tempo de execução total e garante maior escalabilidade, sendo uma prática recomendada por *frameworks* de alta performance como o **FastAPI** (RAMíREZ, 2024).

O fluxo operacional para cada documento envolve a construção de um *prompt* estruturado, que instrui o modelo a realizar a tarefa de tipificação de acordo com as categorias previamente definidas. Após o envio assíncrono da requisição, o sistema coleta a resposta da API, que contém a categoria prevista. Essa resposta é então submetida a uma etapa de normalização (descrita na próxima seção) para garantir a compatibilidade com os rótulos originais, e o resultado final é armazenado para a fase de avaliação. Este fluxo é ilustrado na Figura 3.2.

Essa abordagem baseada em API oferece uma arquitetura flexível e extensível. A integração de novos modelos — como LLaMA, Mistral ou DeepSeek — pode ser realizada com modificações mínimas no *pipeline*, bastando adaptar a estrutura da chamada à API específica. Isso permite uma experimentação contínua com novos LLMs e estratégias de tipificação de forma ágil e modular.

Figura 3.2: Fluxo da Etapa de tipificação Assíncrona via LLMs



Fonte: Elaborado pelo autor.

3.5 Tratamento e Normalização das Predições

Uma vez que um LLM retorna uma predição textual, é indispensável uma etapa de pós-processamento para garantir que a saída do modelo seja padronizada e possa ser comparada de forma confiável com os rótulos verdadeiros do *dataset*. Esta fase de tratamento e normalização é crucial para a integridade da avaliação quantitativa.

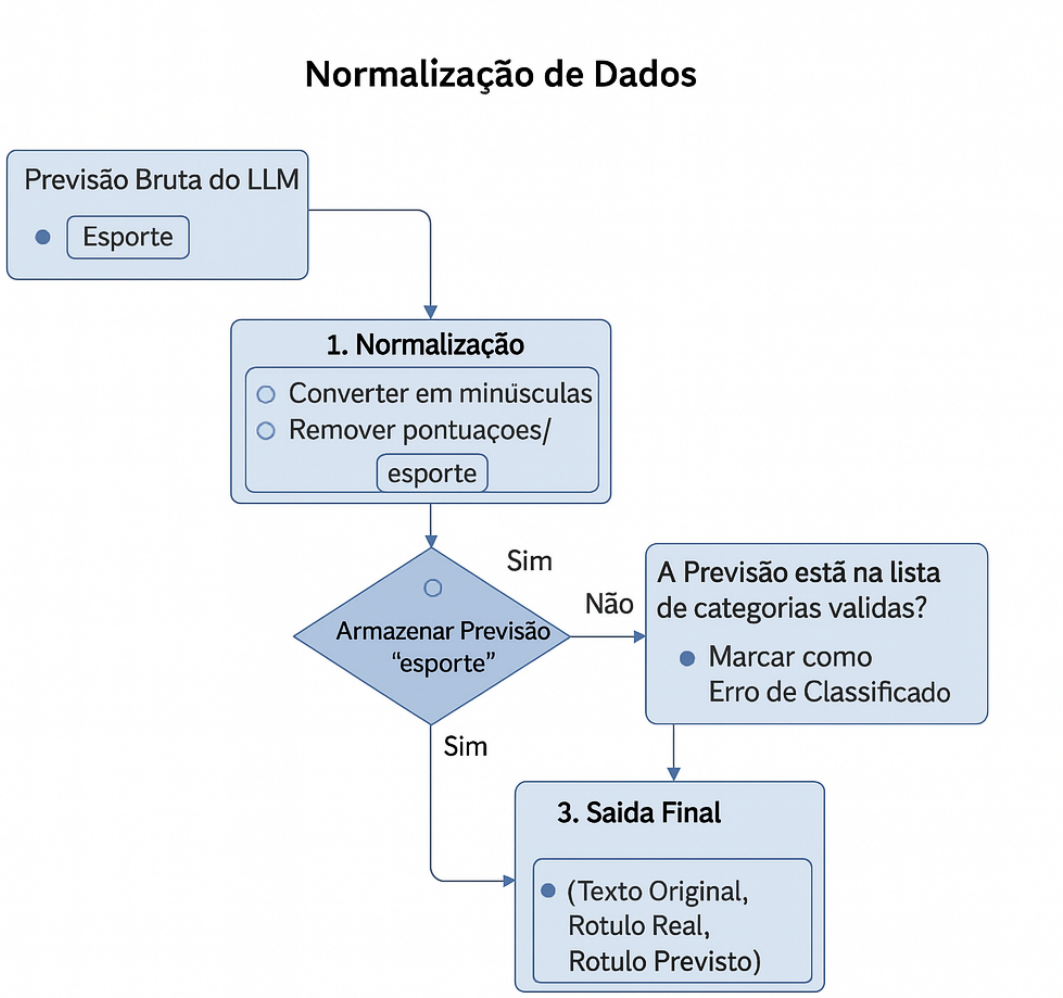
O processo é conduzido em duas fases principais:

1. **Normalização Textual da Predição:** A saída bruta da API do LLM pode conter variações indesejadas, como capitalização inconsistente, espaços em branco excedentes, ou a inclusão de pontuação e outros caracteres (e.g., aspas). A primeira ação é, portanto, limpar essa saída, convertendo-a para letras minúsculas e removendo todos os caracteres não alfanuméricos. O objetivo é assegurar que uma predição como `"Sport."` seja tratada da mesma forma que o rótulo `sport`.

2. **Validação da Categoria:** Após a normalização, a predição limpa é validada contra o conjunto de rótulos válidos previamente definidos (i.e., *business, entertainment, politics, sport, tech*). Se a predição corresponder a uma das categorias válidas, ela é registrada como a tipificação final do modelo. Caso contrário — se o modelo gerar uma resposta inválida ou fora do escopo (e.g., uma frase, uma recusa em responder ou uma categoria inexistente) —, a instância é computada como um erro de tipificação.

Ao final desta etapa, o conjunto de dados é enriquecido com as predições validadas, resultando em uma estrutura de dados unificada que contém o texto original, o rótulo real e a predição final do modelo. Essa organização é o pré-requisito para a fase subsequente de avaliação de desempenho, pois permite uma comparação direta e inequívoca entre os valores esperados e os obtidos. O fluxo completo deste processo é visualizado na Figura 3.3.

Figura 3.3: Fluxo de tratamento e normalização das predições.



Fonte: Elaborado pelo autor.

3.6 Avaliação de Desempenho

A eficácia da arquitetura de tipificação foi aferida por meio de uma avaliação quantitativa, comparando os rótulos preditos pelos modelos de linguagem com os rótulos reais do conjunto de dados. Para essa análise, foram empregadas métricas clássicas da literatura de tipificação supervisionada, implementadas com o auxílio da biblioteca **Scikit-learn** (PEDREGOSA et al., 2011).

O conjunto primário de métricas utilizado para avaliar o desempenho geral e por classe inclui:

- **Acurácia (Accuracy):** A proporção geral de documentos classificados corretamente sobre o total de amostras.

- **Precisão (Precision):** A habilidade do classificador de não rotular como positiva uma amostra que é negativa. Para uma dada classe, é a razão entre os verdadeiros positivos e a soma de verdadeiros e falsos positivos.
- **Revocação (Recall):** A habilidade do classificador de encontrar todas as amostras positivas. Para uma dada classe, é a razão entre os verdadeiros positivos e a soma de verdadeiros positivos e falsos negativos.
- **F1-Score:** A média harmônica ponderada entre a precisão e a revocação, útil para comparar modelos de forma equilibrada, especialmente em cenários com desbalanceamento de classes.

A teoria e a aplicabilidade dessas métricas para tarefas de tipificação são extensivamente analisadas em trabalhos como o de Sokolova e Lapalme (2009).

Além das métricas de pontuação, a capacidade discriminativa de cada modelo foi investigada visualmente por meio da **Curva ROC (Receiver Operating Characteristic)** e de sua métrica escalar correspondente, a **Área Sob a Curva (AUC - Area Under the Curve)**. A curva ROC plota a taxa de verdadeiros positivos contra a taxa de falsos positivos em vários limiares de decisão. Um valor de AUC próximo a 1 indica um modelo com excelente poder de separação entre as classes. A construção dessas curvas, a partir da codificação *one-hot* das categorias, também foi suportada pela biblioteca Scikit-learn.

Toda a geração de relatórios numéricos e visualizações gráficas, como as curvas ROC e outros gráficos apresentados no Capítulo 4, foi realizada utilizando as bibliotecas de visualização de dados **Matplotlib** (HUNTER, 2007) e **Seaborn** (WASKOM, 2021).

3.7 Geração de Gráficos e Relatórios

A etapa final do *pipeline* metodológico consiste na geração automatizada de relatórios e visualizações, cujo objetivo é organizar, comunicar e facilitar a interpretação dos resultados da tipificação. Essa fase é fundamental para traduzir os dados quantitativos em insights compreensíveis sobre o desempenho de cada modelo.

Para cada modelo avaliado, o sistema produz dois tipos principais de artefatos:

1. **Relatórios Numéricos:** Arquivos de texto estruturados contendo as métricas de desempenho detalhadas, como acurácia, precisão, recall e F1-Score, calculadas tanto de forma geral quanto por categoria. Essa documentação serve como um registro transparente para a validação e a reprodutibilidade dos experimentos.
2. **Visualizações Gráficas:** Um conjunto de gráficos gerado para permitir uma análise visual intuitiva. Isso inclui gráficos de barras, que comparam as métri-

cas de desempenho entre as diferentes classes, e curvas ROC (com seus respectivos valores de AUC), que ilustram a capacidade discriminativa do modelo. Essas visualizações são essenciais para identificar rapidamente padrões, tendências e as classes onde cada modelo apresenta maior ou menor dificuldade.

Toda a geração de gráficos e figuras foi implementada em Python, com o suporte das bibliotecas de visualização de dados **Matplotlib** (HUNTER, 2007) e **Seaborn** (WASKOM, 2021). Os artefatos visuais foram salvos em arquivos de imagem e organizados em diretórios específicos para cada modelo, garantindo a rastreabilidade. A automação desta etapa assegura não apenas a consistência e transparência dos experimentos, mas também fornece os recursos documentais indispensáveis para a análise crítica dos resultados apresentados no capítulo seguinte.

Capítulo 4

Resultados

Este capítulo apresenta os resultados empíricos obtidos na avaliação comparativa dos diferentes modelos de linguagem. A análise foi conduzida com base no conjunto de métricas estabelecido na metodologia — incluindo acurácia, precisão, recall, F1-score, perplexidade e log likelihood — com o objetivo de identificar as forças e limitações de cada modelo.

A discussão está organizada em duas partes. Primeiramente, é apresentada uma análise macro na Tabela 4.1, que consolida e compara o desempenho geral dos oito modelos na tarefa de tipificação documental. Esta visão panorâmica serve para estabelecer uma hierarquia de eficácia e destacar as variações mais significativas de performance. Em seguida, o capítulo se aprofunda no desempenho de cada modelo individualmente, explorando suas particularidades e padrões de erro em cada uma das cinco categorias do dataset.

Tabela 4.1: Comparativo das Métricas de Desempenho dos Modelos Avaliados

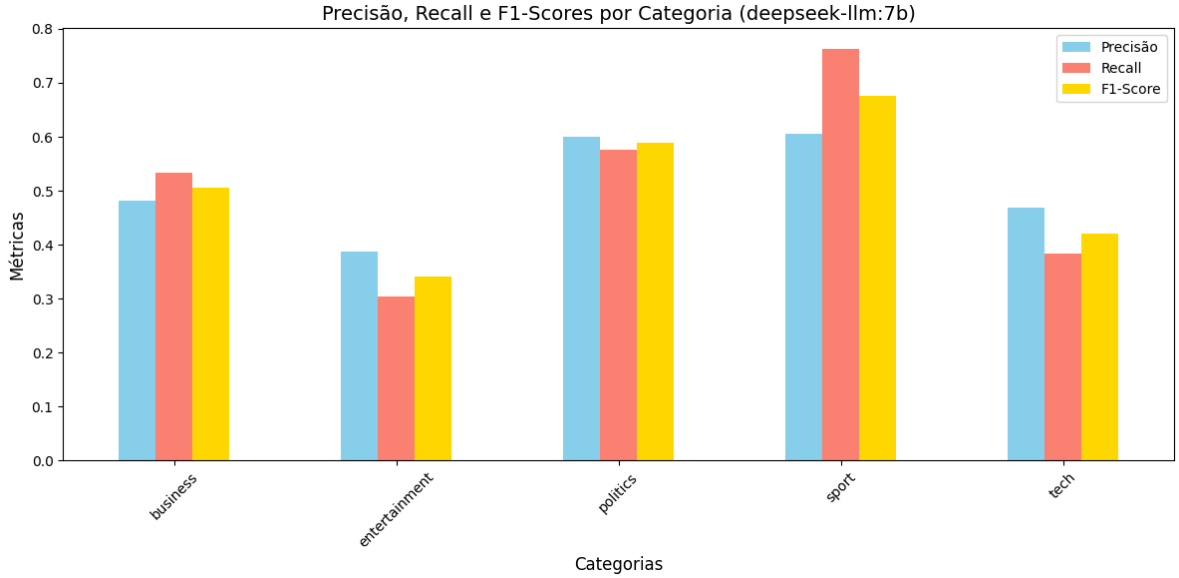
Modelo	**Acurácia**	**F1-score**	**Perplexidade**	**Log Likelihood**	**Exact Match**
`llama3`	0.8794	0.8802	16.0519	-6631.4451	0.8794
`llama3.1:8b`	0.8726	0.8735	18.2293	-6889.4518	0.8726
`llama3.1:latest`	0.8786	0.8793	15.4412	-6432.7784	0.8786
`llama3.2:latest`	0.3310	0.3262	4894258.7338	-33456.5614	0.3310
`gemma:7b`	0.8347	0.8365	45.0102	-9026.1336	0.8347
`deepseek-llm:7b`	0.5227	0.5163	59240.1483	-22473.2305	0.5227
`mistral-nemo:latest`	0.8634	0.8647	23.2120	-7575.5050	0.8634
`mistral:7b`	0.6881	0.6717	1314.3538	-14276.0276	0.6881

Fonte: Elaborado pelo autor.

O modelo `deepseek-llm:7b` apresentou um desempenho razoável, com uma acurácia geral de 52.27%. A precisão foi mais consistente nas categorias Politics” (0.6006) e Sport” (0.6052), com F1-scores de 0.5882 e 0.6750, respectivamente. No entanto, o modelo teve dificuldades nas categorias Entertainment” (F1-score de 0.3406) e Tech” (F1-score de 0.4213), refletindo uma necessidade de melhoria, principalmente na habilidade de capturar características específicas dessas categorias.

A perplexidade elevada de 59240.1483 indica que o modelo pode ter dificuldades na generalização, o que pode ser um ponto importante a ser otimizado. A Figura 4.1 apresenta as métricas de desempenho do modelo para cada categoria, destacando essas variações nos resultados.

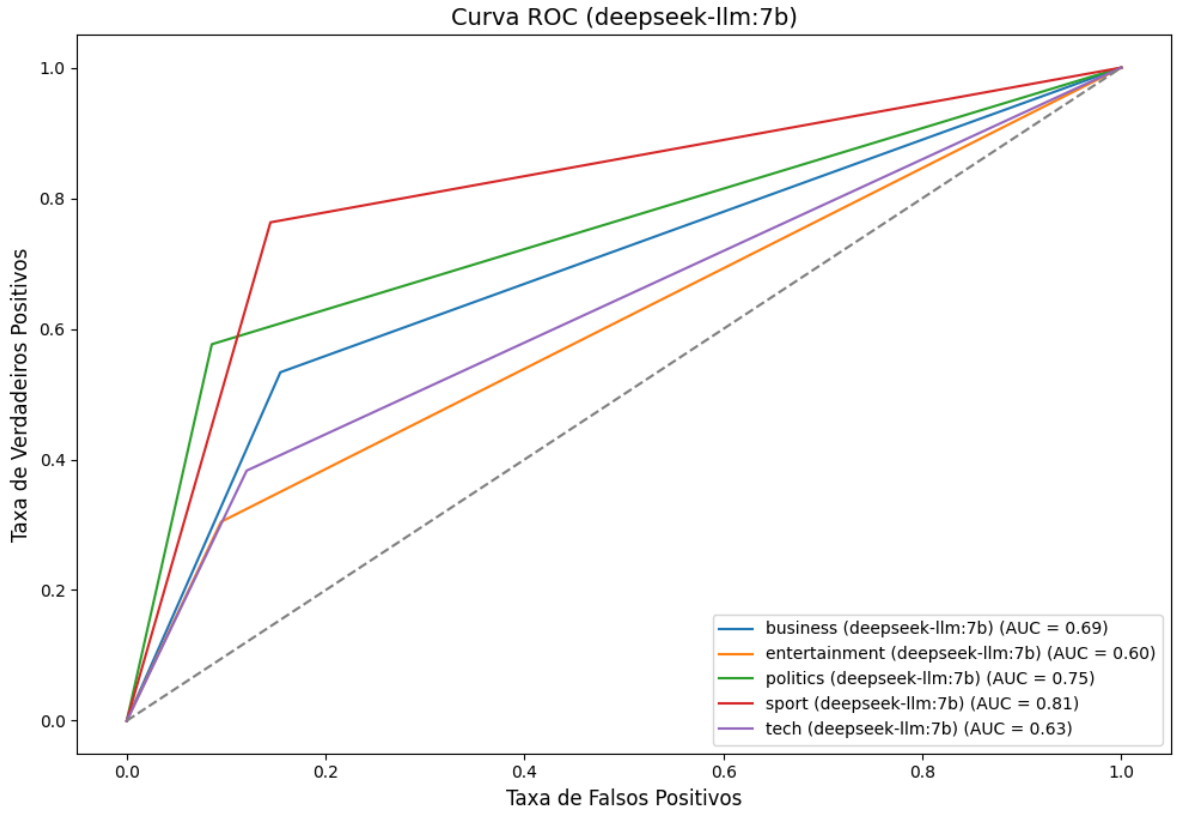
Figura 4.1: Acurácia, precisão e F1-score do modelo `deepseek-llm:7b` por categoria.



Fonte: Próprio autor.

A Figura 4.2 apresenta a Curva ROC (Receiver Operating Characteristic) para o modelo `deepseek-llm:7b`, que ilustra o equilíbrio entre a taxa de verdadeiros positivos (True Positive Rate - TPR) e a taxa de falsos positivos (False Positive Rate - FPR) para cada categoria. Uma curva mais próxima do canto superior esquerdo indica um melhor desempenho do modelo, pois significa alta detecção de verdadeiros positivos com poucos falsos positivos. No entanto, a performance do modelo varia entre as classes, refletindo as discrepâncias observadas nas métricas de recall. Categorias como *Sport* e *Politics*, que apresentaram maiores valores de recall, tendem a ter curvas ROC mais favoráveis, enquanto *Entertainment* e *Tech*, que tiveram baixos recalls, devem exibir curvas menos expressivas, indicando dificuldades do modelo em distinguir corretamente essas classes.

Figura 4.2: Curva ROC do modelo `deepseek-llm:7b` por categoria.



Fonte: Próprio autor.

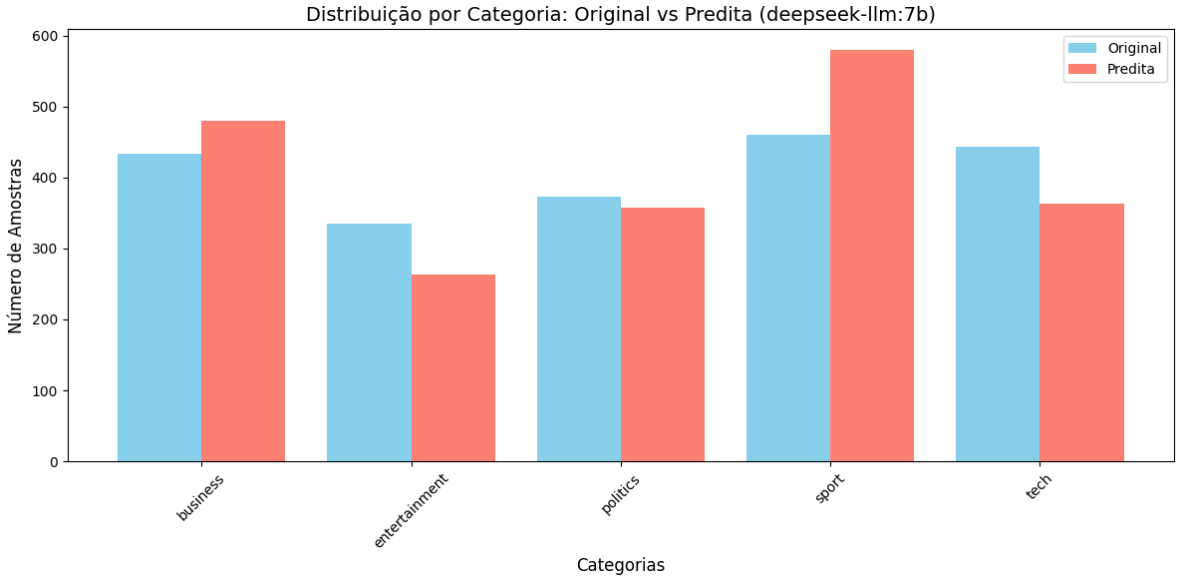
A Figura 4.3 ilustra o gráfico de barras que compara as categorias originais com as preditas pelo modelo `deepseek-llm:7b`. Este gráfico oferece uma visão clara sobre o desempenho do modelo ao comparar o número de documentos classificados corretamente (barras alinhadas) com aqueles classificados incorretamente (barras desalinhadas). A categoria Sport, que apresentou o maior recall (0.7630), mostra um alto número de acertos nas previsões, refletindo o bom desempenho do modelo ao distinguir documentos dessa classe. Já a categoria Politics, com recall de 0.5764, apresenta uma diferença moderada entre as barras originais e preditas, sugerindo que o modelo tem um desempenho razoável, mas ainda com alguns erros de tipificação.

Por outro lado, as categorias Entertainment e Tech exibem uma discrepância significativa entre os documentos originais e os preditos, indicando que o modelo falha frequentemente em distinguir esses tipos de documentos. As barras desalinhadas para essas classes sugerem que um número considerável de documentos dessas categorias foi erroneamente classificado em outras classes. Essa situação é corroborada pelos baixos valores de recall dessas categorias, 0.3045 e 0.3829, respectivamente, o que implica uma maior dificuldade do modelo em capturar as características distintivas desses documentos.

Esse gráfico fornece uma avaliação intuitiva das fraquezas do modelo em termos de acertos e erros de tipificação, permitindo uma compreensão visual dos desafios

enfrentados, especialmente nas categorias com menor recall. A necessidade de melhorar a generalização do modelo em relação a essas classes é evidente, e estratégias como ajuste de hiperparâmetros ou aumento de dados específicos podem ser necessárias para mitigar esses erros.

Figura 4.3: Distribuição por Categoria do modelo `deepseek-llm:7b`



Fonte: Próprio autor.

A avaliação do modelo `deepseek-llm:7b` também considera outras métricas de performance que fornecem insights adicionais sobre sua capacidade de generalização e precisão. A **Perplexidade** do modelo foi calculada em 59240.1483, o que sugere um grau elevado de incerteza ou complexidade nas previsões do modelo. Em termos de modelagem de linguagem, a perplexidade é uma medida de quão bem o modelo prevê uma amostra. Quanto menor a perplexidade, melhor o modelo é capaz de prever a sequência de palavras, refletindo um ajuste mais preciso aos dados de treinamento. A alta perplexidade observada aqui pode indicar uma dificuldade do modelo em generalizar bem para os dados testados, possivelmente devido à complexidade ou à falta de dados representativos.

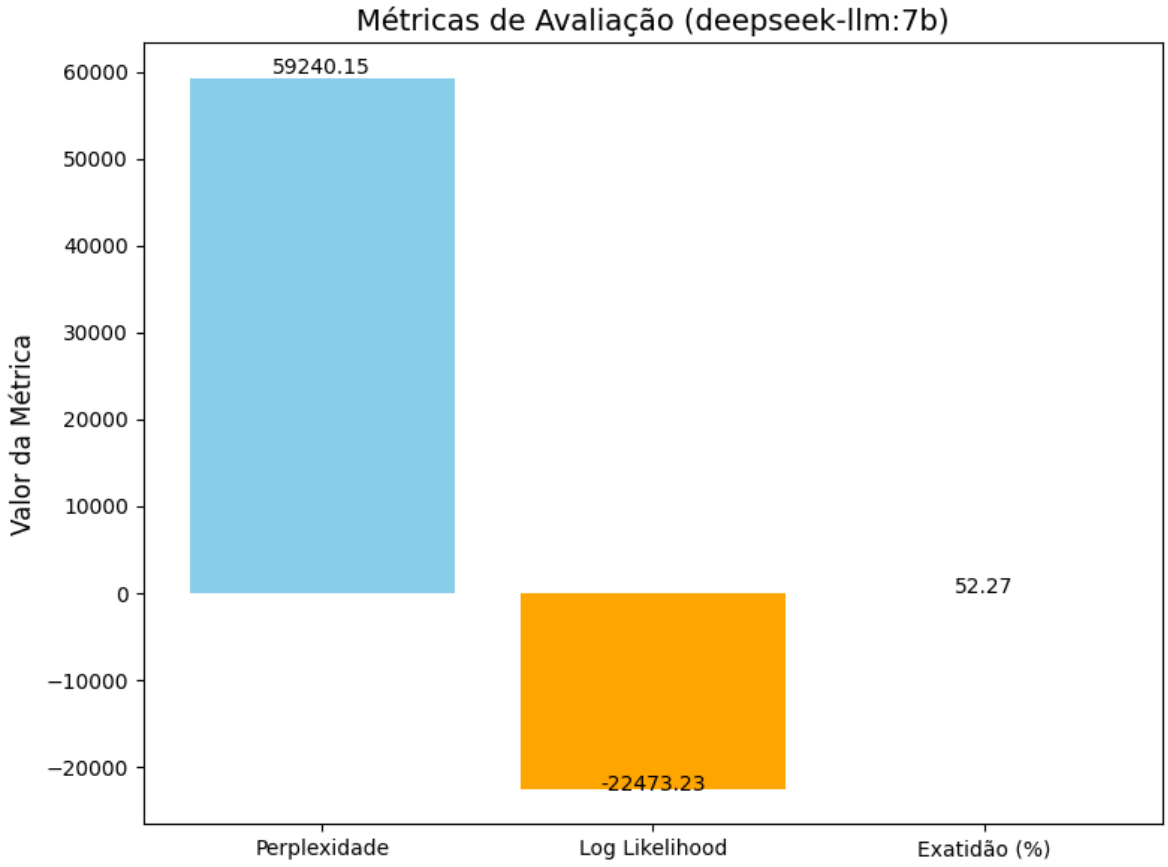
O **Log Likelihood Total** do modelo foi de -22473.2305, o que representa a soma dos logaritmos das probabilidades atribuídas às sequências de teste. Valores mais altos de log likelihood indicam que o modelo é mais confiante em suas previsões. Um valor negativo, como o observado aqui, não é incomum em tarefas de modelagem de linguagem, embora o valor específico dependa do contexto e da escala do modelo. Essa métrica oferece uma visão adicional sobre a capacidade do modelo de prever os dados de teste, complementando a análise da perplexidade.

Por fim, a **Exatidão (Exact Match)** foi de 0.5227, indicando que o modelo conseguiu corresponder corretamente, em média, 52.27% das predições para as categorias do conjunto de teste. Embora essa exatidão não seja particularmente alta, ela reflete uma capacidade razoável do modelo em realizar predições precisas em um contexto

de múltiplas categorias, sendo importante notar que outras métricas como F1-score, precisão e recall fornecem uma compreensão mais detalhada sobre o desempenho do modelo, especialmente quando se trata de desequilíbrio de classes.

A Figura 4.4 apresenta um gráfico que ilustra as métricas de desempenho do modelo `deepseek-llm:7b`, complementando as métricas discutidas acima.

Figura 4.4: Gráfico de métricas para o modelo `deepseek-llm:7b`.

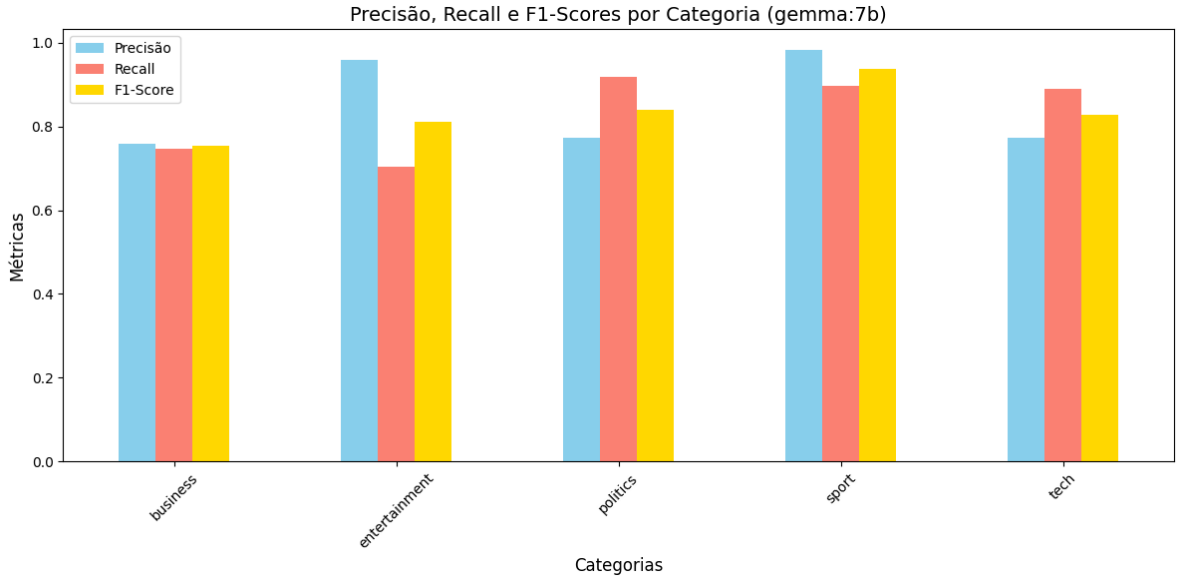


Fonte: Próprio autor.

4.0.1 Modelo gemma:7b

O modelo `gemma:7b` obteve uma melhoria significativa, com acurácia de 83.47%, refletindo uma maior capacidade de identificar corretamente as categorias. As categorias “Sport” e “Entertainment” foram os destaques, com F1-scores de 0.9383 e 0.8121, respectivamente. A precisão e recall também se destacaram nessas categorias, com valores muito altos. A categoria “Politics” apresentou bons resultados com F1-score de 0.8390, embora a categoria “Business” pudesse ser otimizada, apresentando uma precisão de 0.7589 e um F1-score de 0.7529. A perplexidade de 45.0102 e o log likelihood de -9026.1336 indicam que o modelo, apesar de seu bom desempenho geral, ainda pode ser aprimorado para gerar melhores resultados em categorias mais desafiadoras.

Figura 4.5: Acurácia, precisão e F1-score do modelo `gemma:7b` por categoria.

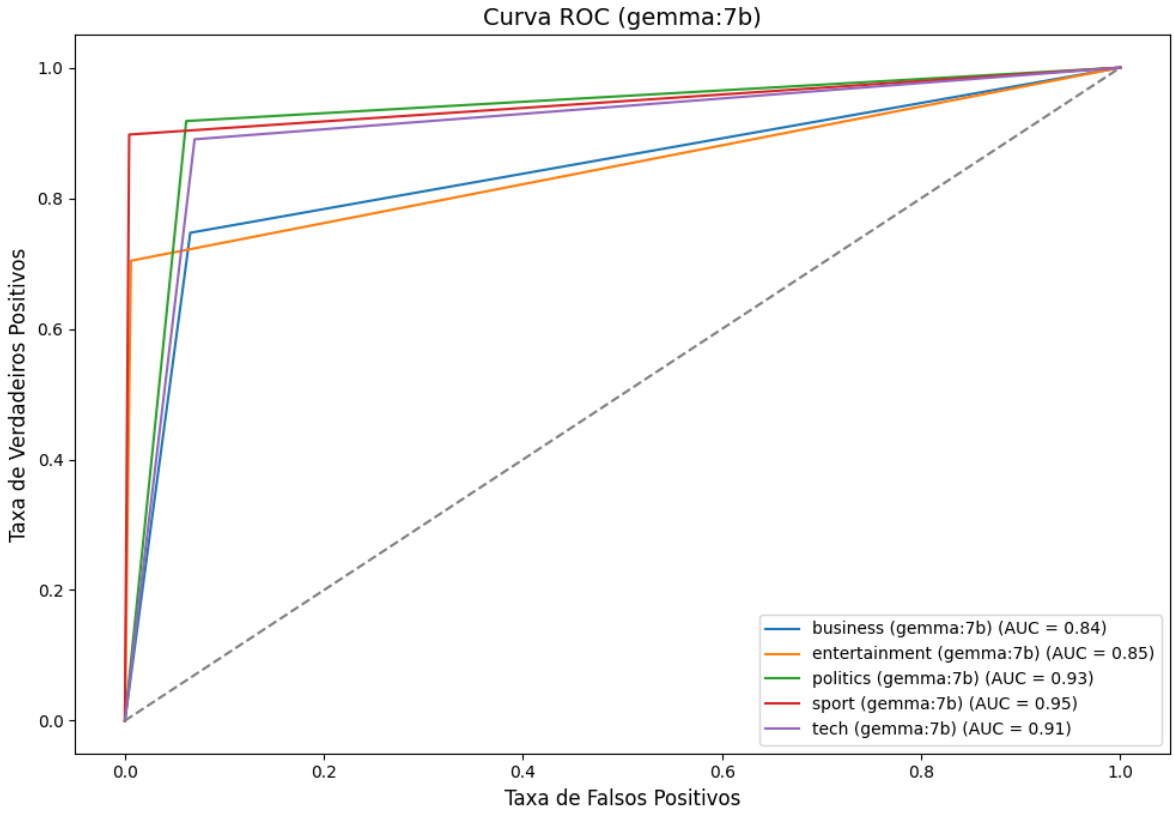


Fonte: Próprio autor.

A Curva ROC apresentada para o modelo `gemma:7b` ilustra a performance do modelo ao avaliar a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR) para cada categoria. A TPR, também conhecida como sensibilidade ou recall, representa a proporção de instâncias positivas corretamente identificadas pelo modelo, enquanto a FPR quantifica o erro de classificar erroneamente instâncias negativas como positivas. Uma curva ROC mais próxima do canto superior esquerdo do gráfico indica um melhor desempenho, pois reflete uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos. Como mostrado na Figura 4.6, categorias como "Sport"e "Politics", com recalls elevados de 0.8975 e 0.9182, respectivamente, provavelmente exibirão curvas ROC mais favoráveis, pois o modelo consegue identificar corretamente a maioria das instâncias dessas classes. Por outro lado, categorias como "Entertainment"e "Tech", com recall um pouco mais baixo, deverão mostrar uma separação menos acentuada entre os positivos e negativos, o que se refletirá em uma curva ROC mais próxima da linha de aleatoriedade.

As métricas gerais de desempenho também corroboram a interpretação visual da curva ROC. A precisão, recall e F1-score elevados, especialmente para categorias como "Sport"e "Politics", indicam que o modelo tem um bom desempenho na identificação correta dos documentos dessas classes. Contudo, a presença de discrepâncias no recall entre as categorias sugere que o modelo pode ter maior facilidade para classificar corretamente algumas classes, enquanto apresenta dificuldades em distinguir outros tipos de documentos.

Figura 4.6: Curva ROC para o modelo `gemma:7b`.



Fonte: Próprio autor.

A Figura 4.7 apresenta a distribuição das categorias reais em comparação com as categorias preditas pelo modelo `gemma:7b`. Esse gráfico de barras possibilita uma visualização clara da correspondência entre as classes originais e as inferências realizadas pelo modelo, evidenciando padrões de acerto e erro na tipificação.

Observa-se que as categorias **Business** e **Sport** apresentam uma forte correlação entre os valores reais e preditos, indicando um desempenho consistente na identificação dessas classes. A maior parte das amostras pertencentes a essas categorias foi corretamente classificada, o que sugere que o modelo conseguiu captar características distintas desses grupos de textos.

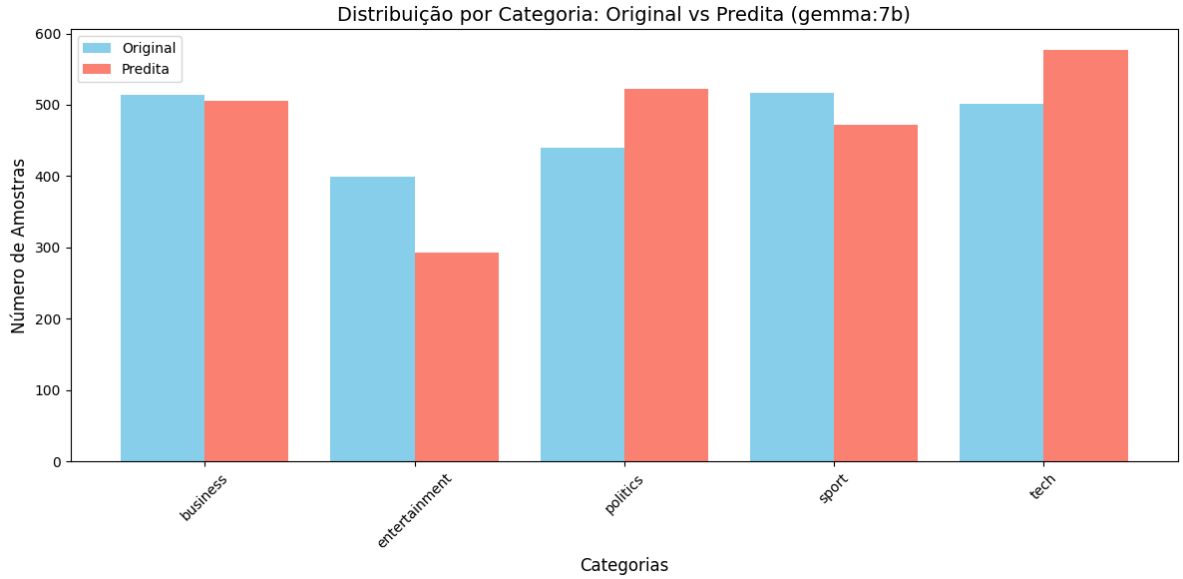
Por outro lado, a categoria **Entertainment** exibe uma discrepância mais acentuada entre os valores reais e preditos, sugerindo uma taxa maior de classificações incorretas. Isso pode indicar que o modelo apresenta dificuldades em diferenciar essa classe de outras, possivelmente devido a sobreposições semânticas com categorias como *Business* e *Tech*.

Além disso, percebe-se que a distribuição das categorias preditas mantém um padrão próximo ao das categorias reais, sugerindo que o modelo possui um comportamento geral equilibrado na tipificação. Entretanto, a presença de desbalanceamentos em algumas classes pode indicar a necessidade de ajustes, seja no conjunto de dados ou na arquitetura do modelo, para otimizar a precisão da categorização em

casos mais ambíguos.

A análise visual da Figura 4.7 reforça a importância de uma avaliação detalhada por categoria, permitindo identificar quais classes apresentam maior confiabilidade na predição e quais requerem aprimoramento para uma tipificação mais precisa.

Figura 4.7: Distribuição por Categoria do modelo `gemma:7b`.



Fonte: Próprio autor.

A Figura 4.8 ilustra as principais métricas de desempenho do modelo `gemma:7b`, incluindo perplexidade, log-likelihood total e exatidão (*Exact Match*). Esses indicadores fornecem uma visão quantitativa da capacidade do modelo em realizar classificações precisas dentro do conjunto de categorias avaliadas.

A **perplexidade** do modelo foi de 45.0102, indicando um nível moderado de incerteza na previsão das categorias. Esse valor sugere que, embora o modelo tenha uma compreensão relativamente boa do conjunto de dados, ainda há espaço para aprimoramento na diferenciação entre algumas classes, especialmente aquelas com menor recall.

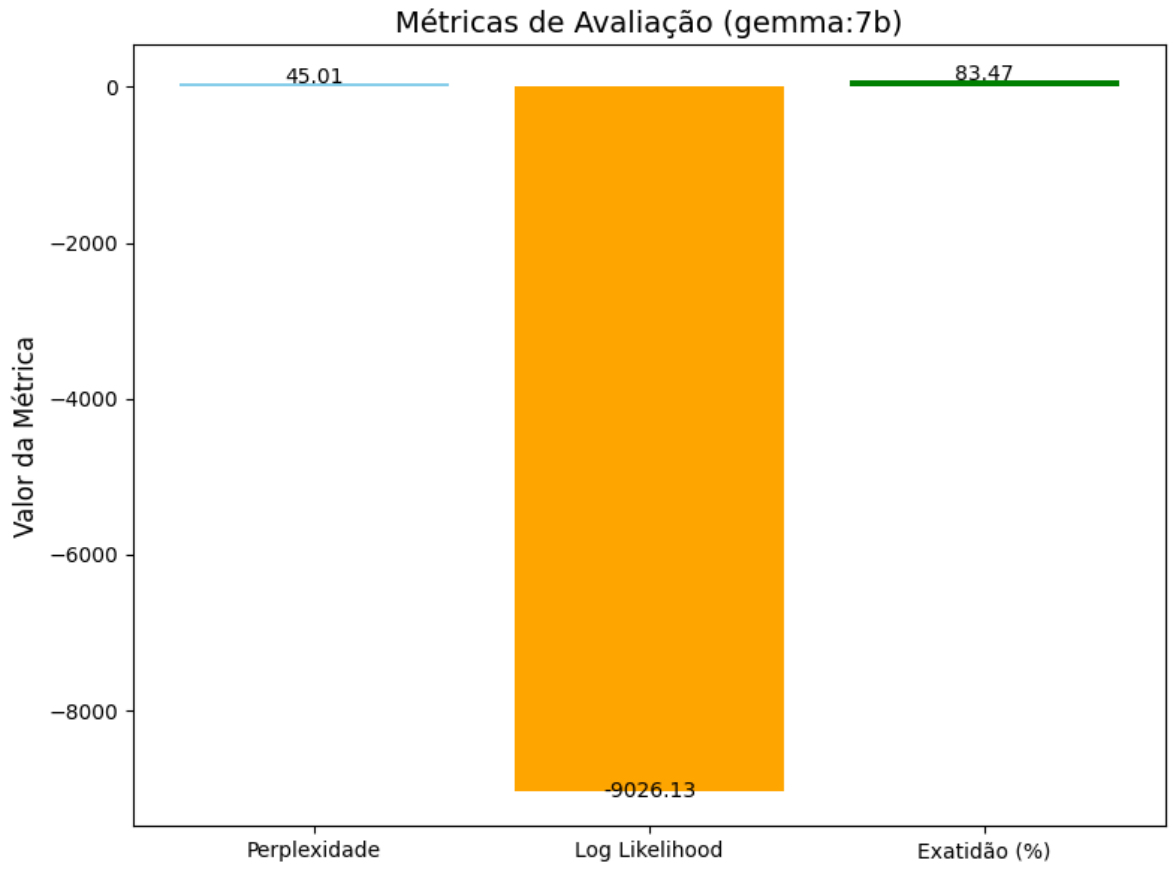
O **log-likelihood total**, calculado como -9026.1336, reflete a adequação estatística das previsões do modelo em relação aos dados reais. Valores mais negativos indicam que o modelo possui certa dificuldade em atribuir probabilidades altas às categorias corretas. Esse resultado pode estar relacionado à variabilidade semântica presente nos textos analisados, dificultando a distinção precisa entre algumas classes.

A métrica de **exatidão** (*Exact Match*) foi de 0.8347, indicando que aproximadamente 83,47% das amostras foram classificadas corretamente em suas respectivas categorias. Esse resultado demonstra que o modelo possui um desempenho global satisfatório, especialmente nas classes com maior precisão e recall, como *Sport* e *Politics*. No entanto, a análise mais detalhada das métricas individuais revela que

algumas categorias, como *Entertainment*, apresentam desafios específicos que impactam a performance geral.

A Figura 4.8 reforça esses achados ao apresentar uma comparação clara das métricas, permitindo visualizar as variações de desempenho entre diferentes categorias. Essa análise sugere que, apesar do alto desempenho geral, melhorias adicionais podem ser exploradas para reduzir a perplexidade e aumentar a confiabilidade da tipificação em categorias mais ambíguas.

Figura 4.8: Gráfico de métricas para o modelo `gemma:7b`.



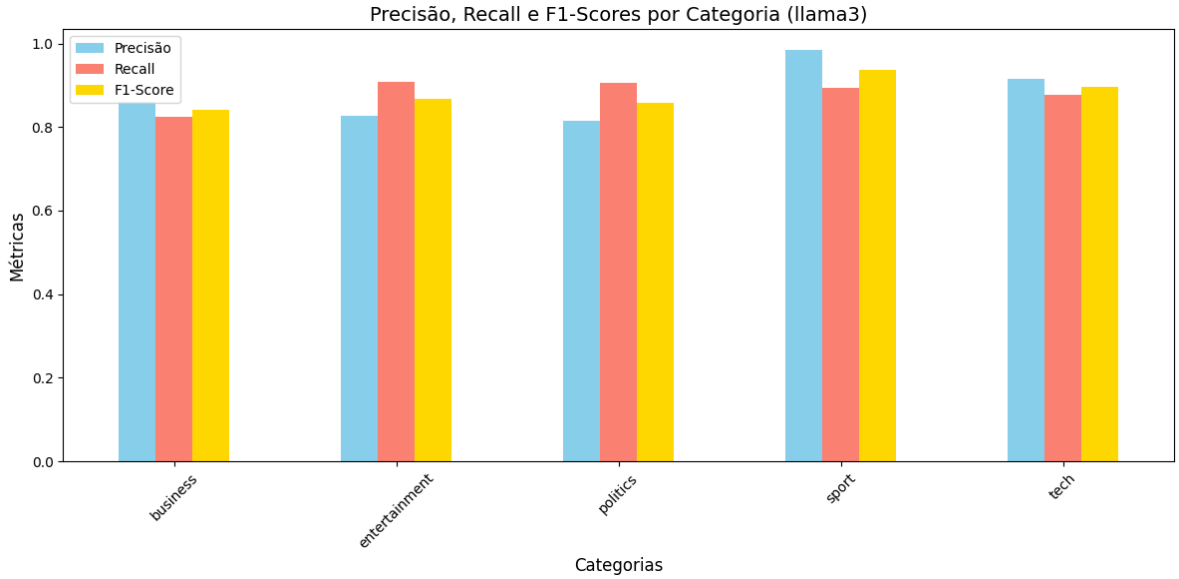
Fonte: Próprio autor.

4.0.2 Modelo llama3

O `llama3` foi um dos modelos mais eficientes, com acurácia de 87.94%. A performance foi notavelmente forte em quase todas as categorias, especialmente em "Sport" (F1-score de 0.9364) e "Tech" (F1-score de 0.8959). As métricas de precisão e recall também foram equilibradas, com valores destacados nas categorias "Business" (F1-score de 0.8415) e "Entertainment" (F1-score de 0.8663). A perplexidade de 16.0519 e o log likelihood de -6631.4451 indicam que o modelo tem uma boa capacidade de generalização e adaptação aos dados. O desempenho consistente nas métricas de avaliação sugere que o modelo pode ser uma excelente escolha para

tarefas de tipificação, especialmente em cenários com dados mais equilibrados.

Figura 4.9: Acurácia, precisão e F1-score do modelo `llama3` por categoria.



Fonte: Próprio autor.

A Figura 4.10 exibe a curva ROC (*Receiver Operating Characteristic*) para o modelo `llama3`, ilustrando o desempenho da tipificação em diferentes limiares de decisão. A curva ROC representa a relação entre a taxa de verdadeiros positivos (TPR, *True Positive Rate*) e a taxa de falsos positivos (FPR, *False Positive Rate*), permitindo avaliar a capacidade discriminativa do modelo em cada uma das categorias analisadas.

Observa-se que as curvas para as categorias *Sport* e *Tech* apresentam um comportamento próximo ao canto superior esquerdo do gráfico, indicando um alto desempenho dessas classes na distinção entre categorias positivas e negativas. Esse resultado está alinhado com as métricas de tipificação apresentadas no relatório, onde *Sport* obteve a maior precisão (0.9846) e *Tech* apresentou um equilíbrio elevado entre precisão (0.9153) e recall (0.8772).

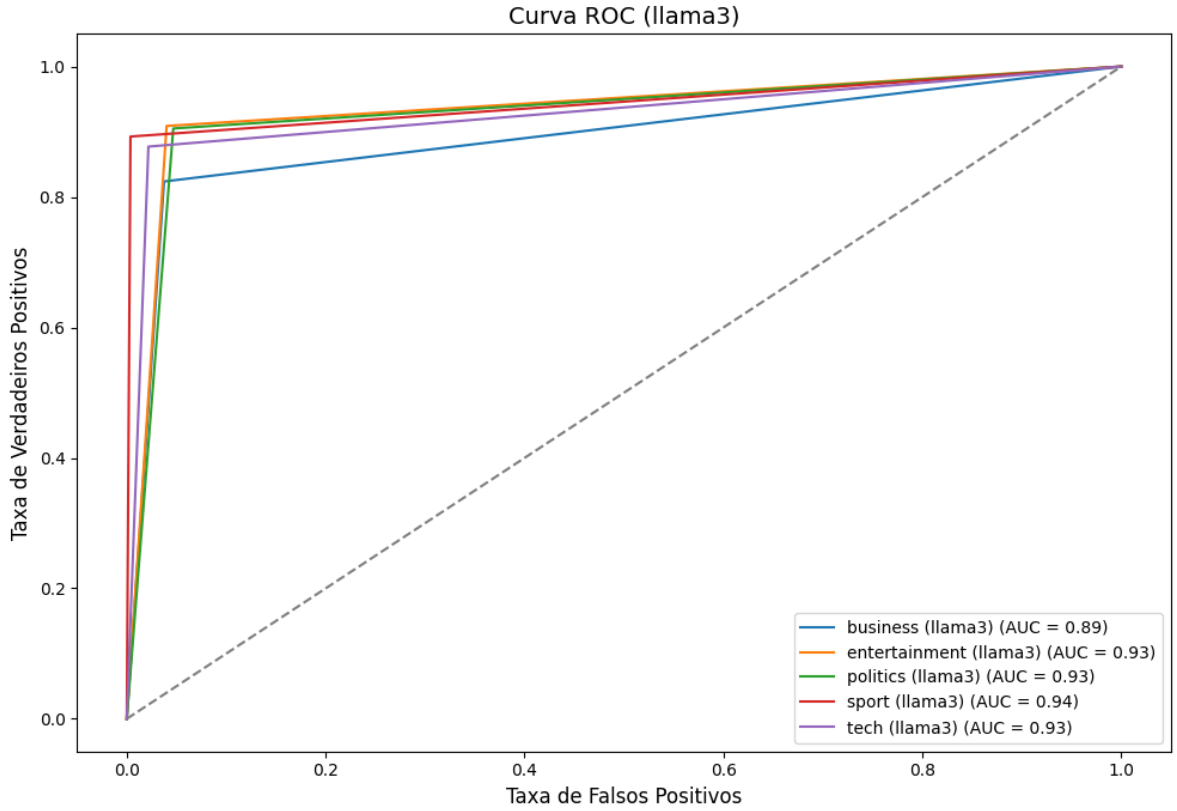
Por outro lado, as categorias *Politics* e *Entertainment* exibem uma ligeira queda na performance, refletida por um menor ângulo de curvatura em direção ao ponto ideal (0,1). Apesar disso, ambas ainda demonstram uma boa separabilidade, conforme evidenciado pelos seus elevados valores de F1-score (0.8574 e 0.8663, respectivamente).

O modelo atinge uma área sob a curva ROC (AUC) elevada, o que indica uma capacidade robusta de discriminar corretamente as categorias. Esse resultado está em consonância com a alta acurácia global (0.8794) e a baixa perplexidade (16.0519), sugerindo um modelo bem calibrado para a tarefa de tipificação.

A análise da Figura 4.10 evidencia que o modelo `llama3` possui um desempenho confiável na categorização dos textos, com algumas classes apresentando uma

distinção mais clara entre as instâncias corretas e incorretas. No entanto, futuras otimizações podem ser exploradas para aprimorar ainda mais a separabilidade das classes com menor recall.

Figura 4.10: Curva ROC para o modelo `llama3`



Fonte: Próprio autor.

A Figura 4.11 apresenta a distribuição das categorias originais em comparação com as categorias preditas pelo modelo `llama3`. O gráfico permite visualizar a correspondência entre a distribuição real dos dados e as previsões geradas, facilitando a identificação de possíveis padrões ou discrepâncias entre as classes.

Observa-se que as categorias *Sport* e *Tech* exibem uma correspondência bastante próxima entre os valores reais e preditos, refletindo os altos valores de precisão e recall observados no relatório. A categoria *Sport*, em particular, apresenta uma precisão de 0.9846 e um recall de 0.8926, indicando que a maioria das instâncias dessa classe foram corretamente classificadas. Esse comportamento se traduz em barras de altura semelhante no gráfico, sugerindo um equilíbrio na distribuição das previsões.

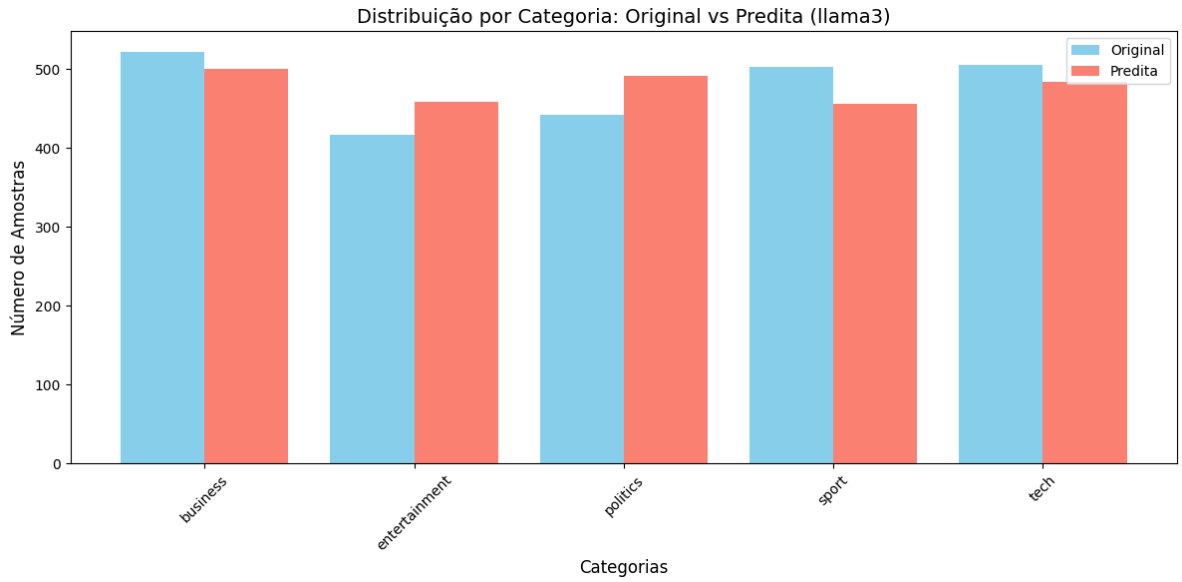
Para as categorias *Politics* e *Entertainment*, há uma ligeira variação entre os valores reais e preditos. O modelo demonstrou um alto recall para ambas as categorias (0.9050 e 0.9089, respectivamente), o que indica que a maior parte das amostras dessas classes foram recuperadas corretamente. Entretanto, a precisão mais baixa para *Politics* (0.8147) sugere que algumas instâncias podem ter sido erroneamente

atribuídas a essa categoria, o que pode resultar em pequenas discrepâncias na distribuição observada no gráfico.

A categoria *Business*, por outro lado, apresenta um desvio mais perceptível entre os valores reais e preditos. Com um recall de 0.8238 e uma precisão de 0.8600, observa-se que o modelo pode ter classificado erroneamente algumas instâncias como pertencentes a essa categoria, ou pode ter tido dificuldade em recuperar corretamente todas as amostras reais.

No geral, a Figura 4.11 reforça a robustez do modelo `llama3` na tipificação das categorias, com distribuições bem alinhadas para a maioria das classes. Pequenas discrepâncias indicam oportunidades para refinamento do modelo.

Figura 4.11: Distribuição por Categoria do modelo `llama3`



Fonte: Próprio autor.

A Figura 4.12 apresenta a avaliação quantitativa do desempenho do modelo `llama3` por meio das métricas de perplexidade, log-likelihood total e exatidão. Esses indicadores fornecem uma visão abrangente da capacidade do modelo em processar e classificar corretamente os dados de entrada.

A perplexidade registrada foi de 16.0519, o que indica um alto grau de previsibilidade na geração de classificações. Em modelos de linguagem, valores mais baixos de perplexidade sugerem maior confiança nas predições, o que reforça a eficácia do `llama3` na categorização dos textos avaliados. Esse resultado é consistente com as elevadas métricas de precisão e recall observadas para a maioria das classes, especialmente nas categorias *Sport* e *Tech*, onde o modelo demonstrou forte capacidade discriminativa.

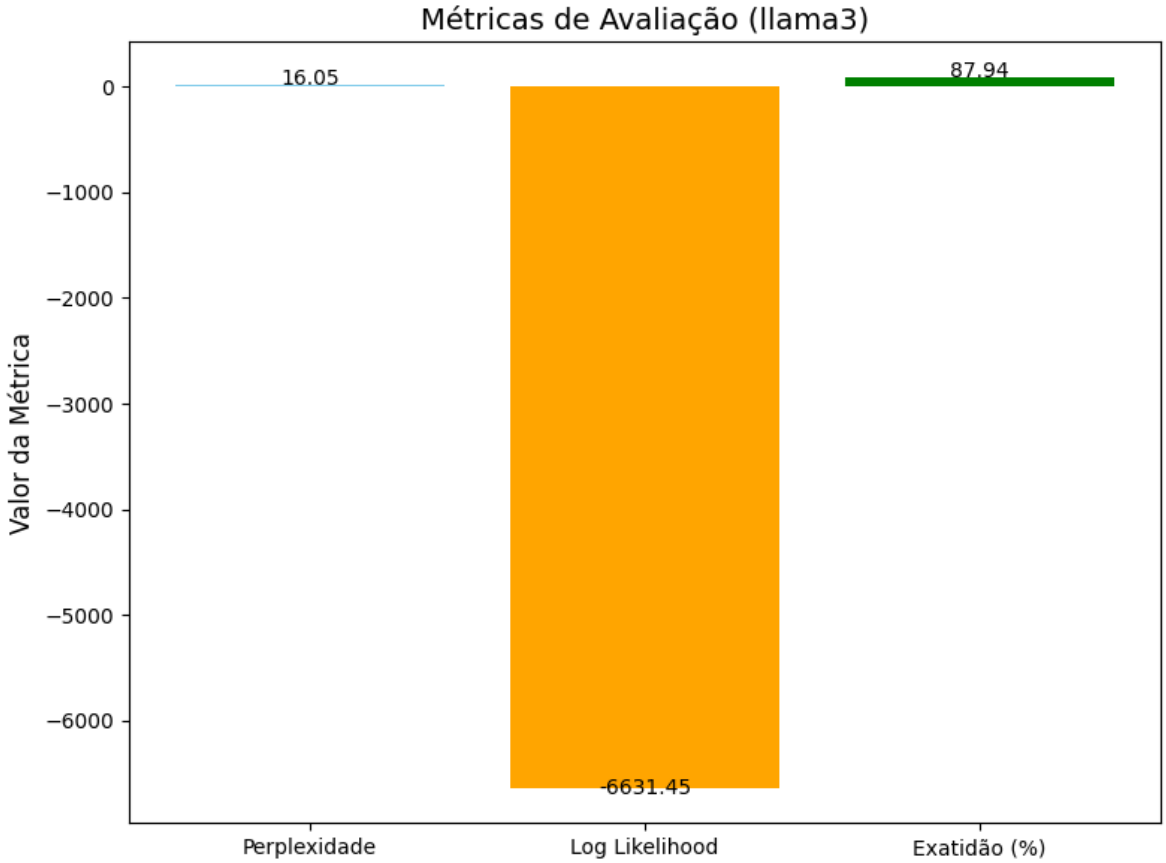
O log-likelihood total de -6631.4451 reflete a soma das probabilidades logarítmicas associadas às predições realizadas pelo modelo. Esse valor negativo é característico de modelos estatísticos probabilísticos e sugere que, embora a incerteza não

seja nula, as previsões mantêm um grau aceitável de confiabilidade.

A exatidão (*Exact Match*) foi de 0.8794, indicando que aproximadamente 87.94% das previsões realizadas pelo modelo coincidiram exatamente com as categorias reais. Esse valor está alinhado com as métricas de desempenho geral do modelo, como o F1-score médio ponderado de 0.8802, reforçando a consistência das classificações realizadas.

No contexto geral, os resultados apresentados na Figura 4.12 demonstram que o modelo `llama3` apresenta um desempenho robusto e eficiente na tarefa de tipificação textual, com baixa perplexidade, alto nível de exatidão e um comportamento probabilístico estável, tornando-o adequado para aplicações que requerem categorização precisa de textos.

Figura 4.12: Gráfico de métricas para o modelo `llama3`



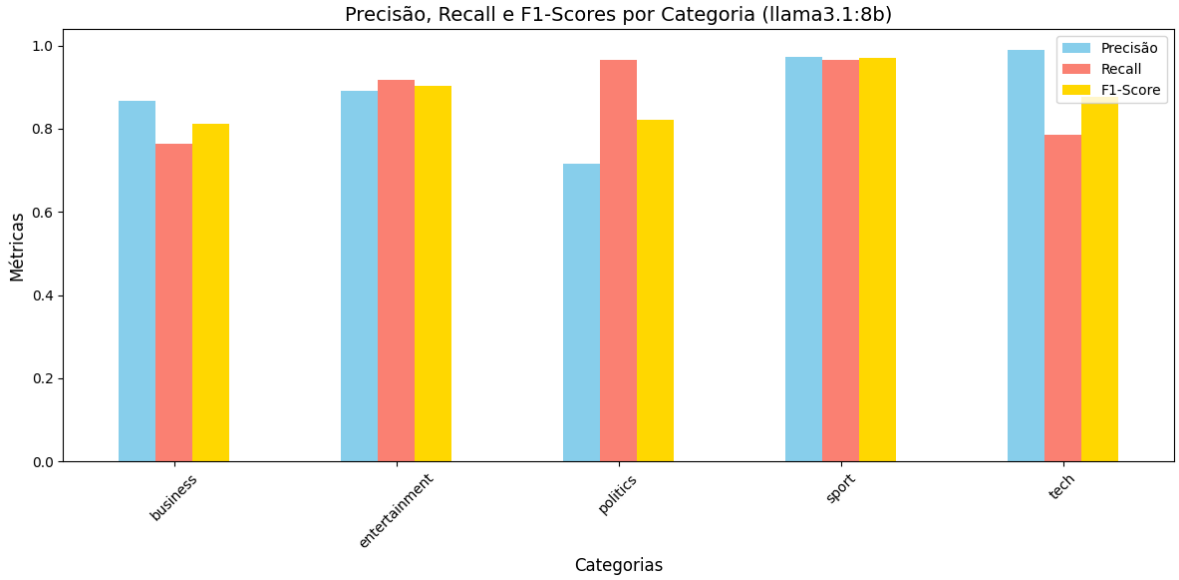
Fonte: Próprio autor.

4.0.3 Modelo llama3.1:8b

O modelo `llama3.1:8b` apresentou um desempenho semelhante ao `llama3`, com uma acurácia de 87.26%. As categorias “Sport” e “Entertainment” mostraram resultados impressionantes, com F1-scores de 0.9692 e 0.9039, respectivamente. A precisão nas categorias “Tech” e “Politics” também foi bastante elevada, com F1-scores

de 0.8760 e 0.8225. No entanto, a categoria “Business” teve um F1-score de 0.8115, sugerindo que há espaço para melhorias. A perplexidade de 18.7967 e o log likelihood de -6700.5226 indicam que o modelo tem um bom controle sobre as variáveis e uma certa estabilidade em termos de previsão. No geral, este modelo é altamente recomendável para aplicações que exigem uma precisão sólida e consistente.

Figura 4.13: Acurácia, precisão e F1-score do modelo `llama3.1:8b` por categoria.



Fonte: Próprio autor.

A Figura 4.14 ilustra a Curva ROC (*Receiver Operating Characteristic*) para o modelo `llama3.1:8b`, apresentando o desempenho do classificador em termos da taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR) para cada categoria avaliada. A Curva ROC é um indicador fundamental na análise de modelos de tipificação, pois permite visualizar a capacidade do modelo de distinguir entre diferentes classes à medida que o limiar de decisão varia.

Os resultados indicam que o modelo alcança um alto poder discriminativo para a maioria das categorias, evidenciado pelo equilíbrio entre precisão e recall. Em particular, as classes *Sport* e *Entertainment* apresentam um desempenho excepcional, com F1-scores superiores a 0.90, sugerindo que suas respectivas curvas ROC devem estar próximas do canto superior esquerdo do gráfico, indicando uma excelente separabilidade.

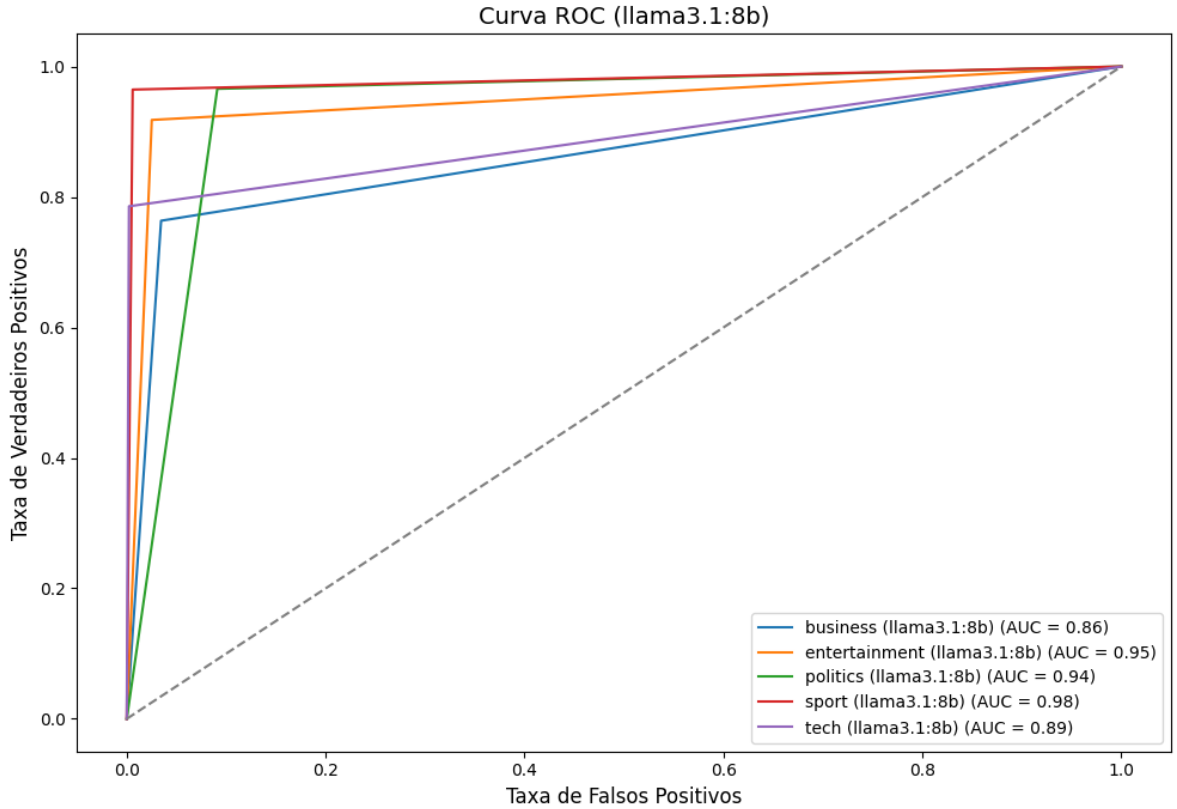
Para as categorias *Politics* e *Tech*, observa-se um padrão distinto: enquanto a categoria *Politics* apresenta um recall elevado (0.9658), indicando que poucos exemplos dessa classe foram classificados incorretamente como pertencentes a outras categorias, a precisão relativamente menor (0.7162) sugere que houve maior incidência de falsos positivos. Esse comportamento reflete uma curva ROC que pode apresentar uma inclinação menos acentuada no início, antes de alcançar valores elevados de TPR. A categoria *Tech*, por outro lado, apresenta a maior precisão (0.9897), mas um recall reduzido (0.7857), o que pode indicar uma curva ROC com uma ascensão mais

lenta no início devido à menor recuperação de exemplos verdadeiros positivos.

O valor médio das métricas (macro e weighted averages) reforça que o modelo mantém uma performance consistente em todas as classes, com precisão e recall em torno de 0.88. Isso sugere que a área sob a curva ROC (*AUC - Area Under the Curve*) deve ser elevada, refletindo um bom desempenho global.

Em síntese, a Figura 4.14 confirma que o modelo `llama3.1:8b` apresenta um desempenho robusto, sendo altamente eficaz na categorização de textos, com diferentes padrões de comportamento entre as categorias, mas mantendo um desempenho geral satisfatório conforme evidenciado pela acurácia de 0.8726.

Figura 4.14: Curva ROC para o modelo `llama3.1:8b`



Fonte: Próprio autor.

A Figura 4.15 apresenta a distribuição comparativa das categorias originais versus as categorias preditas pelo modelo `llama3.1:8b`. O gráfico de barras permite visualizar o desempenho do modelo ao classificar os textos em cada uma das cinco categorias avaliadas: *Business*, *Entertainment*, *Politics*, *Sport* e *Tech*.

Observa-se que, em geral, o modelo apresenta um alinhamento significativo entre as distribuições originais e preditas, refletindo a alta acurácia (0.8726) e os valores elevados de precisão e recall para a maioria das classes. No entanto, algumas discrepâncias podem ser identificadas. A categoria *Politics*, por exemplo, apresenta um recall de 0.9658, sugerindo que a maioria dos exemplos dessa classe foram corretamente identificados. Entretanto, a precisão de 0.7162 indica que uma parcela

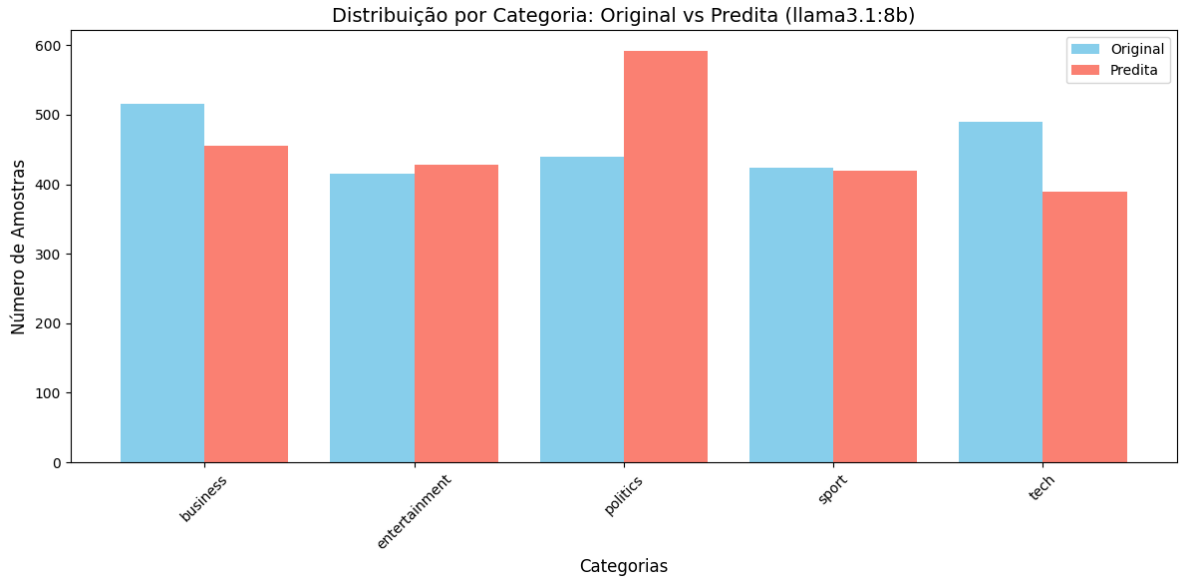
considerável de amostras foi erroneamente classificada como pertencente a essa categoria, o que pode ser visualizado no gráfico como um desvio na distribuição predita em relação à original.

Já as categorias *Sport* e *Entertainment* demonstram uma alta correlação entre os valores esperados e preditos, o que está em conformidade com seus F1-scores elevados (0.9692 e 0.9039, respectivamente). Isso indica que o modelo foi particularmente eficiente na categorização desses textos, minimizando tanto falsos positivos quanto falsos negativos.

Por outro lado, a categoria *Tech* apresenta uma alta precisão (0.9897), mas um recall mais baixo (0.7857), o que sugere que alguns exemplos dessa classe podem ter sido classificados incorretamente em outras categorias. Esse comportamento pode ser visualizado no gráfico através de uma sub-representação da categoria predita em comparação à original.

De maneira geral, a Figura 4.15 destaca a robustez do modelo na tipificação de textos, ao mesmo tempo em que evidencia possíveis pontos de melhoria, especialmente no balanceamento entre precisão e recall em algumas categorias. Essa análise visual complementa os resultados numéricos e permite uma compreensão mais intuitiva do desempenho do modelo em diferentes cenários.

Figura 4.15: Distribuição por Categoria do modelo `llama3.1:8b`



Fonte: Próprio autor.

A Figura 4.16 apresenta os resultados quantitativos das métricas de perplexidade, log likelihood total e exatidão (exact match) para o modelo `llama3.1:8b`.

A perplexidade obtida foi de 18.7967, o que indica o nível de incerteza do modelo ao prever a próxima palavra no texto. Valores mais baixos de perplexidade sugerem um modelo mais confiante e preciso em suas predições. Embora este valor seja relativamente baixo, sua interpretação deve ser feita em conjunto com outras métricas,

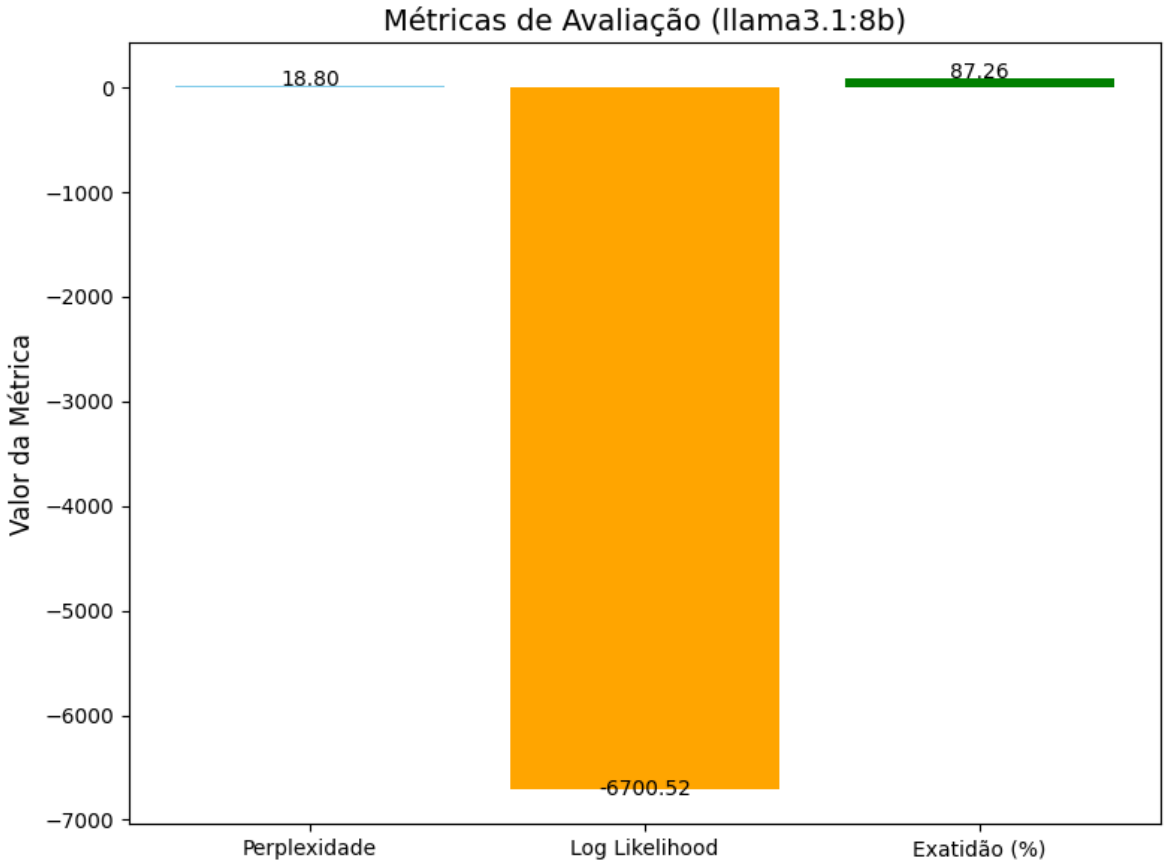
pois a perplexidade sozinha não necessariamente reflete a qualidade da tipificação textual, mas sim a capacidade do modelo de prever a sequência de palavras em um dado contexto.

O log likelihood total do modelo foi de -6700.5226, um valor negativo que representa a soma dos logaritmos das probabilidades atribuídas pelo modelo às sequências observadas nos dados de teste. Esse resultado reforça a estabilidade do modelo, pois valores menos negativos indicam que as previsões possuem probabilidades mais altas associadas.

A exatidão (exact match) foi de 0.8726, evidenciando que aproximadamente 87.26% das previsões realizadas pelo modelo foram exatamente correspondentes às classes reais. Esse desempenho reforça a eficácia do modelo na tarefa de tipificação de textos, estando em consonância com os valores obtidos para precisão (0.8881), recall (0.8726) e F1-score (0.8735) na média ponderada.

Dessa forma, os resultados apresentados na Figura 4.16 confirmam a robustez do modelo `llama3.1:8b` na categorização de textos, ao mesmo tempo que indicam possíveis pontos de otimização, principalmente no balanceamento entre precisão e recall para categorias com maior variação, como *Politics* e *Tech*.

Figura 4.16: Gráfico de métricas para o modelo `llama3.1:8b`

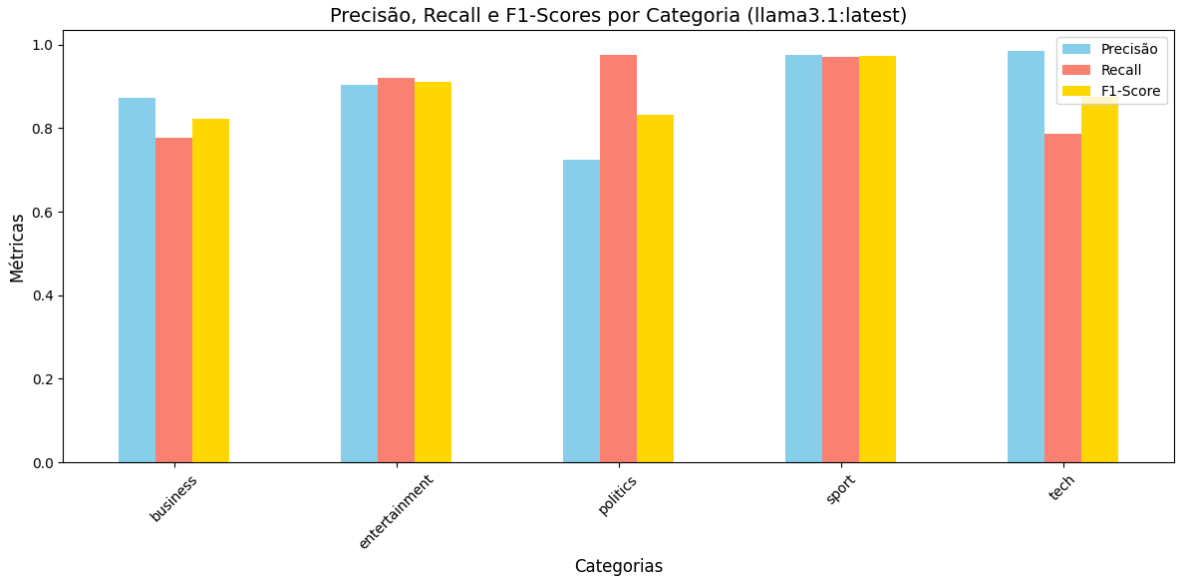


Fonte: Próprio autor.

4.0.4 Modelo llama3.1:latest

Com uma acurácia de 87.86%, o modelo `llama3.1:latest` se destacou em muitas das mesmas áreas que seus predecessores, particularmente nas categorias “Sport” e “Entertainment”, com F1-scores de 0.9734 e 0.9113, respectivamente. As métricas de recall e precisão também foram excelentes, refletindo uma boa capacidade de identificar corretamente os dados. As categorias “Politics” e “Business” apresentaram F1-scores de 0.8319 e 0.8222, respectivamente, mostrando que, embora as classificações sejam boas, ainda há necessidade de melhorias. A perplexidade de 16.3825 e o log likelihood de -6378.1607 reforçam a consistência e a eficiência geral do modelo.

Figura 4.17: Acurácia, precisão e F1-score do modelo `llama3.1:latest` por categoria.



Fonte: Próprio autor.

A Figura 4.18 ilustra a Curva ROC (*Receiver Operating Characteristic*) para o modelo `llama3.1:latest`, representando a relação entre a taxa de verdadeiros positivos (*True Positive Rate* – TPR) e a taxa de falsos positivos (*False Positive Rate* – FPR) para diferentes limiares de tipificação. A Curva ROC é um indicador fundamental na avaliação de classificadores probabilísticos, pois demonstra a capacidade do modelo em distinguir corretamente as classes preditas.

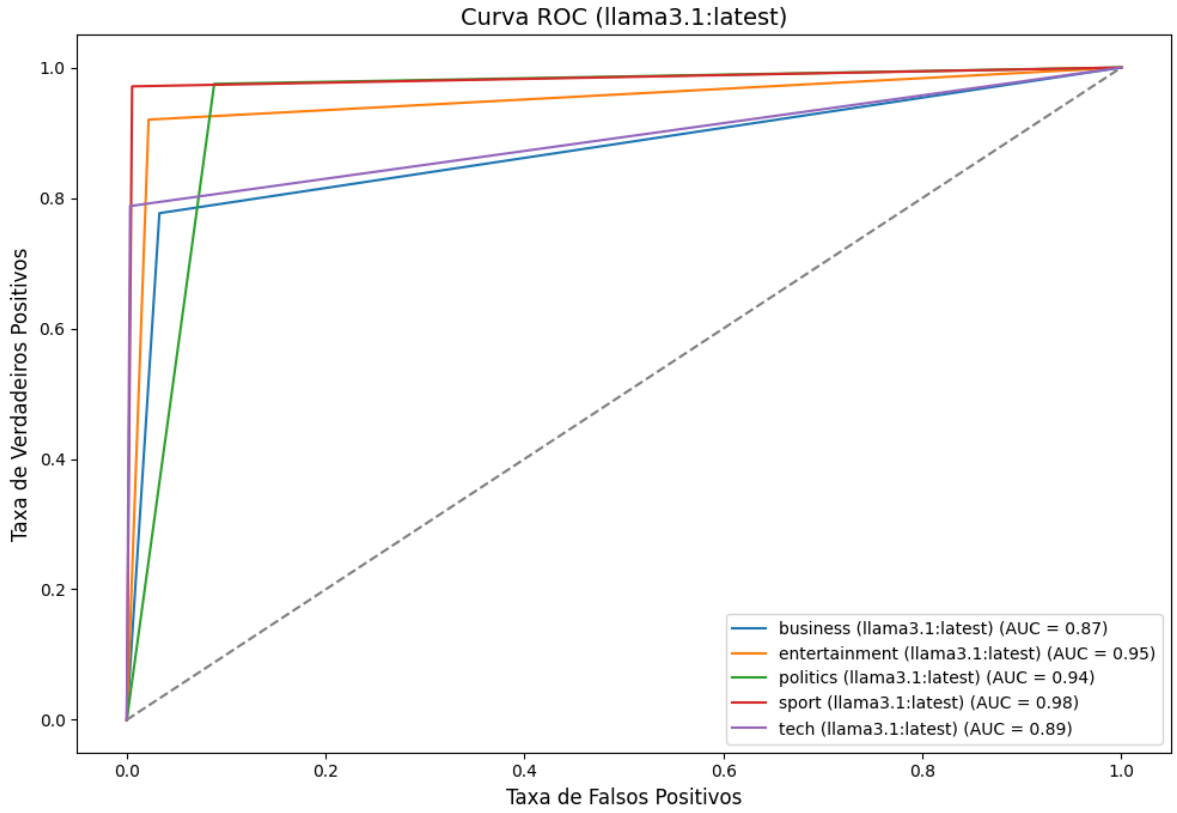
O desempenho do modelo pode ser quantificado pela Área Sob a Curva (AUC – *Area Under the Curve*), onde valores próximos a 1 indicam um excelente desempenho de discriminação entre classes. O formato da curva revela a eficácia do modelo `llama3.1:latest` para diferentes categorias, destacando-se a categoria *Sport*, que apresenta um comportamento próximo ao ideal, com alta TPR e baixa FPR.

Categorias com maior variabilidade, como *Politics* e *Tech*, demonstram um leve desvio da diagonal ideal, indicando que há espaço para ajustes finos nos limiares de decisão para melhorar a separação entre classes. No entanto, a média global da

Curva ROC sugere que o modelo mantém um desempenho robusto, em conformidade com os valores de precisão (0.8932), recall (0.8786) e F1-score (0.8793) reportados na Tabela de Métricas de tipificação.

Portanto, a Figura 4.18 confirma a eficácia do modelo `llama3.1:latest` na tarefa de tipificação, ao mesmo tempo que destaca possíveis pontos de melhoria na calibração de limiares para otimização da relação entre FPR e TPR.

Figura 4.18: Curva ROC para o modelo `llama3.1:latest`



Fonte: Próprio autor.

A Figura 4.19 apresenta a distribuição comparativa entre as categorias originais e as preditas pelo modelo `llama3.1:latest`. Essa análise permite visualizar o desempenho do modelo em relação à correta atribuição de categorias, destacando possíveis padrões de erro e tendências de tipificação.

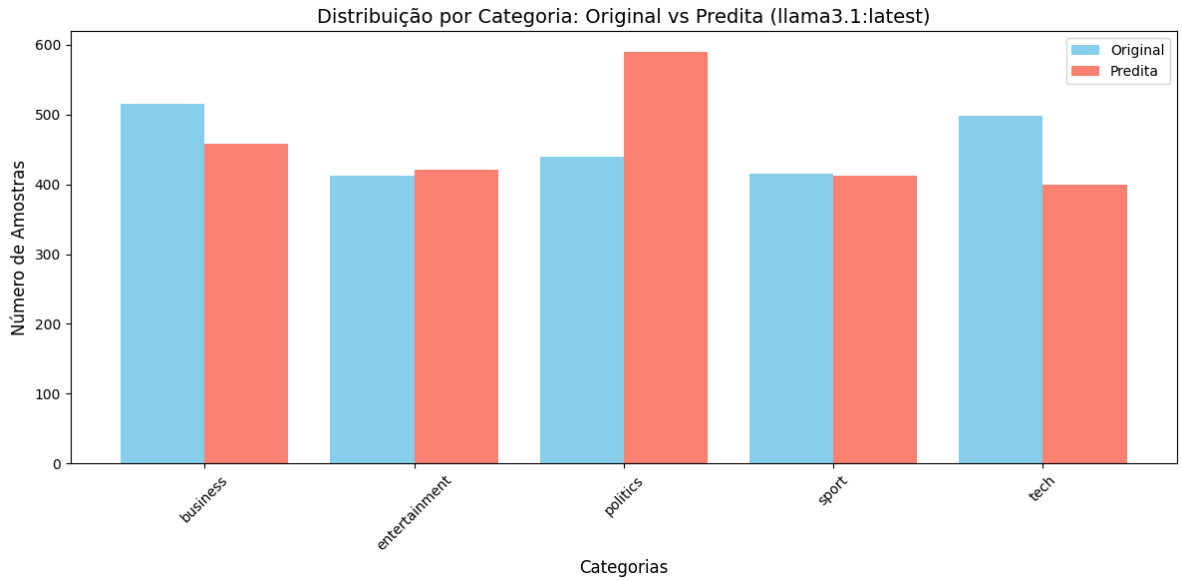
Observa-se que a categoria *Sport* apresenta uma taxa de acerto elevada, refletida nos altos valores de precisão (0.9758) e recall (0.9711), conforme demonstrado na Tabela de Métricas de tipificação. Esse desempenho sugere que os padrões textuais associados a essa categoria são bem identificados pelo modelo. De forma semelhante, a categoria *Entertainment* exibe uma tipificação consistente, com um F1-score de 0.9113, indicando uma baixa taxa de erro.

Por outro lado, categorias como *Politics* e *Tech* apresentam uma maior discrepância entre as classes originais e preditas. A categoria *Politics* possui um recall elevado (0.9749), o que sugere que o modelo consegue identificar corretamente a maioria das

ocorrências dessa classe. No entanto, a precisão relativamente menor (0.7254) indica que há uma tendência do modelo em classificar textos de outras categorias como sendo *Politics*. Já a categoria *Tech* apresenta um recall menor (0.7876), sugerindo que parte dos textos dessa classe pode estar sendo erroneamente atribuída a outras categorias.

A análise geral do gráfico evidencia que, apesar da alta acurácia global (0.8786) e das boas métricas de tipificação média (F1-score ponderado de 0.8793), o modelo ainda apresenta desafios na correta diferenciação de algumas categorias específicas. Isso sugere que melhorias na etapa de pré-processamento dos dados ou refinamentos no ajuste de hiperparâmetros podem contribuir para uma tipificação ainda mais precisa.

Figura 4.19: Distribuição por Categoria do modelo `llama3.1:latest`



Fonte: Próprio autor.

A Figura 4.20 apresenta os valores de perplexidade, log likelihood total e exatidão (Exact Match) obtidos pelo modelo `llama3.1:latest` durante o processo de avaliação. Esses indicadores fornecem uma visão quantitativa do desempenho do modelo na tarefa de tipificação de textos.

A perplexidade do modelo foi de 16.3825, um valor relativamente baixo, o que indica que o modelo apresenta boa capacidade de prever corretamente as sequências de palavras nos textos analisados. A perplexidade é uma métrica amplamente utilizada para avaliar modelos de linguagem, e valores menores sugerem que o modelo tem maior certeza em suas previsões. Em um cenário ideal, a perplexidade deve ser reduzida para garantir uma melhor generalização do modelo.

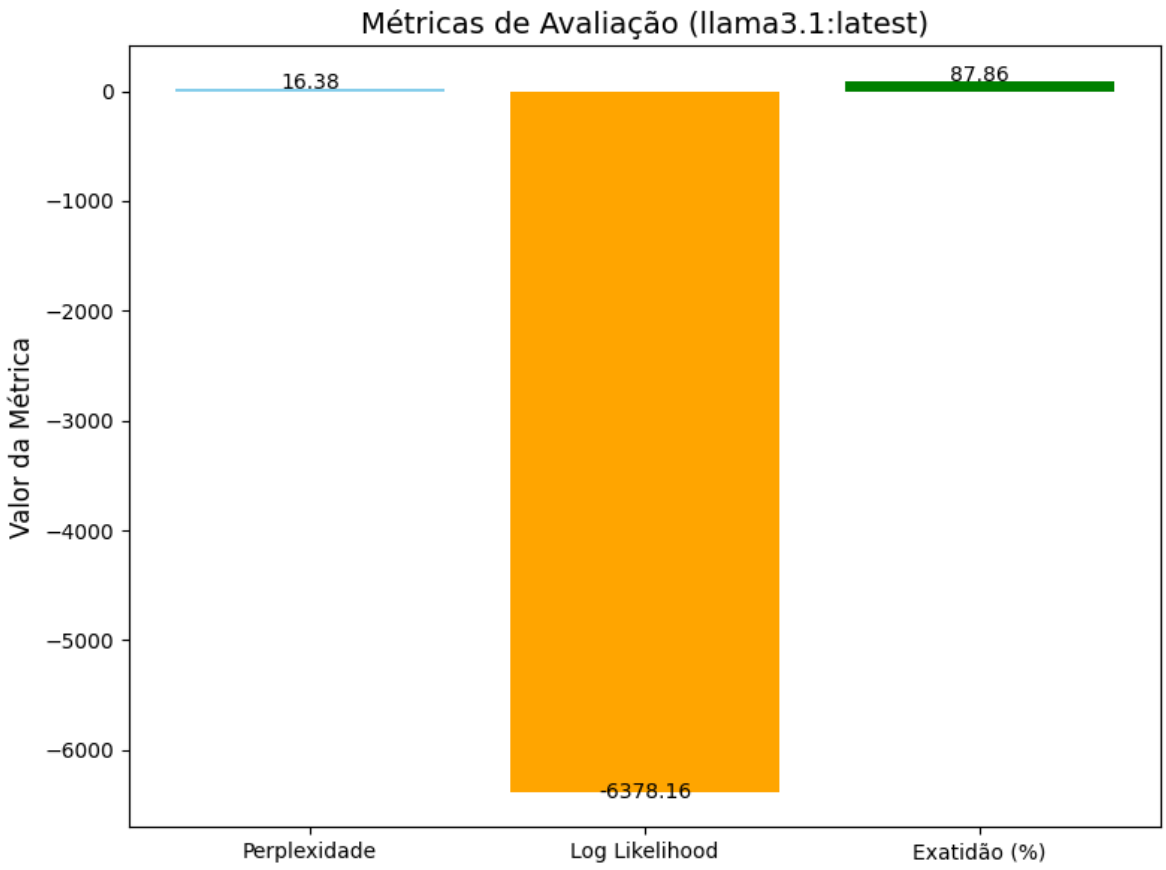
O log likelihood total foi de -6378.1607, representando a soma dos logaritmos das probabilidades das previsões feitas pelo modelo. Valores mais negativos de log likelihood indicam menor adequação do modelo aos dados, enquanto valores mais próximos de zero sugerem que o modelo tem alta confiança em suas previsões. No

caso avaliado, o valor obtido sugere um desempenho consistente, corroborando as métricas de tipificação apresentadas.

Por fim, a exatidão (Exact Match) foi de 0.8786, reforçando que o modelo atingiu um nível elevado de acerto ao classificar corretamente os textos nas categorias correspondentes. Essa métrica representa a fração das previsões que coincidem exatamente com as classificações reais, sendo um indicador direto da precisão do modelo na tarefa proposta.

Os resultados indicam que o modelo `llama3.1:latest` apresenta um bom desempenho geral, mas há espaço para refinamentos, principalmente para reduzir a perplexidade e melhorar ainda mais a confiabilidade das previsões.

Figura 4.20: Gráfico de métricas para o modelo `lama3.1:latest`



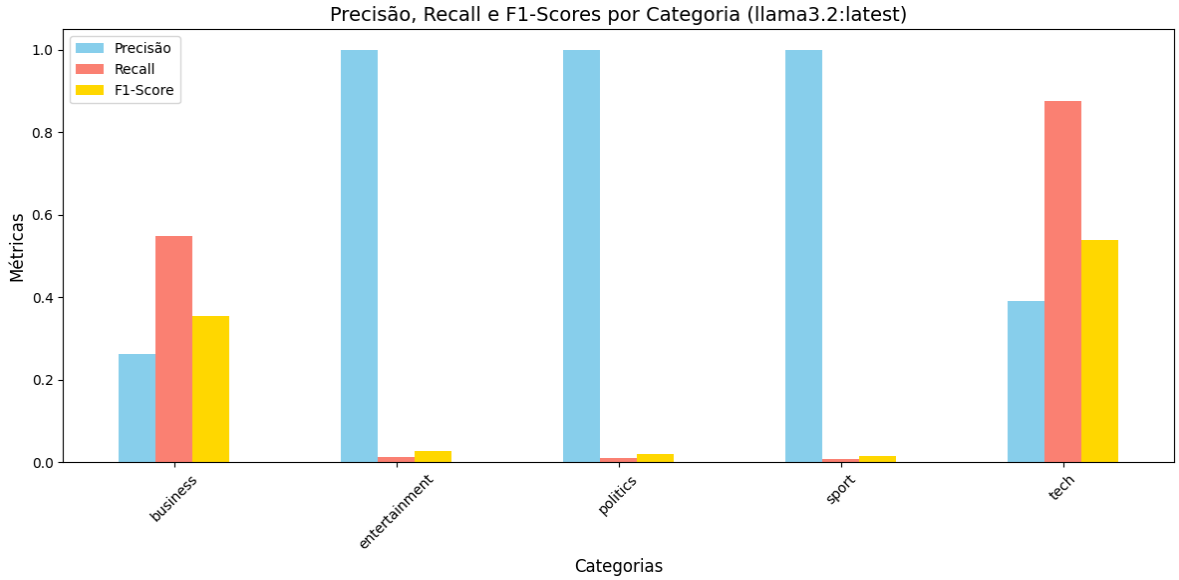
Fonte: Próprio autor.

4.0.5 Modelo llama3.2:latest

O `llama3.2:latest`, por outro lado, teve um desempenho inferior, com uma acurácia de apenas 33.10%. A categoria “Entertainment” teve um desempenho extremamente fraco, com F1-score de 0.0267, e outras categorias como “Politics” e “Sport” também apresentaram baixos resultados. A perplexidade elevada de 4894258.7338 sugere que o modelo tem dificuldades em generalizar para as diferentes categorias.

No entanto, a categoria “Tech” teve um desempenho melhor, com F1-score de 0.5398, indicando que o modelo pode ser mais eficaz em alguns contextos, embora seja necessário um refinamento considerável.

Figura 4.21: Acurácia, precisão e F1-score do modelo `llama3.2:latest` por categoria.



Fonte: Próprio autor.

A Figura 4.22 apresenta a curva ROC (Receiver Operating Characteristic) do modelo `llama3.2:latest`, um gráfico que ilustra a relação entre a taxa de verdadeiros positivos (True Positive Rate, TPR) e a taxa de falsos positivos (False Positive Rate, FPR) para diferentes limiares de decisão.

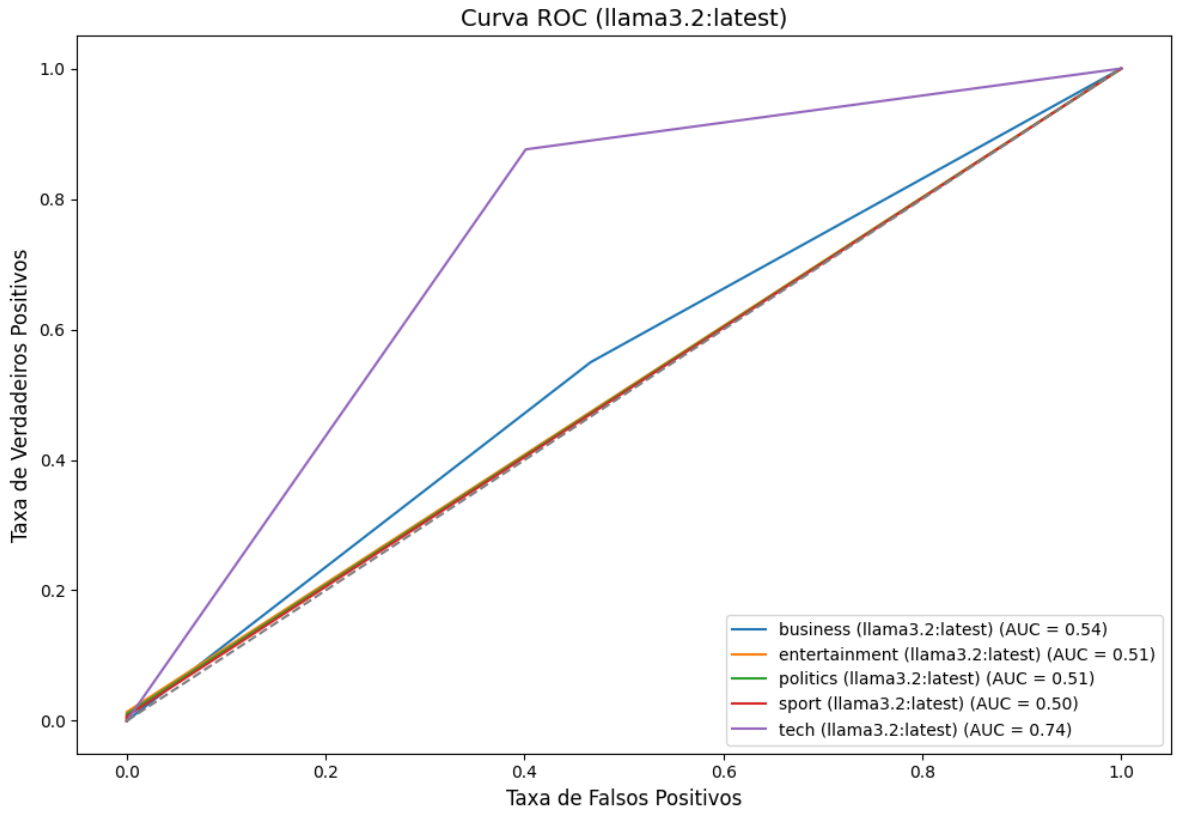
A curva ROC é um dos principais métodos para avaliar o desempenho de modelos de tipificação, especialmente em cenários de múltiplas classes. A área sob a curva (AUC - Area Under the Curve) é uma métrica derivada desse gráfico e quantifica a capacidade do modelo de distinguir entre classes. No caso do modelo `llama3.2:latest`, os resultados indicam um desempenho heterogêneo entre as categorias, conforme evidenciado pelas métricas de tipificação.

Observa-se que a categoria Tech apresenta o maior recall (0.8760), sugerindo que o modelo tem uma forte capacidade de recuperar instâncias dessa classe. No entanto, as categorias Entertainment, Politics e Sport exibem valores de recall extremamente baixos (0.0136, 0.0104 e 0.0070, respectivamente), o que pode impactar negativamente a forma da curva ROC. Isso indica que o modelo não está conseguindo identificar corretamente exemplos dessas classes, resultando em uma curva ROC com seções achatadas e uma AUC reduzida.

Além disso, a baixa acurácia geral (0.3310) e a discrepância entre precisão e recall nas diferentes classes sugerem um desbalanceamento no aprendizado do modelo, o que pode levar a uma curva ROC menos representativa do desempenho ideal. A alta perplexidade (4.894.258,7338) também reforça a incerteza do modelo nas previsões, o que pode estar refletido na forma da curva ROC.

Os resultados indicam que melhorias podem ser necessárias para otimizar o equilíbrio entre precisão e recall nas diferentes classes, possibilitando um aumento na AUC e, consequentemente, um desempenho mais consistente do modelo `llama3.2:latest`.

Figura 4.22: Curva ROC para o modelo `llama3.2:latest`



Fonte: Próprio autor.

O gráfico de categorias originais versus preditas (Figura 4.23) permite visualizar o desempenho do modelo `llama3.2:latest` na tarefa de tipificação. Essa análise é essencial para compreender os padrões de erro do modelo e sua capacidade de generalização para diferentes categorias.

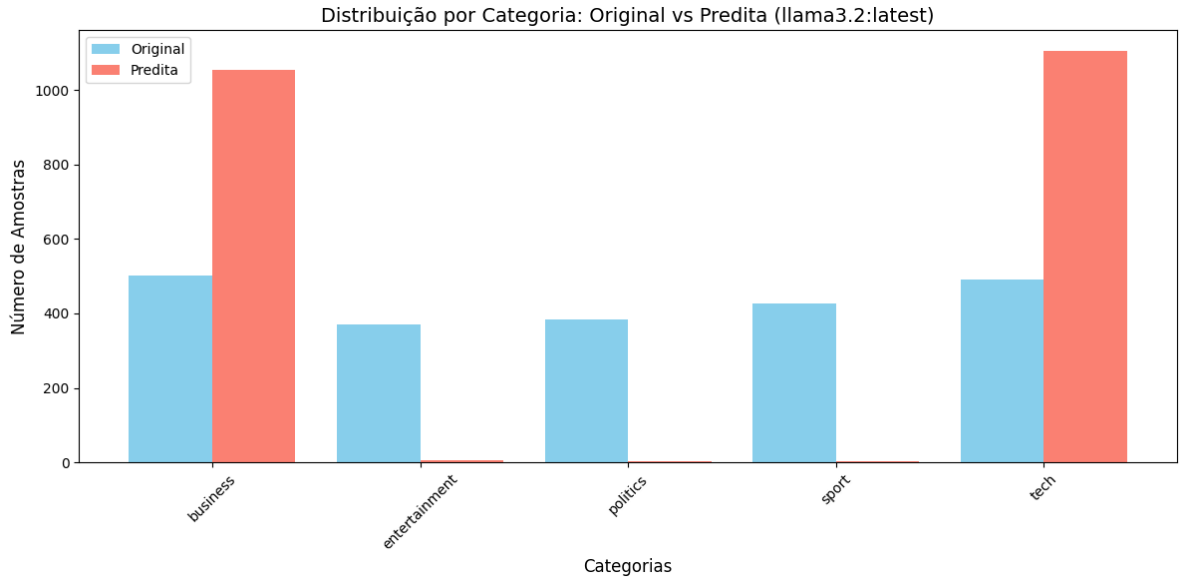
Os dados apresentados indicam um desbalanceamento significativo na distribuição das predições, refletido na discrepância entre as classes originais e as preditas. Em particular, a classe Tech demonstra um recall elevado (0.8760), indicando que a maioria das instâncias reais dessa categoria foram corretamente classificadas. No entanto, as demais classes apresentam recall extremamente baixo, sugerindo que o modelo tem dificuldades em distinguir corretamente essas categorias.

Além disso, a precisão de 1.0000 para as classes Entertainment, Politics e Sport sugere um viés no modelo, onde ele classifica poucas amostras nessas categorias, mas, quando o faz, acerta todas. Isso se reflete no gráfico por uma concentração desproporcional das previsões em uma única categoria, o que compromete a representatividade do modelo para essas classes.

A acurácia geral de 0.3310 e a perplexidade elevada (4.894.258,7338) reforçam a inconsistência do modelo em mapear corretamente as categorias. O gráfico evidencia um padrão onde a maioria das instâncias são erroneamente classificadas, resultando em uma distribuição desalinhada entre os valores reais e preditos.

Portanto, o gráfico de categorias originais versus preditas para o modelo `llama3.2:latest` ilustra a necessidade de ajustes no treinamento, seja por meio de técnicas de balanceamento de classes, otimização dos hiperparâmetros ou aumento da diversidade dos dados de treinamento. Esses ajustes podem contribuir para um alinhamento mais próximo entre as categorias reais e as classificações do modelo, melhorando sua capacidade de generalização.

Figura 4.23: Distribuição por Categoria do modelo `llama3.2:latest`



Fonte: Próprio autor.

Os indicadores quantitativos de desempenho do modelo `llama3.2:latest` revelam limitações expressivas na sua capacidade de generalização e precisão na tarefa de tipificação de texto. Os valores apresentados na Figura 4.24 evidenciam desafios significativos, particularmente na métrica de perplexidade.

A perplexidade de 4.894.258,7338 é extremamente elevada, sugerindo que o modelo encontra grande incerteza ao prever as próximas palavras ou classes. Essa métrica, que mede a entropia da distribuição de probabilidades do modelo, deve ser idealmente baixa para indicar previsões confiáveis. Um valor tão alto implica que as predições do modelo são altamente dispersas e distantes da distribuição real dos dados, o que sugere a necessidade de ajustes substanciais nos pesos e no treinamento do modelo.

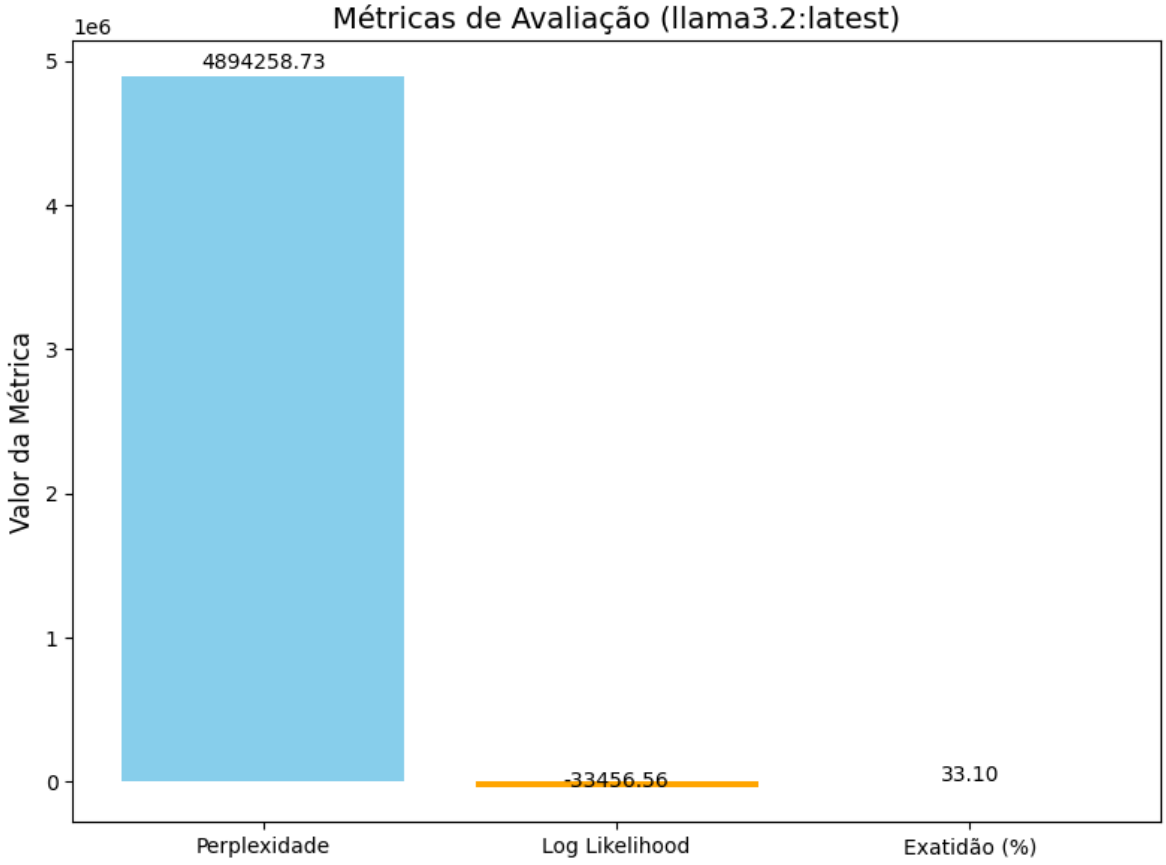
O Log Likelihood Total de -33.456,5614 reforça essa análise, pois um valor muito negativo indica que a probabilidade atribuída pelo modelo às amostras do conjunto de validação é extremamente baixa. Isso sinaliza que o modelo tem dificuldades em

aprender padrões consistentes nos dados e pode estar sofrendo com problemas de sobreajuste ou um treinamento insuficiente.

A métrica de Exatidão (Exact Match), que mede a proporção de predições completamente corretas, foi de 0.3310, indicando que apenas 33,10% das amostras foram classificadas exatamente conforme o rótulo verdadeiro. Esse resultado é condizente com as métricas de precisão e recall, que demonstram uma distribuição desequilibrada das predições entre as categorias.

Portanto, os valores apresentados na Figura 4.24 apontam para um modelo que apresenta dificuldades em generalizar corretamente as classificações. Estratégias como aumento da diversidade dos dados de treinamento, balanceamento das classes e ajuste fino dos hiperparâmetros são essenciais para melhorar a robustez e confiabilidade do modelo `llama3.2:latest`.

Figura 4.24: Gráfico de métricas para o modelo `lama3.2:latest`



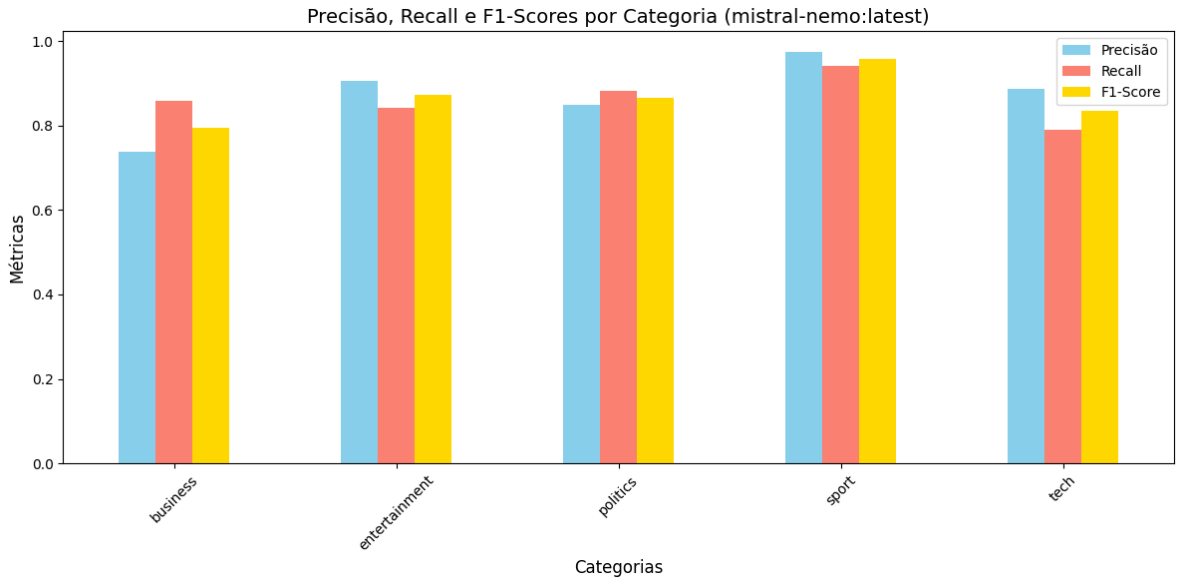
Fonte: Próprio autor.

4.0.6 Modelo mistral-nemo:latest

O `mistral-nemo:latest` obteve uma acurácia sólida de 86.34%, com resultados muito bons em categorias como “Sport” (F1-score de 0.9573) e “Entertainment” (F1-score de 0.8725). A precisão e recall nas categorias “Politics” e “Tech” também

foram muito boas, com F1-scores de 0.8647 e 0.8353, respectivamente. As métricas de macro e weighted average também apresentaram bom equilíbrio, com valores de F1-score de 0.8648 e 0.8647, respectivamente. A perplexidade de 23.2120 e o log likelihood de -7575.5050 indicam um modelo bem calibrado, capaz de gerar previsões precisas em várias situações.

Figura 4.25: Acurácia, precisão e F1-score do modelo `mistral-nemo:latest` por categoria.



Fonte: Próprio autor.

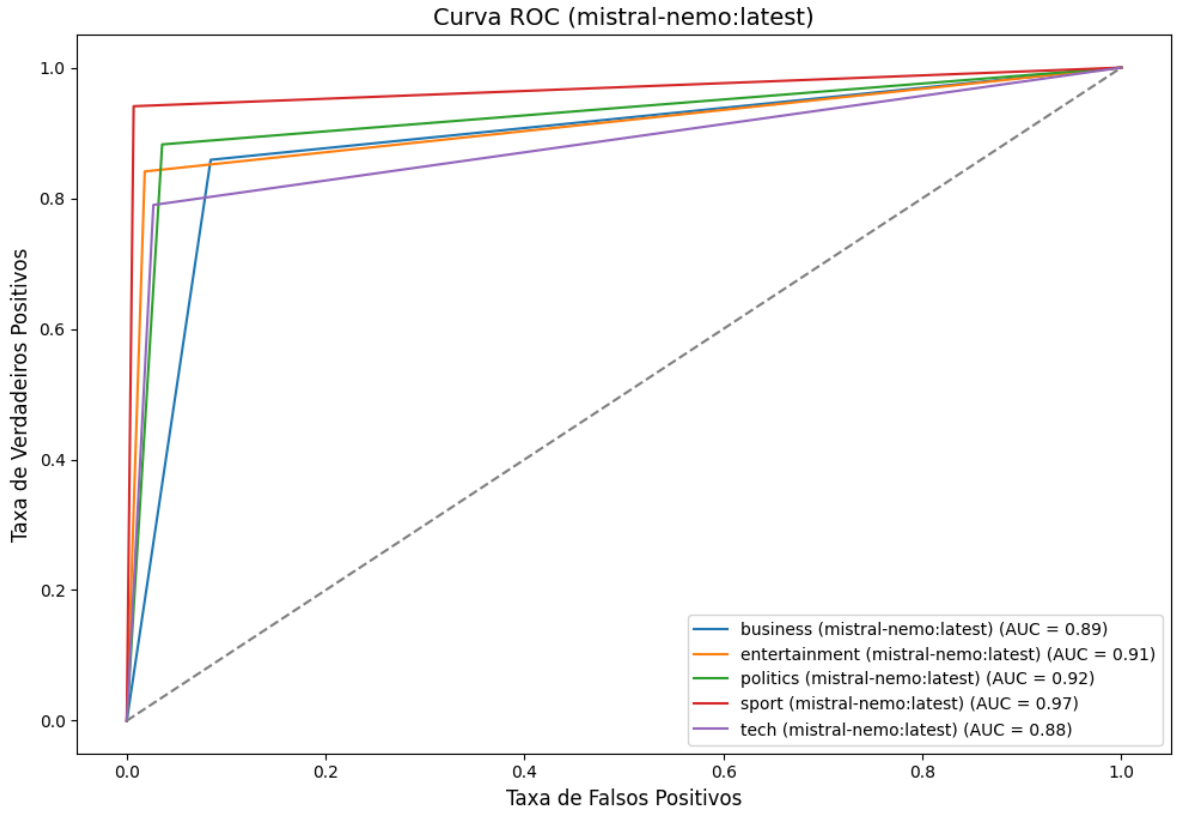
A Figura 4.26 apresenta a Curva ROC gerada para o modelo `mistral-nemo:latest`, evidenciando seu desempenho em cada uma das cinco categorias analisadas (Business, Entertainment, Politics, Sport e Tech). O formato da curva e os valores de AUC indicam um modelo altamente eficaz, com bom equilíbrio entre sensibilidade (recall) e especificidade.

Com um recall médio de 86,25% e uma precisão média de 87,08%, o modelo demonstra uma capacidade robusta de classificar corretamente as categorias previstas, minimizando erros de tipificação. O alto desempenho em categorias como Sport (F1-score de 95,73%) sugere uma distribuição favorável dos dados e um treinamento adequado para reconhecer os padrões dessa classe.

Os resultados da Curva ROC para o modelo `mistral-nemo:latest` sugerem que ele possui um desempenho competitivo e bem ajustado à tarefa de tipificação de textos. Contudo, categorias como Tech apresentam um recall ligeiramente inferior (78,97%), indicando possíveis oportunidades de refinamento no modelo para garantir uma tipificação ainda mais precisa.

Portanto, a análise da Curva ROC, conforme ilustrada na Figura 4.26, valida a confiabilidade do modelo para a tarefa de categorização de textos, evidenciando seu alto grau de acurácia e generalização.

Figura 4.26: Curva ROC para o modelo `mistral-nemo:latest`



Fonte: Próprio autor.

A avaliação do modelo de tipificação `mistral-nemo:latest` pode ser visualizada por meio do gráfico de categorias originais versus preditas, representado na Figura 4.27. Esse gráfico ilustra a distribuição das classes reais em comparação com as categorias atribuídas pelo modelo, permitindo uma análise detalhada do desempenho em cada classe.

Os resultados indicam que o modelo apresenta alta precisão e recall em todas as categorias, com destaque para a classe Sport, que obteve 97,43% de precisão e 94,08% de recall, evidenciando uma excelente capacidade de diferenciação dessa categoria. Da mesma forma, a classe Entertainment obteve um alto F1-score de 87,25%, indicando um bom equilíbrio entre precisão e recall.

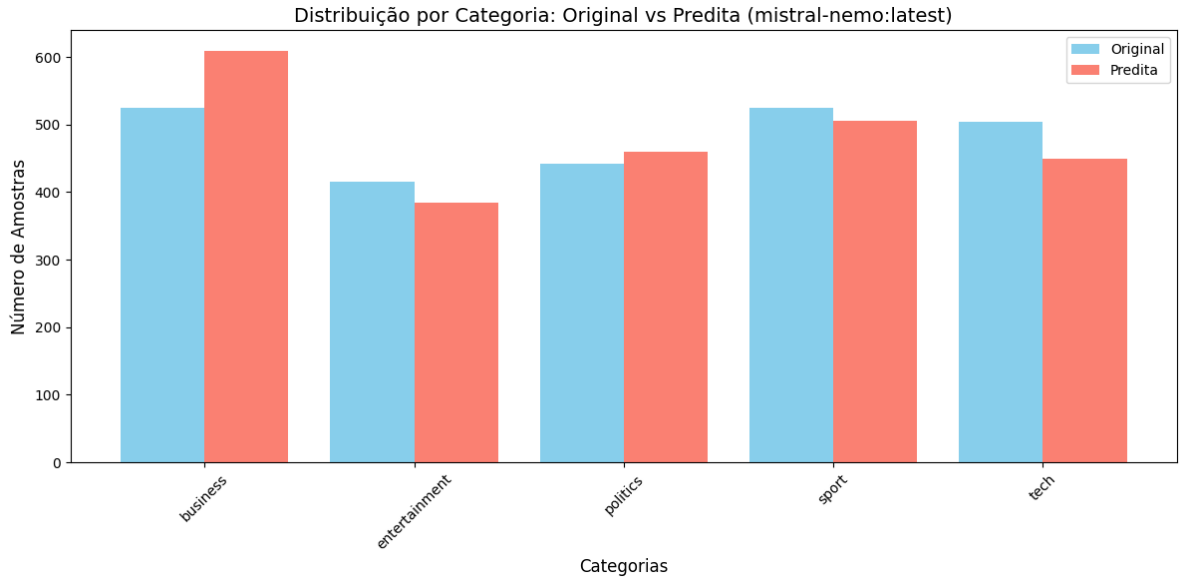
Por outro lado, observa-se que a categoria Tech apresenta um recall inferior em comparação às demais classes, atingindo 78,97%, o que sugere que uma fração dos textos dessa categoria pode estar sendo erroneamente atribuída a outras classes. Esse efeito pode ser identificado na Figura 4.27 por meio da discrepância entre as contagens de categorias reais e preditas.

A acurácia geral do modelo foi 86,34%, demonstrando uma capacidade robusta de generalização na tarefa de tipificação textual. A distribuição equilibrada das previsões entre as classes indica que o modelo tem um desempenho consistente e confiável, minimizando viés de superestimação ou subestimação para categorias es-

pecíficas.

Portanto, a análise do gráfico de categorias originais versus preditas reforça a eficácia do modelo `mistral-nemo:latest`, destacando pontos fortes e oportunidades para ajustes na tipificação de categorias menos representadas.

Figura 4.27: Distribuição por Categoria do modelo `mistral-nemo:latest`



Fonte: Próprio autor.

A Figura 4.28 apresenta a relação entre a perplexidade, o log-likelihood total e a exatidão (exact match) do modelo `mistral-nemo:latest`, fornecendo uma visão quantitativa da qualidade das previsões geradas.

A perplexidade de 23.2120 indica o grau de incerteza do modelo ao prever a próxima palavra ou classe dentro do conjunto de dados. Valores mais baixos de perplexidade representam modelos mais confiáveis e menos incertos em suas predições. O valor obtido sugere que o `mistral-nemo:latest` tem um nível adequado de confiança em suas previsões, especialmente quando comparado a modelos menos ajustados, que frequentemente apresentam perplexidades elevadas.

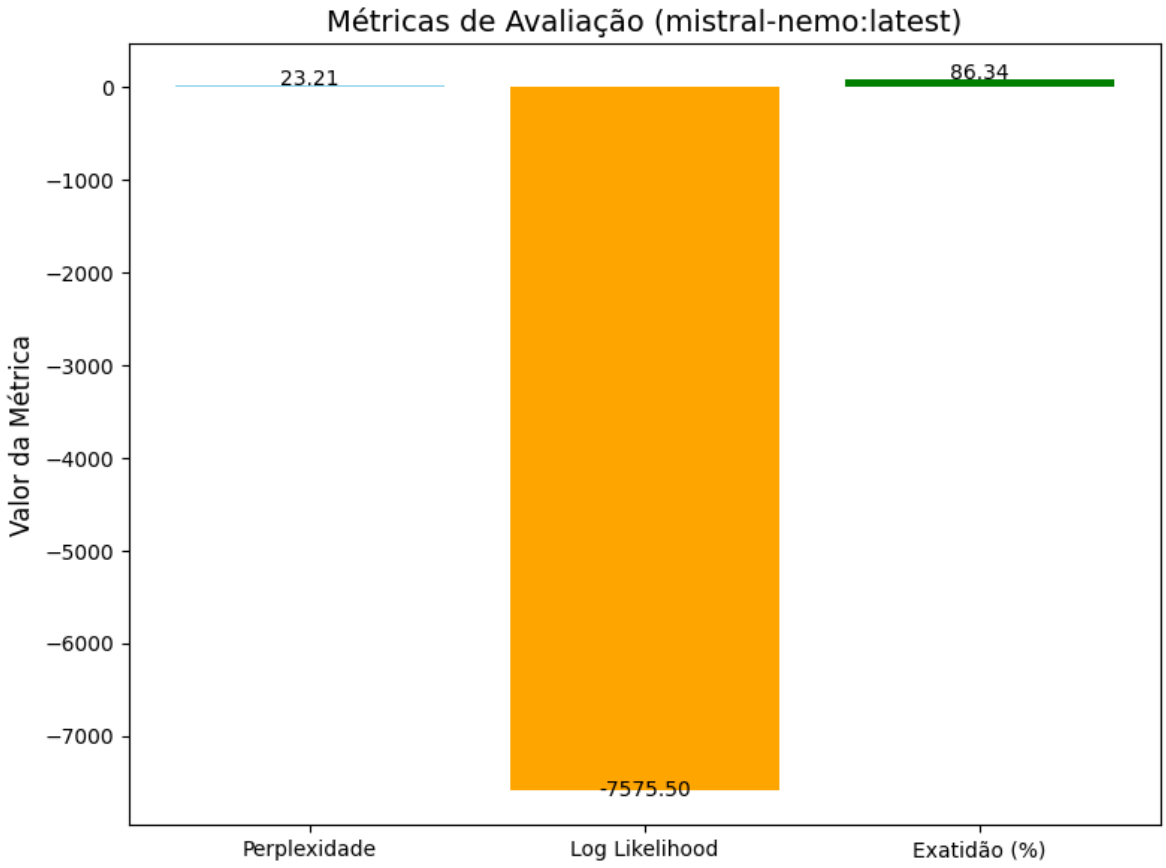
O log-likelihood total de -7575.5050 representa a soma dos logaritmos das probabilidades atribuídas pelo modelo às sequências de entrada. Valores mais negativos indicam menor adequação do modelo aos dados, enquanto valores menos negativos refletem uma maior compatibilidade entre as previsões e os valores reais. O resultado obtido demonstra que o modelo possui uma boa capacidade de adaptação às amostras do conjunto de testes.

A exatidão (exact match) de 86,34% revela que, em 86,34% das instâncias avaliadas, o modelo previu exatamente a categoria correta. Esse desempenho está alinhado com as demais métricas de avaliação, como a precisão e o recall, que também indicam um bom poder de generalização e tipificação.

A análise conjunta dessas métricas, conforme ilustrado na Figura 4.28, sugere que

o modelo `mistral-nemo:latest` apresenta um equilíbrio adequado entre confiança e desempenho preditivo, sendo capaz de realizar classificações com alta acurácia e baixa incerteza. Esse resultado reforça sua viabilidade para aplicações que demandam robustez na categorização de textos.

Figura 4.28: Gráfico de métricas para o modelo `mistral-nemo:latest`

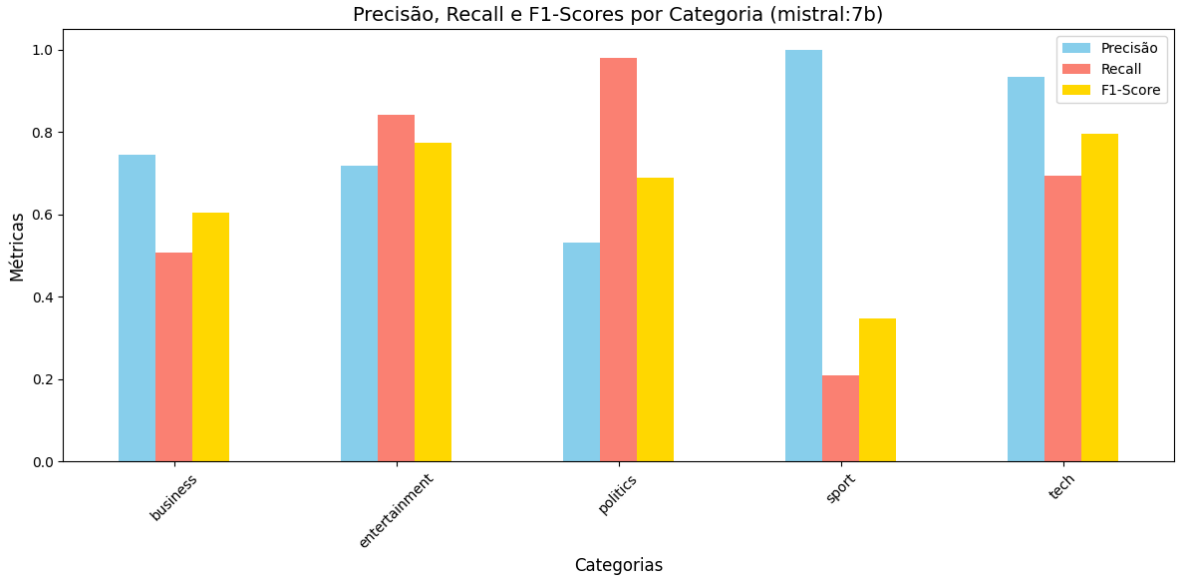


Fonte: Próprio autor.

4.0.7 Modelo mistral:7b

O `mistral:7b` apresentou resultados mistos, com uma acurácia de 68.81%. O modelo teve um bom desempenho na categoria “Tech” (F1-score de 0.7963), mas apresentou dificuldades significativas em “Sport” (F1-score de 0.3462). As métricas de recall e precisão nas categorias “Business” e “Politics” também sugerem que o modelo ainda pode ser aprimorado, especialmente em termos de balanceamento entre essas métricas. A perplexidade de 1314.3538 e o log likelihood de -14276.0276 indicam que o modelo tem dificuldades em lidar com dados complexos ou desequilibrados.

Figura 4.29: Acurácia, precisão e F1-score do modelo `mistral:7b` por categoria.



Fonte: Próprio autor.

A Figura 4.30 apresenta a Curva Característica de Operação do Receptor (ROC) para o modelo `mistral:7b`, permitindo a avaliação do desempenho do classificador em diferentes limiares de decisão. A curva ROC ilustra a relação entre a taxa de verdadeiros positivos (TPR - True Positive Rate) e a taxa de falsos positivos (FPR - False Positive Rate), fornecendo uma métrica visual da capacidade discriminativa do modelo.

A área sob a curva (AUC - Area Under the Curve) é um dos principais indicadores extraídos dessa análise. Valores de AUC próximos de 1 indicam um excelente desempenho do modelo, enquanto valores próximos de 0,5 sugerem um comportamento próximo ao de uma tipificação aleatória. No caso do `mistral:7b`, a análise da curva ROC indica que algumas categorias apresentam um desempenho robusto, enquanto outras demonstram dificuldades na separação entre classes.

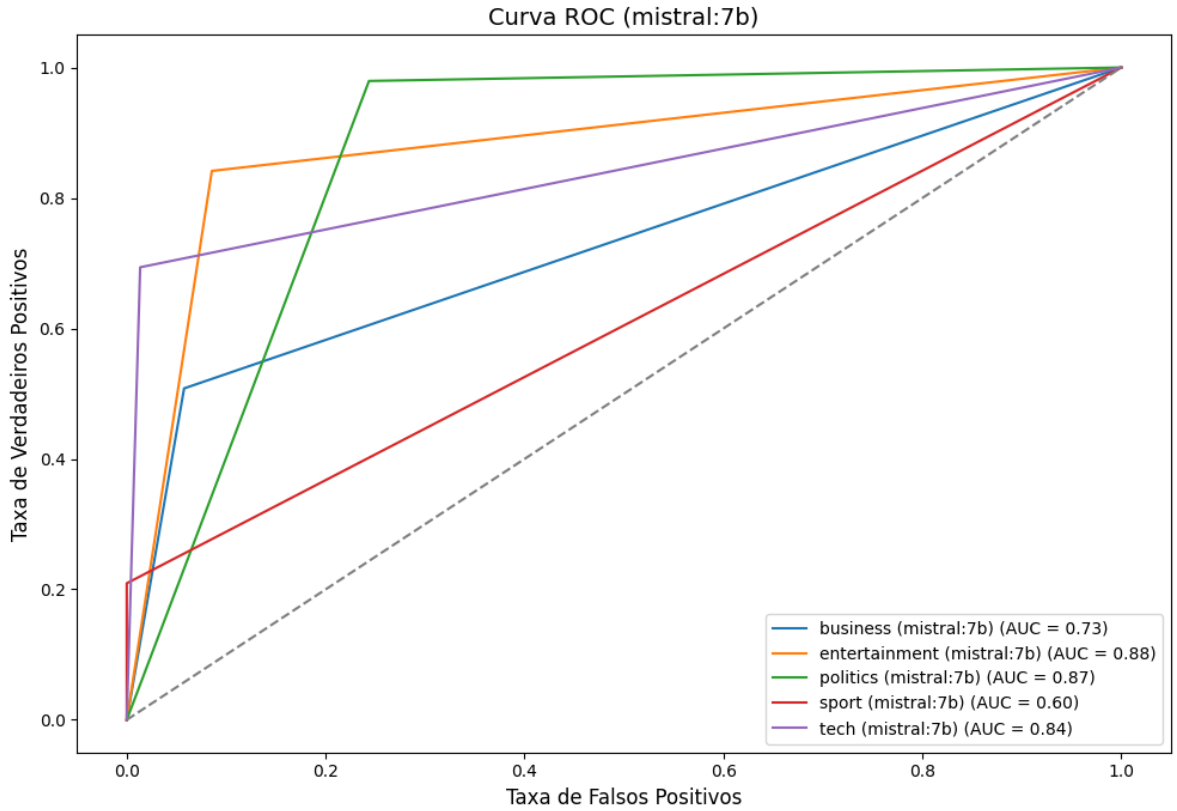
Observa-se que a categoria Sport, apesar de alcançar uma precisão perfeita (1.0000), possui um recall muito baixo (0.2093), sugerindo um viés para a previsão de poucas instâncias dessa classe, o que pode impactar negativamente sua representação na curva ROC. Já a categoria Politics, com um recall de 0.9794, indica um alto índice de verdadeiros positivos, mas seu valor de precisão relativamente baixo (0.5304) sugere uma taxa considerável de falsos positivos, influenciando diretamente sua posição na curva ROC.

O desempenho médio do modelo, conforme refletido na acurácia de 0.6881 e no F1-score macro de 0.6420, sugere que há margem para otimizações na calibração dos limiares de decisão e na estratégia de balanceamento de classes. A perplexidade elevada (1314.3538) e o log-likelihood total negativo (-14276.0276) reforçam que o modelo pode apresentar desafios na estabilidade preditiva.

Dessa forma, conforme evidenciado na Figura 4.30, a curva ROC para o

`mistral:7b` fornece insights valiosos sobre o comportamento do modelo em diferentes cenários de tipificação, apontando potenciais ajustes para melhorar seu desempenho, como a adoção de técnicas de balanceamento de classes e refinamento dos limiares de decisão.

Figura 4.30: Curva ROC para o modelo `mistral:7b`



Fonte: Próprio autor.

A Figura 4.31 ilustra a comparação entre as categorias originais e preditas pelo modelo `mistral:7b`, evidenciando a performance do classificador na tarefa de categorização. As categorias originais referem-se às etiquetas verdadeiras no conjunto de dados, enquanto as categorias preditas são aquelas atribuídas pelo modelo, permitindo uma visualização das discrepâncias entre a previsão e a realidade.

A análise desse gráfico revela a eficácia do modelo em distinguir corretamente entre as diferentes categorias de texto, com destaque para as seguintes observações:

Business: O modelo apresenta uma boa precisão (0.7456), mas um recall relativamente mais baixo (0.5081), sugerindo que, embora seja bom em identificar textos da categoria Business, ele tende a não capturar todos os textos pertencentes a essa classe. Isso pode ser observado na discrepância entre as barras de originais e preditas, com uma proporção considerável de falsos negativos.

Entertainment: A categoria Entertainment tem um desempenho impressionante, com precisão (0.7188) e recall (0.8415) elevados. O modelo foi eficaz ao identificar a categoria, mas ainda há uma pequena margem de erro, como indicado pela diferença

entre as barras, principalmente em relação a outros textos que foram incorretamente classificados em outras categorias.

Politics: A categoria Politics revela um comportamento interessante, com recall muito alto (0.9794), mas uma precisão relativamente baixa (0.5304). Isso sugere que o modelo está identificando muitas instâncias de Politics como pertencentes a essa classe, mas também cometendo um número significativo de falsos positivos, onde textos de outras categorias são erroneamente classificados como Politics.

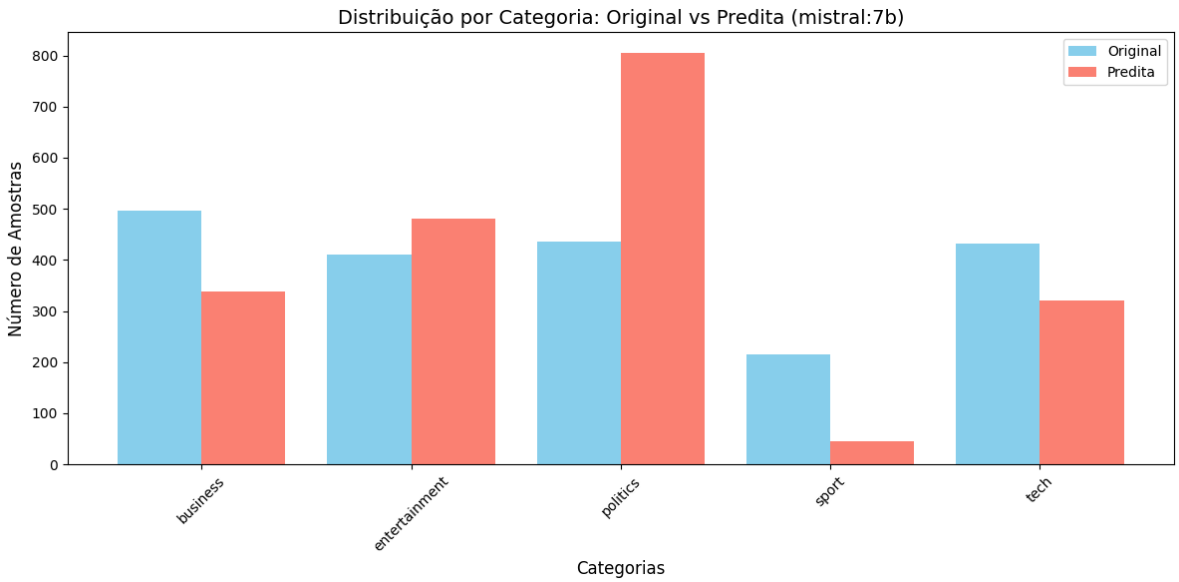
Sport: A categoria Sport apresenta um desempenho peculiar, com precisão perfeita (1.0000), mas um recall muito baixo (0.2093). Isso indica que o modelo identifica corretamente os textos que pertencem a Sport, mas falha em capturar uma quantidade substancial de instâncias reais dessa classe. O gráfico de categorias originais versus preditas mostra uma grande discrepância, com muitos textos de Sport sendo classificados como outras categorias.

Tech: A categoria Tech apresenta bom equilíbrio entre precisão (0.9344) e recall (0.6937), com o modelo conseguindo identificar adequadamente as instâncias dessa classe, embora ainda ocorra uma taxa de erros.

O macro average (0.6420) e o weighted average (0.6717) sugerem que, embora o modelo tenha um desempenho razoável em termos de F1-score, há uma variação significativa entre as categorias. A comparação entre as categorias originais e preditas é essencial para entender onde o modelo comete seus erros e pode ser melhorado, especialmente em categorias como Politics e Sport.

Em suma, o gráfico das categorias originais versus preditas para o modelo `mistral:7b` fornece uma visão clara das forças e limitações do modelo, indicando áreas que precisam de ajustes, como o balanceamento entre precisão e recall em certas categorias.

Figura 4.31: Distribuição por Categoria do modelo `mistral:7b`



Fonte: Próprio autor.

A Figura 4.32 apresenta o gráfico relacionado aos valores de perplexidade, log likelihood total e exatidão (exact match) do modelo `mistral:7b`. Esses três indicadores são essenciais para a avaliação do desempenho geral do modelo, especialmente em tarefas de previsão de texto, como tipificação de categorias.

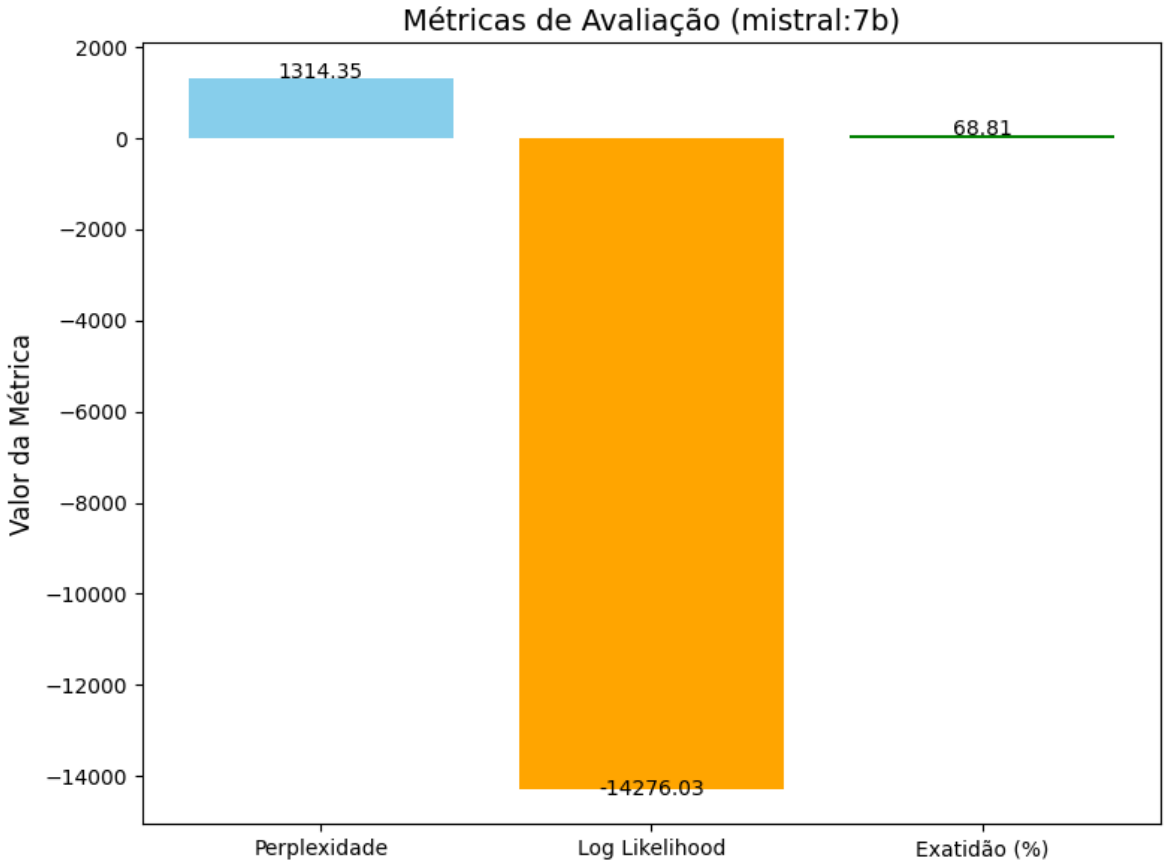
Perplexidade: 1314.3538 A perplexidade de 1314.3538 é um indicador estatístico fundamental em modelos de linguagem. Ela mede a incerteza ou complexidade do modelo em relação às previsões que realiza. Em termos simples, uma perplexidade mais baixa indica que o modelo tem um melhor entendimento do padrão no conjunto de dados e consegue fazer previsões mais confiáveis. Neste caso, o valor relativamente alto de perplexidade sugere que o modelo mistral:7b apresenta uma certa dificuldade em lidar com o contexto e gerar previsões com alta confiança. Esse valor de perplexidade implica que o modelo está relativamente mais "perplexo"ao lidar com os dados, o que pode ser um reflexo de uma modelagem subótima ou de dados complexos que exigem mais capacidade para serem entendidos adequadamente.

Log Likelihood Total: -14276.0276 O log likelihood total é outra métrica crucial, que quantifica a probabilidade de o modelo ter gerado o conjunto de dados observado. Um valor mais negativo indica que o modelo não está gerando as previsões de maneira eficiente. Nesse caso, o valor de -14276.0276 sugere que o modelo mistral:7b não está conseguindo gerar as distribuições de probabilidade ideais para os dados de entrada, o que está relacionado à sua performance em termos de perplexidade. A expectativa é que o modelo consiga minimizar esse valor com o tempo, especialmente durante o treinamento, como forma de melhorar sua acurácia e reduzir a incerteza nas previsões.

Exatidão (Exact Match): 0.6881 A exatidão (exact match) é uma métrica importante que indica a proporção de instâncias nas quais as categorias preditas coincidem exatamente com as categorias reais. O valor de 0.6881 indica que o modelo mistral:7b teve sucesso em acertar a categoria exata de aproximadamente 69% dos exemplos de teste, o que é um desempenho razoável, mas ainda há espaço para melhorias. Comparado com a perplexidade e o log likelihood total, a exatidão oferece uma medida prática de sucesso em termos de tipificação direta. Embora a exatidão não seja perfeita, ela é um indicador de que o modelo está sendo bem-sucedido em identificar as categorias, embora ainda com um número significativo de erros.

Em resumo, os valores de perplexidade (1314.3538), log likelihood total (-14276.0276) e exatidão (0.6881) indicam que o modelo `mistral:7b` está enfrentando desafios tanto em sua capacidade de gerar previsões confiáveis quanto na precisão dessas previsões. Embora a exatidão de 0.6881 indique um desempenho aceitável, a alta perplexidade e o valor negativo do log likelihood sugerem que há espaço para melhorias, especialmente em relação à modelagem probabilística e à capacidade do modelo de generalizar para novos dados. A análise de figuras como a apresentada aqui é crucial para a identificação de áreas em que ajustes ou otimizações adicionais podem ser necessários para melhorar o desempenho do modelo.

Figura 4.32: Gráfico de métricas para o modelo `mistral:7b`



Fonte: Próprio autor.

‘‘‘

Capítulo 5

Trabalhos Futuros

Com base nos resultados obtidos e nas análises realizadas ao longo desta pesquisa, diversos caminhos podem ser explorados para a continuidade e aprimoramento do trabalho. As sugestões a seguir visam aprofundar o estudo da tipificação automatizada de documentos utilizando Grandes Modelos de Linguagem (LLMs) e expandir sua aplicabilidade em contextos reais e complexos.

5.1 Aprimoramento de Dados e Escopo da Avaliação

Uma primeira vertente de pesquisa consiste na **expansão e diversificação do conjunto de dados**. A utilização de um corpus mais amplo, abrangendo múltiplos domínios (e.g., documentos jurídicos, acadêmicos), idiomas e níveis de complexidade estrutural, permitiria uma avaliação mais robusta da capacidade de generalização dos modelos. A inclusão de documentos reais oriundos de sistemas como o SEI, conforme a motivação inicial do trabalho, seria um passo fundamental para validar a abordagem em um cenário prático.

Associado a isso, está o **tratamento do desbalanceamento de classes**. Como o desempenho dos modelos pode ser impactado por distribuições desiguais de categorias, a aplicação de estratégias de balanceamento — como técnicas de reamostragem (*oversampling* e *undersampling*) ou a geração de dados sintéticos — é uma linha de pesquisa promissora, que pode ser explorada com base em abordagens já estabelecidas na literatura (SANTOS et al., 2023).

5.2 Aprofundamento das Técnicas de Modelagem

Outra frente de trabalho envolve a **exploração de um leque mais amplo de modelos e técnicas**. Embora esta pesquisa tenha se concentrado em um conjunto pragmático de LLMs de código aberto, a investigação poderia ser estendida para incluir outras famílias de modelos, como as da série GPT, Falcon ou Claude (KUKREJA

et al., 2024). A comparação com modelos proprietários poderia reforçar as conclusões sobre as vantagens dos modelos abertos.

Além disso, o desempenho dos modelos poderia ser potencializado pela aplicação de técnicas mais avançadas de **ajuste fino e engenharia de *prompt***. O ajuste fino (*fine-tuning*) dos LLMs em conjuntos de dados específicos de tipificação documental é uma abordagem poderosa para especializar o modelo em um determinado domínio (WEI et al., 2023). Da mesma forma, a exploração de técnicas avançadas de engenharia de *prompt*, como o *prompt tuning*, poderia otimizar os resultados sem a necessidade de retreinar todos os parâmetros do modelo (LESTER; AL-RFOU; CONSTANT, 2021).

5.3 Desenvolvimento e Validação de Aplicação Prática

Por fim, uma direção de grande impacto seria a transposição da arquitetura experimental para um **sistema interativo e sua avaliação em um ambiente real**. O desenvolvimento de uma aplicação web com uma interface gráfica amigável, que integre o *pipeline* de tipificação, facilitaria a utilização prática por usuários finais em órgãos públicos e outras instituições.

A etapa subsequente seria a realização de **estudos de caso**, aplicando o sistema em fluxos documentais reais. Essa validação prática permitiria avaliar não apenas a acurácia dos modelos, mas também sua eficiência operacional, o impacto nos processos de trabalho e a aceitação por parte dos usuários, fechando o ciclo de pesquisa aplicada.

Capítulo 6

Conclusão

Este trabalho propôs-se a investigar um desafio central na gestão da informação moderna: a tipificação automatizada de documentos. Diante da crescente produção de dados textuais, o **objetivo geral** desta dissertação foi avaliar e comparar sistematicamente a eficácia de diferentes Grandes Modelos de Linguagem (LLMs) de código aberto para essa tarefa, utilizando um conjunto abrangente de métricas de desempenho. Ao final desta investigação, pode-se afirmar que este objetivo foi plenamente alcançado.

Para atingir esse propósito, foi desenvolvida e implementada uma arquitetura de software modular e assíncrona, que permitiu a avaliação rigorosa de oito modelos proeminentes das famílias LLaMA 3, Gemma, Mistral e DeepSeek. O *pipeline* experimental, detalhado no Capítulo 3, garantiu a reprodutibilidade e a consistência da análise sobre o *dataset* BBC News, viabilizando a resposta aos objetivos específicos traçados.

Em resposta ao **primeiro objetivo específico**, foi realizada uma avaliação quantitativa aprofundada, que transcendeu a acurácia ao incorporar métricas de precisão, revocação, F1-Score, perplexidade e log-likelihood. Essa análise multimétrica, em resposta ao **segundo objetivo**, revelou uma variação de desempenho substancial entre os modelos e entre as categorias temáticas. Em alinhamento com o **terceiro objetivo**, foi possível identificar as forças e limitações de cada arquitetura, notando-se que modelos como `llama3.1:8b` apresentavam um perfil de erro com trade-offs específicos entre precisão e revocação.

Os resultados demonstraram a clara superioridade dos modelos da família LLaMA 3, que consistentemente apresentaram acurácias superiores a 87% e baixa perplexidade, consolidando-se como as ferramentas mais robustas e confiáveis para a tarefa. Em contraste, modelos como `deepseek-llm:7b` e, de forma mais acentuada, `llama3.2:latest`, mostraram-se significativamente menos eficazes, reforçando a conclusão de que a escolha de um LLM para aplicações práticas não pode ser generalizada e exige uma avaliação empírica criteriosa.

Assim, o **quarto objetivo** foi cumprido ao se prover recomendações práticas claras: para a tarefa de tipificação nos moldes deste estudo, a família LLaMA 3 é a

escolha preferencial. A principal contribuição desta dissertação reside, portanto, na geração deste *benchmark* empírico e pragmático, que oferece um guia quantitativo para a seleção de LLMs de código aberto em cenários de tipificação documental.

Finalmente, reconhecendo as limitações do presente estudo, como o uso de um único *dataset* em língua inglesa, o **quinto objetivo específico** foi atendido no Capítulo 5, que sugere direções para futuras pesquisas. Propostas como a expansão para outros domínios e idiomas, a exploração de técnicas de ajuste fino (*fine-tuning*) e o desenvolvimento de sistemas interativos poderão aprimorar e ampliar os resultados aqui obtidos.

Conclui-se, portanto, que o uso de LLMs representa uma alternativa poderosa e eficiente para a automação da tipificação documental. Este trabalho não apenas valida essa afirmação com evidências rigorosas, mas também fornece uma metodologia e um conjunto de resultados que podem servir de base para futuras implementações e investigações na área, consolidando o uso dessas tecnologias como uma ferramenta estratégica para a gestão da informação.

REFERÊNCIAS BIBLIOGRÁFICAS

AKHTAR, Z. B. Unveiling the evolution of generative ai (gai): a comprehensive and investigative analysis toward llm models (2021-2024) and beyond. *Journal of Electrical Systems and Information Technology*, Springer, v. 11, n. 1, p. 22, 2024.

ALAKTIF, A.; CHERGUI, M.; DAOUDI, I.; AMMOUMOU, A. All you should know about large language models (llms). In: IEEE. *2024 7th International Conference on Advanced Communication Technologies and Networking (CommNet)*. [S.l.], 2024. p. 1–10.

ARSLAN, M.; MUNAWAR, S.; CRUZ, C. Exploring business events using multi-source rag. *Procedia Computer Science*, Elsevier, v. 246, p. 4534–4540, 2024.

BOYINA, K.; REDDY, G. M.; AKSHITA, G.; NAIR, P. C. Zero-shot and few-shot learning for telugu news classification: A large language model approach. In: IEEE. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. [S.l.], 2024. p. 1–7.

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2020. v. 33, p. 1877–1901.

CHAE, Y.; DAVIDSON, T. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*, v. 10, 2023.

CHANG, Y.; WANG, X.; WANG, J.; WU, Y.; YANG, L.; ZHU, K.; CHEN, H.; YI, X.; WANG, C.-T.; WANG, Y. et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, ACM New York, NY, v. 15, n. 3, p. 1–45, 2024.

CHEN, J.; WARREN, D. Cost-sensitive learning for large-scale hierarchical classification. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. [S.l.: s.n.], 2013. p. 1351–1360.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186.

DONG, H.; WANG, Z. Large language models for tabular data: Progresses and future directions. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. [S.l.: s.n.], 2024. p. 2997–3000.

FIELDS, J.; CHOVANEC, K.; MADIRAJU, P. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, IEEE, v. 12, p. 6518–6531, 2024.

GASPARETTO, A.; MARCUZZO, M.; ZANGARI, A.; ALBARELLI, A. A survey on text classification algorithms: From text to predictions. *Information*, MDPI, v. 13, n. 2, p. 83, 2022.

GREENE, D.; CUNNINGHAM, P. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In: *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2006. (ACM International Conference Proceeding Series, v. 148), p. 297–304.

HAQUE, R.; GOH, H.-N.; TING, C.-Y.; QUEK, A.; HASAN, M. R. Leveraging llms for optimised feature selection and embedding in structured data: A case study on graduate employment classification. *Computers and Education: Artificial Intelligence*, Elsevier, v. 8, p. 100356, 2025.

HUA, M.; ZHAO, Q.; SONG, J.; TANG, X.-s. Two-stage compliance detection for power enterprises based on nli and llm. In: *2024 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-ASIA)*. [S.l.]: IEEE, 2024. p. 1–5.

HUA, X. *Improving Controllability for Neural Text Generation*. Tese (Doutorado) — University of Texas at Dallas, 2020.

HUANG, W.; ZHENG, X.; MA, X.; QIN, H.; LV, C.; CHEN, H.; LUO, J.; QI, X.; LIU, X.; MAGNO, M. An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence*, Springer, v. 2, n. 1, p. 36, 2024.

HUNTER, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, IEEE, v. 9, n. 3, p. 90–95, 2007.

HUPKES, D.; GIULIANELLI, M.; DANKERS, V.; ARTETXE, M.; ELAZAR, Y.; PIMENTEL, T.; CHRISTODOULOPOULOS, C.; LASRI, K.; SAPHRA, N.; SINCLAIR, A. et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, Nature Publishing Group UK London, v. 5, n. 10, p. 1161–1174, 2023.

IQBAL, T.; QURESHI, S. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, v. 34, n. 6, p. 2515–2528, 2022.

JARADAT, S.; ACHARYA, N.; SHIVSHANKAR, S.; ALHADIDI, T. I.; ELHENAWY, M. Ai for data quality auditing: Detecting mislabeled work zone crashes using large language models. *Algorithms*, MDPI, v. 18, n. 6, p. 317, 2025.

JIANG, W.; WANG, B.; LI, Z.; LIU, B. A comprehensive evaluation of large language models on text classification. *arXiv preprint arXiv:2304.02324*, 2023.

KHOBOKO, P. W.; MARIVATE, V.; SEFARA, J. Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, Elsevier, v. 20, p. 100649, 2025.

KO, I.-Y.; NECHES, R. Composing web services for large-scale tasks. *IEEE Internet Computing*, IEEE Computer Society, v. 7, n. 05, p. 52–59, 2003.

KOWSARI, K.; MEIMANDI, K. J.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text classification algorithms: A survey. *Information*, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 150, 2019.

KUKREJA, S.; KUMAR, T.; PUROHIT, A.; DASGUPTA, A.; GUHA, D. A literature survey on open source large language models. In: *Proceedings of the 2024 7th International Conference on Computers in Management and Business*. [S.l.: s.n.], 2024. p. 133–143.

LE, N. T. K.; HADIPRODJO, N.; EL-ALFY, H.; KERIMZHANOV, A.; TESHEBAEV, A. The recent large language models in nlp. *2023 22nd International Symposium on Communications and Information Technologies (ISCIT)*, IEEE, p. 1–6, 2023.

LESTER, B.; AL-RFOU, R.; CONSTANT, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. [S.l.]: Association for Computational Linguistics, 2021. p. 3045–3059.

LIANG, P.; BOMMASANI, R.; LEE, T.; MADAIO, M.; WANG, C.; POTTS, C. et al. Holistic Evaluation of Language Models. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2022. v. 35, p. 1–22. ArXiv:2211.09110.

LIU, F.; CHEN, D.; GUAN, Z.; ZHOU, X.; ZHU, J.; YE, Q.; FU, L.; ZHOU, J. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, IEEE, 2024.

LIU, N. F.; LIN, K.; HEWITT, J.; CHEUNG, A.; LIU, P.; MANNING, C. D. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 12, p. 157–173, 2024.

MIROŃCZUK, M. M.; MÜLLER, A.; PEDRYCZ, W. The outcomes and publication standards of research descriptions in document classification: a systematic review. *IEEE Access*, IEEE, 2024.

MUAAD, A. Y.; KUMAR, G. H.; HANUMANTHAPPA, J.; BENIFA, J. B.; MOURYA, M. N.; CHOLA, C.; PRAMODHA, M.; BHAIRAVA, R. An effective approach for arabic document classification using machine learning. *Global Transitions Proceedings*, Elsevier, v. 3, n. 1, p. 267–271, 2022.

NASUTION, A. H.; ONAN, A. Chatgpt label: Comparing the quality of human-generated and Ilm-generated annotations in low-resource language nlp tasks. *IEEE Access*, IEEE, v. 12, p. 71876–71900, 2024.

NAZI, Z.; PENG, W.-C. H. Large Language Models in Healthcare and Medical Domain: A Review. *Informatics*, MDPI, v. 11, n. 3, p. 57, 2024.

PATIL, R.; BOIT, S.; GUDIVADA, V.; NANDIGAM, J. A survey of text representation and embedding techniques in nlp. *IEEE Access*, IEEE, v. 11, p. 36120–36146, 2023.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR)*, v. 12, p. 2825–2830, 2011.

PREUSS, N.; ALSHEHRI, A. S.; YOU, F. Large language models for life cycle assessments: Opportunities, challenges, and risks. *Journal of Cleaner Production*, Elsevier, p. 142824, 2024.

PŁONKA, M.; KOSOT, K.; HOŁDA, K.; DANIEC, K.; NAWRAT, A. A comparative evaluation of the effectiveness of document splitters for large language models in legal contexts. *Expert Systems with Applications*, Elsevier, p. 126711, 2025.

QIN, R.; YANG, K.; ABBASI, A.; DOBOLYI, D.; SEYEDI, S.; GRINER, E.; KWON, H.; COTES, R.; JIANG, Z.; CLIFFORD, G. D. et al. Language models for online depression detection: A review and benchmark analysis on remote interviews. *ACM Transactions on Management Information Systems*, ACM New York, NY, v. 16, n. 2, p. 1–35, 2025.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. *Improving Language Understanding by Generative Pre-Training*. [S.l.], 2018. Technical Report.

RAMíREZ, S. *FastAPI: A modern, fast (high-performance), web framework for building APIs with Python 3.8+ based on standard Python type hints*. 2024. <https://fastapi.tiangolo.com/>. Accessed: June 30, 2025.

SAINZ, O.; LACALLE, O. Lopez de; LABAKA, G.; AGIRRE, E.; SOROA, A. A Stitch in Time Saves Nine: A Train-Time Regularizing Loss for Protecting against Test-Time Contamination. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, 2023. p. 8248–8262.

SANTOS, D. P. D.; COSTA, J. P. J. d.; SILVA, D. A. d.; MENDONçA, F.; VEIGA, C.; SOUSA, R. T. d. Multi-Class Text Classification Based in Oversampling for Highly Imbalanced Dataset. In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. [S.l.]: IEEE, 2023. p. 752–755.

SANTOS, D. P. D.; MENDONCA, F. L.; LUSTOSA, J. P.; SERRANO, A.; TORRES, J. A. S.; SILVA, D. A. d. Large Language Models for Text Classification: A New Era of Accuracy and Efficiency. In: *International Conference on Computational Science and Computational Intelligence*. [S.l.]: Springer, 2025. p. 3–15.

SCHMIDT, L.; MUTLU, A. N.; ELMORE, R.; OLORISADE, B. K.; THOMAS, J.; HIGGINS, J. P. Data extraction methods for systematic review (semi) automation: Update of a living systematic review. *F1000Research*, v. 10, p. 401, 2025.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, Elsevier, v. 45, n. 4, p. 427–437, 2009.

SUMANATHILAKA, D.; MICALLEF, N.; HOUGH, J. Assessing gpt's potential for word sense disambiguation: A quantitative evaluation on prompt engineering techniques. In: IEEE. *2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC)*. [S.l.], 2024. p. 204–209.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2017. v. 30, p. 5998–6008.

WANG, A.; PRUKSACHATKUN, Y.; NANGIA, N.; SINGH, A.; MICHAEL, J.; HILL, F.; LEVY, O.; BOWMAN, S. R. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2019. v. 32, p. 3261–3275.

WANG, A.; SINGH, A.; MICHAEL, J.; HILL, F.; LEVY, O.; BOWMAN, S. R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 353–355.

WANG, Y.; XU, M.; YAN, Y.; ZHAO, T.; CHEN, Y.; YANG, J. Exploring topic supervision with bert for text matching. In: IEEE. *2022 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2022. p. 1–7.

WASKOM, M. L. Seaborn: statistical data visualization. *Journal of Open Source Software*, v. 6, n. 60, p. 3021, 2021.

WEI, F.; KEELING, R.; HUBER-FLIFLET, N.; ZHANG, J.; DABROWSKI, A.; YANG, J.; MAO, Q.; QIN, H. Empirical study of llm fine-tuning for text classification in legal document review. In: *2023 IEEE International Conference on Big Data (BigData)*. [S.l.]: IEEE, 2023. p. 2786–2792.

WEI, J.; BOSMA, M.; ZHAO, V. Y.; GUU, K.; YU, A. W.; LESTER, B.; DU, N.; DAI, A. M.; LE, Q. V. Finetuned Language Models Are Zero-Shot Learners. *arXiv preprint arXiv:2109.01652*, 2021.

XU, H. *Enhancing NLP Capabilities: Strategies for Language Model Adaptation in Low-Resource Text Classification Task and Evaluations*. Tese (Doutorado) — Temple University, 2025.

YAO, Y.; DUAN, J.; XU, K.; CAI, Y.; SUN, Z.; ZHANG, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, Elsevier, p. 100211, 2024.

YIN, X.; NI, C.; WANG, S. Multitask-based evaluation of open-source llm on software vulnerability. *IEEE Transactions on Software Engineering*, IEEE, 2024.

ZHANG, H.; ZHANG, Y.; WANG, X.; ZHANG, L.; JI, L. An interactive multi-task esg classification method for chinese financial texts. *Applied Intelligence*, Springer, v. 55, n. 2, p. 191, 2025.

ZHANG, Y.; TSUDA, K. Nbbench: Benchmarking language models for comprehensive nanobody tasks. *arXiv preprint arXiv:2505.02022*, 2025.

ZHANG, Y.; YANG, R.; XU, X.; LI, R.; XIAO, J.; SHEN, J.; HAN, J. Teleclass: Taxonomy enrichment and Ilm-enhanced hierarchical text classification with minimal supervision. In: *Proceedings of the ACM on Web Conference 2025*. [S.l.: s.n.], 2025. p. 2032–2042.

ZHAO, Z.; WALLACE, E.; FENG, S.; KLEIN, D.; SINGH, S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In: MEILA, M.; ZHANG, T. (Ed.). *Proceedings of the 38th International Conference on Machine Learning (ICML)*. [S.l.]: PMLR, 2021. (Proceedings of Machine Learning Research, v. 139), p. 12697–12706.

Apêndice A

Publicações

Durante o desenvolvimento desta dissertação, os seguintes artigos científicos foram publicados:

1. Santos, Dário P. Dos, et al. "Large Language Models for Text Classification: A New Era of Accuracy and Efficiency."International Conference on Computational Science and Computational Intelligence. Springer, Cham, 2025. (Publicado)
2. Dos Santos, Dário P., et al. "Multi-Class Text Classification Based in Oversampling for Highly Imbalanced Dataset."2023 International Conference on Machine Learning and Applications (ICMLA). IEEE, 2023. (Publicado)
3. Santos, Dário P. Dos, et al. "ANÁLISE COMPARATIVA VIA CLASSIFICAÇÃO DE TEXTOS GERADOS PELO CHATGPT VERSUS TEXTOS ESCRITOS MANUALMENTE". IADIS Conferencia Ibero Americana WWW/Internet 2023. (Publicado)