Cartas Indígenas ao Brasil: Classificação Multi-Rótulo

Caio Almeida
Universidade Federal da Bahia

Renata Vieira
Universidade de Évora

Débora Abdalla
Universidade Federal da Bahia

Resumo

Este artigo investiga a classificação automática multi-rótulo de cartas indígenas ao Brasil em categorias temáticas. A partir do acervo digital *Cartas Indígenas ao Brasil*, que constitui um corpus de 871 cartas anotadas em 18 categorias, comparamos três abordagens de classificação: um modelo lexical (TF-IDF + regressão logística), um modelo contextual (BERTimbau-base) e um classificador baseado em grandes modelos de linguagem (LLM). Para lidar com o desbalanceamento do corpus, empregamos estratégias de balanceamento de classes no modelo neural. Os resultados revelam um trade-off entre precisão e recall: o baseline lexical apresenta maior precisão (0,65), enquanto o BERTimbau demonstra maior recall (0,67), especialmente em categorias minoritárias. Ambos alcançam macro-F1 de 0,42, evidenciando que a classificação multi-rótulo neste domínio é uma tarefa desafiadora, em especial devido ao desbalanceamento do corpus e à sobreposição semântica entre categorias. O classificador baseado em LLM atinge alto recall, especialmente em categorias minoritárias, mas tende a superestimar o número de rótulos por documento, reforçando o trade-off entre precisão e cobertura observado nas outras duas abordagens. A análise detalhada por classe revela comportamentos complementares entre os modelos, sugerindo que abordagens híbridas podem superar as limitações individuais de cada método. O corpus e os scripts dos experimentos serão disponibilizados publicamente.

1 Introdução

A classificação automática de textos é uma tarefa fundamental em Processamento de Linguagem Natural, com aplicações que vão desde a organização de documentos até a detecção de notícias falsas, narrativas ou discurso de ódio. No entanto, a maioria dos recursos e modelos disponíveis para português foi desenvolvida com base em textos jornalísticos, redes sociais ou domínios técnicos, deixando lacunas significativas em domínios específicos como textos produzidos por comunidades indígenas.

O acervo *Cartas Indígenas ao Brasil* (Costa, 2013, 2018, 2021, 2019; Costa e Xucuru-Kariri, 2018) reúne documentos que vão do século XVII ao XXI, constituindo uma fonte única para o estudo das vozes indígenas brasileiras. Essas cartas apresentam características que tornam sua classificação automática particularmente desafiadora: vocabulário específico, múltiplos temas interconectados por carta, e distribuição altamente desbalanceada entre categorias.

Nosso artigo investiga a viabilidade de classificar automaticamente essas cartas em categorias temáticas, comparando três abordagens: um modelo lexical clássico (TF-IDF + regressão logística), um modelo neural contextual (BERTimbau) e um classificador baseado em grandes modelos de linguagem (LLM) via *prompting*. Especificamente, buscamos entender qual abordagem é mais adequada para este domínio, como o desbalanceamento extremo entre categorias afeta o desempenho e se os modelos apresentam comportamentos complementares que possam ser explorados.

As contribuições deste trabalho são: (1) uma análise comparativa detalhada de classificadores em um domínio pouco explorado; (2) análise de impatcot de estratégias de balanceamento; (3) evidências de complementaridade entre modelos lexicais, contextuais e baseadas em LLM; e (4) disponibilização pública do corpus, prompts e scripts.

2 Motivação

No Brasil, comunidades indígenas, quilombolas e outros grupos minoritários são alvos frequentes de campanhas de desinformação que aumentam desigualdades sociais e exclusão digital (Arora et al., 2023; UNESCO, 2021), e as tecnologias de IA atuais não são adaptadas para suas realidades (O'Neil,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 708–716

2016). O projeto em que este experimento se insere visa preencher essa lacuna, desenvolvendo ferramentas de IA de código aberto que ofereçam suporte específico para a verificação de narrativas em tais contextos.

Comunidades marginalizadas são especialmente vulneráveis à desinformação devido a fatores como barreiras linguísticas e desigualdades históricas. Mesmo com avanços recentes na representatividade política dessas comunidades no Brasil, ainda há pouca participação em debates relacionados à governança e aplicação de IA (UNESCO, 2021; Prestes et al., 2024).

Nossa pesquisa propõe utilizar o conteúdo das cartas indígenas, estruturando esses dados para análises posteriores, como categorização multi-rótulo. A longo prazo, pretende-se compreender como essas narrativas impactam dinâmicas sociais contemporâneas em redes sociais, especialmente em contextos de desinformação.

3 O Acervo Cartas Indígenas ao Brasil

A base empírica desta pesquisa é o acervo virtual do Projeto "As Cartas dos Povos Indígenas ao Brasil". A plataforma do projeto disponibiliza gratuitamente cartas públicas escritas por indivíduos, povos e organizações indígenas. Essas correspondências são, em sua maioria, destinadas ao Estado brasileiro, representado por seus dirigentes, como Presidentes da República e o Congresso Nacional, bem como à imprensa e à sociedade brasileira (Costa, 2013, 2018, 2021, 2019; Costa e Xucuru-Kariri, 2018).

As cartas abrangem o período da década de 1970 até os dias atuais, mas há registros de outras épocas históricas, como o conjunto de escritos em Tupi dos Potiguara, no século XVII. O acervo foi construído por meio da coleta de documentação dispersa em sites de organizações indigenistas e indígenas, jornais, redes sociais, grupos de e-mail, podcasts, aplicativos de mensagens, catálogos de exposições de arte, livros, artigos científicos e arquivos públicos online. No total, foram analisadas mais de 100 mil páginas virtuais. Ressalta-se que esta é a primeira pesquisa comparativa sobre correspondências indígenas, tratando-as como corpus empírico e tornando-as acessíveis a um público mais amplo. Utilizando uma metodologia qualitativa, foram reunidos mais de mil documentos, classificados segundo quatro critérios baseados na literatura sobre o tema e na definição dos sujeitos da pesquisa:

O primeiro critério foi a disponibilidade gratuita na internet, assegurando que as epístolas tivessem difusão autorizada pelos autores e acesso livre. O segundo critério considerou correspondências endereçadas a destinatários públicos, como representantes do Estado e governos, imprensa e organizações civis, incluindo as próprias organizações indígenas e indigenistas. Também foram incluídos destinatários indígenas, como lideranças, aldeias, territórios, movimentos sociais e povos.

O terceiro critério foi a presença de dialogia no conteúdo, característica essencial para definir um documento como carta, mesmo quando faltavam aspectos formais do gênero epistolar. Assim, documentos finais de eventos e reuniões, ofícios e notas foram incluídos se atendessem às definições teóricas propostas pelos autores.

O quarto critério considerou a identificação dos documentos a partir de elementos caracterizadores de cartas, definidos pelos sujeitos da pesquisa. Muitas vezes, a própria nomenclatura do texto indicava sua natureza epistolar. Entretanto, também foram analisados documentos referenciados por termos correlatos, como "esta memória", "este memorial", "nossa história", "nossa conversa", "mensagem", "manifesto" e "depoimento". Esses critérios foram estabelecidos a partir de leituras teóricas e da prática investigativa, que constantemente questionava se determinado documento poderia ser considerado uma carta.

As cartas foram coletadas e classificadas conforme categorias descritivas de conteúdo no contexto do projeto "As Cartas dos Povos Indígenas ao Brasil". Essa categorização temática prévia foi utilizada nesse trabalho. O material fornece ainda indicação de autoria, data e local de publicação, autoria individual ou coletiva, tipo de destinatário, verbos e expressões que descrevem ações, sentimentos e reflexões dos autores, dos antagonistas ou parceiros dos indígenas e dos destinatários. Também foram analisados os objetivos dos remetentes ao escrever as correspondências, como reclamar, comunicar, denunciar ou pleitear, e os temas centrais abordados. A análise priorizou a relação entre as quatro principais categorias: autoria, destinatário, assunto e contexto de produção, buscando construir uma narrativa sobre as expressões, verbos e modos de dizer utilizados pelos remetentes para expressar sua relação com o Brasil.

4 Trabalhos Relacionados

A classificação multi-rótulo de textos tem sido amplamente estudada em PLN, com abordagens que vão desde métodos clássicos baseados em TF-IDF (Salton e Buckley, 1988) até modelos neurais contextuais. Em português, o BERTimbau (Souza et al., 2020) tem apresentado resultados competitivos em diversas tarefas, incluindo classificação de sentimento e tópicos.

O desbalanceamento de classes é um desafio recorrente em classificação multi-rótulo, especialmente em domínios especializados. Estratégias como oversampling, class weights e focal loss têm sido propostas para mitigar esse problema, e em cenários com múltiplos rótulos por instância, a correlação entre categorias adiciona complexidade, motivando abordagens que modelam explicitamente essas dependências (Tarekegn et al., 2021).

Quanto a textos de comunidades indígenas, há poucos recursos públicos disponíveis para PLN. Os trabalhos existentes tendem a focar em aspectos linguísticos como léxico e morfossintaxe de línguas indígenas (Payne, 2014), não em classificação de textos em português produzidos por essas comunidades. Em paralelo, a literatura sobre *data colonialism* alerta para a marginalização de vozes indígenas em sistemas de IA (Arora et al., 2023; Noble, 2018).

Nosso trabalho contribui para essa literatura ao avaliar sistematicamente classificadores em um corpus de textos indígenas, analisando não apenas métricas agregadas, mas também o comportamento por categoria e a complementaridade entre abordagens.

5 Corpus para Classificação

A partir do acervo descrito na seção anterior, construímos um corpus estruturado para tarefas de classificação automática.

Implementamos um *crawler* em Python que percorre as páginas de busca do acervo e coleta, para cada carta, a URL, o título, o conteúdo principal e os rótulos de assunto fornecidos pela própria plataforma. As categorias temáticas são extraídas e resultam em uma coluna `labels` potencialmente multi-rótulo (por exemplo, *Terra|Violência*).

O conjunto bruto resultante contém 1.168 cartas, das quais 1.159 possuem conteúdo textual. Em uma análise exploratória inicial, observamos:

- Cerca de 24% das cartas não possuem rótulo de assunto, o que inviabiliza seu uso direto em classificação supervisionada.
- Muitas cartas têm múltiplos rótulos (por exemplo *Terra|Violência|Saúde*), refletindo a interseção de pautas.
- Há grande variedade de tamanho de texto (de 4 a mais de 6.000 palavras, com média de 664 e mediana de 536), o que impacta o custo de modelos neurais.

Categoria	Freq.	Categoria	Freq.
Terra	626	Política Indígena	81
Violência	256	Assassinatos	72
Saúde	219	Org. da FUNAI	71
Gestão Terr. e Amb.	191	Mulheres Indígenas	67
Educação	152	Segurança Pública	51
Política Indigenista	106	Bem viver	48
Consulta aos povos	105	Segurança alimentar	37
Meio Ambiente	103	Crianças Indígenas	37
Direitos Culturais	85	Indíg. na política	34

Tabela 1: Frequência das 18 categorias temáticas no corpus multi-rótulo (total de 871 cartas).

Estatística	Mín.	Máx.	Média	Mediana
Palavras por carta	4	6.061	663,9	536,0
Rótulos por carta	1	13	2,69	2,0

Tabela 2: Estatísticas descritivas do corpus de 871 cartas.

Para obter um corpus adequado para classificação supervisionada, adotamos uma estratégia de filtragem em que foram removidas cartas sem rótulo de assunto ou sem conteúdo textual disponível. Os rótulos foram normalizados (remoção de espaços extras e padronização de grafia) e então contabilizamos a frequência de cada rótulo em cenário multi-rótulo, removendo as categorias com menos de 20 ocorrências. Após esses filtros, obtivemos um subconjunto de 871 cartas anotadas com 18 categorias temáticas. A Tabela 1 apresenta a frequência de cada categoria no corpus final.

5.1 Estatísticas descritivas

A Tabela 2 apresenta as estatísticas descritivas do corpus. Os textos variam bastante em tamanho, de 4 a 6.061 palavras, com média de 664 e mediana de 536 palavras. Essa variação reflete a diversidade do acervo, que inclui desde breves manifestos até documentos de assembleias indígenas.

Quanto à distribuição de rótulos, cada carta possui em média 2,69 categorias (mediana de 2), variando de 1 a 13 rótulos. Essa característica multi-

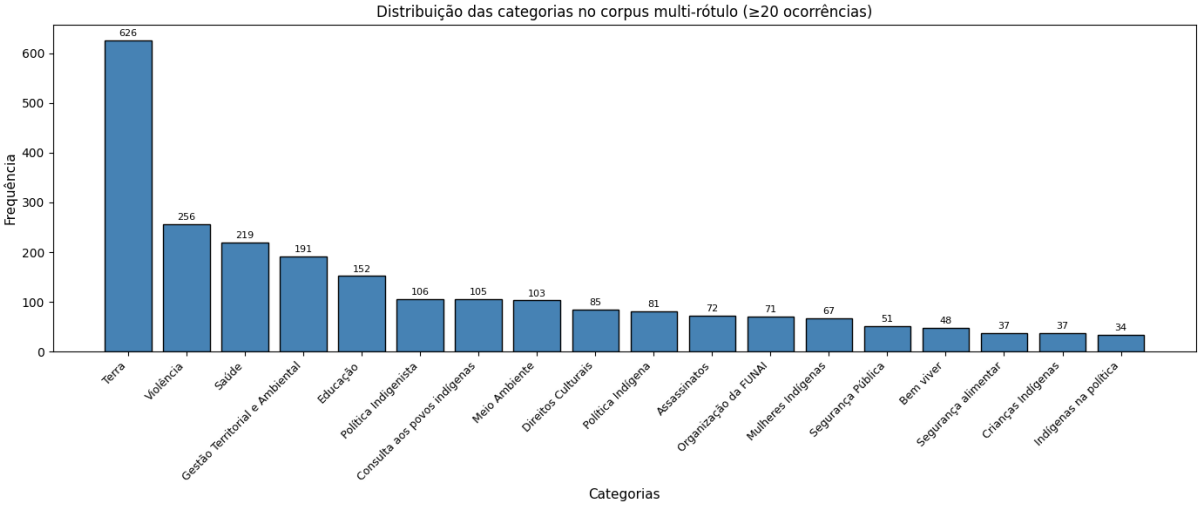

Figura 1: Distribuição das categorias no subconjunto multi-rótulo (18 classes, $\geq$ 20 ocorrências).

Categoria 1	**Categoria 2**	**Freq.**
Terra	Violência	200
Terra	Gestão Terr. e Amb.	169
Terra	Saúde	148
Terra	Educação	97
Terra	Meio Ambiente	95
Educação	Saúde	94
Terra	Política Indigenista	79
Terra	Consulta aos povos	76
Gestão Terr. e Amb.	Política Indigenista	70
Assassinatos	Violência	66

Tabela 3: As dez co-ocorrências mais frequentes entre categorias temáticas.

rótulo reflete o caráter interseccional das pautas indígenas, em que questões territoriais frequentemente se entrelaçam com violência, saúde e direitos culturais. Esses valores indicam que a maioria das cartas aborda mais de um tema simultaneamente, com um número máximo observado de 13 rótulos em um único documento. Essa distribuição reforça a adequação do enquadramento do problema como classificação multi-rótulo, uma vez que abordagens mono-rótulo seriam incapazes de capturar a complexidade temática presente no corpus.

5.2 Co-ocorrências temáticas

A análise de co-ocorrências entre categorias revela padrões importantes sobre a estrutura temática das cartas. A Tabela 3 apresenta os dez pares de categorias mais frequentes.

A categoria *Terra* aparece em 8 das 10 co-ocorrências mais frequentes, confirmando sua centralidade nas pautas indígenas. A co-ocorrência *Terra* + *Violência* (200 casos) é a mais comum, refletindo os conflitos fundiários que presentes na história das comunidades indígenas brasileiras. Outro caso que notamos também é que *Assassinatos* e *Violência* co-ocorrem frequentemente (66 casos), evidenciando a gravidade das denúncias presentes no corpus.

Além disso, observam-se co-ocorrências relevantes entre *Educação* e *Saúde* (94 casos), bem como entre *Saúde* e *Violência* (64 casos), indicando que demandas por serviços básicos e educação frequentemente aparecem articuladas a contextos de conflito e vulnerabilidade social.

Do ponto de vista analítico, essas co-ocorrências sugerem que as categorias não se comportam como semanticamente independentes no corpus analisado, mas refletem relações estruturais entre diferentes dimensões das pautas indígenas. Em termos computacionais, isso indica que modelos que assumem independência entre rótulos podem estar limitados, e que abordagens futuras poderiam explorar explicitamente dependências entre categorias para melhorar a modelagem do problema.

O corpus resultante constitui um importante recurso público de cartas indígenas brasileiras estruturado para tarefas de PLN, preenchendo uma lacuna na disponibilidade de dados textuais produzidos por comunidades indígenas para pesquisa em processamento de linguagem natural. O corpus anotado e os scripts de coleta serão disponibilizados publicamente para fomentar pesquisas futuras.

Classe	Peso	Classe	Peso
Terra	0,39	Política Indígena	9,50
Violência	2,65	Direitos Culturais	10,94
Saúde	2,98	Org. da FUNAI	10,94
Gestão Terr. e Amb.	3,65	Assassinatos	12,24
Educação	4,64	Mulheres Indígenas	13,16
Política Indigenista	6,91	Bem viver	16,91
Consulta aos povos	7,46	Segurança Pública	18,65
Meio Ambiente	7,70	Segurança alimentar	20,75
		Crianças Indígenas	24,38
		Indíg. na política	25,48

Tabela 4: Pesos por classe (*pos_weight*), calculados como razão neg/pos. Classes frequentes têm peso baixo; classes raras têm peso alto.

Época	Loss Tr.	Loss Val.	Micro-F1	Macro-F1
1	1,196	1,263	0,190	0,160
2	1,081	1,221	0,452	0,319
3	1,091	1,170	0,403	0,328
4	0,848	1,153	0,430	0,355
5	0,827	1,127	0,470	0,387
6	0,766	1,153	0,524	**0,400**
7	0,660	1,178	0,524	0,395
8	0,653	1,203	0,533	0,399

Tabela 5: Evolução do treinamento do BERTimbau. O melhor modelo (época 6) foi selecionado pelo macro-F1 na validação. O treinamento foi interrompido na época 8 pelo *early stopping*.

6 Metodologia

Formulamos o problema como classificação multi-rótulo, em que cada carta pode ser associada a uma ou mais categorias temáticas dentre 18 possíveis. Utilizamos sempre o texto completo da carta. Dividimos o corpus em três conjuntos: treino (70%, 609 cartas), validação (15%, 131 cartas) e teste (15%, 131 cartas), com partição aleatória e semente fixa para reprodutibilidade. Cada conjunto de rótulos por carta é representado como um vetor multi-*hot*, utilizando `MultiLabelBinarizer`, de forma que o problema possa ser tratado como um conjunto de 18 tarefas binárias em paralelo.

6.1 Baseline lexical: TF-IDF + regressão logística

Como baseline, empregamos uma abordagem clássica de vetorização TF-IDF com *n*-gramas (1-2) e regressão logística linear em esquema One-vs-Rest, com balanceamento de classes. Para comparação justa com o modelo neural, o baseline foi treinado com os conjuntos de treino e validação combinados.

6.2 Modelo contextual: BERTimbau-base

Para comparação com modelos contextuais, utilizamos o BERTimbau-base em configuração de classificação de sequência com 18 rótulos em regime multi-rótulo. Substituímos a função de perda padrão por `BCEWithLogitsLoss` ponderada por frequência inversa de classe (*pos_weight*), onde o peso de cada classe é calculado como a razão entre exemplos negativos e positivos no conjunto de treino. Com isso pretendemos mitigar o desbalanceamento extremo entre categorias. A Tabela 4 apresenta os pesos calculados para cada classe.

Realizamos *fine-tuning* por até 10 épocas, com *early stopping* baseado no macro-F1 do conjunto de validação (paciência de 2 épocas), *batch* de 8 exemplos, taxa de aprendizado de $2 \cdot 10^{-5}$, *warmup* de 10% das iterações e truncamento em 512 tokens. A Tabela 5 apresenta a evolução das métricas durante o treinamento.

6.3 Experimento com grande modelo de linguagem (LLM)

Para complementar, realizamos um experimento adicional de classificação multi-rótulo utilizando um LLM. O experimento foi conduzido com o mesmo conjunto de 18 categorias (frequência mínima de 20 ocorrências no conjunto supervisionado), mantendo a mesma divisão treino/validação/teste (70/15/15), e avaliando no mesmo conjunto de teste (131 cartas). Utilizamos o modelo GPT 4 (Achiam et al., 2023) com temperatura 0,0. Ressaltamos que este experimento com LLM tem caráter *exploratório* e não busca competir em igualdade de condições com modelos supervisionados, pois a inferência por *prompting* depende de escolhas de instruções e pode variar conforme o modelo e o contexto. Nosso objetivo é utilizá-lo como uma referência de alta cobertura semântica.

A inferência foi realizada por *prompting* (a ser disponibilizado na versão final), com instruções para retornar exclusivamente um objeto JSON no formato `{"labels": [...]}`, restringindo a saída ao vocabulário de 18 rótulos do experimento. Para garantir viabilidade e consistência entre exemplos, o texto de cada carta foi truncado para no máximo 9000 caracteres. Predições fora do vocabulário de rótulos foram descartadas. Custos e reprodutibilidade são limitações inerentes a essa abordagem, razão pela qual a incluímos como análise complementar.

Modelo	Micro-F1	Macro-F1
TF-IDF + LR	0,63	0,42
BERTimbau (otimizado)	0,55	0,42
LLM (GPT 4)	0,51	0,40

Tabela 6: Resultados globais no conjunto de teste (131 cartas, 18 classes, cenário multi-rótulo).

6.4 Métricas de avaliação

Devido ao cenário multi-rótulo, utilizamos precisão, recall e F1 por classe, além de micro-F1 e macro-F1 como métricas agregadas. A micro-F1 captura o desempenho agregado ponderando classes pela frequência, enquanto a macro-F1 mede a média simples da F1 por rótulo, sendo mais sensível ao desempenho em classes minoritárias.

7 Resultados

7.1 Resultados globais

A Tabela 6 resume os resultados globais no conjunto de teste (131 cartas, 18 categorias). O baseline lexical alcança micro-F1 de 0,63 e macro-F1 de 0,42, enquanto o BERTimbau obtém micro-F1 de 0,55 e macro-F1 de 0,42, com macro-F1 equivalente entre estes dois modelos. Os resultados globais obtidos com o LLM revelam micro-F1 de 0,51 e macro-F1 de 0,40.

7.2 Resultados detalhados por classe

As Tabelas 7, 8 e 9 apresentam os resultados detalhados de precisão, recall e F1 para cada uma das 18 categorias, permitindo uma análise mais aprofundada do comportamento de cada abordagem.

7.3 Análise comparativa por classe

A análise detalhada sugere um padrão de complementaridade entre os modelos:

Precisão vs. Recall: O baseline apresenta precisão agregada superior (0,65 vs. 0,46), enquanto o BERTimbau demonstra recall maior (0,67 vs. 0,60). Isso indica que o baseline é mais conservador, errando menos quando prediz, enquanto o BERTimbau captura mais instâncias verdadeiras, ao custo de mais falsos positivos. Os resultados obtidos com o LLM revelam recall elevado (0,73) e precisão mais baixa (0,39), e sugerem tendência à sobre-predição de categorias. Apesar disso, observa-se para o LLM desempenho alto em classes de maior suporte, enquanto categorias com menor suporte no teste tendem a apresentar maior variância.

Categoria	Prec.	Rec.	F1	Sup.
Terra	0,88	0,91	0,90	99
Violência	0,57	0,64	0,60	44
Saúde	0,57	0,67	0,62	30
Gestão Terr. e Amb.	0,67	0,76	0,71	29
Educação	0,70	0,80	0,74	20
Política Indigenista	0,42	0,47	0,44	17
Consulta aos povos	0,50	0,53	0,52	15
Assassinatos	0,43	0,20	0,27	15
Direitos Culturais	0,17	0,07	0,10	14
Mulheres Indígenas	1,00	0,54	0,70	13
Política Indígena	0,44	0,33	0,38	12
Org. da FUNAI	0,40	0,20	0,27	10
Segurança Pública	0,20	0,10	0,13	10
Bem viver	0,25	0,14	0,18	7
Crianças Indígenas	0,00	0,00	0,00	6
Segurança alimentar	0,00	0,00	0,00	4
Indíg. na política	—	—	—	4
Meio Ambiente	—	—	—	12
Micro avg	0,65	0,60	0,63	361
Macro avg	0,46	0,40	0,42	361

Tabela 7: Resultados detalhados do baseline (TF-IDF + Regressão Logística) por categoria. "Sup."= "Suporte"(número de exemplos no teste).

Categorias majoritárias: Os três modelos apresentam bom desempenho em *Terra* (F1 0,90 / 0,87 / 0,89). O baseline supera em *Educação* e *Gestão Territorial*, enquanto o BERTimbau é superior em *Política Indigenista* e o LLM é superior em *Mulheres Indígenas*.

Categorias minoritárias: O BERTimbau demonstra vantagem em várias categorias raras: *Assassinatos*, *Direitos Culturais*, *Política Indígena*, dentre outras. Por outro lado, o baseline é superior em *Mulheres Indígenas*, categoria em que alcança precisão perfeita (1,00).

7.4 Análise de erros

Nossa análise qualitativa dos erros revela padrões distintos entre os modelos:

Over-prediction do BERTimbau: O modelo neural prediz em média 3,99 rótulos por carta, enquanto o baseline prediz 2,53 (*ground truth*: 2,76). Isso explica o alto recall mas baixa precisão do BERTimbau.

Conservadorismo do baseline: O baseline prediz "apenas Terra" em 22 cartas, contra 19 do BERTimbau (*ground truth*: 28). Ambos sub-predizem casos de rótulo único, mas o baseline é um pouco mais conservador.

Padrões de confusão: O BERTimbau frequentemente adiciona rótulos correlacionados. Por exemplo, em uma carta sobre repúdio a ataques paramilitares anotada apenas com *Violência*, o BERTimbau

Categoria	Prec.	Rec.	F1	Sup.
Terra	0,87	0,88	0,87	99
Violência	0,54	0,75	0,63	44
Saúde	0,59	0,63	0,61	30
Gestão Terr. e Amb.	0,66	0,66	0,66	29
Educação	0,65	0,65	0,65	20
Política Indigenista	0,46	0,71	0,56	17
Consulta aos povos	0,36	0,53	0,43	15
Assassinatos	0,27	0,67	0,38	15
Direitos Culturais	0,31	0,29	0,30	14
Mulheres Indígenas	0,36	0,69	0,47	13
Política Indígena	0,35	0,67	0,46	12
Meio Ambiente	0,23	0,58	0,33	12
Org. da FUNAI	0,11	0,20	0,14	10
Segurança Pública	0,17	0,40	0,24	10
Bem viver	0,21	0,43	0,29	7
Crianças Indígenas	0,07	0,33	0,11	6
Indíg. na política	0,25	0,50	0,33	4
Segurança alimentar	0,09	0,25	0,13	4
Micro avg	0,46	0,67	0,55	361
Macro avg	0,36	0,55	0,42	361

Tabela 8: Resultados detalhados do BERTimbau (otimizado) por categoria. "Sup."= "Suporte"(número de exemplos no teste).

Categoria	Prec.	Rec.	F1	Sup.
Terra	0,87	0,91	0,89	99
Mulheres Indígenas	0,68	0,93	0,79	14
Educação	0,64	0,86	0,73	21
Saúde	0,56	0,80	0,66	25
Violência	0,49	0,95	0,64	40
Assassinatos	0,62	0,62	0,62	13
Gestão Territorial e Ambiental	0,27	0,59	0,37	29
Indígenas na política	0,19	0,67	0,30	6
Política Indigenista	0,17	0,94	0,28	18
Segurança Pública	0,27	0,27	0,27	11
Direitos Culturais	0,15	0,88	0,26	16
Meio Ambiente	0,25	0,25	0,25	12
Bem viver	0,50	0,17	0,25	6
Consulta aos povos indígenas	0,14	0,36	0,20	11
Crianças Indígenas	0,15	0,27	0,19	11
Segurança alimentar	0,14	0,20	0,17	5
Organização da FUNAI	0,17	0,14	0,15	7
Política Indígena	0,10	0,25	0,14	8
Micro avg	0,39	0,73	0,51	352
Macro avg	0,35	0,56	0,40	352

Tabela 9: Resultados detalhados do classificador baseado em LLM (GPT 4) por categoria. "Sup."= "Suporte"(número de exemplos no teste).

predisse *Assassinatos*, *Terra* e *Violência*, erros compreensíveis dado o contexto semântico, mas que inflam os falsos positivos.

Super-prediction do LLM: O classificador baseado em LLM apresentou o comportamento mais expansivo: em média 5,05 rótulos por carta, contra 2,69 no *ground truth*. Esse padrão explica o alto recall (0,73) combinado à baixa precisão agregada (0,39). Qualitativamente, o LLM tende a adicionar rótulos "plausíveis" do ponto de vista semântico, mas não anotados como ouro, especialmente em cartas sobre demarcação e contestação de atos normativos. Nesses casos, mesmo quando a anotação privilegia apenas *Terra*, o modelo frequentemente acrescenta *Política Indigenista* e *Direitos Culturais*, o que infla falsos positivos. Em situações mais raras, observa-se também um deslocamento do eixo central, quando o LLM prioriza rótulos correlatos (por exemplo, *Meio Ambiente* e *Gestão Territorial e Ambiental*) e deixa de prever *Terra*, sugerindo sensibilidade a pistas lexicais fortemente ambientais (barragens, rios, floresta) que podem sobrepor o enquadramento temático predominante.

7.5 Exemplos de classificação

Listamos aqui alguns exemplos para ilustrar o comportamento dos modelos, com quatro casos representativos do conjunto de teste.

Exemplo 1: BERTimbau mais preciso. Uma carta do Conselho de Articulação do Povo Guarani do Rio Grande do Sul (fevereiro de 2017) foi anotada apenas com *Terra*. O baseline predisse *Terra* e *Violência*, enquanto o BERTimbau acertou com apenas *Terra*.

Exemplo 2: Baseline mais robusto. Uma carta das lideranças Hunikui (Kaxinawá) anotada apenas com *Saúde* foi classificada corretamente pelo baseline. O BERTimbau predisse *Mulheres Indígenas* e *Saúde*, adicionando um rótulo incorreto.

Exemplo 3: BERTimbau captura mais contexto. Uma carta do VIII Encontro Estadual dos Operadores Indígenas em Direitos, anotada com *Educação*, *Saúde* e *Terra*, foi classificada corretamente pelo BERTimbau. O baseline adicionou erroneamente *Consulta aos povos indígenas*.

Exemplo 4: LLM superprediz rótulos plausíveis. Uma nota de repúdio da Articulação dos Povos Indígenas do Brasil (APIB) sobre mudanças nos procedimentos de demarcação foi anotada apenas com *Terra*. O LLM predisse *Terra*, mas adicionou também *Política Indigenista* e *Direitos Culturais*. Embora esses rótulos sejam semanticamente coerentes com o conteúdo (referências a medidas do governo, Constituição e direitos), eles configuram falsos positivos sob o esquema de anotação, ilustrando a tendência do LLM a ampliar a cobertura temática à custa de precisão.

8 Discussão

8.1 Análise dos modelos

As características do corpus podem explicar os padrões observados:

Marcadores lexicais fortes. Categorias como *Terra*, *Educação* e *Mulheres Indígenas* possuem vocabulário distintivo que favorece o baseline.

Sobreposição semântica. Categorias como *Assassinatos/Violência* e *Política Indígena/Política Indigenista* são semanticamente próximas, causando confusão no BERTimbau.

Desbalanceamento. *Terra* representa 72% das cartas, enquanto *Indígenas na política* aparece em apenas 4%, portanto mesmo com *pos_weight*, o modelo neural tende a sobre-predizer classes raras.

8.2 Potencial para abordagens híbridas

A complementaridade que observamos sugere potencial para abordagens híbridas. Uma estratégia possível seria utilizar o baseline para categorias com vocabulário distintivo (alta precisão) e o BERTimbau para categorias semanticamente complexas onde o recall é mais importante. Os resultados obtidos com o LLM sugerem que este pode ser útil como abordagem de alta cobertura para recuperação de temas, mas que estratégias de calibração (por exemplo, limitação do número de rótulos retornados, critérios adicionais de decisão ou pós-processamento) podem ser necessárias para reduzir falsos positivos em cenários multi-rótulo.

8.3 Limitações do truncamento

O truncamento em 512 tokens é uma limitação importante ao modelo contextual. Se considerarmos que a média de palavras por carta é de 664 e a mediana de 536, uma proporção significativa das cartas é truncada. Cartas longas (algumas excedem 6.000 palavras) perdem informação que pode ser relevante para a classificação.

Estratégias como *chunking* com agregação de predições ou modelos de contexto longo (por exemplo, Longformer, BigBird) podem mitigar essa limitação em trabalhos futuros.

9 Conclusão

Nosso artigo investigou a classificação automática multi-rótulo de cartas indígenas ao Brasil, comparando três abordagens. A análise detalhada por classe revela que dois modelos tem comportamentos complementares: o baseline é superior em categorias com vocabulário distintivo, enquanto o BERTimbau apresenta vantagem em categorias minoritárias e semanticamente complexas. Essa complementaridade sugere que abordagens híbridas, que combinam predições de ambos os modelos ou selecionam o modelo por categoria, podem superar as limitações individuais.

O experimento com LLM confirma sua capacidade de capturar múltiplas dimensões semânticas, resultando em alto recall e bom desempenho em categorias minoritárias. No entanto, o modelo tende a predizer muitos rótulos, com média superior ao número real de categorias por carta, o que impacta negativamente a precisão. Esses resultados reforçam que, no domínio analisado, LLMs se comportam como ferramentas complementares, e não substitutas, aos modelos supervisionados tradicionais.

Os resultados também evidenciam que a classificação multi-rótulo em domínios desbalanceados e com sobreposição semântica permanece um desafio para PLN. A estratégia de balanceamento via *pos_weight* foi essencial para que o BERTimbau alcançasse desempenho competitivo, mas não suficiente para superar o baseline em todas as métricas.

9.1 Limitações

O estudo apresenta limitações importantes. Optamos por manter categorias raras no recorte por sua relevância temática no domínio, mesmo cientes de que o suporte reduzido impõe um teto de desempenho no cenário supervisionado. O desbalanceamento acentuado entre classes resulta em desempenho próximo a zero para categorias raras como *Crianças Indígenas* e *Segurança alimentar*. O truncamento em 512 tokens limita o aproveitamento de cartas longas. Além disso, não realizamos análise de vieses sociolinguísticos ou regionais nos modelos, aspecto especialmente importante em contextos indígenas onde há grande diversidade de povos e regiões.

9.2 Trabalhos Futuros

Pretendemos investigar algumas direções principais. Primeiramente, abordagens híbridas que combinem as predições do baseline e do BERTimbau, seja por votação, stacking ou seleção de modelo por categoria. Pretendemos também avaliar modelos de contexto longo para aproveitar cartas extensas sem truncamento e técnicas de aumento de dados para categorias minoritárias.

Referências

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anmol Arora and 1 others. 2023. Risk and the future of ai: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, 33(3):100478.

Suzane Lima Costa. 2013. O que (ainda) podem as cartas? *Revista de Estudos em Língua e Literatura*, 19:87–98.

Suzane Lima Costa. 2018. As cartas das mulheres indígenas ao brasil. *Estudos Linguísticos e Literários*, 59:109–123.

Suzane Lima Costa. 2019. As cartas dos povos indígenas ao brasil: a construção do arquivo 2000–2015. *Memoria Americana. Cuadernos de Etnohistoria*, 26:94.

Suzane Lima Costa. 2021. Gestos de utopia no sul global: as cartas indígenas para o mundo. *Thomas Project Journal for Utopian Thoughts*, 5:75–89.

Suzane Lima Costa e Rafael Xucuru-Kariri. 2018. Autoria indígena em quinze anos de cartas. *Trabalhos em Linguística Aplicada*, 57:1364–1376.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York, USA.

Doris L Payne. 2014. *Amazonian linguistics: Studies in lowland South American languages*. University of Texas Press.

Edson Prestes, Lutiana Valadares Fernandes Barbosa, Viviane Ceolin Dallasta Del Grossi, Cynthia Picolo Gonzaga de Azevedo, Gustavo Macedo, e Renan Maffei. 2024. Ai and brazil's indigenous populations: A call for participation. Relatório técnico, UNESCO. Acesso em: março de 2025.

Gerard Salton e Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Fábio Souza, Rodrigo Nogueira, e Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Adane Nega Tarekegn, Mario Giacobini, e Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

UNESCO. 2021. Recommendation on the ethics of artificial intelligence. Acesso em: março de 2025.