Análise comparativa da classificação de comentários sobre pessoas trans no YouTube usando grandes modelos de linguagem

Vitor Lucio Giorgio Cardoso de Carvalho, Silas Lima Filho

[1] Universidade Federal do Rio de Janeiro (UFRJ)
Instituto de Computação – Curso de Bacharelado em Ciência da Computação

{vitorlgcc,silaslfilho}@ic.ufrj.br

Resumo. *Este artigo compara quatro abordagens para a classificação de comentários do YouTube direcionados a pessoas trans: um classificador tradicional, uma LLM em zero-shot, uma LLM em few-shot e uma LLM com Geração Aumentada por Recuperação (RAG). Com base em comentários anotados manualmente, o estudo analisa desempenho global, comportamento por classe e padrões recorrentes de erro. Os resultados mostram que a estratégia few-shot apresentou o melhor equilíbrio entre as classes, enquanto o RAG trouxe ganhos limitados no cenário avaliado.*

1. Introdução

A ampliação do acesso à Internet nas últimas décadas transformou a forma como pessoas se informam, se comunicam e participam da vida pública. No Brasil, esse processo ocorreu de maneira expressiva: entre 2005 e 2024, a proporção de domicílios com acesso à rede passou de 13% para 86% [Brasil. Secretaria de Comunicação Social 2024]. Nesse contexto, plataformas digitais passaram a ocupar posição central na circulação de informações, opiniões e disputas simbólicas, com destaque para o YouTube, cujo número de usuários aumentou em quase 300% entre 2008 e 2018 [Ortiz-Ospina 2019].

Se, por um lado, esse ambiente digital amplia oportunidades de expressão e participação, por outro, também favorece a circulação de conteúdos ofensivos, discriminatórios e violentos. Estudos indicam que a dinâmica dessas plataformas pode intensificar a exposição a discursos hostis e normalizar manifestações de intolerância contra grupos minorizados [Comunica Que Muda 2016]. À medida que a interação entre usuários ativos aumenta, o desafio de identificar e moderar comentários com este perfil nocivo torna-se mais desafiador.

No caso particular da população trans, esse cenário adquire gravidade ainda maior. A circulação recorrente de ataques, ofensas e deslegitimações em plataformas digitais contribui para a reprodução de estigmas e para a manutenção de formas de violência simbólica e social. Esse problema se torna alarmante no contexto brasileiro, marcado por altos índices de violência contra pessoas trans [Narcisa and Bonets 2025]. Dessa forma, a análise de manifestações online dirigidas a esse grupo não se limita ao ambiente virtual, mas perpetua um discurso já existente fora das plataformas digitais.

Grandes modelos de linguagem (LLMs) têm se mostrado promissores para tarefas de Processamento de Linguagem Natural devido à sua capacidade de compreender e gerar

linguagem natural a partir de grandes volumes de dados, além de se adaptarem a diferentes tarefas por meio de instruções textuais, chamadas *prompts*. No entanto, em problemas dependentes de contexto, como a identificação de discurso de ódio, seu desempenho depende não apenas da instrução, mas também da qualidade do conteúdo contextual disponível durante a inferência. Nesse cenário, o contexto semântico torna-se importante para tornar a classificação mais consistente. Uma das estratégias mais promissoras para esse fim é a Geração Aumentada por Recuperação (RAG), que complementa a inferência do modelo com informações recuperadas de documentos externos de forma síncrona.

Trabalhos anteriores investigaram a detecção de discurso transfóbico com modelos supervisionados tradicionais, o uso de LLMs em tarefas de classificação textual e a aplicação de RAG para detecção de discurso de ódio. Ainda assim, permanecem poucos estudos que combinem comentários reais de plataformas digitais, transfobia, comparação direta entre *zero-shot*, *few-shot* e RAG, e análise qualitativa de erros em contexto conversacional.

Este trabalho investiga a seguinte questão de pesquisa: como diferentes estratégias de contextualização com LLMs afetam a classificação de comentários transfóbicos, neutros e de apoio em comentários reais do YouTube? Para isso, compara-se de forma exploratória BART-MNLI, Llama em *zero-shot*, Llama em *few-shot* e Llama com RAG. A contribuição do estudo está na análise quantitativa por classe e na análise qualitativa de erros, articulando classificação automática, análise de mídias sociais e impactos sociotécnicos sobre grupos vulnerabilizados.

A seção a seguir relata trabalhos diretamente relacionados ao problema de pesquisa. A seção 3 descreve a obtenção dos dados e o método do experimento. A seção seguinte compreende os resultados obtidos e a discussão dos dados. Ao final, a seção 5 relata os próximos passos e desafios desse trabalho em andamento.

2. Trabalhos Relacionados

Foram priorizados trabalhos próximos ao problema investigado, envolvendo detecção de transfobia ou discurso de ódio, classificação textual com LLMs e uso de RAG para moderação ou classificação textual.

Murakami (2022) investigou a detecção automatizada de discurso transfóbico em português com modelos tradicionais de aprendizado de máquina, estabelecendo um *baseline* nacional relevante para o tema [Murakami 2020]. Chakravarthi (2024) apresentou um conjunto multilíngue de comentários do YouTube para detecção de homofobia e transfobia em inglês, tâmil e tâmil-inglês, além de avaliar modelos de aprendizado de máquina e aprendizado profundo [Chakravarthi 2024]. Tornisiello (2024) comparou *pipelines* com LLMs em configurações *zero-shot*, *few-shot* e RAG para classificação de discurso de ódio, oferecendo uma referência metodológica próxima à estrutura experimental adotada neste artigo [Tornisiello 2024]. Prasannan et al. (2025) propuseram geração de contradiscurso para conteúdos homofóbicos e transfóbicos em malaiala, usando RAG e tradução para lidar com um idioma de baixo recurso [Prasannan et al. 2025]. Olivert-Iserte et al. (2025) investigaram análise de sentimentos voltada à comunidade LGBTQ+ com BERT, LLMs e recuperação contextual em dados do Reddit, aproximando-se do uso combinado de modelos supervisionados, LLMs e contexto externo [Olivert-Iserte et al. 2025].

Este trabalho se diferencia desses estudos por focalizar comentários reais do

YouTube sobre pessoas trans, comparar diretamente BART-MNLI, Llama em *zero-shot*, Llama em *few-shot* e Llama com RAG, e analisar qualitativamente os erros produzidos em comentários curtos, ambíguos e dependentes de contexto conversacional. Assim, embora dialogue com estudos sobre detecção de homofobia/transfobia, LLMs e RAG, a contribuição está em observar como essas estratégias se comportam em um cenário exploratório centrado especificamente na classificação triádica de comentários sobre pessoas trans.

3. Materiais e Método

3.1. Base de dados

Os dados foram coletados com a YouTube Data API v3[1] a partir de vídeos relacionados a pessoas trans. Realizou-se uma busca com termos associados ao escopo da pesquisa, como *transgender*, *trans women*, *trans men* e *trans rights*. A partir dos vídeos retornados, foram selecionados dois vídeos em inglês com maior relevância para compor um estudo exploratório: um vídeo anterior, usado apenas como fonte de exemplos e contexto para as abordagens *few-shot* e RAG, e um vídeo separado, usado como conjunto avaliado.

Não houve treinamento ou ajuste fino dos modelos. Os 483 comentários do vídeo anterior foram usados como conjunto de exemplos, enquanto os 938 comentários do vídeo avaliado foram usados para comparar as predições dos quatro *pipelines*. Todos os comentários foram anotados manualmente em três classes: **ódio** (*hateful*), **apoio** (*supportive*) ou **neutro** (*neutral*). A anotação manual foi utilizada como *ground truth*. A Tabela 1 resume a composição dos dois conjuntos.

Conjunto	Total	Principais	Respostas	Hateful	Neutral	Supportive
Exemplos	483	242	241	190	162	131
Avaliação	938	500	438	412	307	219

Tabela 1. Caracterização dos comentários usados como exemplos e avaliação.

Para reduzir riscos de exposição de dados pessoais, a base final mantém apenas atributos essenciais: identificador do vídeo, identificador do comentário, indicação de resposta, identificador do comentário respondido, texto e rótulo manual quando aplicável. A Tabela 3 apresenta exemplos curtos, apenas para ilustrar o tipo de conteúdo encontrado e alguns cenários de erro.

O foco deste estudo experimental não está em prever subtipos específicos de transfobia, mas em analisar comparativamente quatro fluxos de inferência em tarefas de classificação desse domínio de dados. No caso do BART, categorias semânticas intermediárias foram usadas apenas como rótulos candidatos e posteriormente mapeadas para as três classes finais.

3.2. Estratégia de classificação

A estratégia de inferência consistiu na comparação de quatro *pipelines* aplicados sobre o mesmo conjunto avaliado. A primeira abordagem utiliza BART-MNLI[2] em

[1] `https://developers.google.com/youtube/v3?hl=pt-br`
[2] `https://huggingface.co/facebook/bart-large-mnli`

zero-shot, tratado como *baseline* por ser um classificador textual baseado em inferência semântica. As demais abordagens utilizam Llama 3.1-8B Instruct[3], executado via Ollama[4], com variações na forma de fornecer orientação: apenas instruções, instruções acompanhadas de exemplos e instruções complementadas por contexto recuperado externamente.

Nas abordagens com LLM, a estrutura do *prompt* foi mantida constante sempre que possível. O modelo recebe a definição das três classes, regras para insultos, invalidação, desumanização, sarcasmo depreciativo, apoio explícito e neutralidade, além da descrição do vídeo e do comentário original quando o item analisado é uma resposta. A saída solicitada ao modelo é um objeto JSON contendo o campo `label`, restrito a *hateful*, *neutral* ou *supportive*.

Na configuração *few-shot*, são inseridos até 10 exemplos rotulados de cada classe, extraídos do vídeo anterior. A abordagem RAG utiliza a mesma fonte de exemplos, mas de forma dinâmica: os exemplos e uma *guideline* de classificação formam um *corpus* recuperável, indexado com *embeddings* do modelo `all-MiniLM-L6-v2`[5]. O índice foi construído com *Facebook AI Similarity Search* (FAISS)[6], usando `chunk_size=800`, sobreposição de 200 caracteres e recuperação dos três trechos mais similares (`top_k=3`). A *guideline* contém critérios contextuais e expressões indicativas, como repulsa explícita, invalidação de gênero, desumanização, sarcasmo depreciativo e ataques em respostas. Os trechos recuperados são então adicionados ao *prompt* antes da inferência. O código-fonte utilizado para coleta, classificação e avaliação está disponível publicamente[7].

4. Resultados e discussão

A Tabela 2 apresenta os resultados detalhados por classe e por métrica. De modo geral, a abordagem *few-shot* obteve o melhor desempenho, alcançando *Macro-F1* de 0,594 e acurácia de 0,61. Esse resultado sugere que a inclusão direta de exemplos anotados no *prompt* contribuiu para uma separação mais consistente entre comentários de ódio, apoio e neutralidade. O RAG superou modestamente o *baseline* e a configuração *zero-shot* em *Macro-F1*, mas permaneceu abaixo do *few-shot*, indicando que a recuperação de contexto, na configuração avaliada, não foi suficiente para produzir ganhos consistentes.

Observando-se as métricas por classe, o *baseline* apresentou o desempenho mais limitado, embora tenha mantido valores razoáveis em *hateful*. Esse comportamento sugere que a abordagem reconhece padrões evidentes de hostilidade, mas tem menor capacidade de lidar com dependência contextual. A abordagem *zero-shot* ampliou o *recall* de *hateful*, mas teve o menor F1 para *supportive*. Com a introdução de exemplos anotados, o *few-shot* obteve o melhor equilíbrio global, especialmente em *hateful* e *supportive*. Ainda assim, a classe *neutral* permaneceu instável, com baixo *recall* nessa configuração. O RAG apresentou desempenho próximo ao *zero-shot*, sugerindo que os trechos recuperados podem ter introduzido ruído ou contexto pouco discriminativo para parte dos comentários.

[3] `https://ollama.com/library/llama3.1`
[4] `https://ollama.com/`
[5] `https://www.sbert.net/`
[6] `https://faiss.ai/index.html`
[7] `https://github.com/v-giorgio/short-transphobia-rag-social-network-analysis`

Classe	Métrica	Baseline	Zero-shot	Few-shot	RAG
Hateful	Precisão	0.54	0.53	0.59	0.53
	Recall	0.55	0.75	0.80	0.75
	F1	0.54	0.62	0.68	0.62
Neutral	Precisão	0.44	0.51	0.60	0.55
	Recall	0.53	0.46	0.37	0.43
	F1	0.48	0.48	0.46	0.48
Supportive	Precisão	0.61	0.82	0.70	0.68
	Recall	0.42	0.29	0.60	0.35
	F1	0.50	0.43	0.65	0.46
Macro-F1	–	0.506	0.510	0.594	0.520
Accuracy	–	0.51	0.55	0.61	0.55

Tabela 2. Comparação completa das métricas de classificação por *pipeline*.

4.1. Análise qualitativa

A Figura 1 mostra que os erros não se distribuem de forma uniforme. O *baseline* errou principalmente comentários *hateful* (187 casos), enquanto o *zero-shot* errou mais comentários *supportive* (156 casos). No *few-shot*, a principal fonte de erro foi a classe *neutral* (193 casos), sugerindo que os exemplos aumentaram a sensibilidade do modelo às classes de borda. No RAG, os erros concentraram-se em *neutral* (175) e *supportive* (142), o que reforça a hipótese de que o contexto recuperado nem sempre foi discriminativo.

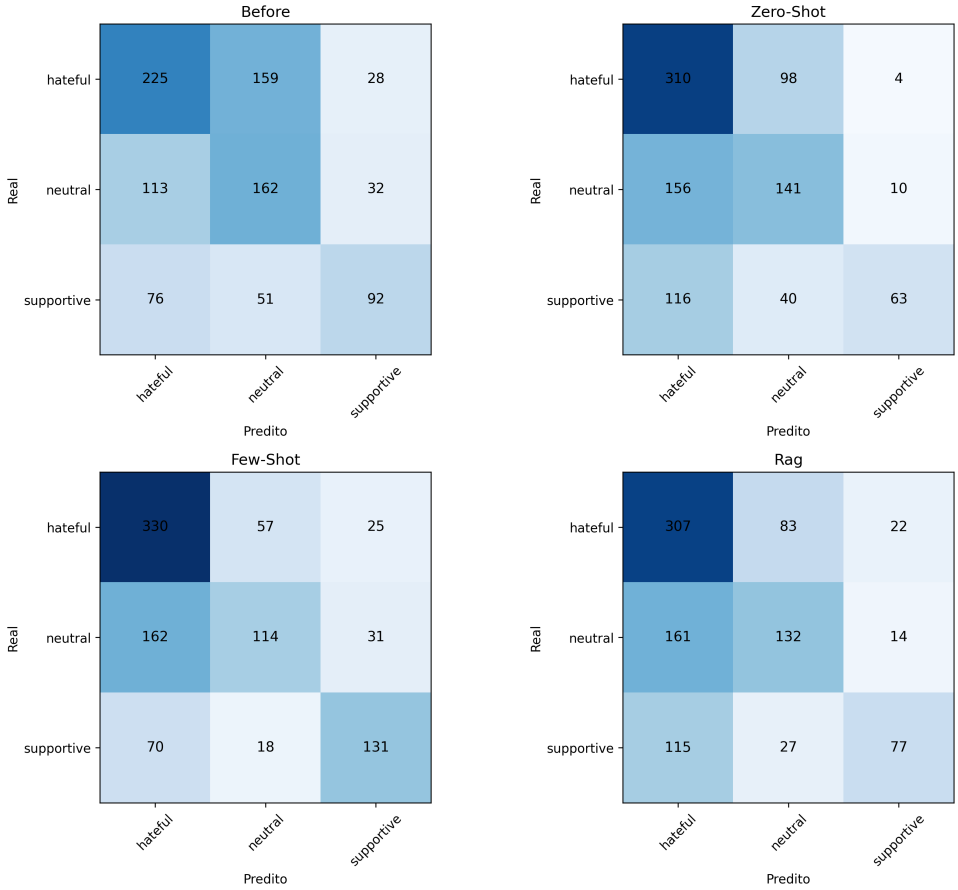

Figura 1. Comparativo das matrizes de confusão das *pipelines* avaliadas.

Comentário	Real	BART	Zero	Few	RAG
"gross"	hateful	neutral	hateful	hateful	hateful
"They should be illegal"	hateful	neutral	hateful	hateful	hateful
"For all the straight guys..."	supportive	hateful	hateful	supportive	hateful
"bro, be quiet... let her be her"	supportive	neutral	supportive	hateful	hateful
"And people wonder why fathers aren't around"	hateful	neutral	hateful	hateful	hateful
"I feel bad for the dad..."	hateful	neutral	hateful	supportive	supportive

Tabela 3. Exemplos curtos e cenários representativos de erro entre *pipelines*.

A análise dos exemplos indica padrões recorrentes. Comentários muito curtos, como *"gross"*, dependem fortemente do contexto do vídeo para serem interpretados como ataque à pessoa trans retratada. Em outros casos, o *baseline* tende a neutralizar ataques indiretos, como reclamações sobre pronomes ou comentários sobre ser pai de uma pessoa trans, enquanto as LLMs se arriscam mais na classe *hateful*. Também há respostas de apoio que exigem compreender a *thread*: em *"bro, be quiet... let her be her"*, apenas o *zero-shot* identificou o sentido de defesa, enquanto perguntas retóricas mais longas foram melhor capturadas pelo *few-shot*. Esses casos reforçam que parte dos erros depende menos de vocabulário ofensivo explícito e mais de contexto conversacional.

Esses resultados indicam que melhorias futuras devem tratar explicitamente os tipos de erro observados. No entanto, a análise ainda é limitada pelo tamanho reduzido do conjunto avaliado, pela ausência de ajuste fino em um modelo de referência supervisionado, como BERT, e pela necessidade de uma documentação mais sistemática do *corpus* e das regras usadas no RAG.

5. Conclusão

Este trabalho apresentou uma comparação exploratória de quatro estratégias para a classificação de comentários reais do YouTube sobre pessoas trans: BART-MNLI, Llama em *zero-shot*, Llama em *few-shot* e Llama com RAG. Os resultados indicam que a configuração *few-shot* obteve o melhor equilíbrio entre as classes, enquanto o RAG apresentou ganhos mais modestos, sugerindo que somente a inclusão de contexto recuperado não garante melhora quando o material recuperado não é suficientemente útil. O estudo avalia um conjunto pequeno de comentários e não inclui uma comparação com outro modelo de aprendizado supervisionado. Ainda assim, os resultados evidenciam padrões relevantes de erro em comentários curtos, ambíguos e dependentes de contexto conversacional, indicando caminhos concretos para expandir o estudo.

Como próximos passos, pretende-se construir um *dataset* em português, anotado em uma ferramenta própria por múltiplos anotadores, com registro de critérios de decisão e cálculo de concordância. Também será ampliada a comparação com modelos supervisionados, como MNB e BERT, para avaliar se abordagens ajustadas ao domínio superam estratégias baseadas apenas em inferência por *prompt*. No caso do RAG, pretende-se documentar melhor o *corpus* recuperável, incorporar mais documentos e combinar diferentes fontes, incluindo contexto adicional sobre os vídeos e materiais explicativos sobre transfobia. Por fim, serão avaliados outros LLMs além do Llama, permitindo verificar se os padrões observados neste estudo se mantêm em modelos e configurações distintas.

Referências

Brasil. Secretaria de Comunicação Social (2024). População do brasil chega a 212,6 milhões de habitantes, aponta ibge. 29 ago. 2024. Atualizado em 30 ago. 2024. Disponível em: https://www.gov.br/secom/pt-br/assuntos/noticias/2024/08/populacao-do-brasil-chega-a-212-6-milhoes-de-habitantes-aponta-ibge. Acesso em: 29 mar. 2026.

Chakravarthi, B. R. (2024). Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18:49–68. Disponível em: https://link.springer.com/article/10.1007/s41060-023-00400-0. Acesso em: 16 maio 2026.

Comunica Que Muda (2016). Dossiê intolerâncias: visíveis e invisíveis no mundo digital. São Paulo: Nova/sb. Disponível em: https://www.comunicaquemuda.com.br/dossie/intolerancia-nas-redes/. Acesso em: 29 mar. 2026.

Murakami, L. (2020). Detecção automática de discurso de ódio online: a transfobia no twitter. In *Encontro Virtual da ABCiber*. Disponível em: https://abciber.org.br/simposios/index.php/virtualabciber/virtual2020/paper/view/1008/468. Acesso em: 29 mar. 2026.

Narcisa, T. and Bonets, V. (2025). Brasil é o país que mais mata pessoas trans e travestis, aponta dossiê. CNN Brasil, Belém, 27 jan. 2025. Atualizado em 11 abr. 2025. Disponível em: https://www.cnnbrasil.com.br/nacional/brasil-e-o-pais-que-mais-mata-pessoas-trans-e-travestis-aponta-dossie/. Acesso em: 29 mar. 2026.

Olivert-Iserte, M., Serras, F., Civit, M., and Gonzàlez-Agirre, A. (2025). Pld at homo-lat 2025: Enhancing dialectal sentiment analysis through contextual retrieval and translation. In *Proceedings of HOMO-LAT 2025*, volume 4098 of *CEUR Workshop Proceedings*. Disponível em: https://ceur-ws.org/Vol-4098/HOMOLAT2025_paper1.pdf. Acesso em: 16 maio 2026.

Ortiz-Ospina, E. (2019). A ascensão das mídias sociais. Our World in Data. Disponível em: https://ourworldindata.org/rise-of-social-media. Acesso em: 29 mar. 2026.

Prasannan, P., Kumaresan, P. K., Rajiakodi, S., Subalalitha, C. N., and Chakravarthi, B. R. (2025). Counter-speech generation for homophobic and transphobic social media content in malayalam. *Social Network Analysis and Mining*, 15:87. Disponível em: https://link.springer.com/article/10.1007/s13278-025-01507-x. Acesso em: 16 maio 2026.

Tornisiello, V. R. (2024). Explorando llms abertos para classificação de discurso de ódio em jogos online. 60 p. Monografia (MBA em Inteligência Artificial e Big Data). Disponível em: https://bdta.abcd.usp.br/directbitstream/b0376907-ca64-4b77-a711-f1f2b526e35f/. Acesso em: 29 mar. 2026.