Classificação de discursos sobre pessoas trans no YouTube com Geração Aumentada por Recuperação

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. *Discurso de ódio contra pessoas trans ataca identidades de gênero marginalizadas e pode se propagar rapidamente em ambientes digitais. Apesar dos avanços em modelos de linguagem, detectar discursos transfóbicos em português ainda envolve escassez de dados anotados e ambiguidade discursiva. Este trabalho classifica comentários sobre pessoas trans, coletados de vídeos jornalísticos do YouTube, em três categorias: intolerante, apoio e neutro. Avaliamos abordagens zero-shot, few-shot e com geração aumentada por recuperação (RAG), usando um modelo generativo local e documentos do domínio. Como referência supervisionada, realizamos fine-tuning do BERTimbau e comparamos os resultados por macro-F1 e F1 por classe. A melhor abordagem foi o modelo generativo com RAG baseado em documentos externos.*

Abstract. *Hate speech against transgender people targets marginalized gender identities and can spread rapidly in digital environments. Despite advances in language models, detecting transphobic discourse in Brazilian Portuguese remains challenging due to limited annotated data and discursive ambiguity. This work classifies comments about transgender people, collected from journalistic YouTube videos, into three categories: hateful, supportive, and neutral. We evaluate zero-shot, few-shot, and retrieval-augmented generation (RAG) approaches using a local generative model and domain-related documents. As a supervised baseline, we fine-tune BERTimbau and compare the approaches using macro-F1 and class-wise F1. The best-performing approach was the generative model with document-based RAG.*

1. Introdução

Pessoas trans e travestis continuam expostas a formas persistentes de hostilidade no debate público brasileiro. Somente em 2025, foram registrados 80 assassinatos de pessoas trans e travestis no Brasil [Associação Nacional de Travestis e Transexuais 2026]. Esse cenário de violência não se limita ao espaço físico, visto que, em ambientes digitais, debates sobre direitos trans e identidade de gênero também se manifestam em comentários, respostas e interações públicas, tornando plataformas como o YouTube espaços relevantes para investigar como discursos de apoio, neutralidade e rejeição aparecem sobre pessoas trans.

Comentários em vídeos jornalísticos são particularmente relevantes porque reagem a eventos públicos concretos e com uma grande diversidade de opiniões. Diferentemente de postagens isoladas, esses comentários costumam estar associados a uma notícia,

a uma pessoa mencionada, a uma decisão institucional ou a um episódio de violência. Assim, a interpretação do comentário pode depender tanto do texto escrito pelo usuário quanto do contexto do vídeo ao qual ele responde e de contexto externo ao vídeo.

Este trabalho parte do problema de classificar automaticamente a atitude discursiva de comentários sobre pessoas trans em português brasileiro. A tarefa é desafiadora porque comentários transfóbicos nem sempre aparecem como ofensas explícitas, visto que podem ocorrer por invalidação de identidade, ironia, transfobia disfarçada ou aparente neutralidade. Ao mesmo tempo, comentários negativos sobre um evento noticiado não são necessariamente comentários contra pessoas trans. Essa ambiguidade torna importante avaliar se informações contextuais externas ajudam modelos de linguagem a tomar decisões mais consistentes.

Diante disso, investigamos se a adição de contexto por geração aumentada por recuperação (RAG) contribui para a classificação de comentários sobre pessoas trans no YouTube [Lewis et al. 2020]. Para isso, comparamos abordagens generativas com e sem recuperação de contexto a uma referência supervisionada baseada em BERTimbau [Souza et al. 2023], usando comentários anotados em três categorias de atitude discursiva: intolerante, apoio e neutro. A análise busca observar se transcrições e documentos relacionados ao domínio fornecem pistas úteis para comentários cujo sentido depende do evento jornalístico, de referências implícitas ou de conhecimento social sobre pessoas trans.

A próxima seção apresenta trabalhos relacionados ao tema, incluindo detecção de discurso de ódio, transfobia e classificação textual em mídias digitais. A Seção 3 descreve os materiais e métodos, incluindo coleta, anotação, modelos utilizados e métricas de avaliação. A Seção 4 discute os resultados obtidos e compara as estratégias avaliadas. Por fim, a Seção 5 apresenta as conclusões e aponta limitações e próximos passos da pesquisa.

2. Trabalhos Relacionados

A detecção automática de discurso de ódio é amplamente estudada em PLN, inclusive para o português brasileiro. Conjuntos como HateBR e ToLD-Br contribuíram para a disponibilidade de dados anotados em português, contemplando comentários dirigidos a diferentes grupos sociais [Vargas et al. 2025, Leite et al. 2020]. Esses recursos são importantes como referência metodológica, mas diferem do recorte deste trabalho por tratarem de plataformas e alvos mais amplos. No caso de comentários sobre pessoas trans, a classificação depende de distinções mais específicas entre rejeição de identidade, apoio e neutralidade..

Trabalhos recentes investigam homofobia e transfobia em comentários do YouTube, incluindo o conjunto multilíngue de 15.141 comentários proposto por Chakravarthi et al. [Chakravarthi et al. 2021]. Esse recorte foi posteriormente explorado em tarefas compartilhadas com modelos baseados em Transformers, como variações do BERT [Chakravarthi 2024]. Apesar da relevância desses estudos, eles não se concentram no português brasileiro nem no uso de contexto externo associado aos vídeos.

Outros estudos analisam a transfobia no YouTube para além da classificação supervisionada, como Channon e Mathieson (2025), que destacam conteúdo transfóbico

em comentários do YouTube que podem escapar da moderação automatizada [Channon and Mathieson 2025]. No Brasil, Barros e Silva mostram como comentários associados ao caso Nikolas Ferreira mobilizam discursos científicos e religiosos para naturalizar a exclusão de identidades trans [Barros and Silva 2025]. Esses trabalhos reforçam a importância do contexto discursivo, mas não avaliam estratégias computacionais de recuperação de contexto para apoiar a classificação.

Por fim, a Geração Aumentada por Recuperação (RAG) tem sido usada para incorporar contexto adicional a modelos de linguagem durante a inferência [Lewis et al. 2020]. Em moderação de conteúdo, tarefas de classificação podem ser formuladas como problemas de RAG, apoiadas por políticas, critérios ou conhecimento contextual recuperado [Willats et al. 2025]. Este artigo explora essa abordagem na classificação de comentários sobre pessoas trans no YouTube em português brasileiro, usando contexto dos vídeos e documentos auxiliares indexados vetorialmente, uma combinação ainda pouco explorada na literatura.

3. Materiais e Métodos

O conjunto de dados foi coletado a partir de vídeos jornalísticos do YouTube relacionados a pessoas trans, considerando canais como BBC News Brasil, CNN Brasil, Folha de S.Paulo, G1/Globo, GloboNews, Metrópoles, Record News, SBT e UOL. A coleta utilizou a YouTube Data API v3[1], por meio dos pontos de acesso `search.list`, `videos.list` e `commentThreads.list`, com buscas separadas por canal, limite de até 50 vídeos por canal e até 100 comentários por vídeo. As consultas usaram termos associados ao debate público sobre pessoas trans, como “trans”, “transexual”, “transgênero”, “travesti”, “nome social”, “ideologia de gênero”, “cota trans” e “transfobia”. Após limpeza, normalização e curadoria manual para remoção de vídeos irrelevantes, duplicados ou tangenciais ao tema, 158 vídeos e 6.072 comentários candidatos foram organizados para anotação.

A anotação foi realizada por meio de uma ferramenta própria de *crowdsourcing*, hospedada na Vercel[2], e desenvolvida para apresentar comentários aos anotadores e armazenar as respostas em uma instância Supabase[3], acessível pela URL[4]. O projeto foi submetido à apreciação do Comitê de Ética via Plataforma Brasil[5], ainda sem conclusão até a finalização deste trabalho, e adotou boas práticas éticas, incluindo consentimento informado, anonimização, minimização de dados pessoais e possibilidade de interrupção da participação pelos anotadores a qualquer momento. Participaram do processo seis pessoas cis e duas pessoas trans. Cada comentário recebeu rótulos para atitude discursiva, ofensividade e aspectos discursivos; entretanto, esta versão do artigo considera apenas a tarefa principal de atitude discursiva, com as classes intolerante, apoio e neutro. Como nem todos os comentários candidatos foram anotados a tempo para esta versão, os experimentos utilizaram apenas a porção consolidada do conjunto. O *dataset* final empregado nos experimentos contém 956 comentários, provenientes de 124 vídeos. Os rótulos finais foram produzidos a partir das anotações disponíveis, com o critério de ter sido anotado

[1] `https://developers.google.com/youtube/v3/docs`
[2] `https://vercel.com/`
[3] `https://supabase.com/`
[4] `https://annotation-system-wine.vercel.app/`
[5] `https://plataformabrasil.saude.gov.br/`

por pelo menos uma pessoa trans ou por duas pessoas cis, e agregando as respostas por regras de maioria ponderada, com maior peso para a decisão da pessoa trans, uma vez que é a pessoa mais envolvida ao domínio analisado.

O fluxo metodológico está sintetizado na Figura 1. Primeiramente, o conjunto consolidado foi dividido usando *split* aleatório estratificado em treino, validação e teste, com 668, 144 e 144 comentários, respectivamente, seguindo uma partição aproximada de 70/15/15. O mesmo conjunto de teste foi usado em todas as comparações, e os exemplos de *few-shot* foram selecionados apenas do treino, evitando vazamento de informação. Foram avaliados o BERTimbau com *fine-tuning* supervisionado e um modelo generativo local da família Qwen3 14B, nas estratégias *zero-shot*, *few-shot* e RAG. Optou-se pelo primeiro por ser um modelo *encoder-only*, referência para tarefas de classificação textual, com dados em português. O segundo foi escolhido por mostrar-se um dos mais avançados, até o momento da pesquisa, disponibilizados pela ferramenta Ollama[6]. Os *prompts* seguiram uma estrutura comum, com papel do modelo, instruções da tarefa, definição dos rótulos e retorno em JSON. No RAG, foram incorporados contextos recuperados de transcrições dos vídeos coletados e documentos relacionados ao domínio. As fontes dos documentos foram a `Wikipedia`, `ANTRA` e `gov.br`, através de consultas baseadas em entidades e palavras-chave extraídas dos títulos dos vídeos. Os documentos recuperados e divididos em *chunks* foram transformados em *embeddings* com o modelo `BAAI/bge-m3`[7] e persistidos com o auxílio de um índice vetorial FAISS[8]. A avaliação utilizou macro-F1 como métrica principal, complementada por F1 por classe, precisão e recall.

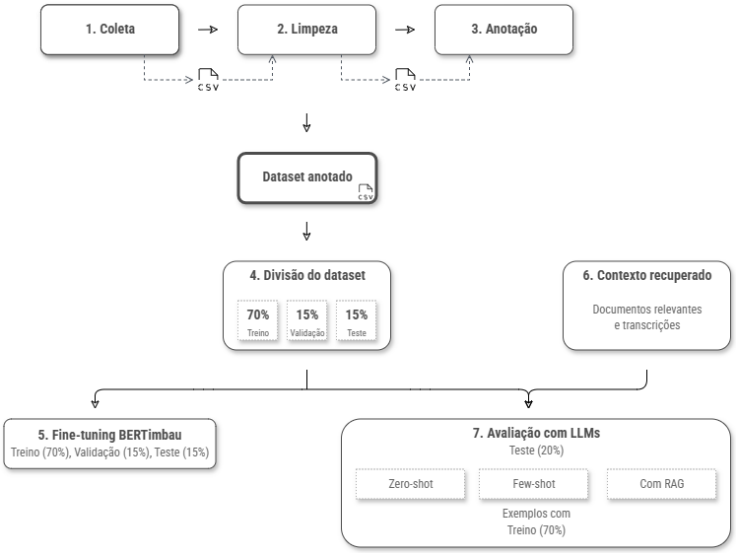

Figura 1. Fluxo metodológico geral da coleta, anotação, divisão do dataset, treinamento e avaliação dos modelos.

4. Resultados e Discussão

A avaliação priorizou o F1 por classe e o macro-F1. O F1 sintetiza precisão e recall em uma única medida, sendo definido como $2PR/(P + R)$. Já o macro-F1 corresponde à

[6]`https://ollama.com/`
[7]`https://huggingface.co/BAAI/bge-m3`
[8]`https://github.com/facebookresearch/faiss/wiki`

média simples do F1 das classes, atribuindo o mesmo peso a intolerante, apoio e neutro, o que é adequado para comparar o desempenho em cenários com possível desbalanceamento entre classes. A Tabela 1 resume os resultados da tarefa de atitude discursiva, cujas métricas variam de zero a um, com valores maiores indicando melhor desempenho.

Abordagem	Precisão	Recall	Macro F1	F1-I	F1-A	F1-N
BERTimbau	0.441	0.444	0.442	0.463	0.370	0.492
Qwen ZS	0.636	0.552	0.549	0.395	0.622	0.630
Qwen FS	0.657	0.618	0.611	0.648	0.694	0.489
Qwen RAG-T	0.614	0.594	0.592	0.647	0.667	0.462
Qwen RAG-D	0.641	0.615	**0.618**	0.647	0.667	0.542
Qwen RAG-C	0.629	0.616	0.616	0.631	0.680	0.537

Tabela 1. Desempenho das abordagens na classificação de atitude discursiva. ZS e FS indicam, respectivamente, as abordagens *Zero-Shot* e *Few-Shot*. RAG-T, RAG-D e RAG-C indicam, nesta ordem, a abordagem RAG com transcrições, documentos e ambos combinados. F1-I, F1-A e F1-N indicam o F1 das classes intolerante, apoio e neutro.

O melhor resultado geral foi obtido pelo Qwen com RAG baseado em documentos externos, com macro-F1 de 0.618, seguido pelo RAG combinado, com 0.616, e pelo *few-shot*, com 0.611. Todas as configurações com Qwen superaram o BERTimbau neste recorte experimental, embora essa diferença deva ser interpretada com cautela devido ao tamanho limitado do conjunto supervisionado. Em termos de precisão e recall, o *few-shot* apresentou os maiores valores agregados, com precisão de 0.657 e recall de 0.618, indicando maior proporção de acertos entre suas predições e maior recuperação dos exemplos reais das classes. Ainda assim, sua macro-F1 ficou ligeiramente abaixo das variantes com RAG baseado em documentos e RAG combinado, sugerindo que a distribuição do desempenho entre as classes foi mais equilibrada nessas configurações. Quanto ao F1 por classe, o *zero-shot* teve melhor desempenho em neutro (0.630), mas baixo desempenho em intolerante (0.395). O *few-shot* obteve o melhor F1 para apoio (0.694), enquanto as variantes com RAG mantiveram melhor equilíbrio entre intolerante e neutro.

Entre as estratégias com recuperação, o uso de documentos externos apresentou o melhor equilíbrio entre as classes. A configuração apenas com transcrições superou o *zero-shot*, mas ficou abaixo das configurações com documentos, sugerindo que a transcrição do vídeo nem sempre contém o contexto mais discriminativo para classificar o comentário. No geral, dado o conjunto limitado de dados e o uso de somente uma LLM, os resultados quantitativos das abordagens RAG indicaram que a combinação de fontes pode ser útil, especialmente na identificação da classe intolerante, que é a mais crítica socialmente. Contudo, há espaço para experimentos com outros modelos, com um *dataset* maior e analisando outras nuances do texto, como a ofensividade e aspectos do discurso.

A análise qualitativa da Tabela 2 sugere três padrões principais. Em E1 e E2, o *zero-shot* tende a tratar ironia e invalidação de identidade como neutralidade, especialmente quando o ataque não aparece em forma de insulto explícito. Em E3, exemplos no *prompt* e contexto recuperado ajudam a capturar apoio breve. A classe neutra mostrou-se instável, visto que em E4, as abordagens RAG acertaram, porém em E5, a abordagem combinada acertou, enquanto com transcrições errou. Uma possível interpretação é que críticas terminológicas ou a políticas públicas podem ser confundidas com ataque a pessoas trans. Os resultados indicam benefício inicial do contexto recuperado, mas também mostram que ele pode alucinar quando o material recuperado enfatiza o tema geral em

ID	Contexto	Comentário	Real	Predições
E1	Thammy e diversidade	"até Thammy tem barba..."	I	ZS: N; FS/RAG: I
E2	Gênero além do binário	"meu gênero é helicóptero..."	I	ZS: N; FS/RAG: I
E3	Casal transgênero	"o que vale é o amor"	A	ZS: N; FS/RAG: A
E4	Declaração de nascimento	"parturiente é a mulher..."	N	FS: I; RAG: N
E5	Cota trans na Unicamp	"tem que entrar por mérito..."	N	FS/RAG-D: I; RAG-C: N

Tabela 2. Exemplos qualitativos de acertos e erros recorrentes. I, A e N indicam, respectivamente, intolerante, apoio e neutro.

vez da atitude do comentário.

5. Conclusão

Os resultados sugerem que a adição de contexto por RAG pode beneficiar a classificação de comentários sobre pessoas trans, especialmente quando há escassez de dados anotados para treinamento supervisionado. Neste recorte experimental, as estratégias com Qwen superaram o BERTimbau ajustado no conjunto disponível, e a recuperação de documentos externos obteve o melhor macro-F1. Esse resultado indica que informações contextuais externas podem complementar os sinais presentes no comentário e no título do vídeo, sobretudo em casos de ironia, apoio em poucas palavras e invalidação indireta. Ao mesmo tempo, a instabilidade observada na classe neutra mostra o quanto é difícil distinguir comentários ambíguos de atitudes indiretamente intolerantes, especialmente para transfobia.

Como continuidade deste trabalho, pretende-se ampliar o conjunto anotado e avaliar outros modelos generativos e supervisionados. Serão exploradas as demais dimensões já previstas no processo de anotação, incluindo a classificação de ofensividade do comentário, relevante para contrastar casos ambíguos em que a atitude discursiva não é explicitamente hostil, e a identificação de aspectos do discurso, como religião, política e negação de identidade de gênero. Este trabalho busca contribuir para novas pesquisas voltadas à identificação de discursos transfóbicos em redes sociais digitais.

Origem da pesquisa Este trabalho surgiu no contexto de um projeto de pesquisa em análise de redes sociais, voltado à investigação de discursos de ódio em ambientes digitais, e evoluiu para um Trabalho de Conclusão de Curso em andamento no Instituto de Computação da Universidade Federal do Rio de Janeiro. Devido à natureza sensível dos comentários e ao fato de o conjunto de dados ainda integrar um TCC em andamento, os dados anotados não serão disponibilizados publicamente nesta etapa. A disponibilização futura de versões anonimizadas ou agregadas será avaliada após a conclusão da pesquisa, considerando aspectos éticos, legais e de reprodutibilidade.

Declaração sobre uso de IA. No desenvolvimento deste trabalho, o Claude foi empregado para auxiliar na geração e revisão de código das *pipelines* experimentais, enquanto o GPT foi utilizado como apoio à revisão gramatical do texto. Todo o conteúdo científico, incluindo decisões metodológicas, análise dos resultados e verificação das referências bibliográficas, permaneceu sob responsabilidade integral dos autores humanos, em conformidade com o Código de Conduta para Autores da SBC.

Referências

Associação Nacional de Travestis e Transexuais (2026). Dossiê: Assassinatos e violências contra travestis e transexuais brasileiras em 2025. https://antrabrasil.org/wp-content/uploads/2026/01/dossie-antra-2026.pdf. Disponível em: https://antrabrasil.org/wp-content/uploads/2026/01/dossie-antra-2026.pdf. Acesso em: 18 maio 2026.

Barros, B. R. G. and Silva, D. d. C. P. (2025). Projeções escalares da transfobia em interações digitais: o caso do deputado nikolas ferreira no Youtube. *Revista do GELNE*, 27(2):e40199. DOI: https://doi.org/10.21680/1517-7874.2025v27n1ID40199.

Chakravarthi, B. R. (2024). Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, 18:49–68. DOI: https://doi.org/10.1007/s41060-023-00400-0.

Chakravarthi, B. R., Priyadharshini, R., Ponnusamy, R., Kumaresan, P. K., Sampath, K., Thenmozhi, D., Thangasamy, S., Nallathambi, R., and McCrae, J. P. (2021). Dataset for identification of homophobia and transophobia in multilingual YouTube comments. arXiv preprint arXiv:2109.00227. DOI: https://doi.org/10.48550/arXiv.2109.00227.

Channon, L. and Mathieson, N. (2025). Automated detection of mainstreamed transphobic content on YouTube. *Bulletin of Applied Transgender Studies*, 4(1–3):41–75. DOI: https://doi.org/10.57814/49jz-0663.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.aacl-main.91.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. DOI: https://doi.org/10.48550/arXiv.2005.11401.

Souza, F., Nogueira, R., and Lotufo, R. (2023). BERT models for brazilian portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing*, 149:110901. DOI: https://doi.org/10.1016/j.asoc.2023.110901.

Vargas, F., Carvalho, I., Pardo, T. A. S., and Benevenuto, F. (2025). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456. DOI: https://doi.org/10.1017/nlp.2024.18.

Willats, R., Pennington, J., Mohan, A., and Vidgen, B. (2025). Classification is a RAG problem: A case study on hate speech detection. arXiv preprint arXiv:2508.06204. DOI: https://doi.org/10.48550/arXiv.2508.06204.