Proposta recebida em Março 2026 e aceite para publicação em Junho 2026.

Um Comitê de Classificadores para Anotação Automática de Linguagem Tóxica em Português sob Escassez de Dados

An Ensemble of Classifiers for Automatic Annotation of Toxic Language in Portuguese under Data Scarcity

Francisco Assis Ricarte Neto
Instituto Federal do Piauí

Rafael Torres Anchiêta
Instituto Federal do Maranhão

Raimundo Santos Moura
Universidade Federal do Piauí

Pedro de Alcântara dos Santos Neto
Universidade Federal do Piauí

André Macedo Santana
Universidade Federal do Piauí

Resumo

Mensagens com linguagem tóxica configuram um problema recorrente nas redes sociais, evidenciando a necessidade urgente de métodos automáticos eficazes para mitigar seu impacto. Em geral, tais abordagens dependem de grandes volumes de dados rotulados, cuja construção é onerosa e demorada, além de demandar considerável esforço humano no processo de anotação. Para enfrentar esse desafio, este trabalho propõe um comitê de classificadores voltado à anotação automática de linguagem tóxica em português, concebido para operar com um número reduzido de dados previamente anotados. O comitê integra três estratégias complementares: um método semi-supervisionado baseado em grafos heterogêneos, uma abordagem de aprendizagem *few-shot* e um método fundamentado em *Retrieval-Augmented Generation*, ambos baseados em grandes modelos de linguagem. A proposta é avaliada em múltiplos *corpora*, considerando conjuntos originais e filtrados por concordância total entre anotadores. Os resultados indicam que o comitê apresenta desempenho competitivo, superando o melhor método individual em até 2% em cenários de maior equilíbrio entre os classificadores constituintes e mantendo desempenho comparável nos demais, além de preservar concordância moderada a substancial com os rótulos originais, evidenciando seu potencial para a construção de recursos linguísticos anotados sob escassez de dados.

Palavras chave

linguagem tóxica; anotação automática; comitê de classificadores

Abstract

Messages containing toxic language are a recurring problem on social media, highlighting the urgent need for effective automatic methods to mitigate their impact. Most existing approaches rely on large volumes of annotated data, which are costly, time-consuming, and highly labor-intensive. To address this challenge, this work proposes an ensemble of classifiers for the automatic annotation of toxic language in Portuguese, designed to operate under limited labeled data. The ensemble integrates three complementary strategies: a semi-supervised method based on heterogeneous graphs, a *few-shot* learning approach, and a *Retrieval-Augmented Generation* method, both grounded in large language models. The proposal is evaluated across multiple *corpora*, considering both their original versions and subsets filtered by total inter-annotator agreement. The results indicate that the ensemble exhibits competitive performance, surpassing the best individual method by up to 2% in scenarios of greater balance among the constituent classifiers and maintaining comparable performance in the remaining ones, while preserving moderate to substantial agreement with the original labels, demonstrating its potential for constructing annotated linguistic resources under data scarcity.

Keywords

toxic language; automatic annotation; ensemble of classifiers

1. Introdução

As redes sociais constituem poderosas ferramentas virtuais de interação humana, conectando pessoas em escala global e viabilizando a troca de informações, a expressão de opiniões e o debate de ideias. Entretanto, esses ambientes também têm favorecido a disseminação de diferentes formas de linguagem tóxica, por meio de mensagens que extrapolam os limites da liberdade de expressão e comprometem a quali-

dade da interação social. Esse tipo de conteúdo caracteriza-se pelo uso de linguagem inadequada, incluindo expressões profanas, insultos, ameaças e declarações ofensivas dirigidas a indivíduos ou grupos (Founta et al., 2018). Além das manifestações explícitas, a toxicidade pode ocorrer de forma mais sutil, por meio de comportamentos como comentários rudes, observações desrespeitosas ou ataques implícitos, frequentemente associados ao discurso de ódio ou a outras práticas comunicativas abusivas (Cjadams et al., 2017).

A relevância social da toxicidade em ambientes digitais tem sido amplamente discutida na literatura recente (Fortuna & Nunes, 2018; Poletto et al., 2020; Jahan & Oussalah, 2023). Embora existam marcos legais em diversos países que a tipificam como crime, como ocorre no Brasil (Brasil, 1989), além de políticas de moderação que preveem a remoção de conteúdos e, em casos mais graves, a exclusão de usuários das plataformas (X, 2026; Youtube, 2026), a linguagem tóxica permanece um problema de difícil contenção nos ambientes digitais. Diante do crescimento contínuo do número de usuários e do volume massivo de mensagens que exigem monitoramento, torna-se imprescindível o desenvolvimento de soluções automáticas para a detecção dessa linguagem. Do ponto de vista computacional, o problema tem sido tradicionalmente abordado como uma tarefa de classificação de texto, baseada majoritariamente em técnicas de Aprendizado de Máquina Supervisionado e de Aprendizagem Profunda (Jahan & Oussalah, 2023; Badjatiya et al., 2017). Tais abordagens pressupõem a disponibilidade de grandes volumes de dados anotados, o que constitui um problema, uma vez que os maiores esforços no desenvolvimento de recursos linguísticos são direcionados à língua inglesa (Poletto et al., 2020), enquanto línguas secundárias, como o português, permanecem subrepresentadas. Embora alguns *corpora* em português tenham sido propostos (de Pelle & Moreira, 2017; Nascimento et al., 2019; Fortuna et al., 2019; Vargas et al., 2022; Leite et al., 2020; Trajano et al., 2024), a disponibilidade de conjuntos de dados anotados em larga escala ainda é limitada.

A construção manual de *corpora* é um processo oneroso e demorado, agravado, no caso da linguagem tóxica, pela subjetividade interpretativa inerente à tarefa, que frequentemente requer o reconhecimento de manifestações implícitas por meio de recursos pragmáticos da linguagem, bem como por influências socioculturais e ideológicas dos próprios anotadores, aspectos que podem comprometer a consistência e a confiabilidade dos rótulos atribuídos (Aroyo et al., 2019; Hettiachchi et al., 2023). Adicionalmente, a anotação manual de textos tóxicos envolve impactos humanos significativos, uma vez que a exposição contínua a conteúdos violentos, abusivos ou perturbadores pode acarretar prejuízos psicológicos aos avaliadores[1].

Em resposta a essas limitações, a anotação automática de dados textuais tem emergido como uma alternativa promissora para a construção escalável de recursos linguísticos (Wang et al., 2021). Nesse contexto, abordagens semi-supervisionadas (Santos et al., 2022) e estratégias de *pseudo-labeling* (Dirting et al., 2022) vêm sendo exploradas para a construção e ampliação automática de *corpora* voltados à detecção de toxicidade. Mais recentemente, grandes modelos de linguagem (LLMs) têm sido empregados nessa tarefa (Gilardi et al., 2023; Tan et al., 2024), viabilizando inferências com poucos ou mesmo sem exemplos rotulados (*few-shot* e *zero-shot*). Contudo, tais métodos geralmente se restringem a um único *corpus* ou a um domínio específico e, em muitos casos, não avaliam sistematicamente a confiabilidade dos rótulos artificiais gerados. Além disso, o elevado custo computacional dos LLMs ainda representa um obstáculo relevante à adoção ampla dessas abordagens.

Diante desse cenário, este trabalho propõe um comitê constituído por três métodos complementares: (i) um método semi-supervisionado baseado em grafos heterogêneos, que explora relações estruturais entre textos e unidades linguísticas para a propagação de rótulos; (ii) uma estratégia fundamentada em aprendizagem *few-shot* com grandes modelos de linguagem; e (iii) um método baseado em *Retrieval-Augmented Generation* (RAG), que incorpora exemplos previamente anotados por meio de mecanismos de recuperação semântica. As decisões produzidas pelos três métodos são combinadas por meio de votação majoritária, constituindo o comitê proposto. Essa estratégia de agregação tem como objetivo aumentar a robustez do processo de anotação automática e reduzir a influência de vieses específicos de cada método, explorando a complementaridade entre abordagens fundamentadas em paradigmas distintos, cujos benefícios se manifestam de forma mais pronunciada em cenários de maior equilíbrio entre os métodos constituintes. A metodologia foi concebida para operar sob restrições de dados e em cenários de mudança de domínio, com o objetivo central de

[1] https://olhardigital.com.br/2023/01/19/pro/chatgpt-openai-explorou-trabalhadores-quenianos-para-identificar-conteudos-ilegais-e-criminosos/

gerar rótulos automáticos confiáveis para a construção e ampliação de *corpora* anotados. A proposta parte de um único *corpus* curado, composto por 1.400 instâncias (Neto et al., 2024), utilizado como base de treinamento, e é avaliada em *corpora* amplamente empregados na literatura (Vargas et al., 2022; Fortuna et al., 2019; Leite et al., 2020), com volumes significativamente superiores. Esse desenho experimental permite investigar se um volume reduzido de dados anotados é suficiente para sustentar a geração automática de rótulos confiáveis em cenários de domínio cruzado.

Para avaliar, de forma sistemática, a robustez dessa estratégia sob diferentes condições de anotação, a metodologia é examinada em dois cenários experimentais complementares. No primeiro, o comitê é avaliado com base nos *corpora* em suas versões originais, preservando os rótulos atribuídos pelos anotadores humanos e, consequentemente, as divergências interanotador inerentes ao processo de anotação manual. No segundo cenário, os conjuntos de dados são filtrados para incluir apenas instâncias com concordância plena entre os anotadores, resultando em subconjuntos com menor divergência anotativa e, presumivelmente, maior consistência nos rótulos atribuídos. A comparação entre esses cenários permite analisar, de forma controlada, em que medida a consistência das anotações de referência influencia o desempenho, a concordância e a estabilidade dos métodos de anotação automática.

Nesse enquadramento, o presente trabalho organiza sua investigação em torno de quatro questões de pesquisa, que orientam tanto o desenho experimental quanto a análise dos resultados:

QP1 Entre os classificadores individuais que compõem o comitê, semi-supervisionado baseado em grafos heterogêneos, aprendizagem *few-shot* e *Retrieval Augmented Generation*, qual apresenta melhor desempenho na anotação automática de linguagem tóxica sob escassez de dados rotulados?

QP2 A combinação desses métodos por meio de votação majoritária, constituindo um comitê de classificadores, resulta em desempenho superior ao do melhor método individual nos diferentes *corpora* avaliados?

QP3 Como o desempenho dos classificadores automáticos, incluindo o comitê, varia ao considerar apenas instâncias com concordância plena entre os anotadores, mantendo as mesmas configurações experimentais, e o que isso indica sobre a consistência das anotações de referência?

QP4 Em que medida o grau de concordância entre os métodos individuais do comitê está associado aos ganhos obtidos pela estratégia de votação majoritária nos diferentes *corpora* e cenários experimentais avaliados?

A partir das investigações conduzidas em torno dessas questões, destacam-se como principais contribuições deste trabalho os seguintes aspectos:

- Proposta de um comitê de classificadores para a anotação automática de linguagem tóxica em português, baseada na integração de três paradigmas complementares, semi-supervisionado em grafos heterogêneos, aprendizagem *few-shot* e *Retrieval-Augmented Generation*, combinados por votação majoritária.
- Avaliação sistemática do comitê em cenário de domínio cruzado, com treinamento realizado em um único *corpus* curado de pequeno volume e aplicação a três *corpora* de avaliação com volumes substancialmente superiores e domínios distintos, permitindo verificar a capacidade de generalização da abordagem sob escassez de dados rotulados.
- Análise comparativa em dois cenários de anotação, versões originais dos *corpora* e subconjuntos filtrados por concordância plena entre anotadores, evidenciando o impacto da consistência das anotações de referência sobre o desempenho dos métodos automáticos avaliados.
- Avaliação de equidade no comportamento dos classificadores em relação a instâncias contendo termos identitários, complementando a análise de desempenho com indicadores de equidade fundamentados na literatura (Borkan et al., 2019; Dixon et al., 2018).
- Disponibilização da metodologia e dos resultados como contribuição para a expansão de recursos linguísticos anotados em português, língua ainda sub-representada no contexto de pesquisas em detecção e anotação automática de linguagem tóxica.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta a fundamentação conceitual sobre toxicidade e fenômenos correlatos; a Seção 3 discute os trabalhos relacionados; a Seção 4 detalha os *corpora* utilizados; a Seção 5 descreve a metodologia proposta; a Seção 6 apresenta a configuração experimental, os resultados e a discussão; e, por fim, a Seção 7 discute as conclusões e direções para trabalhos futuros.

2. Toxicidade e Conceitos Correlatos

Na literatura, é possível encontrar uma variedade de definições para linguagem tóxica e seus correlatos, que dependem da perspectiva teórica sob a qual esse fenômeno é estudado. No contexto da sociolinguística, a linguagem tóxica é compreendida em sua dimensão performativa, na qual palavras ofensivas atuam como ações capazes de causar danos ou incitar à violência no mundo real (Butler, 2021). Nessa perspectiva, o significado da toxicidade não é determinado apenas pela estrutura léxica ou sintática do texto, mas também pela compreensão pragmática que articula os participantes do discurso e a história discursiva, configurando-a como um ato e não apenas como conteúdo. Em contraste, na literatura derivada do Processamento de Linguagem Natural, as definições acerca da linguagem tóxica tratam esse fenômeno de forma operacional, voltada à classificação automática de textos, organizando-o em manifestações específicas, como discurso de ódio, agressividade, linguagem ofensiva e *cyberbullying* (Fortuna & Nunes, 2018). Embora cada uma dessas categorias apresente particularidades relevantes para a tarefa de anotação, é frequente a sobreposição entre seus elementos característicos, o que pode introduzir ambiguidade no processo de anotação. O discurso de ódio refere-se especificamente a ataques baseados em características identitárias, como raça, gênero ou orientação sexual (Schmidt & Wiegand, 2017; Waseem & Hovy, 2016), enquanto a agressividade caracteriza-se pela intenção deliberada de agredir, sem necessariamente se vincular a marcadores identitários (Zampieri et al., 2019). A linguagem ofensiva, por sua vez, manifesta-se por meio de expressões obscenas ou depreciativas, sejam explícitas ou veladas (Fortuna & Nunes, 2018; Vargas et al., 2022), ao passo que o *cyberbullying* distingue-se pela frequência e persistência das agressões direcionadas a uma vítima específica (Mladenović et al., 2021). Essa sobreposição, somada à possibilidade de manifestações implícitas baseadas em ironia ou sarcasmo, pode comprometer a consistência dos *corpora* produzidos e, consequentemente, o desempenho dos modelos treinados a partir deles.

Nesse enquadramento teórico, o presente trabalho adota a definição proposta por Poletto et al. (2020), que organiza os diferentes fenômenos de linguagem abusiva em uma estrutura conceitual na qual a **toxicidade** é compreendida como o conceito mais abrangente, englobando manifestações como discurso de ódio, agressividade e linguagem ofensiva, conforme ilustrado na Figura 1. A adoção dessa estrutura justifica-se em razão de sua compatibilidade com a maioria dos *corpora* disponíveis em português, cujos esquemas de anotação variam quanto à granularidade, e por permitir que o método proposto seja aplicado a diferentes domínios sem a necessidade de remapeamento conceitual, o que favorece sua aplicação em cenários de mudança de domínio. Reconhece-se, contudo, que essa escolha implica uma simplificação, uma vez que o tratamento de todas as manifestações ofensivas sob um único rótulo binário não captura subtipos específicos de toxicidade. Assim, ao longo deste artigo, toda manifestação de caráter ofensivo é tratada como linguagem tóxica, respeitadas as particularidades conceituais inerentes a cada categoria correlata.

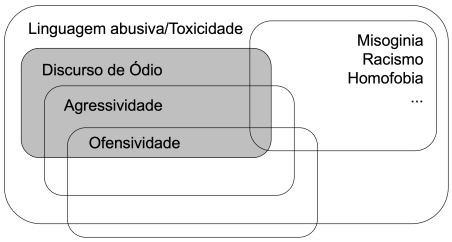

Figura 1: Estrutura conceitual de toxicidade e fenômenos de linguagem abusiva relacionados, adaptada de Poletto et al. (2020).

3. Trabalhos Relacionados

O combate ao conteúdo ofensivo em redes sociais tem motivado numerosos esforços acadêmicos voltados ao desenvolvimento de métodos automáticos para a detecção de linguagem tóxica (Fortuna & Nunes, 2018; Poletto et al., 2020). No contexto da língua portuguesa, diferentes abordagens têm sido investigadas, incluindo estratégias baseadas em léxicos (Pelle et al., 2018; Vargas et al., 2021), métodos tradicionais de Aprendizagem de Máquina (de Pelle & Moreira, 2017; Leite et al., 2020), modelos de *Deep Learning* (Soto et al., 2019) e classificadores fundamentados em arquiteturas *transformers* (Frediani et al., 2025). Apesar dos avanços alcançados, tais abordagens permanecem fortemente dependentes da disponibilidade de grandes volumes de dados anotados para treinamento, validação e avaliação. Mais recentemente, grandes modelos de linguagem também passaram a ser explorados na detecção de toxicidade, alcançando resultados expressivos (Oliveira et al., 2023, 2024; Assis et al., 2024). Contudo, essas abordagens ainda enfrentam limitações associadas aos elevados custos computacionais e aos desafios de escalabilidade, especialmente em cenários de aplicação em larga escala.

Embora os avanços na detecção automática de linguagem tóxica sejam expressivos, a dependência de grandes volumes de dados anotados permanece um dos principais gargalos da área. A anotação manual é extremamente onerosa, além de demandar muito tempo em sua realização. Além disso, está sujeita a vieses socioculturais e à subjetividade interpretativa inerente à tarefa, o que pode comprometer a consistência e a qualidade dos dados e, consequentemente, afetar o desempenho dos métodos de detecção de toxicidade (Aroyo et al., 2019; Hettiachchi et al., 2023). Diante desse cenário, cresce o interesse por estratégias de anotação automática voltadas à criação e à ampliação de *corpora* linguísticos, com o objetivo de reduzir custos e acelerar o processo de anotação.

Entre essas estratégias, destacam-se abordagens baseadas em auto-treinamento, em que um classificador inicialmente treinado com um conjunto reduzido de dados anotados é utilizado para rotular automaticamente novas instâncias, posteriormente incorporadas de forma iterativa ao processo de treinamento. Nessa linha, Saifullah et al. (2024) propõem um método para anotação automática de discurso de ódio em comentários do YouTube em indonésio, combinando representações *TF–IDF* e *Word2Vec* em um esquema de meta-vetorização. De maneira semelhante, Alsafari & Sadaoui (2021) exploram o auto-treinamento no contexto da língua árabe, avaliando diferentes combinações de classificadores e adotando limiares elevados de confiança para incorporar instâncias automaticamente rotuladas. Apesar de demonstrarem potencial para a expansão de *corpora*, tais abordagens permanecem fortemente dependentes do desempenho inicial do modelo e dos critérios de seleção, o que pode resultar na propagação de erros e na amplificação de vieses ao longo do processo iterativo.

Avançando nessa direção, Santos et al. (2022) propõem uma abordagem que integra auto-treinamento, propagação de rótulos por similaridade semântica e um classificador baseado em *Generative Adversarial Networks* associado ao BERT, organizados em um esquema iterativo. A metodologia transfere conhecimento de *corpora* previamente anotados, originalmente coletados em outros domínios, para anotar automaticamente um *corpus* não rotulado em um cenário *cross-domain*. Os resultados indicam que, embora a estratégia seja eficaz para ampliar o volume de dados anotados automaticamente, o simples aumento do conjunto de treinamento não garante melhorias consistentes no desempenho, podendo, inclusive, introduzir ruído e vieses no processo de anotação. Em outra vertente semi-supervisionada, Neto et al. (2024) investigam uma abordagem baseada em grafos heterogêneos para a anotação automática de linguagem tóxica. O estudo configura-se como uma investigação inicial, restrita a um único *corpus* e a uma configuração específica de *embeddings* estáticas, não explorando de forma sistemática cenários de mudança de domínio nem os desafios associados à detecção de formas implícitas de toxicidade, como ironia e sarcasmo, o que evidencia lacunas ainda existentes na modelagem da subjetividade linguística em português.

Estratégias de *pseudo-labeling* associadas a modelos supervisionados baseados em arquiteturas *transformers*, em especial as derivadas do BERT (Devlin et al., 2019), têm sido amplamente exploradas na literatura. Nessa direção, Suryawanshi et al. (2020) utilizam um conjunto reduzido de dados anotados manualmente para treinar diferentes classificadores, observando desempenho superior nas arquiteturas baseadas em BERT. O modelo selecionado é então empregado para rotular automaticamente um volume maior de dados não anotados, que passam a compor um conjunto expandido, utilizado em uma nova etapa de treinamento. De forma semelhante, Dirting et al. (2022) propõem uma abordagem também baseada em BERT para a classificação multi-rótulo da severidade do discurso de ódio, utilizando *pseudo-labeling* para anotar dados provenientes de múltiplas plataformas sociais. Embora esses estudos reportem melhorias consistentes em métricas tradicionais, como precisão, cobertura e *F-measure*, permanecem limitadas as análises acerca da confiabilidade dos rótulos automaticamente gerados e dos efeitos cumulativos da possível propagação de erros ao longo do processo iterativo de anotação.

Outros estudos investigam estratégias fundamentadas na similaridade semântica e no uso explícito de recursos linguísticos para a anotação automática de dados tóxicos. Phanomtip et al. (2021), por exemplo, exploram a combinação de dados rotulados e não rotulados na detecção de *cyberbullying* em tweets, empregando um classificador supervisionado baseado em SVM com representações *TF–IDF*, bem como métodos baseados em similaridade do cosseno. Os resultados indicam que a incorporação de instâncias automaticamente rotuladas por meio de um classificador supervisionado contribui para a melhoria do desempenho, enquanto abordagens fundamentadas exclusivamente na similaridade apresentam desempenho inferior. Em outra linha, Pelosi et al. (2017) propõem uma metodologia

para a anotação automática de linguagem ofensiva em italiano baseada em léxicos especializados e regras linguísticas, com ênfase na identificação de expressões tabu. Apesar dos resultados promissores, ambas as abordagens revelam forte dependência de recursos linguísticos específicos e enfrentam limitações na adaptação a novos domínios.

Após essas abordagens predominantemente automáticas, outras propostas buscam mitigar as limitações de escalabilidade e dependência de recursos específicos por meio de estratégias híbridas. Nesse sentido, abordagens semi-automáticas baseadas no paradigma *Human-in-the-Loop* têm sido propostas como forma de equilibrar custo e qualidade na anotação de dados textuais. Botella-Gil et al. (2024) integram técnicas de *Active Learning* a um processo de pré-anotação automática conduzido por modelos de aprendizado profundo, cujas predições são posteriormente validadas por especialistas humanos. Aplicada à construção de um *corpus* em espanhol composto por mensagens violentas, a metodologia proporciona uma redução significativa no tempo de anotação manual, mantendo níveis competitivos de desempenho na detecção automática. Ainda assim, esse tipo de abordagem permanece dependente da disponibilidade contínua de anotadores humanos, o que pode comprometer sua escalabilidade em cenários de grande volume de dados.

Em síntese, embora a literatura registre avanços significativos no emprego de estratégias automáticas e semi-supervisionadas para a anotação de linguagem tóxica, persistem lacunas relevantes. Grande parte dos estudos são abordagens isoladas, avaliadas em cenários controlados e frequentemente restritas a um único *corpus* ou domínio específico, com análises limitadas quanto à capacidade de generalização entre contextos distintos. Ademais, as investigações que avaliam sistematicamente a confiabilidade dos rótulos gerados ainda são escassas, especialmente em tarefas caracterizadas por um elevado grau de subjetividade interpretativa, como a identificação de toxicidade implícita, de ironia e de sarcasmo. Permanecem igualmente pouco exploradas estratégias que articulem múltiplos paradigmas de anotação automática e que sejam avaliadas em diferentes níveis de ruído de anotação e em cenários de domínio cruzado. No contexto da língua portuguesa, tais limitações tornam-se ainda mais evidentes, em razão da escassez de recursos linguísticos anotados e da predominância de modelos supervisionados cuja capacidade de generalização é insuficientemente avaliada.

4. *Corpora*

Para o desenvolvimento e a avaliação da metodologia proposta, foram utilizados quatro *corpora* linguísticos de linguagem tóxica em língua portuguesa: Toxic-BR (Neto et al., 2024), ToLD-BR (Leite et al., 2020), HateBR (Vargas et al., 2022) e HLPHSD (Fortuna et al., 2019). A seleção desses conjuntos de dados buscou contemplar diferentes contextos de coleta, esquemas de anotação e plataformas de origem, permitindo uma avaliação mais abrangente da robustez e da capacidade de generalização da abordagem proposta em cenários de mudança de domínio.

O *corpus* Toxic-BR é composto por 1.400 textos em língua portuguesa extraídos da rede social *X* (anteriormente *Twitter*), coletados durante o segundo turno das eleições presidenciais brasileiras de 2022, período marcado por intensa polarização política e por frequente troca de ofensas entre apoiadores dos candidatos (Neto et al., 2024). Inicialmente, os dados foram automaticamente anotados de forma binária (*tóxico* e *não tóxico*) por meio da *Perspective API* (Lees et al., 2022) e do modelo GPT-3 (Brown et al., 2020); posteriormente, essas anotações passaram por um processo de curadoria manual conduzido por especialistas humanos, que avaliaram a qualidade das classificações automáticas e corrigiram inconsistências, visando aumentar a consistência e a confiabilidade dos rótulos finais.

Além do Toxic-BR, este trabalho utiliza o ToLD-BR, um *corpus* composto por 21.000 *tweets* manualmente anotados em sete categorias, incluindo racismo, linguagem obscena, insulto, xenofobia, LGBTQ+fobia, misoginia e conteúdos não tóxicos (Leite et al., 2020). O processo de anotação envolveu 48 colaboradores, organizados em grupos de três avaliadores responsáveis por aproximadamente 1.500 textos cada, o que permitiu múltiplas avaliações por instância. Adicionalmente, os autores disponibilizam uma versão binária do conjunto, na qual 9.255 mensagens originalmente associadas às categorias de toxicidade foram agrupadas na classe tóxica, enquanto as demais foram rotuladas como não tóxicas.

De maneira complementar, o *corpus* HateBR é constituído por 7.000 textos em português, coletados a partir de publicações em perfis políticos da rede social *Instagram* (Vargas et al., 2022). As anotações contemplam diferentes esquemas, incluindo classificação binária (ofensivo vs. não ofensivo), níveis graduais de ofensividade e categorias específicas de discurso de ódio, como sexismo, homofobia, intolerância religiosa e racismo. Um diferencial desse conjunto reside no

fato de ter sido integralmente anotado por três especialistas com elevada formação acadêmica, o que contribui para maior consistência e qualidade nas rotulações.

Por fim, o *corpus* HLPHSD também reúne textos oriundos da antiga rede social *Twitter*, totalizando 3.882 *tweets* neutros e 1.788 *tweets* rotulados como discurso de ódio (Fortuna et al., 2019). Quando identificado conteúdo ofensivo, são atribuídos rótulos hierárquicos adicionais correspondentes a categorias específicas, como sexismo, homofobia e racismo. Para a tarefa de classificação binária, cada *tweet* foi avaliado por três anotadores, o que permitiu analisar o grau de concordância entre os avaliadores humanos.

A Tabela 1 sintetiza a distribuição de instâncias tóxicas e não tóxicas em cada *corpus*, bem como a média de *tokens* por texto. De modo geral, observa-se que a classe não tóxica é a mais frequente em todos os conjuntos, o que reflete tanto a distribuição natural de conteúdos em ambientes digitais quanto os desafios inerentes à coleta e à anotação de mensagens explicitamente ofensivas. Além disso, nota-se que os *corpora* oriundos da rede social *X* apresentam, em média, textos mais longos do que os do HateBR, possivelmente em razão de diferenças no estilo discursivo e na dinâmica de interação entre as plataformas.

Corpus	**Tóxicos**	**Não Tóxicos**	**Origem**	**#Tokens**
Toxic-BR	615	785	*X*	49.47
ToLD-BR	9.255	11.745	*X*	30.22
HateBR	3.500	3.500	*Instagram*	24.31
HLPHSD	1.788	3.882	*X*	37.43

Tabela 1: Distribuição das instâncias por classe, origem dos textos, bem como a média do número de *tokens* por instância nos *corpora* em suas versões originais.

Além das características de distribuição apresentadas, os *corpora* analisados apresentam níveis distintos de concordância interanotador, conforme relatado nos respectivos trabalhos. Tais variações, no entanto, não são interpretadas neste estudo como evidência direta de maior ou menor ambiguidade entre os conjuntos de dados, uma vez que podem decorrer de diferenças nos protocolos de anotação, nas diretrizes adotadas, no perfil dos anotadores ou nos procedimentos de validação dos rótulos. Considerando, contudo, que a identificação de linguagem tóxica constitui uma tarefa intrinsecamente subjetiva, instâncias com concordância total entre os anotadores são, neste trabalho, tratadas como casos de menor divergência interanotador no contexto específico de cada *corpus*. Nesse sentido, realizou-se uma etapa adicional de filtragem, mantendo-se apenas os textos para os quais houve consenso absoluto quanto ao rótulo atribuído. A Tabela 2 apresenta a distribuição das instâncias após esse procedimento, evidenciando, como efeito direto, um aumento do desbalanceamento entre as classes e uma redução das médias de *tokens* dos *corpora* filtrados. O uso específico de cada conjunto na avaliação experimental é detalhado na Seção 6.

Corpus	**Tóxicos**	**Não Tóxicos**	**#Tokens**
ToLD-BR	1.453	11.571	21.12
HateBR	2.382	3.298	18.17
HLPHSD	507	1.957	19.75

Tabela 2: Distribuição de instâncias tóxicas e não tóxicas, bem como a média do número de *tokens* por texto, nos *corpora* após a filtragem por concordância total.

5. Metodologia

Esta Seção descreve a metodologia desenvolvida para a anotação automática de linguagem tóxica. O comitê de classificadores integra três abordagens complementares que compartilham o mesmo objetivo de atribuição automática de rótulos de toxicidade, mas exploram princípios distintos de modelagem e inferência: um método semi-supervisionado responsável pela propagação de rótulos em grafos heterogêneos; um classificador baseado em LLM que realiza inferência orientada por exemplos; e outro classificador que conduz a inferência enriquecida por mecanismos de recuperação semântica. Essa integração tem como objetivo aumentar a robustez do processo de anotação automática e reduzir a influência de vieses específicos de cada método, explorando a complementaridade entre abordagens com princípios distintos de modelagem e inferência.

A Figura 2 apresenta a visão geral da metodologia. O processo inicia-se com textos brutos, submetidos a pré-processamento e normalização. O texto resultante é então encaminhado aos três métodos do comitê, que realizam, de forma independente, a classificação binária das instâncias como *tóxicas* ou *não tóxicas*. As predições são agregadas por meio de votação majoritária, o que define o rótulo final. Os textos automaticamente anotados podem ser utilizados na construção ou na ampliação de *corpora* de linguagem tóxica. Embora tradicionalmente associadas à detecção, as abordagens são empregadas sob a perspectiva da anotação automática.

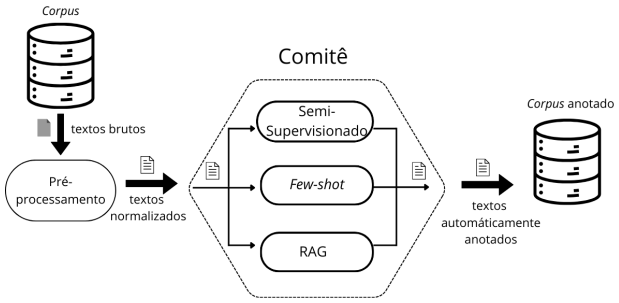

Figura 2: Visão geral da metodologia com um comitê de classificadores para a anotação automática de textos tóxicos.

Nas Subseções seguintes, detalham-se a etapa de pré-processamento (Subseção 5.1), os métodos que compõem o comitê (Subseções 5.2, 5.3 e 5.4) e a estratégia de combinação das predições (Subseção 5.5).

5.1. Pré-processamento

A etapa de pré-processamento consiste na normalização do conteúdo textual utilizado pelos métodos do comitê. Para isso, foram removidas menções a *URLs*, indicadores de *retweets* (RT), referências a usuários e *emojis*, e foram descartados os textos compostos exclusivamente por *emojis*. Além disso, considerando a predominância de linguagem coloquial em ambientes digitais, caracterizada por abreviações e variações ortográficas frequentes, aplicou-se a ferramenta Enelvo (Costa Bertaglia & Volpe Nunes, 2016) para a normalização textual. Esse procedimento contribui para tornar os dados mais homogêneos e, consequentemente, mais adequados ao processamento pelos classificadores do comitê.

5.2. Método Semi-Supervisionado

O método semi-supervisionado adotado neste trabalho foi inspirado na proposta de (Neto et al., 2024), que modela textos tóxicos por meio de grafos heterogêneos e emprega um algoritmo de propagação de rótulos para a anotação automática de linguagem tóxica. Para sua integração ao comitê, a abordagem foi estendida com a incorporação de *word embeddings* contextuais e com a aplicação conjunta do algoritmo original de propagação de rótulos e de uma estratégia alternativa, permitindo examinar diferentes dinâmicas de propagação em grafos, aspectos não explorados no estudo original.

O método semi-supervisionado está estruturado em três etapas principais: (i) modelagem do grafo, (ii) regularização e (iii) classificação, as quais são descritas a seguir.

5.2.1. Modelagem do grafo

Grafos constituem estruturas amplamente empregadas na representação de dados e têm recebido atenção crescente na última década, sendo aplicados com êxito em diversas tarefas, como a modelagem de tópicos e a desambiguação de nomes, frequentemente com resultados promissores (King et al., 2014). Em particular, grafos heterogêneos possibilitam a integração de informações estruturais, expressas por arestas que conectam nós de diferentes tipos com conteúdos não estruturados associados a cada entidade, resultando em modelos mais expressivos e semanticamente enriquecidos (Zhang et al., 2019). Sua principal contribuição reside na explicitação das relações entre entidades distintas, o que favorece uma representação mais informativa e contextualizada em múltiplos domínios de aplicação.

Com base nesses princípios, a primeira etapa do método semi-supervisionado consiste na modelagem dos textos em um grafo heterogêneo, no qual cada instância é representada por nós correspondentes à sentença textual, aos *tokens* que a compõem e a um nó adicional denominado Toxicidade. Esse nó não representa o rótulo final da instância, mas atua como uma característica auxiliar incorporada à estrutura do grafo, contribuindo exclusivamente para a modelagem relacional durante o processo de regularização, sem ser utilizado como variável supervisionada nem como rótulo de saída do método. O valor associado ao nó Toxicidade corresponde a um índice de ofensividade estimado com base nos termos potencialmente ofensivos presentes no texto, identificados por meio do léxico MOL (Vargas et al., 2021). O grau de toxicidade de cada termo é definido com o auxílio da *Perspective API* (Lees et al., 2022), e o valor final da toxicidade do texto é calculado como a soma dos índices atribuídos aos termos ofensivos identificados.

No que se refere às arestas do grafo, estas são não direcionadas e ponderadas, estabelecendo conexões exclusivamente entre nós do tipo sentença e nós do tipo *token*, sem ligações diretas entre sentenças nem entre *tokens*. As conexões entre nós de sentença e o nó de Toxicidade, por sua vez, não recebem pesos e são utilizadas apenas para representar relações estruturais na modelagem do grafo. O peso de cada aresta que conecta um nó *sentença* a um nó *token* é calculado a partir da média dos valores do vetor de *embedding* associado ao termo correspondente, de modo que cada aresta passe a refletir um valor escalar que representa a contribuição semântica do *token* no contexto da sentença. Adicionalmente, nós de **sentença** compartilham nós de

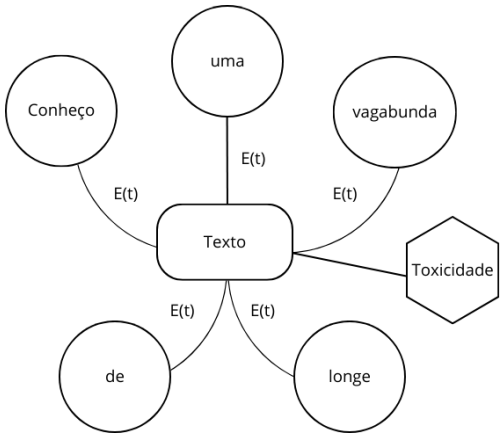

Figura 3: Exemplo da estrutura do grafo modelada para a sentença "*Conheço uma vagabunda de longe!*".

token sempre que o mesmo termo ocorre em diferentes textos, permitindo que sentenças distintas se conectem indiretamente por meio desses *tokens* compartilhados e favorecendo a propagação relacional de informação ao longo da estrutura do grafo. A Figura 3 ilustra a configuração completa do grafo heterogêneo construído a partir de um exemplo de texto tóxico, evidenciando os nós de sentença, *tokens* e Toxicidade, bem como as arestas ponderadas que conectam sentenças e *tokens*. Diferentemente da proposta original, a atribuição dos pesos das arestas foi realizada com base em *embeddings* contextuais extraídos do modelo BERTimbau (Souza et al., 2020), em sua versão pré-treinada, aplicado diretamente aos dados brutos, sem ajuste fino ou engenharia de *prompts*. Em contraste com *embeddings* distribucionais estáticos, representações contextuais, como as geradas pelo BERTimbau, capturam variações semânticas dependentes do contexto, produzindo vetores distintos para cada ocorrência de um mesmo *token* (Smith, 2020).

5.2.2. Regularização

A etapa de regularização é responsável pela extração de características a partir da estrutura do grafo, configurando-se como um processo de classificação transdutiva ou semi-supervisionada. Seu objetivo consiste em determinar um conjunto de rótulos que satisfaça simultaneamente duas propriedades: (i) consistência com os dados previamente rotulados e (ii) coerência com a topologia do grafo, sob a premissa de que nós vizinhos tendem a compartilhar rótulos semelhantes (Rossi, 2015).

Na proposta original de Neto et al. (2024), empregou-se exclusivamente o algoritmo *Gaussian Fields and Harmonic Functions* (GFHF) apresentado por Zhu et al. (2003). No presente trabalho, além do GFHF, incorpora-se também o algoritmo *Local and Global Consistency* (LGC) (Zhou et al., 2004), ampliando a análise para diferentes estratégias de propagação de rótulos. Ambos os métodos propagam rótulos no grafo heterogêneo com base na similaridade entre nós conectados; contudo, diferem no tratamento de nós previamente rotulados. Enquanto o GFHF mantém os rótulos iniciais fixos ao longo de todo o processo, o LGC permite sua atualização iterativa, conferindo maior flexibilidade ao mecanismo de propagação.

Para a execução dos algoritmos de regularização, é necessário definir um conjunto inicial de nós rotulados, que serve como ponto de partida para a classificação transdutiva. Por exemplo, ao selecionar 20% dos nós como pré-rotulados, pode-se assegurar que 10% das instâncias de cada classe sejam escolhidas aleatoriamente. A inferência dos rótulos dos nós não rotulados é realizada com base na similaridade com seus vizinhos mais próximos, ponderada pelos pesos das arestas do grafo. Como resultado, os algoritmos de regularização produzem, para cada nó, uma representação vetorial induzida pela estrutura relacional do grafo, conforme ilustrado na Tabela 3. Nessa tabela, o campo **ID** identifica o nó correspondente à instância textual; os campos **Valor 1** e **Valor 2** representam suas coordenadas; e o campo **Rótulo** indica a classe inferida pelo regularizador, sendo **1** para textos tóxicos e **0** para textos não tóxicos.

ID	Valor 1	Valor 2	Rótulo
440	0.03221	0.01447	1
702	0.05716	0.02365	1
6437	0.0029	0.00011	0
6464	-30.09728	-1.11603	0

Tabela 3: Exemplo de saída do algoritmo de regularização LGC.

5.2.3. Classificação

Na etapa de classificação, as coordenadas produzidas pelo processo de regularização (ver Tabela 3) são utilizadas como atributos de entrada para algoritmos de Aprendizagem de Máquina Supervisionada.

5.3. Método baseado em aprendizagem *few-shot*

A abordagem de aprendizagem *few-shot* orienta a inferência do LLM por meio da inclusão explícita de exemplos previamente anotados no *prompt*. Para cada conjunto de textos a ser classificado, são selecionadas quatro instâncias do *corpus* de treinamento, duas de cada classe, incorporadas como referências para a decisão sobre novas entradas e mantidas fixas ao longo de todo o processo, assegurando uniformidade na instrução fornecida ao modelo. A escolha dessa configuração preserva o comprimento do *prompt* dentro de limites compatíveis com modelos de menor porte (Brown et al., 2020), além de que a distribuição balanceada entre categorias contribui para reduzir o viés de instrução (da Silva Oliveira et al., 2024; Szcz et al., 2025). Os exemplos selecionados, juntamente com o texto-alvo, estruturam o *prompt* conforme o formato apresentado no Exemplo 5.3, que é então submetido ao modelo, que realiza a inferência e retorna a classificação binária da instância como *tóxico* ou *não tóxico*.

> **Exemplo (Prompt do Comitê).** Você é um especialista em detecção de toxicidade em textos. A linguagem tóxica é caracterizada por qualquer forma de expressão fortemente indelicada, rude ou ofensiva, incluindo palavrões ou declarações que demonstrem desrespeito a indivíduos ou grupos.
>
> Sua tarefa é analisar o texto fornecido e classificá-lo como `"tóxico"` ou `"nao_tóxico"`, considerando apenas o conteúdo textual apresentado e os exemplos fornecidos.
>
> **Saída:**
>
> - Responda exclusivamente em JSON.
> - Não produza texto adicional.
>
> **Formato da saída:**
> {`"label": «tóxico|nao_tóxico>"`}
> **Exemplos:** {exemplos}
> **Texto a classificar:** "{texto}"

5.4. Método baseado em *Retrieval-Augmented Generation*

O método baseado em *Retrieval-Augmented Generation* (RAG) estende a abordagem *few-shot* ao substituir a seleção fixa de exemplos por um mecanismo automático de recuperação semântica (Lewis et al., 2020). Nessa estratégia, o texto-alvo é inicialmente convertido em uma representação vetorial e utilizado como consulta para recuperar os k exemplos mais semelhantes a partir de uma base indexada construída com o conjunto de treinamento. A indexação e a busca vetorial são realizadas por meio da biblioteca FAISS[2] em conjunto com o modelo de *embeddings* *paraphrase-multilingual-MiniLM-L12-v2*[3], selecionado por seu suporte multilíngue, incluindo o português, e pelo equilíbrio entre a qualidade de representação semântica e a eficiência computacional. Em seguida, os exemplos recuperados, juntamente com seus respectivos rótulos, são incorporados à construção do *prompt*, preservando o mesmo formato adotado no método *few-shot*. O *prompt* resultante é posteriormente encaminhado ao LLM para a etapa de inferência, na qual o rótulo de toxicidade é atribuído ao texto-alvo. O valor de $k = 4$ foi mantido para permitir uma comparação controlada entre as estratégias *few-shot* e RAG.

5.5. Estratégia de Combinação do Comitê

Após a execução independente dos três métodos descritos nas subseções anteriores, suas predições são agregadas por meio de votação majoritária. A predição final é definida pela Equação 1:

$$\hat{y} = \arg\max_{j} \sum_{t=1}^{T} d_{t,j} \tag{1}$$

em que $d_{t,j} = 1$ se o classificador t prediz a classe j, e $d_{t,j} = 0$ caso contrário. Essa formulação explicita que a classe selecionada é a que acumula o maior número de votos entre os classificadores do comitê. Essa estratégia de aprendizagem baseada em comitês é amplamente reconhecida por sua capacidade de aumentar a robustez e a estabilidade das predições ao combinar classificadores com comportamentos distintos (Zhou, 2012).

6. Experimentos e Resultados

Esta Seção apresenta as configurações experimentais adotadas para a avaliação do comitê de classificadores e para a análise dos resultados obtidos, com ênfase no desempenho da classe tóxica, por se tratar do caso mais crítico no contexto da detecção e da anotação automática de linguagem abusiva. Para essa finalidade, o desempenho do comitê e de seus métodos constituintes é avaliado por meio da métrica *F-measure* e do coeficiente kappa de Cohen. A *F-measure* foi escolhida por corresponder à média harmônica

[2] https://python.langchain.com/docs/integrations/vectorstores/faiss/

[3] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

entre precisão e cobertura, permitindo uma avaliação equilibrada entre a proporção de instâncias corretamente classificadas e a capacidade do modelo de recuperar casos relevantes. Complementarmente, o coeficiente Kappa de Cohen é empregado para mensurar o grau de concordância entre dois avaliadores, considerando tanto a concordância observada quanto a esperada ao acaso, o que resulta em uma estimativa mais realista da confiabilidade dos rótulos atribuídos; sua interpretação permite caracterizar níveis de concordância que variam de fraca a quase perfeita (Landis & Koch, 1977). Neste trabalho, o Kappa é calculado com base na concordância entre os rótulos originais dos *corpora* e os atribuídos automaticamente pelos métodos do comitê.

6.1. Configuração Experimental

Os métodos que compõem o comitê foram avaliados de forma independente, adotando-se o *corpus* Toxic-BR como base de treinamento e os *corpora* ToLD-BR, HateBR e HLPHSD exclusivamente como conjuntos de teste. Ressalta-se que o Toxic-BR reúne 1.400 instâncias (Neto et al., 2024), enquanto os conjuntos de avaliação variam entre 5.670 e 21.000 textos (Fortuna et al., 2019; Vargas et al., 2022; Leite et al., 2020). Essa assimetria configura um regime de treinamento com baixo volume de dados em relação aos cenários de avaliação em larga escala, permitindo examinar a capacidade de generalização do comitê a partir de um conjunto reduzido de exemplos previamente curados. Consequentemente, o desenho experimental permite analisar o comportamento dos métodos em diferentes domínios, plataformas de coleta e esquemas de anotação, sem necessidade de ajustes específicos para cada *corpus*. Ademais, como linhas de base comuns a todas as abordagens, considerou-se uma configuração *zero-shot*, com o propósito de mensurar o patamar de desempenho obtido exclusivamente por meio de *prompting*, sem o fornecimento de exemplos rotulados, e um classificador BERTimbau (Souza et al., 2020) submetido a ajuste fino no *corpus* Toxic-BR, permitindo situar os resultados do comitê em relação a um método supervisionado baseado em *transformers* (Wolf et al., 2020).

Com base nesse desenho experimental, definiram-se dois cenários de avaliação. No primeiro, os métodos foram aplicados às versões originais dos *corpora*, conforme apresentado na Tabela 1. No segundo, utilizaram-se apenas os subconjuntos compostos por instâncias com concordância total entre os anotadores humanos, visando reduzir o impacto do ruído de anotação e investigar seus efeitos sobre o desempenho dos métodos (ver Tabela 2). Cabe ressaltar que esses dois cenários desempenham papéis complementares. As QP1 e QP2 são respondidas com base no cenário original. A QP3 é examinada exclusivamente no cenário filtrado, por tratar do efeito da concordância plena entre anotadores sobre o desempenho dos métodos. A QP4 é analisada em ambos os cenários, o que permite verificar como o grau de concordância entre os métodos do comitê se comporta entre os dois regimes de definição do rótulo.

No método semi-supervisionado baseado em grafos heterogêneos, a construção do grafo integra instâncias do conjunto de treinamento (Toxic-BR) como exemplos iniciais para a aplicação dos algoritmos de regularização, viabilizando a propagação de rótulos nos *corpora* de teste (ToLD-BR, HateBR e HLPHSD) e permitindo a exploração de relações estruturais entre textos provenientes de diferentes domínios. Conforme discutido em (Neto et al., 2024), a definição de um volume adequado de rótulos iniciais é fundamental para assegurar estatísticas mais estáveis na identificação de linguagem tóxica e maior alinhamento com os conjuntos de avaliação. Assim, foram investigadas proporções de 10%, 20% e 30% de instâncias pré-rotuladas. As representações vetoriais obtidas ao final da etapa de regularização foram empregadas como atributos de entrada para três classificadores supervisionados, implementados com a biblioteca *Scikit-Learn* (Pedregosa et al., 2011): um *Multi-Layer Perceptron* (MLP)[4], uma *Support Vector Machine* (SVM) treinada por meio do *SGDClassifier*[5] e um modelo de *Gradient Boosting* (GB), implementado com o *HistGradientBoostingClassifier*[6]. Como a seleção das instâncias pré-rotuladas ocorre aleatoriamente, diferentes execuções podem gerar variações na estrutura do grafo e, consequentemente, nos resultados obtidos. Por essa razão, cada configuração experimental foi executada 10 vezes, considerando ambos os algoritmos de regularização, GFHF e LGC.

Já para os métodos baseados em grandes modelos de linguagem, incluindo o método de linha de base *zero-shot*, a estratégia *few-shot* e a abordagem fundamentada em *Retrieval-Augmented Generation*, optou-se por modelos de porte reduzido, considerando as limitações computaci-

[4]Configurado com 100 unidades na camada oculta, função de ativação ReLU, otimizador Adam, taxa de aprendizado de 0,001 e máximo de 300 iterações.

[5]Empregada com núcleo linear.

[6]Utilizado com parâmetros padrão da biblioteca, sem ajuste adicional.

onais disponíveis e o propósito de viabilizar a anotação automática em cenários de custo restrito. Nesse contexto, foram empregados os modelos multilíngues Qwen2.5-7B (Yang et al., 2025) e Granite3.3-8B (IBM Research, 2025), utilizados sem ajuste fino e explorados exclusivamente por meio de engenharia de *prompt*. Os experimentos com LLMs foram conduzidos localmente em uma máquina equipada com GPU *NVIDIA GeForce RTX 3060* (12 GB), CPU *Intel(R) Core(TM) i5-10400F* (2,90 GHz) e 128 GB de memória RAM. Para tanto, a implementação foi realizada com a biblioteca *Ollama*[7], que permite a execução eficiente de modelos de linguagem em ambiente local.

Como linha de base adicional baseada em aprendizado supervisionado, avaliou-se o modelo BERTimbau *Base* (Souza et al., 2020) submetido a ajuste fino no *corpus* Toxic-BR. O modelo foi configurado para classificação binária, com uma camada linear aplicada ao vetor do token [CLS], e treinado por 3 épocas, com tamanho de batch de 16, utilizando a biblioteca *transformers* da *Hugging Face* (Wolf et al., 2020)[8]. Manteve-se o mesmo cenário de domínio cruzado adotado para os demais métodos, com treinamento exclusivo no Toxic-BR e a avaliação nos *corpora* ToLD-BR, HateBR e HLPHSD, sem qualquer ajuste fino adicional nos conjuntos de avaliação.

6.2. Resultados Globais dos Classificadores

Esta Subseção apresenta uma visão geral do desempenho dos métodos avaliados, fornecendo uma base comparativa para a discussão detalhada nas Subseções seguintes. A Tabela 4 reúne os melhores resultados obtidos por cada classificador individual nos três *corpora* de avaliação, em termos de *F-measure* para a classe tóxica e do coeficiente Kappa de Cohen, considerando os *corpora* em suas versões originais.

De modo geral, os resultados apresentados na Tabela 4 indicam que os classificadores baseados em LLMs apresentam desempenho superior ao do classificador semi-supervisionado e ao classificador baseado em BERTimbau com ajuste fino na maioria dos cenários avaliados, sendo essa diferença mais pronunciada nos *corpora* HateBR e HLPHSD, enquanto, no ToLD-BR, os resultados das abordagens são mais próximos entre si. Entre as estratégias baseadas em LLMs, o classificador *few-shot* obteve os melhores resultados individuais em dois dos três *corpora* avaliados (HateBR e HLPHSD), além de apresentar desempenho equivalente ao do RAG no ToLD-BR e de apresentar ganhos em relação aos classificadores de linha de base, tanto *zero-shot* quanto BERTimbau com ajuste fino, nos três *corpora*.

Corpus	Método	Configuração	F	κ
ToLD-BR	BERT	BERTimbau *ajuste-fino*	0,72	0,45
	zero-shot	Granite3.3-8B	0,73	0,47
	semi	LGC_10%_SVM	0,72	0,46
	few-shot	Granite3.3-8B	**0,74**	**0,48**
	RAG	Qwen2.5-7B	**0,74**	0,45
HateBR	BERT	BERTimbau *ajuste-fino*	0,82	0,66
	zero-shot	Qwen2.5-7B	0,86	0,71
	semi	LGC_30%_MLP	0,37	0,22
	few-shot	Qwen2.5-7B	**0,87**	**0,73**
	RAG	Qwen2.5-7B	0,85	0,70
HLPHSD	BERT	BERTimbau *ajuste-fino*	0,56	0,37
	zero-shot	Qwen2.5-7B	0,62	0,38
	semi	LGC_10%_SVM	0,41	0,06
	few-shot	Qwen2.5-7B	**0,63**	**0,41**
	RAG	Qwen2.5-7B	0,61	0,39

Tabela 4: Melhores resultados obtidos por cada método nos *corpora* avaliados em suas versões originais, em termos de *F-measure* (F) e coeficiente Kappa de Cohen (κ). A configuração adotada em cada caso é indicada na coluna **Configuração**. Valores em negrito indicam o melhor resultado global em cada *corpus*.

O classificador RAG, por sua vez, apresenta desempenho comparável ao *zero-shot* em parte dos cenários e superior ao BERTimbau, mantendo comportamento competitivo entre as abordagens baseadas em LLM. Esses resultados são observados em configurações que utilizam exemplos rotulados no *prompt*, seja por seleção fixa (*few-shot*) ou por recuperação semântica (RAG), nos cenários de melhor desempenho. No caso do classificador semi-supervisionado, observa-se desempenho competitivo apenas no *corpus* ToLD-BR, com resultados próximos aos das abordagens baseadas em LLM, enquanto nos demais *corpora* os valores de *F-measure* e Kappa são significativamente inferiores, evidenciando maior sensibilidade do método à variação de domínio.

Quanto à **QP1**, o classificador *few-shot* apresentou o melhor desempenho individual na anotação automática de linguagem tóxica nos *corpora* avaliados, com o resultado mais expressivo observado no *corpus* HateBR no cenário original.

[7] https://ollama.com/

[8] Foram utilizados a taxa de aprendizado de (5×10^{-5}), aquecimento inicial de 500 passos e decaimento de peso de 0,01.

6.3. Análise da Anotação Automática por Classificador

As subseções a seguir detalham o comportamento de cada classificador constituinte do comitê, com ênfase em aspectos que não são diretamente capturados pela Tabela 4, como padrões de variação entre as configurações, sensibilidade à mudança de domínio e diferenças entre os modelos de linguagem avaliados. Os resultados das melhores configurações do método semi-supervisionado e das abordagens baseadas em LLMs, detalhados pelos modelos Qwen2.5-7B e Granite3.3-8B para cada método e *corpora*, são apresentados nos Apêndices A e B (Tabelas 11 e 12).

6.3.1. Linhas de Base: Zero-Shot e BERTimbau

Dois métodos foram adotados como linhas de base para comparação com os classificadores do comitê. O primeiro, o método *zero-shot*, emprega os mesmos modelos de LLMs e a mesma estrutura de *prompt* adotados nas abordagens *few-shot* e RAG, distinguindo-se apenas pela ausência de exemplos rotulados. O segundo, o classificador BERTimbau com ajuste fino no Toxic-BR, representa uma abordagem supervisionada baseada em *transformers*, permitindo situar os resultados do comitê em relação a um paradigma amplamente consolidado na literatura para detecção de toxicidade (Leite et al., 2020; Oliveira et al., 2024; Salles et al., 2025).

No que se refere ao *zero-shot*, os modelos Qwen2.5-7B e Granite3.3-8B apresentam desempenho semelhante entre si nos três *corpora* avaliados. Já o BERTimbau permanece sistematicamente abaixo das demais abordagens avaliadas, incluindo o *zero-shot*, com diferença mais pronunciada no HLPHSD, conforme indicado na Tabela 4, comportamento esperado em um cenário de domínio cruzado no qual o ajuste fino foi conduzido em um único *corpus* de domínio específico.

6.3.2. Classificador Semi-Supervisionado

Para o método semi-supervisionado, as melhores configurações por *corpus* foram selecionadas com base no desempenho obtido em cada conjunto, a partir das combinações descritas na Subseção 6.1. O algoritmo LGC com classificador SVM e 10% de dados pré-rotulados apresentou os melhores resultados nos *corpora* ToLD-BR e HLPHSD, enquanto a combinação entre LGC, classificador MLP e 30% de dados pré-rotulados obteve o melhor desempenho no *corpus* HateBR. Cumpre destacar que todas as melhores configurações partiram do algoritmo LGC, evidenciando que, no contexto de domínio cruzado e de identificação de toxicidade aqui investigado, a possibilidade de atualização iterativa dos rótulos pré-anotados se mostrou mais eficaz do que a manutenção fixa adotada pelo GFHF, ainda que os ganhos do LGC nos *corpora* HateBR e HLPHSD permaneçam aquém de patamares satisfatórios. Essas configurações foram adotadas como referência para os experimentos subsequentes.

Considerando essas configurações, observa-se que o classificador semi-supervisionado apresenta desempenho mais elevado no *corpus* ToLD-BR, com *F-measure* de 0,72 e níveis de concordância moderados. Nos *corpora* HateBR e HLPHSD, por sua vez, os valores de desempenho são inferiores, com redução tanto na *F-measure* quanto nos níveis de concordância. Em comparação aos demais classificadores avaliados, o semi-supervisionado mostra-se competitivo no ToLD-BR, com resultados próximos aos do BERTimbau, do *zero-shot* e das abordagens baseadas em LLMs com exemplos rotulados. Nos *corpora* HateBR e HLPHSD, contudo, fica sistematicamente abaixo de todos eles. Esse comportamento ocorre em cenários de mudança de domínio entre o conjunto de treinamento e os dados avaliados, condição associada à redução da concordância com as anotações humanas nesses *corpora*. Enquanto no ToLD-BR os valores de Kappa indicam concordância moderada, nos demais *corpora* o nível de concordância varia de fraco a razoável, evidenciando limitações do método na reprodução dos padrões de rotulagem em cenários de domínio cruzado.

6.3.3. Classificadores baseados em LLMs: few-shot e RAG

As abordagens *few-shot* e RAG apresentam comportamentos próximos entre si nos três *corpora* avaliados quando se considera o desempenho agregado dos dois métodos, com ganhos discretos em relação à linha de base *zero-shot* e desempenho consistentemente superior ao do BERTimbau com ajuste fino, em particular nos *corpora* HateBR e HLPHSD. No ToLD-BR, ambos os métodos apresentam desempenho equilibrado, com concordância moderada em relação aos rótulos humanos. No HateBR, observa-se o melhor desempenho global do estudo, alcançado pelo *few-shot* com o modelo Qwen2.5-7B, com concordância substancial. No HLPHSD, o desempenho de ambas as abordagens é mais modesto, sugerindo que a incorporação de exemplos no *prompt* é menos eficaz nesse domínio.

No que se refere às nuances entre os modelos, o Qwen2.5-7B apresenta melhor desempenho na configuração *few-shot* do que na configuração RAG nos três *corpora* avaliados. Em contrapartida, com o Granite3.3-8B, observa-se o padrão inverso nos *corpora* HateBR e HLPHSD, em que a configuração RAG supera o *few-shot* com o mesmo modelo, ainda que, no ToLD-BR, o *few-shot* permaneça superior. Esse resultado evidencia que as diferenças entre as duas abordagens não são uniformes entre os modelos e os *corpora* avaliados. De modo geral, as diferenças entre *few-shot* e RAG permanecem marginais, indicando que ambas as estratégias produzem inferências semelhantes quando aplicadas ao mesmo modelo e ao mesmo *corpus*.

6.4. Desempenho do Comitê de Classificadores

A avaliação do comitê de classificadores no contexto da **QP2** baseia-se na comparação entre o desempenho obtido por votação majoritária e o do melhor classificador individual em cada *corpus*, conforme apresentado na Tabela 5 para os *corpora* em suas versões originais.

De maneira geral, os resultados evidenciam comportamentos distintos do comitê nos diferentes *corpora* avaliados. Conforme apresentado na Tabela 5, no *corpus* ToLD-BR, observa-se um ganho de desempenho de 2% na *F-measure* para a classe tóxica, indicando que a combinação dos métodos por votação majoritária superou o melhor resultado individual. Por sua vez, o coeficiente Kappa de Cohen de 0,51 indica um nível moderado de concordância entre as predições do comitê e as anotações humanas, o que é compatível com os valores observados para os métodos individuais nesse *corpus*.

Em contrapartida, nos *corpora* HateBR e HLPHSD, o comitê não supera o método *few-shot* no cenário original, apresentando reduções marginais de *F-measure* de -2% e -1%, respectivamente (ver Tabela 5). Esse comportamento pode ser explicado pela presença de um método individual dominante, o *few-shot*, cujo desempenho superior não é compensado pela votação majoritária quando os demais métodos divergem sistematicamente de suas predições. Em cenários em que há forte disparidade de desempenho entre os métodos do comitê, a votação majoritária tende a introduzir ruído proveniente dos classificadores mais fracos, resultando em uma degradação marginal do resultado agregado (Zhou, 2012).

Em relação à **QP2**, os resultados indicam que o comitê supera o melhor método individual de forma consistente apenas no *corpus* ToLD-BR, cenário caracterizado por maior equilíbrio entre os resultados dos métodos individuais e por melhor desempenho do comitê em *F-measure* e Kappa de Cohen. Nos demais *corpora*, o comitê mantém um patamar competitivo sem sofrer degradação significativa, evidenciando robustez suficiente para operar em diferentes domínios, ainda que a estratégia de votação majoritária se mostre mais eficaz quando as métricas dos classificadores são próximas entre si. A fim de ampliar a perspectiva avaliativa, as métricas complementares de acurácia, precisão, cobertura da classe tóxica e *macro-F1*, referentes às melhores configurações dos métodos constituintes, bem como às do comitê, são apresentadas no Apêndice D (Tabela 13). Essa estratégia permite uma avaliação mais abrangente do comportamento das abordagens, sem comprometer a clareza e a objetividade da Subseção de resultados.

6.5. Impacto da Concordância Total na Anotação Automática

A avaliação do impacto da concordância total entre anotadores sobre a anotação automática, em resposta à **QP3**, baseia-se na comparação dos resultados obtidos pelos métodos nos *corpora* em suas versões originais e nos subconjuntos compostos exclusivamente por instâncias com consenso absoluto entre os avaliadores humanos. Para essa análise, adotaram-se a *F-measure* e o coeficiente Kappa de Cohen como indicadores de desempenho e de confiabilidade dos rótulos gerados.

Em razão do acentuado desbalanceamento entre as classes nos *corpora* filtrados, conforme indicado na Tabela 2, o conjunto com o maior número de instâncias foi submetido a uma subamostragem balanceada para igualar o número de exemplos tóxicos e não tóxicos. Mantêm-se, portanto, as mesmas configurações experimentais adotadas anteriormente, diferenciando-se apenas pela utilização desses subconjuntos, o que permite analisar com maior precisão o efeito das divergências entre as anotações sobre os resultados. A Tabela 6 apresenta os resultados obtidos por cada método nos *corpora* originais e filtrados, bem como as variações relativas decorrentes da filtragem. Para cada método e *corpus*, os resultados do cenário original são adotados como referência.

Corpus	Método	Configuração	F-measure	(±%)	Kappa	(±%)
ToLD-BR	*few-shot*	Granite3.3-8b	0.74	ref.	0.48	ref.
	Comitê		**0.76**	**+2%**	**0.51**	**+3%**
HateBR	*few-shot*	Qwen2.5-7b	**0.87**	ref.	**0.73**	ref.
	Comitê		0.85	-2%	0.72	-1%
HLPHSD	*few-shot*	Qwen2.5-7b	**0.63**	ref.	0.41	ref.
	Comitê		0.62	-1%	**0.43**	**+3%**

Tabela 5: Comparação entre o melhor método individual e o comitê por *corpus*, em *F-measure* e Kappa de Cohen, com variação percentual (±%) em relação ao método individual.

		Original (ref.)		**Filtrado**		**Variação**	
Corpus	**Método**	**F**	κ	**F**	κ	**±%F**	**±%κ**
ToLD-BR	BERT	0,72	0,45	0,80	0,54	+8%	+9%
	zero-shot	0,73	0,47	0,80	0,55	+7%	+8%
	semi	0,72	0,46	0,79	0,56	+7%	+10%
	few-shot	0,74	0,48	0,81	0,56	+7%	+8%
	RAG	0,74	0,45	0,80	0,50	+6%	+5%
	Comitê	0,76	0,51	**0,82**	**0,58**	+6%	+7%
HateBR	BERT	0,83	0,66	0,86	0,74	+3%	+8%
	zero-shot	0,86	0,71	0,88	0,74	+2%	+3%
	semi	0,37	0,22	0,44	0,27	+7%	+5%
	few-shot	0,87	0,73	**0,89**	0,78	+2%	+5%
	RAG	0,85	0,70	0,88	0,75	+3%	+5%
	Comitê	0,85	0,72	**0,89**	**0,79**	+4%	+7%
HLPHSD	BERT	0,56	0,37	0,73	0,54	+17%	+17%
	zero-shot	0,62	0,38	0,82	0,61	+20%	+23%
	semi	0,41	0,06	0,38	0,23	−3%	+17%
	few-shot	0,63	0,41	**0,83**	**0,64**	+20%	+23%
	RAG	0,61	0,39	0,82	0,63	+21%	+24%
	Comitê	0,62	0,43	0,82	0,61	+20%	+18%

Tabela 6: *F-measure* (F) e Kappa de Cohen (κ) nos cenários original (referência) e filtrado por concordância total, com variação percentual (±%).

De forma consistente, a filtragem por concordância total beneficia a maioria dos métodos avaliados nos três *corpora*, incluindo tanto as linhas de base *zero-shot* e BERTimbau com ajuste fino quanto os classificadores constituintes do comitê, representados pelo método semi-supervisionado baseado em grafos heterogêneos e pelas abordagens baseadas em LLMs com exemplos rotulados. Como esse padrão de melhoria é observado em métodos fundamentados em paradigmas distintos, os resultados sugerem que os ganhos obtidos estão associados à maior consistência das instâncias de referência, e não a características específicas de cada abordagem. Esse padrão também indica que parte substancial das classificações incorretas no cenário original estava relacionada a instâncias com divergência entre os anotadores, e não necessariamente a limitações intrínsecas dos métodos.

Quanto à intensidade desses ganhos, observam-se variações de modesta a expressiva conforme o *corpus* considerado. No HateBR, que já apresenta elevado nível de concordância interanotador (Kappa de Fleiss = 0,70), os métodos avaliados registram ganhos entre 2% e 7% em *F-measure*. No ToLD-BR, os ganhos mostram-se uniformes entre os métodos, situando-se entre 6% e 8%. No HLPHSD, que apresenta o menor índice de concordância interanotador (Kappa de Fleiss = 0,17 na tarefa binária), os métodos alcançam ganhos de até 21% em *F-measure* e de 24% em Kappa de Cohen.

No que se refere ao método semi-supervisionado, observam-se ganhos de desempenho nos *corpora* ToLD-BR e HateBR após a filtragem. No ToLD-BR, esses ganhos são mais evidentes, enquanto, no HateBR, o método permanece aquém das abordagens baseadas em LLMs. No *corpus* HLPHSD, por sua vez, o método registra uma leve queda na

F-measure (-3%), apesar de um discreto ganho no coeficiente Kappa de Cohen, que mantém um nível razoável de concordância com as anotações humanas. Este desempenho do método semi-supervisionado nos três *corpora* reflete um cenário que combina mudança de domínio e redução do volume de dados disponíveis após a filtragem e a subamostragem balanceada. Essas condições relacionam-se à sensibilidade do método à estrutura do grafo heterogêneo, que depende da coocorrência de termos entre instâncias rotuladas e não rotuladas. Nesse contexto, a redução do conjunto de dados e as diferenças de domínio tendem a alterar esses padrões de coocorrência, o que se reflete em variações na estrutura do grafo em cenários de menor desempenho do método.

De maneira geral, esses resultados indicam que a filtragem por concordância total eleva o desempenho dos métodos individuais a um patamar já bastante consolidado, no qual os ganhos adicionais em *F-measure* obtidos pela agregação por votação majoritária tornam-se naturalmente mais discretos. Cumpre destacar, contudo, que o comitê apresenta o maior Kappa de Cohen no cenário filtrado nos *corpora* ToLD-BR e HateBR, mantendo-se em patamar competitivo no HLPHSD, evidenciando que a agregação contribui para um alinhamento consistente com as anotações humanas, mesmo em condições em que os métodos individuais já apresentam desempenho elevado.

No que se refere à **QP3**, os resultados indicam que todos os métodos avaliados, incluindo o comitê, apresentam desempenho consistentemente superior quando aplicados a instâncias com concordância total entre os anotadores, mantendo as mesmas configurações experimentais adotadas no cenário original. Ainda que as diferenças entre os *corpora* possam decorrer de múltiplos fatores, como protocolos de anotação, diretrizes adotadas e perfil dos anotadores, os ganhos observados após a filtragem evidenciam que a consistência das anotações de referência desempenha papel central na avaliação de métodos automáticos de identificação de linguagem tóxica, tarefa marcada por elevado grau de subjetividade. Sob as mesmas condições experimentais, subconjuntos compostos por instâncias com maior consenso entre anotadores associam-se a melhores níveis de desempenho em todos os paradigmas avaliados.

6.6. Análise Qualitativa dos Erros e Avaliação de Equidade

Esta Subseção apresenta uma análise qualitativa dos resultados obtidos nos dois cenários experimentais, com ênfase no desempenho do comitê. Adicionalmente, avalia-se o grau de viés associado ao conteúdo tóxico por meio de uma lista pré-definida de termos de identidade, adaptada de (Dixon et al., 2018; Kennedy et al., 2020), que foi traduzida e ampliada manualmente para contemplar as especificidades do contexto sociocultural brasileiro (ver Apêndice C). Para tanto, a Tabela 7 apresenta as matrizes de confusão do comitê para os *corpora* avaliados, permitindo identificar os padrões de erro em cada cenário. Nesse contexto, a distinção entre falsos positivos e falsos negativos assume um papel central. Falsos positivos correspondem à atribuição indevida do rótulo de toxicidade a textos não tóxicos, ou seja, à interpretação equivocada de ofensa em conteúdos legítimos; por sua vez, falsos negativos referem-se à não identificação de instâncias efetivamente tóxicas, permitindo que conteúdos potencialmente prejudiciais permaneçam sem rotulação adequada. Em aplicações de detecção automática de toxicidade, esse tipo de erro tende a ser particularmente crítico, pois permite que conteúdos ofensivos ou prejudiciais continuem circulando nas plataformas sem qualquer sinalização ou intervenção. A fim de analisar os padrões de erro, as matrizes de confusão apresentam, além das contagens absolutas, percentuais normalizados por linha (classe verdadeira), permitindo observar a proporção relativa de acertos e erros em cada classe.

De forma complementar, a Tabela 8 apresenta exemplos textuais extraídos dos *corpora* originais, acompanhados dos rótulos atribuídos pelos anotadores humanos, das predições individuais dos métodos que compõem o comitê e da decisão final agregada. Nessa tabela, o valor **1** corresponde à classe **tóxico**, enquanto o valor **0** corresponde à classe **não tóxico**. A análise desses exemplos contribui para uma compreensão mais aprofundada das fragilidades ainda presentes no comitê, particularmente em casos que envolvem toxicidade implícita, ambiguidade contextual ou uso irônico da linguagem.

No *corpus* ToLD-BR, observa-se que o comitê apresenta, no cenário original, um número expressivo de erros de classificação, com predominância de falsos positivos, que correspondem a 36,4% das instâncias não tóxicas classificadas incorretamente como tóxicas, enquanto os falsos negativos representam 11,3% das instâncias tóxicas não identificadas pelo modelo, con-

Classe Verdadeira	Classe Predita	
	Tóxico	**Não tóxico**
Tóxico	**8213 (88.7%)**	1042 (11.3%)
Não tóxico	4273 (36.4%)	**7472 (63.6%)**

(a) comitê – ToLD-BR (original)

Classe Verdadeira	Classe Predita	
	Tóxico	**Não tóxico**
Tóxico	**1380 (95.0%)**	73 (5.0%)
Não tóxico	554 (38.1%)	**899 (61.9%)**

(b) comitê – ToLD-BR (filtrado)

Classe Verdadeira	Classe Predita	
	Tóxico	**Não tóxico**
Tóxico	**2863 (81.8%)**	637 (18.2%)
Não tóxico	358 (11.1%)	**2863 (88.9%)**

(c) comitê – HateBR (original)

Classe Verdadeira	Classe Predita	
	Tóxico	**Não tóxico**
Tóxico	**2092 (87.8%)**	290 (12.2%)
Não tóxico	207 (8.7%)	**2175 (91.3%)**

(d) comitê – HateBR (filtrado)

Classe Verdadeira	Classe Predita	
	Tóxico	**Não tóxico**
Tóxico	**1220 (68.2%)**	568 (31.8%)
Não tóxico	903 (23.3%)	**2979 (76.7%)**

(e) comitê – HLPHSD (original)

Classe Verdadeira	Classe Predita	
	Tóxico	**Não tóxico**
Tóxico	**421 (83.0%)**	86 (17.0%)
Não tóxico	87 (17.2%)	**420 (82.8%)**

(f) comitê – HLPHSD (filtrado)

Tabela 7: Matrizes de confusão do comitê de classificadores para os *corpora* ToLD-BR, HateBR e HLPHSD, considerando os conjuntos em sua versão original e após a filtragem por concordância total entre anotadores. Os valores indicam contagens absolutas, com percentuais normalizados por linha (classe verdadeira) apresentados entre parênteses.

forme evidenciado na matriz de confusão da Tabela 7(a). Esse padrão indica maior sensibilidade dos classificadores do comitê à classe tóxica, em detrimento da especificidade na identificação de conteúdos não tóxicos. No cenário filtrado por concordância total (ver Tabela 7(b)), verifica-se uma redução substancial dos erros, com apenas 554 falsos positivos (38,1%) e 73 falsos negativos (5,0%), evidenciando o impacto positivo da remoção de instâncias com divergência anotativa. Ainda assim, a persistência de um número elevado de falsos positivos sugere que a dificuldade não decorre exclusivamente das divergências interpretativas na anotação, mas também da identificação inadequada de termos ofensivos empregados em contextos não agressivos, como em situações de autodepreciação ou humor. Nesses casos, os classificadores tendem a priorizar o sinal lexical da palavra ofensiva, em detrimento da interpretação pragmática do texto. Tal situação pode ser observada no exemplo iv) da Tabela 8, "*nooo mano kkk eu sou muito burro to bolado*", em que o termo "burro" conduz todos os métodos que compõem o comitê a uma interpretação literal de toxicidade. Por outro lado, os falsos negativos tendem a ocorrer em sentenças nas quais a ofensa é veiculada de forma implícita ou metafórica, exigindo inferência pragmática para a correta identificação da toxicidade, como no exemplo ii) "*eu pensava que era uma vaca a burra meu pai*", em que o insulto é direcionado a um terceiro sem o uso de marcadores convencionais de discurso ofensivo.

Em contraste com o padrão observado no *corpus* ToLD-BR, em que os falsos positivos predominam, o *corpus* HateBR apresenta uma inversão na distribuição dos erros. Conforme a matriz de confusão do comitê (Tabela 7(c)), observa-se maior incidência de falsos negativos (637), o que corresponde a 18,2% das instâncias tóxicas não identificadas pelo comitê. Ainda assim, a proporção geral de erros é inferior à observada no ToLD-BR, o que indica que o comitê identifica com maior clareza a manifestação da toxicidade no HateBR. Quando consideradas apenas as instâncias com concordância total entre os anotadores (Tabela 7(d)), observa-se uma redução consistente em ambos os tipos de erro, com 12,2% de falsos negativos e 8,7% de falsos positivos, reforçando o efeito positivo da remoção de instâncias com divergência nos rótulos. No plano qualitativo, os exemplos (vi) e (vii) da Tabela 8 ilustram erros associados ao emprego de ironia e sarcasmo, em que textos rotulados como não tóxicos são interpretados pelo comitê como ataques ou desqualificações, o que contribui para a ocorrência de falsos positivos. Já os exemplos viii) e x) evidenciam a situação inversa, em que a toxicidade é expressa de forma indireta e depende de conhecimento contextual, o que favorece falsos negativos quando o teor ofensivo não se manifesta de forma lexicalmente explícita.

Corpus	Exemplo textual	Rótulo original	semi	few	RAG	Comitê
ToLD-BR	i) sou a favor de invadir o stf e não chegarmos lá de mãos vazias....	1	0	1	1	1
	ii) eu pensava que era uma vaca a burra meu pai	1	1	0	0	0
	iii) gente eu não sou feia parem de me tratar assim	0	0	0	1	0
	iv) nooo mano kkk eu sou muito burro to bolado	0	1	1	1	1
	v) eu também tiro foto com aviados e simpatizo mas isso não quer dizer nada se eu pudesse ainda metia uma surra	1	0	1	1	1
HateBR	vi) jesus é cada uma desse cidadão é de assustar	0	0	1	1	1
	vii) é sério que estão comemorando a invasão dos nazistas caralho decepção	0	1	1	1	1
	viii) mais verdadeira que nota de 3 reais	1	0	1	0	0
	ix) hipócrita	1	0	1	1	1
	x) faz um programa com o lula. sugestão para o nome: sol quadrado	1	0	0	0	0
HLPHSD	xi) só matando amiga	1	0	1	1	1
	xii) falou populista já sei que é da turma fake news e retardado	0	1	1	1	1
	xiii) tem mão de homem ou seja com calos	1	1	0	0	0
	xiv) os "refugiados" e a classe política traidora	0	1	1	1	1
	xv) abri a janela tava uma gorda no apartamento da frente olhando pra baixo e chorando acho que caiu a empadinha dela	1	0	0	0	0

Tabela 8: Exemplos textuais com rótulos humanos, predições dos métodos individuais (semi, few, RAG) e decisão do comitê. Os valores 1 e 0 indicam classe tóxico e não tóxico, respectivamente.

No *corpus* HLPHSD, o padrão de erro no cenário original revela predominância de falsos positivos (903) em relação aos falsos negativos (568), conforme indicado na Tabela 7(e). Ao se considerar o subconjunto filtrado por concordância total entre anotadores (Tabela 7(f)), observa-se uma redução expressiva e praticamente equilibrada dos erros, com 86 falsos negativos e 87 falsos positivos, sugerindo que parcela significativa das classificações incorretas no cenário original pode estar relacionada a instâncias com divergência interanotador. Em termos qualitativos, o exemplo xiii) "*tem mão de homem ou seja com calos*" ilustra um caso de toxicidade sutil, na medida em que pressupõe que a masculinidade estaria condicionada à presença de determinados traços físicos, como calos nas mãos, implicando que a ausência desses traços descaracterizaria o sujeito como homem. Por outro lado, os exemplos xii) e xiv) indicam possíveis inconsistências na rotulagem manual. No primeiro caso, verifica-se uma ofensa direta, na medida em que o termo pejorativo retardado" é empregado para qualificar o interlocutor no contexto do debate em que o texto foi coletado, o que justificaria sua classificação como tóxico. Já no segundo, a associação do substantivo entre aspas "refugiados" à figura de traidor pode igualmente configurar um ataque ou uma desqualificação implícita. Em ambos os casos, o comitê classificou os textos como tóxicos, sugerindo que parte dos erros atribuídos ao sistema reflete divergências na rotulagem realizada pelos anotadores.

Para além da análise binária dos textos tóxicos, é importante observar que determinados termos ou expressões podem representar formas específicas de toxicidade direcionadas a indivíduos ou grupos com base em atributos como aparência física, orientação sexual ou raça, o que pode introduzir desafios adicionais para os classificadores na identificação desse fenômeno. No exemplo xv), "*abri a janela tava uma gorda no apartamento da frente olhando pra baixo e chorando acho que caiu a empadinha dela*", a ofensa dirige-se à aparência física, configurando um caso de gordofobia. Mesmo quando a intenção depreciativa pode ser inferida do contexto, os classificadores não conseguem associar o termo "gorda" a essa manifestação específica de toxicidade, classificando o texto como não tóxico. De maneira semelhante, o exemplo v) "*eu também tiro foto com aviados e simpatizo mas isso não quer dizer nada se eu pudesse ainda metia uma surra*" apresenta uma variação lexical ("aviados") potencialmente empregada para camuflar a ofensa. Nesse caso, os classificadores baseados em LLMs conseguem interpretar o contexto discursivo e inferir

que a toxicidade está associada a um ataque de natureza homofóbica, enquanto o método semi-supervisionado falha nessa detecção. Termos com baixa conectividade no grafo heterogêneo, como a variação lexical "aviados", dispõem de menos evidências para a propagação de rótulos, conforme já discutido na Subseção 6.5.

Para complementar a análise qualitativa com uma avaliação quantitativa de disparidades, empregaram-se métricas de avaliação de equidade (Borkan et al., 2019), combinando três métricas baseadas na área sob a curva característica de operação, conhecida como AUC (do inglês *Area Under the Curve*): (i) *Subgroup AUC*, que mensura a capacidade do modelo de discriminar entre tóxico e não tóxico dentro do conjunto de instâncias identitárias; (ii) *Background Positive, Subgroup Negative* (BPSN), que avalia a sensibilidade a falsos positivos no grupo, isto é, textos não tóxicos do grupo classificados como tóxicos por associação com marcadores identitários; e (iii) *Background Negative, Subgroup Positive* (BNSP), que avalia a sensibilidade a falsos negativos, ou seja, textos tóxicos do grupo não identificados pelo modelo. Em todas essas métricas, valores mais próximos de 1 indicam menor disparidade. A análise foi conduzida de forma agregada, considerando todas as instâncias que contêm ao menos um termo da lista identitária como um único grupo (ver Apêndice C), em coerência com o tratamento binário de toxicidade adotado em todo o trabalho.

Corpus	Método	n	Sub. AUC	BPSN	BNSP
ToLD-BR	semi	1560	0,67	0,66	0,76
	few-shot	1560	0,68	0,69	0,75
	RAG	1560	0,68	0,68	0,73
HateBR	semi	570	0,62	0,61	0,64
	few-shot	570	0,79	0,77	0,88
	RAG	570	0,78	0,77	0,87
HLPHSD	semi	1251	0,63	0,51	0,68
	few-shot	1251	0,70	0,68	0,75
	RAG	1251	0,68	0,67	0,73

Tabela 9: Resultados das métricas de equidade (*Subgroup AUC*, BPSN e BNSP) para as instâncias contendo termos identitários nos *corpora* avaliados. Valores próximos de 1 indicam maior equidade e ausência de viés entre instâncias identitárias e não identitárias.

A análise das métricas apresentadas na Tabela 9 revela comportamentos distintos entre os classificadores do comitê e os *corpora* avaliados. No ToLD-BR, os três classificadores apresentam métricas uniformes, com *Subgroup AUC* em torno de 0,68, BPSN entre 0,66 e 0,69 e BNSP entre 0,73 e 0,76. Os valores de BNSP, ligeiramente superiores aos de BPSN, indicam maior estabilidade dos métodos na identificação de toxicidade efetiva em instâncias com termos identitários do que na distinção entre textos não tóxicos do grupo e textos tóxicos do restante do *corpus*. Nos *corpora* HateBR e HLPHSD, esse padrão uniforme dá lugar a uma discrepância acentuada entre os classificadores. Os métodos baseados em LLMs se mantêm em patamares elevados, com *Subgroup AUC* entre 0,68 e 0,79, BPSN entre 0,67 e 0,77 e BNSP entre 0,73 e 0,88. O valor elevado de BNSP sugere baixa propensão a falsos negativos em instâncias identitárias. O método semi-supervisionado, em contrapartida, registra valores substancialmente inferiores, com destaque para o BPSN de 0,51 no HLPHSD, próximo ao desempenho aleatório. Esse desequilíbrio entre as métricas BPSN e BNSP no método semi-supervisionado revela que o classificador aprende a associação entre os marcadores de identidade e os rótulos tóxicos no *corpus* de treinamento, configurando um caso típico de correlação espúria, em vez de aprender propriedades semânticas da toxicidade em si.

De modo geral, as análises realizadas nesta Subseção indicam que o comitê de classificadores enfrenta dificuldades na identificação de toxicidade implícita, como ironia, sarcasmo e ofensas dependentes de contexto, que frequentemente demandam conhecimento adicional para serem corretamente interpretadas. A presença de termos comumente associados a contextos ofensivos (e.g., "burro", "nazista") contribui para o aumento da taxa de falsos positivos, na medida em que os métodos tendem a supervalorizar termos isolados em detrimento do significado contextual. Ademais, as métricas de equidade mostram que vieses presentes nos dados de treinamento podem se propagar aos rótulos gerados automaticamente. Esse achado é particularmente relevante no contexto deste trabalho, cuja contribuição central reside em uma metodologia para anotação automática de *corpora* de linguagem tóxica, uma vez que potenciais vieses identificados na fase de anotação podem se propagar aos recursos linguísticos produzidos por meio dessa estratégia e às aplicações que venham a utilizá-los. A análise de equidade conduzida nesta Subseção contribui justamente para mapear esse problema, fornecendo subsídios para o uso responsável da abordagem proposta na construção de *corpora* anotados. A mitigação dessas disparidades por meio de estratégias de *debiasing*, balanceamento de instâncias identitárias no *corpus* de treinamento e análise de sensibilidade a *prompts* constitui uma direção prioritária para trabalhos futuros (Navigli et al., 2023; Dixon et al., 2018).

6.7. Concordância entre os Métodos do Comitê

Com o objetivo de aprofundar a avaliação do comitê de classificadores, esta seção analisa o grau de concordância entre os métodos constituintes com base em suas classificações nos *corpora* avaliados. A Tabela 10 apresenta os valores do coeficiente Kappa de Cohen calculados para cada par de métodos, considerando os cenários original e filtrado por concordância total. Essa métrica permite quantificar o grau de alinhamento entre as decisões dos classificadores e fornece uma estimativa mais robusta da similaridade entre suas predições, complementando a avaliação de desempenho apresentada nas Subseções anteriores.

Cenário	*Corpus*	semi vs *few*	semi vs RAG	*few* vs RAG
Original	ToLD-BR	0.66	0.63	0.72
	HateBR	0.24	0.24	0.84
	HLPHSD	0.20	0.20	0.76
Filtrado	ToLD-BR	0.69	0.61	0.82
	HateBR	0.24	0.23	0.77
	HLPHSD	0.24	0.21	0.82

Tabela 10: Concordância entre os métodos constituintes do comitê, mensurada pelo coeficiente Kappa de Cohen para cada par de abordagens, nos cenários original e filtrado por concordância total.

Considerando os corpora em sua versão original, no ToLD-BR observa-se concordância substancial entre todos os pares de métodos (Kappa entre 0,63 e 0,72), padrão consistente com o desempenho equilibrado relatado na Subseção 6.2, o que ajuda a explicar os ganhos do comitê nesse corpus por meio de votação majoritária.

Nos *corpora* HateBR e HLPHSD, as diferenças de desempenho entre os métodos tornam-se mais acentuadas, acompanhadas de uma mudança nos valores de concordância entre os pares de classificadores. A concordância entre os métodos baseados em LLMs permanece elevada, com Kappa de 0,84 no HateBR e de 0,76 no HLPHSD, indicando forte alinhamento entre os rótulos gerados por essas abordagens. Por sua vez, a concordância no método semi-supervisionado é consideravelmente inferior, situando-se entre 0,20 e 0,24 em ambos os *corpora*. Esse contraste reforça que a métrica de concordância entre os métodos acompanha o comportamento de desempenho observado nesses *corpora*. As abordagens baseadas em LLMs apresentam maior alinhamento entre os rótulos gerados e os melhores resultados em *F-measure*, enquanto o método semi-supervisionado, cujos rótulos divergem dos demais, apresenta também o desempenho mais baixo. Essa observação indica que a votação majoritária tende a refletir o alinhamento do subgrupo mais coeso, que, nesses *corpora*, corresponde às abordagens baseadas em LLMs, com os ganhos do comitê limitados pela divergência nos rótulos do método semi-supervisionado.

No cenário filtrado por concordância total, as observações permanecem, em linhas gerais, consistentes com as do cenário original. A concordância entre o método semi-supervisionado e as abordagens baseadas em LLMs mantém-se substancial apenas no ToLD-BR, permanecendo fraca nos demais *corpora*. Um aspecto relevante, contudo, é o aumento no alinhamento entre os métodos baseados em LLMs após a filtragem. No ToLD-BR, o coeficiente Kappa entre *few-shot* e RAG evoluiu de 0,72 para 0,82, indicando concordância quase perfeita. Uma tendência semelhante é observada no HLPHSD, em que o valor passa de 0,76 para 0,82.

Quanto à **QP4**, os resultados indicam que o grau de concordância entre os classificadores do comitê está positivamente associado aos ganhos obtidos pela estratégia de votação majoritária. No ToLD-BR, em que se observa concordância substancial entre todos os pares de métodos, o comitê obteve ganhos consistentes em relação ao melhor método individual. Nos *corpora* HateBR e HLPHSD, a forte divergência entre o método semi-supervisionado e as abordagens baseadas em LLMs limitou os benefícios da agregação, resultando em desempenho equiparável ou marginalmente inferior ao do melhor classificador isolado. Em conjunto, esses achados sugerem que a eficácia da votação majoritária depende não apenas da qualidade individual dos métodos combinados, mas também do grau de alinhamento entre os rótulos por eles gerados. Nesse sentido, a investigação de estratégias de agregação adaptativas, como votação ponderada ou *stacking*, que incorporem o grau de concordância entre os classificadores constituintes como critério de ponderação, representa um desdobramento natural desta investigação.

7. Conclusão

Este trabalho investigou o uso de um comitê de classificadores para a anotação automática de linguagem tóxica em português, considerando cenários de escassez de dados rotulados e de mudança de domínio. A proposta integra três abordagens complementares funda-

mentadas em paradigmas distintos: um método semi-supervisionado baseado em grafos heterogêneos, uma estratégia de inferência *few-shot* com grandes modelos de linguagem e um método baseado em *Retrieval-Augmented Generation*. Essas abordagens foram combinadas por meio de votação majoritária e avaliadas em múltiplos *corpora*, abrangendo diferentes domínios e plataformas de coleta.

De forma consolidada, os resultados indicam que o comitê proposto constitui uma estratégia metodologicamente consistente e adequada para a anotação automática de linguagem tóxica. O comitê supera os classificadores de linha de base, *zero-shot* e BERTimbau com ajuste fino, na maioria dos cenários avaliados, e mantém desempenho competitivo mesmo quando, isoladamente, não supera o melhor método individual. Em cenários caracterizados por maior equilíbrio de desempenho e alinhamento entre os métodos constituintes, como observado no ToLD-BR, o comitê obtém ganhos consistentes, com melhoria de até 2% na *F-measure* e aumento dos coeficientes Kappa de Cohen. Em *corpora* caracterizados por maior divergência entre os métodos constituintes do comitê, como observado no HateBR e no HLPHSD, o desempenho do comitê mantém-se próximo ao do melhor método individual.

A análise comparativa entre os cenários experimentais, considerando os *corpora* em sua versão original e após a filtragem por concordância total, evidencia que o comitê apresenta o maior coeficiente Kappa de Cohen no cenário filtrado nos *corpora* ToLD-BR e HateBR, mantendo-se em patamar competitivo no HLPHSD, o que sugere que a votação majoritária se beneficia de decisões individuais mais alinhadas. Cumpre destacar que tais resultados foram obtidos a partir de um único *corpus* curado, composto por 1.400 instâncias, utilizado como base para treinamento e inferência, enquanto a avaliação do comitê foi conduzida em conjuntos substancialmente maiores, provenientes de diferentes domínios e plataformas. Esse delineamento experimental indica o potencial da abordagem para ampliar, de forma escalável, os recursos anotados, mesmo em contextos com limitada disponibilidade de dados rotulados.

Complementarmente, a análise qualitativa dos erros e a avaliação de equidade por grupo identitário evidenciaram que o comitê ainda enfrenta dificuldades na identificação de toxicidade implícita, como ironia, sarcasmo e ofensas dependentes de contexto, e que vieses presentes nos dados de treinamento podem se propagar aos rótulos gerados automaticamente. Esse achado é particularmente relevante no contexto deste trabalho, cuja contribuição central reside em uma metodologia para construção ou ampliação de *corpora* anotados, uma vez que dados com vieses comprometem não apenas a qualidade dos recursos produzidos, mas também a das aplicações que venham a utilizá-los.

Entre as limitações deste estudo, destacam-se a utilização de um único *corpus* curado como base de treinamento, o emprego de modelos multilíngues de pequeno porte, sem ajuste fino específico ao domínio, e a ausência de mecanismos de interpretabilidade que permitam explicar, de forma sistemática, as decisões individuais dos classificadores que compõem o comitê. Como trabalhos futuros desta investigação, pretende-se ampliar o treinamento do comitê para múltiplos *corpora*, diversificando os domínios textuais utilizados como base e explorando o uso de modelos de linguagem de maior porte, potencialmente mais aptos a apreender nuances semânticas e subjetivas. Adicionalmente, a incorporação de outras estratégias de agregação, como votação ponderada e *stacking*, bem como o desenvolvimento de mecanismos sensíveis ao contexto discursivo e de adaptação de domínio (Jiang & Zubiaga, 2023; Safikhani & Broneske, 2025), configuram uma direção promissora para o aprimoramento da robustez e da capacidade de generalização da metodologia proposta. No que concerne à avaliação de equidade, a investigação de estratégias de mitigação de vieses, o balanceamento de instâncias identitárias no *corpus* de treinamento e a análise de sensibilidade a *prompts* constituem direções prioritárias (Navigli et al., 2023; Dixon et al., 2018). Por fim, a incorporação de técnicas de interpretabilidade *post-hoc* (Ribeiro et al., 2016), com avaliação da fidelidade das explicações (Jacovi & Goldberg, 2020; Salles et al., 2025), permitirá explicitar os critérios subjacentes às decisões dos classificadores e contribuirá para a construção de *corpora* anotados com maior transparência metodológica.

Referências

Alsafari, Safa & Samira Sadaoui. 2021. Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence* 35(15). 1621–1645. 10.1080/08839514.2021.1988443

Aroyo, Lora, Lucas Dixon, Nithum Thain, Olivia Redfield & Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discus-

sions. Em *Companion Proceedings of The 2019 World Wide Web Conference*, 1100–1105. 10.1145/3308560.3317083

Assis, Gabriel, Annie Amorim, Jonnathan Carvalho, Mariza Ferro, Daniel de Oliveira, Daniela Vianna & Aline Paes. 2024. Explorando técnicas de aprendizado em modelos de linguagem para classificação de discurso de Ódio e ofensivo em Português. *Linguamática* 16(2). 91–113. 10.21814/lm.16.2.446

Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta & Vasudeva Varma. 2017. Deep learning for hate speech detection in Tweets. Em *26th International Conference on World Wide Web Companion*, 759–760. 10.1145/3041021.3054223

Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain & Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. Em *World Wide Web Conference (WWW)*, 491–500. 10.1145/3308560.3317593

Botella-Gil, Beatriz, Robiert Sepúlveda-Torres, Alba Bonet-Jover, Patricio Martínez-Barco & Estela Saquete. 2024. Semi-automatic dataset annotation applied to automatic violent message detection. *IEEE Access* 12. 19651–19664. 10.1109/ACCESS.2024.3361404

Brasil, Presidência da República. 1989. Lei nº 7.716, de 5 de janeiro de 1989. Acesso em: jul. 2025.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. Em *Advances in Neural Information Processing Systems (NeurIPS)*, 1877–1901.

Butler, Judith. 2021. *Discurso de ódio: Uma política do performativo*. São Paulo: Editora Unesp

Cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, Nithum & Will Cukierski. 2017. Toxic comment classification challenge. Kaggle.

Costa Bertaglia, Thales Felipe & Maria das Graças Volpe Nunes. 2016. Exploring word embeddings for unsupervised textual user-generated content normalization. Em *2nd Workshop on Noisy User-generated Text (WNUT)*, 112–120.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

Dirting, Bakwa Dunka, Gloria A. Chukwudebe, Euphemia Chioma Nwokorie & Ikechukwu Ignatius Ayogu. 2022. Multi-label classification of hate speech severity on social media using BERT model. Em *4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, 1–5. 10.1109/NIGERCON54645.2022.9803164

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain & Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. Em *AAAI/ACM Conference on AI, Ethics, and Society*, 67–73. 10.1145/3278721.3278729

Fortuna, Paula & Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Survey* 51(4). 1–30. 10.1145/3232676

Fortuna, Paula, João Rocha da Silva, Juan Soler-Company, Leo Wanner & Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. Em *3rd Workshop on Abusive Language Online*, 94–104. 10.18653/v1/W19-3510

Founta, Antigoni, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos & Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. Em *International AAAI Conference on Web and Social Media (ICWSM)*, 10.1609/icwsm.v12i1.14991

Frediani, João Otávio Rodrigues Ferreira, Gabriel Lino Garcia, Pedro Henrique Paiola, Leandro Aparecido Passos, João Paulo Papa & Aparecido Nilceu Marana. 2025. Hate speech detection in Portuguese using BERTimbau. Em *Progress in Pattern Recognition,*

Image Analysis, Computer Vision, and Applications, 244–255

Gilardi, Fabrizio, Meysam Alizadeh & Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120(30). 10.1073/pnas.2305016120

Hettiachchi, Danula, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora Salim & Mark Sanderson. 2023. How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets. Em *Conference on Human Computation and Crowdsourcing*, 38–50. 10.1609/hcomp.v11i1.27546

IBM Research. 2025. IBM Granite-3.3 language models GitHub repository. Apache-2.0 licensed repository for Granite 3.3 language models, including 8B and 2B variants. Accessed: 2026-02-18.

Jacovi, Alon & Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4198–4205. 10.18653/v1/2020.acl-main.386

Jahan, Md Saroar & Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing* 546. 126232. 10.1016/j.neucom.2023.126232

Jiang, Aiqi & Arkaitz Zubiaga. 2023. SexWEs: Domain-aware word embeddings via cross-lingual semantic specialisation for chinese sexism detection in social media. Em *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 447–458. 10.1609/icwsm.v17i1.22159

Kennedy, Brendan, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani & Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. Em *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5435–5442. 10.18653/v1/2020.acl-main.483

King, Ben, Rahul Jha & Dragomir R. Radev. 2014. Heterogeneous networks and their applications: Scientometrics, name disambiguation, and topic modeling. *Transactions of the Association for Computational Linguistics (TACL)* 2. 1–14. 10.1162/tacl_a_00161

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159–174. 10.2307/2529310

Lees, Alyssa, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler & Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. Em *28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3197–3207. 10.1145/3534678.3539147

Leite, João Augusto, Diego Silva, Kalina Bontcheva & Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. Em *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) and the 10th International Joint Conference on Natural Language Processing*, 914–924. 10.18653/v1/2020.aacl-main.91

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel & Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Em *34th International Conference on Neural Information Processing Systems (NIPS)*,

Mladenović, Miljana, Vera Ošmjanski & Staša Vujičić Stanković. 2021. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys* 54(1). 1–42. 10.1145/3424246

Nascimento, Gabriel, Flavio Carvalho, Alexandre Martins da Cunha, Carlos Roberto Viana & Gustavo Paiva Guedes. 2019. Hate speech detection using brazilian imageboards. Em *25th Brazillian Symposium on Multimedia and the Web*, 325–328. 10.1145/3323503.3360619

Navigli, Roberto, Simone Conia & Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality* 15(2). 1–21. 10.1145/3597307

Neto, Francisco R., Rafael Anchiêta, Raimundo Moura & André Santana. 2024. Abordagem semi-supervisionada para anotação de linguagem tóxica. Em *Anais do XIII Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*, 116–129. 10.5753/brasnam.2024.2965

Oliveira, Amanda, Thiago Cecote, Pedro Silva, Jadson Gertrudes, Vander Freitas & Eduardo

Luz. 2023. How good is ChatGPT for detecting hate speech in Portuguese? Em *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 103–112.

Oliveira, Amanda, Pedro H. Silva, Valéria Santos, Gladston Moreira, Vander L. Freitas & Eduardo J. Luz. 2024. Toxic text classification in Portuguese: Is LLaMA 3.1 8B all you need? Em *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 57–66.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pelle, Rogers, Cleber Alcântara & Viviane P. Moreira. 2018. A classifier ensemble for offensive text detection. Em *24th Brazilian Symposium on Multimedia and the Web*, 237–243. 10.1145/3243082.3243111

de Pelle, Rogers Prates & Viviane P Moreira. 2017. Offensive comments in the Brazilian Web: a dataset and baseline results. Em *Anais do VI Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*, 510–519. 10.5753/brasnam.2017.3260

Pelosi, Serena, Alessandro Maisto, Pierluigi Vitale & Simonetta Vietri. 2017. Mining offensive language on social media. Em *4th Italian Conference on Computational Linguistics (CLiC-it)*, 260–264.

Phanomtip, Aekachai, Thaiyathorn Sueb-in & Sirion Vittayakorn. 2021. Cyberbullying detection on Tweets. Em *18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 295–298. 10.1109/ECTI-CON51831.2021.9454848

Poletto, Fabio, Valerio Basile, Manuela Sanguinetti, Cristina Bosco & Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* 55. 477–523. 10.1007/s10579-020-09502-8

Ribeiro, Marco, Sameer Singh & Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 97–101. 10.18653/v1/N16-3020

Rossi, Rafael Geraldeli. 2015. *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Tese de Doutoramento.

Safikhani, Parisa & David Broneske. 2025. AutoML meets hugging face: Domain-aware pretrained model selection for text classification. Em *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 466–473. 10.18653/v1/2025.naacl-srw.45

Saifullah, Shoffan, Rafał Drężewski, Felix Andika Dwiyanto, Agus Sasmito Aribowo, Yuli Fauziah & Nur Heri Cahyana. 2024. Automated text annotation using a semi-supervised approach with meta vectorizer and machine learning algorithms for hate speech detection. *Applied Sciences* 14(3). 10.3390/app14031078

Salles, Isadora, Francielle Vargas & Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. Em *31st International Conference on Computational Linguistics (COLING)*, 6659–6669.

Santos, Raquel Bento, Bernardo Cunha Matos, Paula Carvalho, Fernando Batista & Ricardo Ribeiro. 2022. Semi-supervised annotation of portuguese hate speech across social media domains. Em *11th Symposium on Languages, Applications and Technologies (SLATE)*, 11:1–11:14. 10.4230/OASIcs.SLATE.2022.11

Schmidt, Anna & Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. Em *5th International Workshop on Natural Language Processing for Social Media*, 1–10. 10.18653/v1/W17-1101

da Silva Oliveira, Amanda, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas & Eduardo José da Silva Luz. 2024. Toxic speech detection in Portuguese: A comparative study of large language models. Em *15th International Conference on Computational Processing of Portuguese (PROPOR)*, 108–116.

Smith, Noah A. 2020. Contextual word representations: putting words into computers. *Communications of the ACM* 63(6). 66–74. 10.1145/3347145

Soto, Claver, Gustavo Nunes & José Gomes. 2019. Avaliação de técnicas de word embedding na tarefa de detecção de discurso de ódio. Em *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 1020–1031. 10.5753/eniac.2019.9354

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28

Suryawanshi, Shardul, Mihael Arcan & Paul Buitelaar. 2020. NUIG at SemEval-2020 task 12: Pseudo labelling for offensive content classification. Em *14th Workshop on Semantic Evaluation*, 1598–1604. 10.18653/v1/2020.semeval-1.208

Szcz Aleksander, esny, Maciej Markiewicz, Łukasz Radliński & Przemysław Kazienko. 2025. Leveraging positional bias of LLM in-context learning with class-few-shot and MajMin alternating ordering. Em *Computational Science – ICCS 2025*, 54–62

Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng & Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 930–957. 10.18653/v1/2024.emnlp-main.54

Trajano, Douglas, Rafael H Bordini & Renata Vieira. 2024. OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation* 58. 1263–1289. 10.1007/s10579-023-09657-0

Vargas, Francielle, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo & Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em *13th Language Resources and Evaluation Conference (LREC)*, 7174–7183.

Vargas, Francielle, Fabiana Goés, Isabelle Carvalho, Frabrício Benevenuto & Thiago A S Pardo. 2021. Contextual-lexicon approach for abusive language detection. Em *International Conference on Recents Advances in Natural Language Processing (RANLP)*, 1438–1447.

Wang, Shuohang, Yang Liu, Yichong Xu, Chenguang Zhu & Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. Em *Findings of the Association for Computational Linguistics: EMNLP*, 4195–4205. 10.18653/v1/2021.findings-emnlp.354

Waseem, Zeerak & Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. Em *NAACL Student Research Workshop*, 88–93. 10.18653/v1/N16-2013

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

X. 2026. The x rules. Acesso em: fev. 2026.

Yang, An, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang & Zihan Qiu. 2025. Qwen2.5 technical report. ArXiv [cs.CL]. 10.48550/arXiv.2412.15115

Youtube. 2026. Hate speech policy. Acesso em: fev. 2026.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra & Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 1415–1420. 10.18653/v1/N19-1144

Zhang, Chuxu, Dongjin Song, Chao Huang, Ananthram Swami & Nitesh V. Chawla. 2019. Heterogeneous graph neural network. Em *25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 793–803. 10.1145/3292500.3330961

Zhou, Dengyong, Olivier Bousquet, Thomas Navin Lal, Jason Weston & Bernhard Schölkopf.

2004. Learning with local and global consistency. Em *Advances in Neural Information Processing Systems (NeurIPS)*, 321–328.

Zhou, Zhi-Hua. 2012. *Ensemble methods: Foundations and algorithms*. CRC Press

Zhu, Xiaojin, Zoubin Ghahramani & John Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. Em *20th International Conference on International Conference on Machine Learning*, 912–919.

Apêndices

A. Resultados detalhados do método semi-supervisionado

Corpus	Regularização	% Dados	Classificador	F-measure
ToLD-BR	GFHF	30%	MLP	0.64
	LGC	10%	SVM	**0.72**
HateBR	GFHF	30%	MLP	0.36
	LGC	30%	MLP	**0.37**
HLPHSD	GFHF	30%	MLP	0.26
	LGC	10%	SVM	**0.41**

Tabela 11: Melhores desempenhos do método semi-supervisionado por *corpus* e algoritmo de regularização, em termos de *F-measure*. Para cada caso, apresenta-se a configuração que obteve o melhor resultado.

B. Resultados detalhados dos classificadores baseados em LLMs

			Original		Filtrado	
Corpus	Método	Modelo	F	κ	F	κ
ToLD-BR	*zero-shot*	Qwen2.5-7B	0,73	0,41	0,74	0,44
		Granite3.3-8B	0,73	0,47	0,80	0,55
	few-shot	Qwen2.5-7B	0,74	0,44	0,79	0,49
		Granite3.3-8B	0,74	0,48	0,81	0,56
	RAG	Qwen2.5-7B	0,74	0,45	0,78	0,45
		Granite3.3-8B	0,73	0,47	0,80	0,50
HateBR	*zero-shot*	Qwen2.5-7B	0,86	0,71	0,87	0,73
		Granite3.3-8B	0,75	0,56	0,82	0,64
	few-shot	Qwen2.5-7B	0,87	0,73	0,89	0,78
		Granite3.3-8B	0,80	0,64	0,86	0,74
	RAG	Qwen2.5-7B	0,85	0,70	0,87	0,73
		Granite3.3-8B	0,83	0,67	0,88	0,75
HLPHSD	*zero-shot*	Qwen2.5-7B	0,62	0,38	0,82	0,61
		Granite3.3-8B	0,58	0,40	0,70	0,40
	few-shot	Qwen2.5-7B	0,63	0,41	0,83	0,64
		Granite3.3-8B	0,59	0,40	0,80	0,62
	RAG	Qwen2.5-7B	0,61	0,39	0,82	0,63
		Granite3.3-8B	0,60	0,39	0,81	0,61

Tabela 12: Desempenho dos classificadores baseados em LLMs com os modelos Qwen2.5-7B e Granite3.3-8B nos *corpora* avaliados, em termos de *F-measure* (F) e coeficiente Kappa de Cohen (κ), nos cenários dos *corpora* originais e filtrados.

C. Lista de Termos Identitários

A lista de termos utilizada para identificar instâncias com conteúdo identitário foi construída por tradução e adaptação para o português brasileiro da lista proposta por Dixon et al. (2018); Kennedy et al. (2020), com adição de termos relacionados ao contexto sociopolítico brasileiro, incluindo jargões associados ao debate político nacional do período de coleta dos *corpora*. Os termos estão organizados por categoria para facilitar a leitura, mas foram tratados de forma agregada nos experimentos, isto é, uma instância é considerada identitária se contém ao menos um termo de qualquer categoria.

Política petralha, petista, mortadela, bolsominion, bozo, bolsonarista, esquerdopata, fascista, nazista, comunista, gado, lulista, esquerda, direita, tucano

Gênero e Sexualidade mulher, lésbica, gay, bissexual, transgênero, trans, queer, lgbt, homossexual, bicha, viado, sapatão, travesti, fêmea, noiva, esposa, mãe

Raça e Etnia negro, preto, pardo, branco, africano, índio, afro-americano, latino, asiático, indígena, cigano, macaco, senzala, escravo

Religião muçulmano, judeu, islã, islâmico, alá, judaico, cristão, católico, evangélico, clero, maomé, religião

Deficiência e Idade cego, surdo, mudo, paralisado, deficiente, autista, retardado, velho, idoso, jovem, adolescente

D. Métricas complementares de desempenho

			Acurácia		Precisão		Cobertura		F-measure		Macro F-measure	
Corpus	**Método**	**Configuração**	**Orig.**	**Filt.**	**Orig.**	**Filt.**	**Orig.**	**Filt.**	**Orig.**	**Filt.**	**Orig.**	**Filt.**
ToLD-BR	BERT	BERTimbau FT	0.72	0.77	0.64	0.71	0.84	0.90	0.72	0.80	0.72	0.77
	zero-shot	Granite3.3-8B	0.73	0.73	0.66	0.66	0.83	0.83	0.73	0.73	0.73	0.73
	semi	LGC_10%_SVM	0.73	0.78	0.66	0.76	0.81	0.81	0.72	0.79	0.73	0.78
	few-shot	Granite3.3-8B	0.74	0.78	0.66	0.72	0.83	0.93	0.74	0.81	0.74	0.78
	RAG	Granite3.3-8B / Qwen	0.73	0.75	0.64	0.68	0.85	0.96	0.74	0.80	0.73	0.74
	Comitê	-	0.75	0.79	0.66	0.72	0.89	0.95	0.76	0.82	0.75	0.78
HateBR	BERT	BERTimbau FT	0.83	0.87	0.86	0.90	0.79	0.83	0.83	0.86	0.83	0.87
	zero-shot	Qwen2.5-7B	0.86	0.87	0.84	0.83	0.88	0.94	0.86	0.88	0.85	0.87
	semi	LGC_30%_MLP	0.61	0.64	0.97	0.96	0.23	0.28	0.37	0.44	0.54	0.59
	few-shot	Qwen2.5-7B	0.87	0.89	0.86	0.87	0.87	0.92	0.87	0.89	0.87	0.89
	RAG	Granite3.3-8B / Qwen	0.83	0.87	0.86	0.84	0.80	0.92	0.83	0.88	0.84	0.87
	Comitê	-	0.86	0.90	0.88	0.91	0.82	0.88	0.85	0.89	0.86	0.90
HLPHSD	BERT	BERTimbau FT	0.73	0.77	0.58	0.89	0.55	0.62	0.56	0.73	0.68	0.76
	zero-shot	Qwen2.5-7B	0.70	0.81	0.51	0.77	0.79	0.87	0.62	0.82	0.68	0.80
	semi	LGC_10%_SVM	0.53	0.61	0.69	0.95	0.47	0.24	0.41	0.38	0.42	0.55
	few-shot	Qwen2.5-7B	0.72	0.82	0.54	0.81	0.76	0.85	0.63	0.83	0.70	0.82
	RAG	Qwen2.5-7B	0.72	0.81	0.54	0.78	0.69	0.87	0.61	0.82	0.70	0.81
	Comitê	-	0.74	0.83	0.58	0.83	0.68	0.83	0.62	0.83	0.71	0.83

Tabela 13: Comparação do desempenho dos métodos nos *corpora* ToLD-BR, HateBR e HLPHSD nos cenários original (*Orig.*) e filtrado (*Filt.*), considerando acurácia, precisão, cobertura, *F-measure* e *Macro F-measure*.