Community Notes no Brasil: consenso, polarização e aceitabilidade algorítmica

Davi Machado da Rocha

https://doi.org/10.1590/SciELOPreprints.15996

Submetido em: 2026-04-29
Postado em: 2026-08-11 (versão 1)
(AAAA-MM-DD)

Community Notes no Brasil: consenso, polarização e aceitabilidade algorítmica

Davi Machado da Rocha
Aluno Especial ICMC USP / UFSCar, São Carlos, SP, Brasil.
ORCID: https://orcid.org/0009-0004-3326-6881

Abril de 2026

Community Notes in Brazil: Consensus, Polarization, and Algorithmic Acceptability

Resumo

Este artigo investiga como o Community Notes opera no recorte lusófono a partir do *snapshot* público de 7 de abril de 2026, do qual extraímos um corpus de 142.448 notas em língua portuguesa. Articulando filtragem linguística com fastText, embeddings multilíngues, modelagem de tópicos com BERTopic, reconhecimento de entidades nomeadas e classificação interpretável de utilidade via SHAP, o trabalho mostra que 82% das notas em português permanecem em NEEDS_MORE_RATINGS e que essa suspensão se distribui de modo sistematicamente desigual entre tópicos. A análise sustenta a tese de que a utilidade no Community Notes é mais bem entendida como *aceitabilidade algorítmica* do que como acurácia factual, e o sistema, como infraestrutura de governança algorítmica da verdade pública.

Palavras-chave: Community Notes, moderação de conteúdo, *bridging-based ranking*, PLN, modelagem de tópicos, SHAP, desinformação, Brasil.

Abstract

This article investigates how Community Notes operates in the Portuguese-language context using the public snapshot from April 7, 2026, from which we extract a corpus of 142,448 notes in Portuguese. Combining language filtering with fastText, multilingual embeddings, topic modeling with BERTopic, named entity recognition, and interpretable helpfulness classification via SHAP, the study shows that 82% of Portuguese-language notes remain in NEEDS_MORE_RATINGS and that this suspension is distributed unevenly across topics. The analysis argues that helpfulness in Community Notes is better understood as *algorithmic acceptability* than as factual accuracy, and that the system operates as an infrastructure for the algorithmic governance of public truth.

Keywords: Community Notes, content moderation, *bridging-based ranking*, natural language processing, topic modeling, SHAP, misinformation, Brazil.

1 Introdução

Nos últimos anos, as grandes plataformas de mídia social passaram a redistribuir uma parte do trabalho de moderação que historicamente cabia a equipes internas e a parceiros de *fact-checking*. O movimento mais visível dessa transição é o Community Notes (antigo Birdwatch), implementado pelo X — anteriormente Twitter — como infraestrutura de checagem colaborativa: usuários voluntários escrevem notas que adicionam contexto a publicações potencialmente enganosas, outros usuários avaliam essas notas, e um algoritmo decide quais delas se tornam públicas. A decisão não é tomada pela maioria. É tomada por um modelo de fatoração de matrizes que estima posições latentes de avaliadores e de notas, e prioriza aquelas que conseguem aceitação ampla mesmo entre grupos com perspectivas distintas — uma família de algoritmos hoje conhecida como *bridging-based ranking* (Wojcik et al., 2022; Ovadya e Thorburn, 2023).

Esse desenho transforma o Community Notes em algo que excede a categoria *fact-checking*. Trata-se de uma infraestrutura de governança algorítmica da verdade pública: um sistema em que a publicação de uma correção depende não apenas de sua acurácia factual, mas de sua capacidade de atravessar divisões entre avaliadores. Como mostraram Allen, Martel e Rand (2022), a partidarização incide pesadamente sobre as próprias avaliações; como mostraram Chuai et al. (2024), o sistema tende a chegar tarde demais para conter a viralização inicial. As notas que conseguem ser publicadas formam, portanto, um subconjunto bastante específico — e o que fica do lado de fora desse subconjunto, mantido em estado de NEEDS_MORE_RATINGS, é frequentemente mais informativo do que o que entra.

Há uma chave interpretativa útil para enquadrar o que se segue. Se a verdade pública é, em termos práticos, aquilo que sobrevive ao confronto livre entre perspectivas divergentes — como propõe uma tradição liberal de filosofia política (Rorty, 2007) —, então o Community Notes pode ser lido como uma tentativa de institucionalizar exatamente esse processo. A pergunta empírica que organiza este trabalho é mais modesta e mais precisa: como esse mecanismo funciona, quem participa efetivamente, o que ele aprova — e o que ele mantém em suspensão.

A literatura empírica sobre o mecanismo concentra-se quase inteiramente no contexto anglófono e no eixo bipartidário norte-americano. Há boas razões para isso: foi nesse contexto que o sistema foi concebido, e foi nele que circularam os primeiros dados públicos de magnitude analisável. Mas esse recorte deixa em aberto uma pergunta de natureza empírica que importa para qualquer agenda comparativa de governança algorítmica: como o mesmo desenho se comporta quando submetido a um ecossistema político, linguístico e institucional distinto daquele em que foi pensado? O Brasil oferece um caso particularmente forte para essa pergunta. É um país com polarização afetiva intensa e com um sistema multipartidário cujo eixo de divisão não se reduz à oposição binária assumida pelos modelos formais de *bridging*. É também um espaço em que a desinformação política tem ocupado lugar central no debate público sobre integridade eleitoral, plataformas e responsabilidade institucional.

Este artigo investiga, sobre essa base, como o Community Notes opera no recorte lusófono. A análise é construída sobre o *snapshot* público de 7 de abril de 2026, do qual extraímos — após detecção de idioma com fastText `lid.176.bin` aplicada ao texto da nota e *threshold* de confiança de 0,80 — um corpus de 142.448 notas em língua portuguesa, cerca de 5,7% do universo global no momento da coleta. A esse corpus articulamos os *ratings* correspondentes (≈9,9 milhões), o

status mais recente de cada nota e uma amostra hidratada de 839 *tweets*-fonte recuperados via API do X em um desenho estratificado por controvérsia, *batSignals* e amostragem aleatória. As decisões de filtragem, a integridade referencial dos *joins* e os critérios de validação da detecção de idioma estão documentados em *pipeline* reprodutível. Sobre esse corpus, executamos ainda o algoritmo oficial de fatoração de matrizes do X, produzindo as coordenadas latentes de notas e avaliadores que tornam a polarização diretamente observável como quantidade calculada.

Três ângulos empíricos articulam a análise. O primeiro é modal: 82% das notas em português permanecem em NEEDS_MORE_RATINGS — apenas 11,2% atravessam o filtro de consenso e são publicadas como úteis (CRH). A imagem resultante é a de um sistema cujo estado predominante não é a decisão, mas a suspensão dela. Curiosamente, essa taxa de publicação supera a observada no recorte anglófono (8,7%), o que sugere que o sistema brasileiro produz consenso com mais frequência relativa, ainda que sobre um volume ordens de magnitude menor. O segundo ângulo é temático: a taxa de publicação varia de forma pronunciada entre assuntos. Tópicos de domínio operacional bem delimitado — apostas *online*, golpes financeiros, vídeos gerados por inteligência artificial — atravessam o filtro com relativa fluidez; controvérsias politicamente carregadas acumulam notas em suspensão em proporções sistematicamente acima da média do corpus. O terceiro ângulo, e talvez o mais revelador, emerge das matrizes de polaridade produzidas pela execução do algoritmo oficial de fatoração sobre o recorte lusófono: as avaliações inscrevem uma geometria latente que torna observável, como quantidade calculada, a dimensão de divisão sobre a qual o consenso opera — ou não. Esses três planos — suspensão modal, assimetria temática e polarização latente — são os eixos por que este artigo examina o Community Notes no Brasil.

As contribuições do trabalho organizam-se em três frentes. *Empiricamente*, oferecemos a primeira análise em larga escala do Community Notes em português, com cobertura temporal estendida e magnitude superior à dos estudos sobre o piloto anglófono. *Metodologicamente*, descrevemos um *pipeline* reprodutível que articula PLN multilíngue, modelagem de tópicos assistida por LLM, NER com correferência e classificação interpretável de utilidade via SHAP — para responder qual forma de evidência, de linguagem, de tema e de polarização consegue atravessar o filtro algorítmico de consenso em português. *Teoricamente*, propomos tratar a utilidade de uma nota como aceitabilidade algorítmica — distinta, mas relacionada, à acurácia factual —, o que abre o sistema à análise como mecanismo concreto de governança da verdade *online*. Os artefatos que permitem reprodução e inspeção pública dos resultados são descritos na declaração de disponibilidade de dados de pesquisa.

O restante do artigo está organizado da seguinte forma. A Seção 2 situa o trabalho na literatura sobre Community Notes, *bridging-based ranking*, governança de plataformas e PLN aplicado a desinformação. A Seção 3 posiciona o trabalho em relação à lacuna identificada. A Seção 4 descreve o *snapshot*, os procedimentos de detecção de idioma e a integridade referencial entre as tabelas. A Seção 5 apresenta o *pipeline* analítico. A Seção 6 reporta os resultados em três planos: panorama temático, infraestrutura de evidência e classificador interpretável de utilidade. As Seções 7 e 8 discutem as implicações e concluem.

2 Trabalhos relacionados

A literatura sobre o Community Notes — e seu antecessor imediato, o Birdwatch — comporta duas leituras. A mais direta enquadra o sistema como uma variante de *fact-checking* terceirizada para a multidão, e mede seu desempenho contra verificadores profissionais. A leitura que adotamos aqui é mais ampla: trata o mecanismo como uma infraestrutura de governança algorítmica da verdade pública, na qual a moderação deixa de ser um ato discreto, executado por equipes internas ou parceiros externos, e passa a depender de um ciclo distribuído em que usuários produzem notas, outros usuários as avaliam, e um algoritmo decide quais delas atravessam o filtro de consenso entre grupos potencialmente divergentes. Essa segunda leitura organiza o que se segue: cada bloco da literatura é apresentado não como subárea isolada, mas como um ângulo do mesmo objeto sociotécnico.

2.1 Mecanismo, evidência e os limites do *crowdsourcing*

A primeira análise sistemática do Birdwatch é a de Pröllochs (2022), que mapeia o uso da plataforma ainda em sua fase piloto. O trabalho mostra que as notas serviam predominantemente para sinalizar conteúdos enganosos — por erro factual, ausência de contexto ou apresentação de alegações não verificadas como fato — e que a percepção de utilidade estava associada ao uso de fontes confiáveis. Um achado particularmente relevante para nosso caso: *tweets* de autores com grande influência tendem a produzir menor consenso entre avaliadores, sugerindo que a polêmica em torno do emissor contamina o julgamento da nota antes mesmo do exame da alegação.

Wojcik et al. (2022) descrevem a peça técnica que sustenta o sistema. O ranqueamento das notas combina dados de avaliação com um modelo de fatoração de matrizes que estima, para cada avaliador e para cada nota, posições latentes em um espaço de baixa dimensionalidade. Notas selecionadas para publicação são aquelas cujo intercepto (utilidade) é alto e cujo fator de polaridade aponta para aceitação ampla — uma instância empírica do que a literatura chamará de *bridging-based ranking*. É essa lógica que torna possível observar, em dados públicos, não apenas qual nota foi publicada, mas em que dimensão de divergência ela se inseria.

Saeed et al. (2022) introduzem a tensão fundamental do modelo. Comparando o Birdwatch com verificadores especialistas, mostram que a multidão oferece escala e velocidade, mas nem sempre produz resultados consistentes ou acionáveis. Há tipos de conteúdo em que *crowd* e *expert* convergem; há outros em que divergem de modo sistemático. Wirtschafter e Majumder (2023) levam adiante essa avaliação cética, apontando limites estruturais do *crowdsourcing* para identificar conteúdos verdadeiramente nocivos. Borenstein et al. (2025) sintetizam o estado atual desse debate: as notas comunitárias, ainda que valiosas, dificilmente substituem o trabalho de verificadores profissionais, sobretudo nos casos em que a verdade pública depende de domínio técnico ou acesso a fontes restritas.

2.2 Partidarização e o problema do consenso

Se o ranqueamento depende de avaliadores humanos, a identidade política desses avaliadores importa. Allen, Martel e Rand (2022) mostram que usuários do Birdwatch desafiam mais frequentemente o conteúdo de adversários políticos e julgam como menos úteis as contribuições

de contrapartidários. O ponto não é trivial: indica que a partidarização não opera apenas no nível do conteúdo verificado — opera também no julgamento do julgamento. Bozarth et al. (2023), em estudo paralelo sobre moderação no Reddit, encontram padrão coerente: comunidades *online* tendem a confiar mais em sinais da própria base do que em verificadores externos, mesmo quando essa preferência implica reduzir a qualidade epistêmica da decisão.

Esses achados têm uma consequência metodológica direta. Não basta examinar o que está escrito numa nota; é preciso examinar quem a avaliou, em que distribuição, e em que tópico. É justamente aí que entram as matrizes de polaridade publicadas pelo X. Elas tornam observável, no plano dos dados, aquilo que a literatura sobre partidarização vinha tratando majoritariamente em escala experimental.

2.3 *Bridging systems* e o ranqueamento por consenso

A discussão sobre algoritmos que privilegiam aceitação cruzada entre grupos vem ganhando densidade conceitual. Ovadya e Thorburn (2023) propõem o termo *bridging systems* para descrever sistemas de alocação de atenção que premiam conteúdos capazes de atravessar divisões sociais — uma resposta direta ao modelo dominante, em que o ranqueamento por engajamento tende a amplificar conteúdo que reforça identidades de grupo. Thorburn, Polukarov e Ventre (2024) formalizam o argumento: *rankings* baseados em engajamento homogêneo produzem, sob hipóteses razoáveis, *sorting* social; *rankings* baseados em engajamento diverso podem mitigá-lo.

A evidência empírica mais robusta desse efeito vem de Piccardi et al. (2025), que conduziram um experimento de campo no X durante a campanha presidencial norte-americana de 2024. Manipular a presença de conteúdo de animosidade interpartidária no *feed* dos participantes alterou medidas de polarização afetiva por margens estatisticamente significativas — demonstração de que mudanças no ranqueamento têm consequências mensuráveis sobre disposições políticas. Em registro complementar, Törnberg et al. (2023) usam simulações com agentes baseados em LLMs para mostrar que algoritmos do tipo *bridging* produzem conversas menos tóxicas e mais cooperativas do que os algoritmos por engajamento.

Para os fins deste trabalho, essa literatura justifica tratar as matrizes de polaridade do Community Notes não como artefato técnico, mas como objeto empírico privilegiado: nelas, o consenso *cross-partisan* deixa de ser um ideal normativo para se tornar uma quantidade calculada.

2.4 Eficácia, lentidão e a janela da viralização

Chuai et al. (2024) examinam, com *Difference-in-Differences* e *Regression Discontinuity Design*, se a implantação do Community Notes reduziu o engajamento com *tweets* enganosos no X/Twitter. O resultado é prudente: embora o volume de notas tenha crescido, não há evidência robusta de redução significativa no engajamento com conteúdos enganosos, sugerindo que o sistema atua tarde demais em relação à janela inicial de viralização. Esse achado dialoga diretamente com a observação central deste trabalho: a maior parte das notas brasileiras permanece em suspensão de consenso, e a lentidão do sistema não é apenas operacional — é estrutural ao desenho de *bridging-based ranking*, que exige acúmulo de avaliações cruzadas para produzir uma decisão de visibilidade.

2.5 Plataformas como infraestrutura de governança

Um quarto eixo da literatura, mais amplo, oferece o enquadramento teórico para o que vem a seguir. Gillespie (2018), Klonick (2018), Gorwa (2019) e Gorwa, Binns e Katzenbach (2020) firmaram a leitura das plataformas como instituições privadas de governo da fala pública — entidades que produzem regras, decidem exceções e operam infraestruturas de visibilidade que se tornaram constitutivas da esfera pública contemporânea. Nessa chave, o Community Notes não desfaz o poder da plataforma: redistribui o trabalho avaliativo, mas mantém a decisão final de visibilidade dependente de critérios proprietários e escolhas de *design* que continuam internas à empresa. Estudos comparados, como o de Zhao e Hu (2025) sobre o sistema do Weibo, mostram que arquiteturas de moderação distribuída assumem formatos muito distintos em diferentes contextos políticos e regulatórios — reforçando a importância de não assumir o desenho do Community Notes como universal.

2.6 PLN aplicado à desinformação: do que detecta ao que se explica

A literatura de PLN sobre desinformação, embora tangencial ao núcleo do nosso argumento, é metodologicamente relevante. Trabalhos como Ayoub, Yang e Zhou (2021), Men e Mariano (2024) e Patel et al. (2025) mostram que arquiteturas baseadas em BERT, DistilBERT e RoBERTa, combinadas com técnicas de explicabilidade como SHAP e LIME, oferecem desempenho competitivo em tarefas de detecção de *fake news* sem abrir mão da interpretabilidade local. Esse é o terreno em que se apoia parte do nosso *pipeline* — em particular, a classificação de utilidade das notas e a inspeção dos sinais que pesam nessa decisão.

Há, porém, uma diferença importante de escopo. A maior parte da literatura de detecção opera em torno de uma classificação binária (verdadeiro/falso) sobre o conteúdo do *post*. Aqui, a tarefa não é detectar desinformação, mas prever a aceitabilidade algorítmica de uma nota — quais traços linguísticos, de fonte e de contexto fazem com que ela atravesse o filtro de consenso. É uma deslocação que retira do PLN o papel de árbitro da verdade e o coloca como ferramenta de inspeção da própria infraestrutura de decisão.

2.7 A lacuna brasileira

Há literatura abundante sobre desinformação no Brasil, com forte concentração em ciclos eleitorais e na atuação de agências brasileiras de *fact-checking*. O que ainda é escasso é a evidência sistemática sobre o Community Notes em português: não dispomos, até onde sabemos, de análises em larga escala sobre quais disputas factuais mobilizam o sistema no Brasil, sobre as entidades políticas e institucionais que reaparecem nas notas, sobre os formatos de evidência que circulam, ou sobre como o consenso *cross-partisan* opera num ambiente político multipartidário marcado por uma polarização afetiva que não se reduz ao eixo bipartidário norte-americano. É nessa lacuna que o presente artigo se insere.

3 Posicionamento

A literatura existente explica como o Community Notes funciona, como a partidarização afeta a avaliação das notas, e por que o sistema tende a chegar tarde demais para conter a viralização. O que ela não endereça é o que acontece quando esse mecanismo é submetido a um ecossistema político, linguístico e institucional distinto daquele em que foi desenhado. Em vez de tomar o Community Notes como intervenção contra desinformação, este artigo o examina como infraestrutura de governança algorítmica da verdade: um sistema em que a publicação de uma correção depende não apenas de sua acurácia factual, mas de sua capacidade de atravessar divisões entre avaliadores com perspectivas distintas. A pergunta que organiza o trabalho — que formas de verdade pública conseguem se tornar consensuais num país politicamente polarizado e com sistema multipartidário? — desloca o debate de "o sistema funciona?" para uma leitura do que o sistema bloqueia, do que publica, e do que mantém em suspensão.

A novidade não está em aplicar PLN a uma nova base, mas em usar essa base para observar, no concreto, quais formas de evidência, de linguagem, de tema e de polarização conseguem atravessar o filtro algorítmico de consenso no Brasil. Essa combinação torna as decisões de modelagem observáveis e auditáveis sem deslocar o foco do artigo: explicar como o filtro de consenso opera no recorte brasileiro.

4 Dados

4.1 O sistema Community Notes e o *snapshot* analisado

O Community Notes é o sistema de moderação colaborativa do X. Quando um usuário considera que uma publicação é potencialmente enganosa, pode escrever uma nota contextual indicando o problema e, se possível, apontando uma fonte. Outros usuários avaliam essa nota como HELPFUL, SOMEWHAT_HELPFUL ou NOT_HELPFUL. Um algoritmo de fatoração de matrizes — descrito em detalhe por Wojcik et al. (2022) — combina os *ratings* em duas estimativas latentes por nota. Para cada par (nota n, avaliador r), o modelo estima uma nota de utilidade prevista:

$$\hat{u}_{nr} = \mu + b_n + b_r + \mathbf{f}_n^\top \mathbf{f}_r, \tag{1}$$

onde μ é a média global, b_n e b_r são os interceptos de nota e avaliador, e $\mathbf{f}_n, \mathbf{f}_r \in \mathbb{R}^d$ são os vetores de polaridade. Uma nota é publicada quando $b_n \geq \tau_u$ e $\|\mathbf{f}_n\| \leq \tau_p$ — isto é, quando sua utilidade percebida é alta e sua posição no espaço de divisão é central o suficiente para não receber avaliações sistematicamente opostas de grupos diferentes. As demais notas permanecem em NEEDS_MORE_RATINGS (NMR), são marcadas como CURRENTLY_RATED_NOT_HELPFUL (CRNH), ou recebem outros estados intermediários.

O *snapshot* público, divulgado pelo X em formato TSV particionado, contém cinco tabelas relacionais: *notes* (notas escritas pelos colaboradores), *ratings* (avaliações individuais), *note_status_history* (histórico de *status* algorítmico), *user_enrollment* (cadastro dos avaliadores) e *note_requests* (pedidos de nota — *batSignals*). Os dados não trazem coluna de idioma, e a granularidade temporal é dada por *timestamps* em milissegundos. Para este trabalho, utilizamos o *snapshot* de 7 de abril de 2026, que contém aproximadamente 2,6 milhões de notas globais e

190 milhões de *ratings*.

4.2 *Pipeline* de ingestão e integridade referencial

A ingestão dos dados foi feita em três camadas. A camada *raw* preserva os arquivos TSV originais com *checksums* MD5 e SHA-256, sem qualquer transformação. A camada *bronze* faz a leitura em DuckDB com `union_by_name=true` e mantém todos os campos como `VARCHAR`. A camada *silver* aplica conversões tipadas e remove registros órfãos — *ratings* e eventos de *status* que referenciam `noteId` ausentes da tabela *notes*. No *snapshot* analisado, registros órfãos representavam 11,36% dos *ratings* e 12,66% dos eventos de *status*, reflexo do fato de que o *dump* público não cobre a totalidade das notas que já existiram no sistema. Essa limitação está registrada em `metadata/orphan_analysis.parquet`, integrante do *dataset* público.

4.3 Detecção de idioma e construção do recorte lusófono

Como o *dump* não inclui anotação de idioma, o recorte em português foi construído por classificação automática do texto da nota. Antes da inferência, o texto passou por limpeza implementada diretamente em DuckDB: remoção de URLs por expressão regular, normalização de *whitespace* e truncamento em 2.000 caracteres. A detecção usou o modelo fastText `lid.176.bin` da Meta (Joulin et al., 2016a, 2016b), capaz de classificar 176 idiomas em escala de milhões de notas em poucos minutos.

A seleção do recorte português aplicou *threshold* de confiança de 0,80, calibrado por amostragem manual estratificada em três faixas (0,80–0,90, 0,90–0,95, 0,95–1,00), com 100 exemplos por faixa anotados quanto ao idioma efetivo. O recorte resultante contém 142.448 notas em português, aproximadamente 5,7% do universo global no *snapshot* — proporção compatível com a participação de outras línguas românicas no *corpus*, mas várias ordens de magnitude inferior ao volume do inglês.

4.4 Acoplamento dos *ratings* e do estado mais recente

Os 10 *shards* de *silver_ratings* foram lidos diretamente do Hugging Face via protocolo `hf://`, com `SEMI JOIN` contra os 142.448 `noteId` em português executado *shard* a *shard* em DuckDB — mantendo o pico de memória em torno de 1,5 GB e permitindo retomada por *shard* em caso de falha. O resultado são 9,9 milhões de *ratings* associados ao recorte lusófono. Para o histórico de *status*, aplicamos `ROW_NUMBER() OVER (PARTITION BY noteId ORDER BY timestampMillisOfCurrentStatus DESC)` sobre o subconjunto português, produzindo o estado mais recente de cada nota.

4.5 Hidratação amostral dos *tweets*-fonte

O *snapshot* público registra apenas o `tweetId` da publicação alvo da nota; o texto do *tweet* não está disponível. Para uma fração do *corpus*, recuperamos o texto via *endpoint* `GET /2/tweets` da API v2 do X em um desenho estratificado: 400 *tweets* com maior número de notas (estrato A — controvérsia), 300 presentes na tabela `note_requests` (estrato B — *batSignals*) e 250 selecionados aleatoriamente (estrato C — base de comparação). Dos 950 IDs solicitados, 839 *tweets* foram recuperados (88,3%); os 111 restantes retornaram erro de acesso, sinalizando *tweets*

deletados, suspensos ou de contas privadas. A amostra hidratada é, por construção, enviesada para controvérsia — o que é desejável para o argumento do trabalho, dado que justamente as notas associadas a esses *tweets* concentram a tensão analítica do sistema.

4.6 Documento analítico

Para a etapa de modelagem temática, cada nota foi convertida em um documento composto pela concatenação do texto da nota e do texto do *tweet*-fonte (quando hidratado), no formato `[NOTA] {summary} [TWEET] {tweet_text}`. Esse formato preserva os dois lados da disputa factual e melhora a coesão semântica dos *embeddings* em casos em que o texto da nota, isoladamente, é breve ou alusivo. Documentos com menos de 30 caracteres foram descartados. Após esse filtro, restaram 142.051 notas válidas para análise. A distribuição de *status* é fortemente assimétrica: 116.741 notas em Pendente (NMR, 82,2%), 15.857 Publicadas (CRH, 11,2%), 4.423 Não publicadas (CRNH, 3,1%) e 5.427 em outros estados (3,8%).

5 Métodos

O *pipeline* analítico organiza-se em torno de quatro tarefas que se articulam progressivamente. A Tarefa 1 (modelagem de tópicos) distribui o *corpus* em estruturas temáticas interpretáveis, revelando a assimetria de consenso entre assuntos. A Tarefa 2 (reconhecimento de entidades nomeadas) mapeia os atores, instituições e fontes presentes nas notas; seu *pipeline* é descrito junto com os resultados na Seção 6.3. A Tarefa 3 (classificação de utilidade e interpretabilidade) treina e inspeciona modelos que aprendem o que a comunidade chama de utilidade — e onde essa aprendizagem diverge do rótulo. A Tarefa 4 (artefatos de dados e infraestrutura pública) produz os recursos que tornam a pesquisa reprodutível e auditável; sua documentação está na Seção 4 e nos artefatos públicos. Esta seção detalha as Tarefas 1 e 3.

Ferramentas e modelos de PLN. As referências técnicas centrais são explicitadas aqui para evitar ambiguidade metodológica: a detecção de idioma usa fastText `lid.176.bin` (Joulin et al., 2016a, 2016b); os *embeddings* semânticos usam `intfloat/multilingual-e5-large-instruct`, com vetores de 1024 dimensões (Wang et al., 2024); a modelagem temática é implementada com BERTopic (Grootendorst, 2022), articulando UMAP, HDBSCAN e c-TF-IDF; o reconhecimento de entidades combina GLiNER multilíngue (Zaratiana et al., 2024) com expressões regulares determinísticas para padrões formais; e a rotulagem dos tópicos usa Marco-Mini-Instruct, citado como modelo da família Marco-MoE (Jiang et al., 2026).

Tarefa 1 — Modelagem temática

A modelagem segue o *pipeline* esquematizado na Figura 1, organizado em seis etapas: representação textual, calibração de hiperparâmetros, *clustering*, redução de *outliers*, rotulagem estruturada com LLM e macroagregação hierárquica.

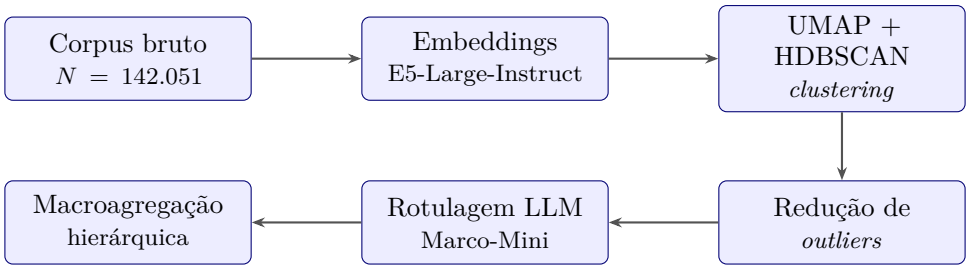

Figura 1: *Pipeline* de modelagem temática: do *corpus* bruto à análise final em seis etapas encadeadas — representação por *embeddings* multilíngues, redução UMAP/*clustering* HDBSCAN, redução de *outliers*, rotulagem estruturada por LLM local e macroagregação hierárquica.

5.0.1 Representação textual

Os documentos analíticos foram convertidos em vetores semânticos de 1024 dimensões pelo modelo `intfloat/multilingual-e5-large-instruct` (Wang et al., 2024) — um *encoder* multilíngue de 560 M de parâmetros, treinado por *contrastive learning* e capaz de receber instruções no estilo `Instruct: ...\n Query: ...`. Em nosso caso, a instrução adotada foi: *"Identifique o assunto factual em disputa neste texto, ignorando aspectos procedurais sobre a plataforma ou sobre como as notas funcionam."* A escolha não é estética: parte significativa do *corpus* contém formulações que descrevem o próprio Community Notes ("esta nota não acrescenta contexto", "o *tweet* não precisa de nota") — material que, sem essa orientação, formaria *clusters* artificiais em torno de meta-comentários sobre a plataforma. A instrução desloca o vetor para o assunto da disputa, não para a forma do julgamento.

A inferência foi feita em GPU (Tesla T4) com *batches* de 32 documentos, comprimento máximo de 512 *tokens*, *mean pooling* mascarado pela *attention mask* e normalização L2. Os vetores resultantes formam a matriz $E \in \mathbb{R}^{142051 \times 1024}$, persistida em disco para evitar recomputação nas etapas seguintes.

5.0.2 Calibração de hiperparâmetros

Antes do *fit* definitivo, executamos um *grid search* sobre os principais hiperparâmetros do par UMAP/HDBSCAN, avaliando cada configuração por três métricas: o *Density-Based Clustering Validation* (DBCV), a fração de *outliers* e o tamanho do maior *cluster*. A calibração tem propósito específico: evitar a formação de megaclusters espúrios — *clusters* que absorvem dezenas de milhares de documentos e correspondem, na prática, a ruído agregado e não a um tópico real. A configuração escolhida foi `n_neighbors = 15`, `n_components = 10`, `min_dist = 0,0` e métrica cosseno para o UMAP; `min_cluster_size = 30` e `min_samples = 5` para o HDBSCAN. Os parâmetros do UMAP estão acima dos *defaults* do BERTopic (`n_components = 5`) por estabilizarem o espaço reduzido para o volume do *corpus*; os do HDBSCAN foram escolhidos para que tópicos de baixa frequência ainda atravessem o limiar de detecção sem inflar excessivamente a contagem total.

Tabela 1: Hiperparâmetros do *pipeline* de modelagem temática. As escolhas foram orientadas pelo *grid search* descrito em 5.0.2, com inspeção visual da hierarquia para o corte em macrotemas (5.0.6).

Componente	**Parâmetro**	**Valor**	**Observação**
UMAP	n_neighbors	15	balanceia estrutura local e global
UMAP	n_components	10	acima do *default* (5) — mais estabilidade
UMAP	min_dist	0,0	favorece adensamento de *clusters*
UMAP	metric	cosseno	coerente com *embeddings* normalizados
HDBSCAN	min_cluster_size	30	piso para tópico válido
HDBSCAN	min_samples	5	conservador a *outliers*
CountVectorizer	ngram_range	$(1, 2)$	uni- e bigramas
CountVectorizer	min_df	5	termo presente em ≥ 5 documentos
KeyBERTInspired	top_n_words	10	palavras-chave por tópico
MMR	diversity	0,3	diversificação moderada
reduce_outliers	threshold	0,975	similaridade mínima para realocar
Linkage hierárquico	method	average	sobre cosseno entre tópicos
Corte do dendrograma	t	0,05	escolhido por inspeção visual

5.0.3 *Clustering* e representação dos tópicos

A fase de *clustering* foi implementada com a biblioteca BERTopic (Grootendorst, 2022) na configuração modular: redução dimensional pelo UMAP calibrado, agrupamento por HDBSCAN, vetorização por `CountVectorizer` com *stopwords* combinadas (lista padrão do `nltk` para português, acrescida de termos genéricos do domínio como "nota", "*tweet*", "usuário", "*post*", e dos marcadores `[nota]`/`[tweet]`), e cálculo do c-TF-IDF para representação por palavras. Para mitigar o efeito conhecido de o c-TF-IDF privilegiar termos extremos, adicionamos duas representações secundárias: KeyBERTInspired, que reordena as palavras candidatas por similaridade com o *embedding* do tópico, e *Maximal Marginal Relevance* (MMR) com `diversity = 0,3`.

O resultado dessa fase, antes da redução de *outliers*, foi uma partição em 898 tópicos válidos e 57.509 documentos não atribuídos (40,5% do *corpus* marcados como -1) — percentual compatível com o que a literatura tem reportado para *corpora* curtos e linguisticamente heterogêneos, mas que exige intervenção para não introduzir viés sistemático na estimativa de relação entre tópico e *status* de publicação.

5.0.4 Redução de *outliers*

A redução foi feita pelo método `reduce_outliers` da BERTopic na estratégia *embeddings*, que recalcula, para cada documento previamente classificado como *outlier*, a similaridade cosseno com o centroide de cada tópico, realocando-o ao tópico mais próximo desde que a similaridade ultrapasse 0,975. Esse limiar favorece realocações conservadoras: documentos de fronteira só são reabsorvidos se houver alta proximidade semântica com o tópico de destino. Após a redução, os *outliers* caíram de 57.509 (40,5%) para 47.162 (33,2%), com 10.347 documentos recuperados. O número de tópicos ativos permaneceu em 898 — sinal de que a redução não fundiu *clusters*, apenas absorveu fronteiras sub-representadas.

5.0.5 Rotulagem estruturada com LLM local

A rotulagem dos 898 tópicos foi feita pelo modelo Marco-Mini-Instruct, da família Marco-MoE (Jiang et al., 2026), em quantização Q6_K, executado localmente via `llama.cpp` com contexto de 8.192 *tokens*. Para cada tópico, o modelo recebeu suas dez palavras-chave principais (KeyBERTInspired e MMR concatenadas), três documentos representativos truncados em 200 caracteres, e um *prompt* em português orientando para o contexto sociopolítico brasileiro.

Duas escolhas de *design* merecem registro. A primeira é a inserção, no *prompt*, de uma restrição explícita contra rótulos genéricos do tipo "desinformação", "manipulação" ou "verificação" — termos que descrevem a tarefa do Community Notes mas não o assunto da nota; sem essa restrição, o LLM tende a homogeneizar a saída. A segunda é o uso de uma gramática GBNF que força a saída a um JSON estruturado com três campos: `contexto_central` (uma frase descrevendo a dinâmica entre *tweet* e nota), `categoria_ampla` (uma categoria de alto nível, escolhida de um conjunto fixo) e `rotulo_curto` (rótulo legível de até seis palavras). A gramática elimina por construção a possibilidade de saída malformada e dispensa rotinas de *retry*.

A título de exemplo: o tópico T9, agrupando documentos sobre repercussão de turnê internacional, recebeu o rótulo *Turnê de Taylor Swift* na categoria Entretenimento; o T7, sobre alegações relativas à eleição venezuelana, recebeu *Maduro ditador* na categoria Política; o T2, sobre cotas e mudanças de admissão no ensino superior, foi rotulado *Cotas e Cortes no Ensino Superior* em Educação. A inspeção qualitativa de 50 tópicos selecionados aleatoriamente indica coerência semântica entre rótulo, palavras-chave e documentos representativos em 47 casos.

5.0.6 Macroagregação hierárquica

Os 898 tópicos individuais oferecem granularidade fina, mas dificultam a leitura agregada do *corpus*. Para sintetizar, agregamos os tópicos em macrotemas via *linkage* aglomerativo aplicado às distâncias cosseno entre os *embeddings* de tópico calculados pelo BERTopic (centroides ponderados pelo c-TF-IDF). O método de ligação adotado foi o *average*, que tende a produzir agrupamentos balanceados em volume. O dendrograma resultante foi inspecionado visualmente, e o corte foi fixado em distância 0,05 — limiar propositalmente conservador, para preservar tópicos de pequeno porte semanticamente distintos como macrotemas autônomos; um corte alternativo em 0,5 produziria 14 macrotemas mais agregados, a custo de perda de granularidade. O resultado são 38 macrotemas, dos quais os 15 maiores concentram cerca de 85% dos documentos atribuídos.

A rotulagem dos macrotemas foi feita por uma segunda passagem do mesmo LLM, recebendo como entrada os campos `contexto_central` e `rotulo_curto` produzidos na etapa anterior — não as palavras-chave brutas. Esse desenho — alimentar o segundo passo com a interpretação do primeiro — evita que cada estágio re-derive do zero o significado do *cluster*, perdendo o contexto construído anteriormente. A gramática GBNF do segundo LLM exige dois campos: `rotulo_macro` (até quatro palavras) e `categoria_dominante` (uma de treze categorias predefinidas, com adição da categoria *Procedural* para macrotemas centrados na própria plataforma).

Tarefa 3 — Classificação de utilidade e interpretabilidade

5.0.7 Classificação de utilidade — desenho experimental

Para complementar a leitura temática do *corpus*, treinamos um classificador binário cujo objetivo é prever, a partir do texto de uma nota, se ela seria considerada útil pela comunidade — isto é, se atravessaria o filtro de consenso e atingiria o estado CURRENTLY_RATED_HELPFUL (CRH). O alvo não é qualidade editorial nem acurácia factual, e sim aceitabilidade algorítmica: o que o modelo aprende é o que uma amostra de avaliadores dispostos em diferentes posições ideológicas tipicamente concordou em rotular como útil. Essa distinção é central para qualquer leitura dos resultados e é retomada na Seção 6.5. Uma segunda tarefa operacional — prever se a nota seria publicada ou não, incluindo notas ainda em NEEDS_MORE_RATINGS — é tratada como referência comparativa ao fim desta seção.

Definição das duas tarefas. Dois problemas foram tratados separadamente. A **tarefa central** — *helpful* vs. *not_helpful* — opera sobre o subconjunto de 19.843 notas em português que chegaram a um *status* definido (CRH ou CRNH), com classe positiva de 78,3% e 17.542 *tweets* únicos como grupos. A pergunta é condicional: dado que a comunidade chegou a uma decisão, ela considerou a nota útil? Uma **tarefa operacional de referência** — *publicada* vs. *não publicada* — opera sobre as 142.051 notas válidas, com classe positiva de 10,9%; aqui o estado NEEDS_MORE_RATINGS é tratado como negativo, misturando no rótulo dois fenômenos distintos (rejeição e ausência de avaliação). O desenho cuidadoso de validação, a análise de interpretabilidade e o estudo de casos foram reservados à tarefa central; a tarefa operacional é mantida como referência comparativa e como mecanismo de ranqueamento de notas ainda sem decisão.

Engenharia de variáveis tabulares. A representação de cada nota inclui, além do texto, um vetor de 20 variáveis tabulares construído por pareamento entre nota e *tweet*-fonte. As variáveis se dividem em três famílias. As *estruturais* capturam propriedades superficiais do texto: número de caracteres, *tokens* e sentenças da nota e do *tweet*, comprimento médio por sentença, quantidade de URLs, números, exclamações e interrogações, e a razão entre comprimento da nota e do *tweet*. As *lexicais* marcam padrões discursivos por expressões regulares: presença de abertura notarial ("é falso", "não é verdade", "na verdade"), de marcadores de opinião, de declaração de desnecessidade ("nenhuma nota necessária"), de menção a figuras políticas — Lula, Bolsonaro, Moraes, Trump, Maduro, entre outros — e de menção a instituições — STF, TSE, PF, Banco Central, Anvisa, IBAMA, INSS, entre outras. As *semânticas leves* medem proximidade entre nota e *tweet* via similaridade de Jaccard sobre *tokens* e razão de novidade lexical. Uma 21ª variável categórica, `tema_modelo`, registra o macrotema atribuído pelo *pipeline* da Seção 5.0.6.

Validação cruzada e métricas. A validação foi feita por *k-fold* cruzado com agrupamento por `tweet_id` — uma escolha não negociável para este *corpus*, dado que múltiplas notas frequentemente cobrem o mesmo *tweet*-fonte, e dividi-las entre treino e teste produziria vazamento por concordância temática. Para as trilhas de *baseline* lexical, utilizou-se `GroupKFold(5)`; para os modelos contextuais, `StratifiedGroupKFold(5, shuffle=True, random_state=42)`, que

adicionalmente preserva a proporção de classes em cada dobra. Como os dois conjuntos de trilhas partem das mesmas notas mas usam *splits* diferentes, a comparação direta exige cautela — conforme explicitado na Tabela 6.

As métricas reportadas seguem o conjunto consolidado em estudos de classificação desbalanceada: ROC-AUC para discriminação geral, PR-AUC para discriminação com ênfase na classe positiva, PR-AUC da classe minoritária para detecção de notas não úteis, F1 e *balanced accuracy* em *threshold* 0,5, e *Matthews Correlation Coefficient* (MCC) como métrica robusta a desbalanço. Para cada métrica, reportamos a estimativa OOF global e o desvio entre as cinco dobras.

5.1 Quatro famílias de modelos

Trilha 1 — TF-IDF + Regressão Logística (*baseline* lexical). A trilha 1 é deliberadamente simples e serve como referência. O texto da nota é vetorizado por TF-IDF com unigramas e bigramas (`min_df=5`), concatenado às 20 variáveis tabulares e à codificação de `tema_modelo`. O classificador é uma regressão logística com `class_weight='balanced'`. Três variantes foram avaliadas: A1 (Nota apenas), A2 (Explicativa), que adiciona o TF-IDF do *tweet*-fonte, e A3 (Operacional), que adiciona as colunas `coreNoteIntercept` e `coreNoteFactor1` extraídas do próprio algoritmo do X. A variante A3 contém *target leakage* deliberado e existe apenas como teto operacional: ela responde o quanto o sistema do X é redutível ao seu próprio *output*, não o quanto o texto da nota prevê o resultado.

Trilha 2 — Empilhamento com Qwen *frozen*. A trilha 2 usa o modelo `Qwen/Qwen3-Embedding-4B` como gerador de *embeddings* semânticas sem ajuste à tarefa. As notas passam pelo Qwen com instrução em inglês e *last-token pooling*; o vetor resultante, de 2.560 dimensões, é normalizado e tratado como *feature* densa. Três sinais alimentam um meta-classificador de regressão logística: a probabilidade de uma regressão sobre TF-IDF, a probabilidade de uma regressão sobre os *embeddings* Qwen e o vetor tabular. As probabilidades dos *base learners* são geradas por validação cruzada interna de três dobras, agrupada por *tweet*, antes de alimentarem o meta-modelo.

Trilha 3 — *Fine-tuning* do Qwen via LoRA (FT-Solo). A trilha 3 adapta o Qwen3-Embedding-4B à tarefa via LoRA, com hiperparâmetros `r=16`, `alpha=32`, `dropout=0,05` e `target_modules="all-linear"` — totalizando aproximadamente 33 milhões de parâmetros treináveis sobre os 4 bilhões do *backbone*. O treino seguiu três épocas com `batch_size=16`, taxa de aprendizado 1e−4, `weight_decay=0,01`, *warmup ratio* 0,1, precisão mista em `bfloat16` na A100 e *gradient checkpointing*. A perda foi `BCEWithLogitsLoss` com `pos_weight` calculado por dobra. Uma partição interna via `StratifiedGroupKFold(10)` monitorou validação e reteve o melhor *checkpoint* por `val_auc`. Ao fim, para cada uma das cinco dobras externas obteve-se um *adapter* LoRA, uma cabeça linear, probabilidades OOF e *embeddings* OOF de 2.560 dimensões — todos persistidos em disco para inspeção e reuso.

Trilha 4 — GBDT sobre probabilidades OOF. A trilha 4 testa se um modelo não-linear consegue combinar sinais textuais, semânticos e tabulares melhor do que uma regressão logística. O classificador é o XGBoost, que captura interações entre variáveis — por exemplo, casos em que

o efeito de uma nota curta depende simultaneamente da presença de URL, do macrotema e da probabilidade atribuída pelo FT-Solo. Todas as variáveis foram construídas de modo *fold-safe*: as tabulares dependem apenas do texto; `tema_modelo_te` é recalculado em cada dobra; e as probabilidades OOF são sempre previsões fora da dobra da observação correspondente.

As *embeddings* densas do FT-Solo foram deliberadamente excluídas como *features* diretas do GBDT. Como foram produzidas por *adapters* diferentes em cada dobra, podem carregar parte da identidade do *adapter* — um vazamento sutil que árvores são particularmente capazes de explorar. A probabilidade `proba_ft_oof`, por sua vez, já resume o sinal supervisionado extraído dessas *embeddings* pela cabeça linear, com menor risco de exploração indevida. Duas variantes foram avaliadas: **Min**, com tabulares + `proba_ft_oof`; e **Full**, com tabulares + `proba_e2_oof`, `proba_es_frozen_oof` e `proba_ft_oof`.

6 Resultados

6.1 Estrutura geral do *corpus*

O *fit* final do *pipeline* produziu 898 tópicos, 47.162 documentos não atribuídos (33,2%) e 94.889 documentos atribuídos a algum tópico (66,8%). O número de macrotemas, após o corte hierárquico em 0,05, foi de 38. Essa estrutura — quase 900 tópicos finos agregados em 38 blocos amplos — é coerente com a hipótese de que o *corpus* comporta uma camada de eventos (tópicos) e uma camada de agendas (macrotemas), e é nessa segunda camada que a leitura sociopolítica se torna mais legível. A Tabela 2 lista os 15 maiores macrotemas por volume de documentos.

Tabela 2: Quinze maiores macrotemas do *corpus* em português, ordenados por volume de documentos. A coluna *Tópicos* indica quantos tópicos finos foram aglutinados em cada macrotema. Os percentuais são calculados sobre o total de 142.051 documentos válidos.

ID	Macrotema	Categoria	Tópicos	Documentos	%
0	Sátira clara	Humor	73	12.090	8,5%
1	Crise política	Política	123	9.833	6,9%
2	Sátira de humorista	Humor	74	7.552	5,3%
3	Economia e política	Economia	72	6.864	4,8%
4	Cultura pop e política	Política	45	6.400	4,5%
5	Cotas e cortes no ensino superior	Educação	47	6.370	4,5%
6	Esporte e arbitragem	Esporte	48	5.474	3,9%
7	Apostas *online*	Golpes/fraudes	61	4.700	3,3%
8	Vídeos gerados por IA	Tecnologia	32	4.344	3,1%
9	Crimes contra a mulher	Segurança	35	3.525	2,5%
10	Vacinação e saúde pública	Saúde	37	2.856	2,0%
11	Conflitos no Oriente Médio	Política	24	2.729	1,9%
12	Geopolítica internacional	Política	19	2.618	1,8%
13	Clima e meio ambiente	Meio Ambiente	25	2.339	1,6%
14	Fraude no INSS e benefícios	Economia	19	2.037	1,4%

Três observações merecem registro. Primeiro, os dois macrotemas mais volumosos do *corpus* (Sátira clara e Sátira de humorista) totalizam quase 14% do material e indicam que parte significativa do uso brasileiro do Community Notes ocorre em torno de conteúdo cômico ou irônico — frequentemente sobre comentários satíricos sinalizados por usuários que leram o *tweet* ao pé da letra. Esse achado, embora lateral à agenda principal de combate à desinformação, é

informativo sobre os limites empíricos do que conta como fato em disputa no sistema. Segundo, a temática propriamente política se distribui por pelo menos quatro macrotemas (Crise política, Cultura pop e política, Conflitos no Oriente Médio, Geopolítica internacional), totalizando cerca de 15% do *corpus*. Terceiro, agendas de saúde pública — historicamente centrais na literatura sobre desinformação no Brasil — aparecem com volume relativamente modesto (2,0%), o que pode sinalizar tanto deslocamento de plataforma para essas disputas quanto reflexo do recorte temporal do *snapshot*.

6.2 Tópicos × *status* de publicação

Cruzar macrotemas com a distribuição de *status* finais revela que a assimetria em favor de NEEDS_MORE_RATINGS não é uniforme: é geograficamente concentrada no espaço semântico. Apostas *online*, golpes financeiros e vídeos gerados por IA atravessam o filtro com regularidade acima da média — a estrutura da alegação admite ancoragem objetiva, verificável contra registros de regulação, metadados de imagem ou bases públicas de entidades. *Crise política* e *Cultura pop e política* acumulam taxas acima de 85%, com vários tópicos finos ultrapassando 90%; notas factualmente sólidas reprovam. A Figura 2 torna esse padrão visível como geometria: tópicos de alta densidade política aglutinam-se num bloco denso; domínios operacionais derivam para as bordas do espaço, semanticamente isolados. Separação espacial que coincide, ponto a ponto, com separação de taxas de consenso.

O eixo temporal (Figura 3) revela o que o corte transversal não captura. Volume de notas não é constante: tópicos emergem, têm picos, decaem. *Participante do BBB*, *VAR e mão no gol* — janelas de atenção circunscritas a episódios midiáticos pontuais, que se dissolvem tão rápido quanto aparecem. Diferente de *Identificar vídeos de IA*, que a partir de 2024 exibe crescimento contínuo, sem pico discreto, sem declínio: tendência. Quando o domínio é objetivamente delimitado e retorna com regularidade, o sistema aprende a produzir consenso.

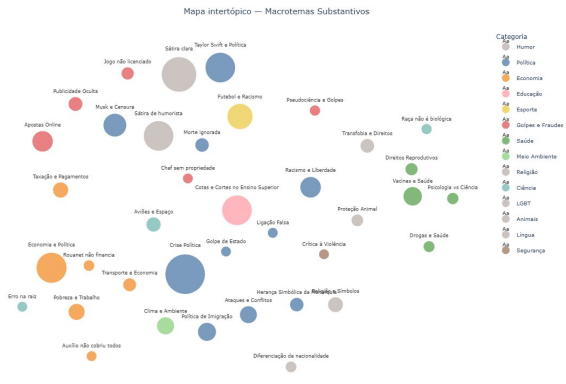

Figura 2: Mapa intertópico após redução de *outliers*. Tópicos de alta densidade política aglutinam-se na metade superior do espaço; domínios operacionais (apostas, tecnologia) ocupam as bordas — padrão que coincide com a assimetria de taxas de publicação.

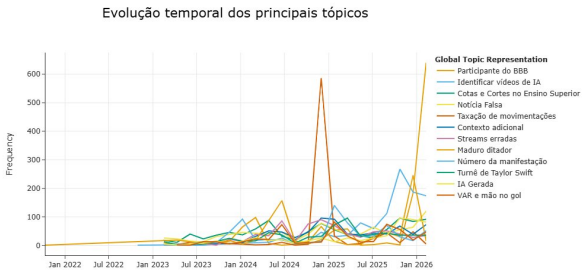

Figura 3: Evolução temporal do volume de notas por tópico (agregado mensal). Picos isolados como *Participante do BBB* contrastam com o crescimento sustentado de *Identificar vídeos de IA* a partir de 2024.

6.3 Infraestrutura de evidência — entidades, atores e fontes

Tópicos descrevem sobre o que as notas falam. O que segue descreve quem é nomeado dentro delas — quais figuras, quais instituições, quais fontes externas aparecem como âncoras de evidência. Não é a mesma operação: tópicos agregam padrões semânticos; entidades agregam padrões referenciais. Operam sobre unidades distintas — documentos versus menções — e a combinação das duas camadas é o que torna possível, na Seção 7, discutir o sistema como infraestrutura de governança, e não apenas como mecanismo de checagem.

6.3.1 *Pipeline* de extração

A extração de entidades segue arquitetura híbrida, combinando o modelo *zero-shot* GLiNER multilíngue (Zaratiana et al., 2024) com expressões regulares para padrões formais. GLiNER tem boa *performance* em classes semânticas abertas (pessoa, organização, programa público, veículo de mídia), mas sub-detecta padrões com sintaxe regular (URL, data, valor monetário, texto de lei). *Regex* tem precisão próxima de um em padrões formais, mas não capta categorias semânticas. A combinação trata cada classe pelo método em que é mais confiável.

A taxonomia define 21 tipos de entidade, dos quais 15 são extraídos por GLiNER e 6 por *regex*. Cada tipo recebe um *threshold* específico, calibrado por inspeção qualitativa de amostras estratificadas, variando de 0,40 (programas públicos) a 0,55 (eventos políticos, processos judiciais, fontes citadas), sobre um piso global de 0,35.

Tabela 3: Taxonomia de entidades. Os tipos extraídos por GLiNER recebem limiar específico calibrado por amostragem; os extraídos por *regex* são considerados deterministicamente válidos. O *papel padrão* distingue entre menções (objetos do enunciado) e fontes ou evidências (referências externas que ancoram a checagem).

Tipo (PT)	Categoria (en, GLiNER)	Limiar	Papel padrão
ATOR_POLITICO	politician	0,45	menção
PESSOA	person	0,45	menção
PESSOA_PUBLICA	public figure	0,45	menção
PARTIDO	political party	0,50	menção
ORGANIZACAO	organization	0,45	menção
ORGAO_PUBLICO	government agency	0,45	menção
ORGAO_JUDICIARIO	court	0,50	menção
ORGAO_SEGURANCA	police or security force	0,50	menção
PROGRAMA_PUBLICO	government program	0,40	menção
LOCAL	location	0,45	menção
EVENTO_POLITICO	political event	0,55	menção
OPERACAO_POLICIAL	police operation	0,55	menção
VEICULO_MIDIA	news media outlet	0,45	fonte ou evidência
PLATAFORMA_DIGITAL	digital platform	0,45	menção
FONTE_CITADA	cited source	0,55	fonte ou evidência
URL_DOMINIO	(*regex*)	—	fonte ou evidência
DATA	(*regex*)	—	fonte ou evidência
VALOR_MONETARIO	(*regex*)	—	fonte ou evidência
ESTATISTICA	(*regex*, percentuais)	—	fonte ou evidência
LEI_NORMA	(*regex*)	—	fonte ou evidência
PROCESSO_JUDICIAL	(*regex*)	—	fonte ou evidência

Três detalhes técnicos do *pipeline* merecem registro. Primeiro, notas com mais de 1.500

caracteres são processadas via *sliding window* com sobreposição de 200 caracteres, evitando truncamento silencioso de menções no fim do texto — uma fonte conhecida de viés em *pipelines* de NER aplicados a *corpora* heterogêneos. Segundo, quando uma entidade GLiNER e uma entidade *regex* coincidem em mais de 50% de seus *spans* (IoU $\geq$ 0,5), prevalece a *regex*. Terceiro, para cada menção o *pipeline* produz um campo `texto_canonico` que normaliza variações ortográficas, permitindo agregação correta de menções equivalentes. A extração processou as 142.051 notas válidas em aproximadamente 100 minutos sobre GPU Tesla T4. O resultado bruto foi de 536.373 entidades; após a camada de curadoria descrita a seguir, o *corpus* consolidado contém 510.845 menções distribuídas por 126.805 notas únicas, média de 4,0 menções por nota.

6.3.2 Curadoria pós-extração: correferência e limpeza

A extração bruta apresenta dois tipos de ruído sistemático. O primeiro é a fragmentação por *aliases*: o mesmo ator político aparece em formas distintas (“Lula”, “Lula da Silva”, “Luiz Inácio”, “presidente Lula”), e cada forma é contabilizada separadamente. O segundo é a presença de descritores genéricos classificados como entidades (“o presidente”, “a ministra”, “o brasileiro”), que inflam a contagem sem corresponder a referentes específicos.

Para tratar esses dois problemas, o *pipeline* aplica uma camada de pós-curadoria construída em oito *rounds* iterativos de inspeção qualitativa. A camada consolida quatro mecanismos: dicionários canônicos para atores políticos e pessoas públicas, reclassificações para corrigir tipos errados, regras contextuais para casos de ambiguidade (por exemplo, “Choquei” como conta no X versus o adjetivo), e listas de descritores genéricos a remover.

Após a curadoria, o *corpus* consolidado de entidades exibe os indicadores da Tabela 4. As principais consolidações tornam-se observáveis: Lula aparece em 5.679 menções; Jair Bolsonaro, em 5.750; Donald Trump, em 2.008. Sem a camada de correferência, esses referentes apareceriam fragmentados em dezenas de variantes ortográficas.

Tabela 4: Resumo do *corpus* de entidades após a camada de curadoria. As 142.051 notas válidas para análise temática reduzem-se a 126.805 notas com pelo menos uma entidade extraída — a unidade de análise efetiva da camada referencial. As menções a pessoas agrupam os tipos `PESSOA`, `ATOR_POLITICO` e `PESSOA_PUBLICA`.

Indicador	Valor	Observação
Notas únicas com entidades	126.805	unidade de análise da camada referencial
Entidades extraídas	510.845	após correferência e limpeza
Tipos de entidade	21	15 via GLiNER + 6 via *regex*
Entidades por nota	4,0	média geral
Menções a pessoas	94.099	política, pública ou genérica
Notas com URL	91,5%	evidência externa ostensivamente exibida

6.3.3 Estrutura geral do *corpus* de entidades

Três padrões se confirmam, um surpreende. URLs: 174.625 menções, 34,2% do total — a presença de fonte ostensivamente exibida não é adorno, é traço estrutural, e o classificador da Seção 6.6 vai confirmar isso na escala preditiva. Atores e instituições: somando `PESSOA`, `ATOR_POLITICO`,

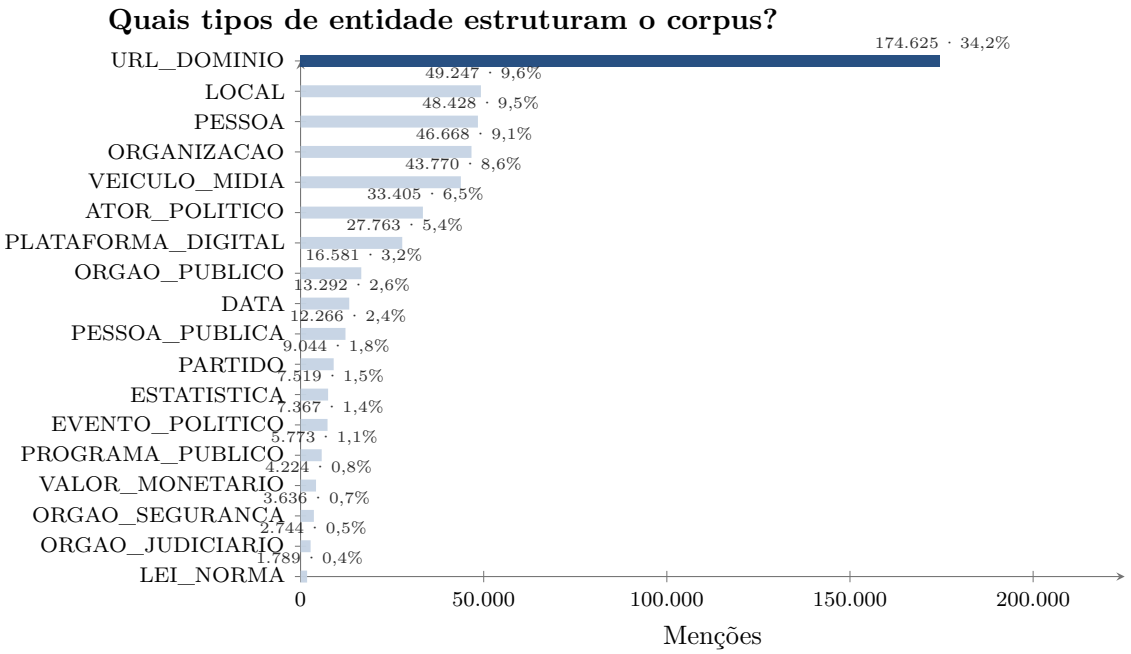

Figura 4: *Top* 18 tipos de entidade por volume de menções, após correferência consolidada.

`PESSOA_PUBLICA`, `ORGAO_PUBLICO` e `PARTIDO`, mais de 119 mil menções. Veículos de mídia tradicional: 43.770 — indicação de que a infraestrutura de evidência das notas brasileiras se apoia, em parte substancial, no jornalismo institucional.

O menos óbvio é o contraste nas classes de evidência formal. Datas (13.292), estatísticas (7.519) e valores monetários (4.224) aparecem com volume consistente; leis (1.789) e órgãos judiciários (2.744) ficam uma ordem de magnitude abaixo — embora sejam, conceitualmente, formas de referência mais robustas. A correção padrão opera por apontamento de número, data ou fonte de imprensa. Referência ao texto de lei ou processo: exceção, não regra.

6.3.4 Concentração de menções e topologia de cauda longa

A distribuição das menções dentro de cada tipo segue um padrão de cauda longa pronunciado. A Tabela 5 reporta, para os cinco tipos mais densos em entidades distintas, quantos referentes únicos absorvem 80% das menções do tipo — uma operacionalização simples da regra 80/20 aplicada ao *corpus* de entidades.

Tabela 5: Concentração das menções por tipo de entidade. *Cauda concentrada*: `ATOR_POLITICO`, `ORGAO_PUBLICO`, `VEICULO_MIDIA` e `PARTIDO` têm menos de 11% dos referentes únicos absorvendo 80% das menções. *Cauda longa*: em `PROGRAMA_PUBLICO`, são necessários cerca de um terço dos programas distintos para alcançar o mesmo limiar.

Tipo	Entidades únicas	Que explicam 80%	% sobre o total
`ATOR_POLITICO`	4.179	257	6,1%
`VEICULO_MIDIA`	3.827	224	5,9%
`PARTIDO`	1.088	97	8,9%
`ORGAO_PUBLICO`	2.181	222	10,2%
`PROGRAMA_PUBLICO`	1.742	588	33,8%

A cauda concentrada em atores políticos, partidos e veículos de mídia é coerente com o que se

sabe sobre a esfera pública digital: alguns poucos nomes monopolizam a atenção, e a checagem reflete essa monopolização. A cauda mais distribuída em programas públicos conta uma história diferente — sugere que a checagem desce a um nível de granularidade institucional que a maior parte dos outros tipos não acompanha: notas sobre Bolsa Família, Auxílio Brasil, Pé-de-Meia, Minha Casa Minha Vida e dezenas de outros programas circulam com volume comparável.

6.3.5 Vizinhança discursiva: o que aparece junto

A análise mais informativa da camada de entidades vem da co-ocorrência: quais entidades aparecem juntas em uma mesma nota. Filtrando entidades com pelo menos cinco notas associadas (5.537 entidades distintas), o *pipeline* conta 63.587 pares distintos de co-ocorrência. A leitura empírica desses pares é o que permite mapear, em escala observacional, as coalizões discursivas em torno das quais as notas se organizam.

A leitura agregada do *corpus* de menções (Figuras 5 e 6) já dá o desenho geral das centralidades: na esfera política propriamente dita, Bolsonaro e Lula dividem a posição central com volumes praticamente equivalentes; na esfera não-política, a ocupação se distribui entre figuras empresariais (Elon Musk), de entretenimento (Taylor Swift, Anitta) e de eventos midiáticos pontuais (caso Choquei).

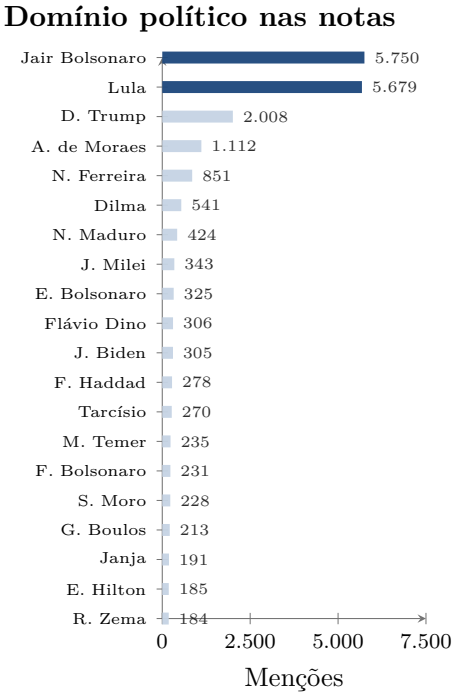

Figura 5: *Top* 20 atores políticos.

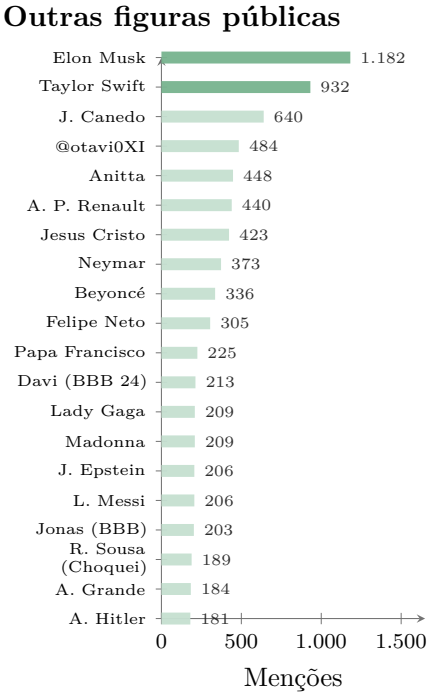

Figura 6: *Top* 20 figuras públicas não-políticas.

No plano de atores, Lula lidera (747 co-ocorrências), seguido por Trump (182), Alexandre de Moraes (114), Michel Temer (107), Dilma Rousseff (91): a vizinhança de Bolsonaro no *corpus* é composta, quase inteiramente, pelo mesmo eixo polarizador que o cerca no debate público. No institucional, STF (239) e PF (230) aparecem com volumes quase idênticos — reflexo da centralidade das disputas em torno do Estado de Direito no período coberto. Na mídia, g1 (349), *cnnbrasil* (228) e *estadão* (85) são os veículos co-citados com mais frequência, revelando

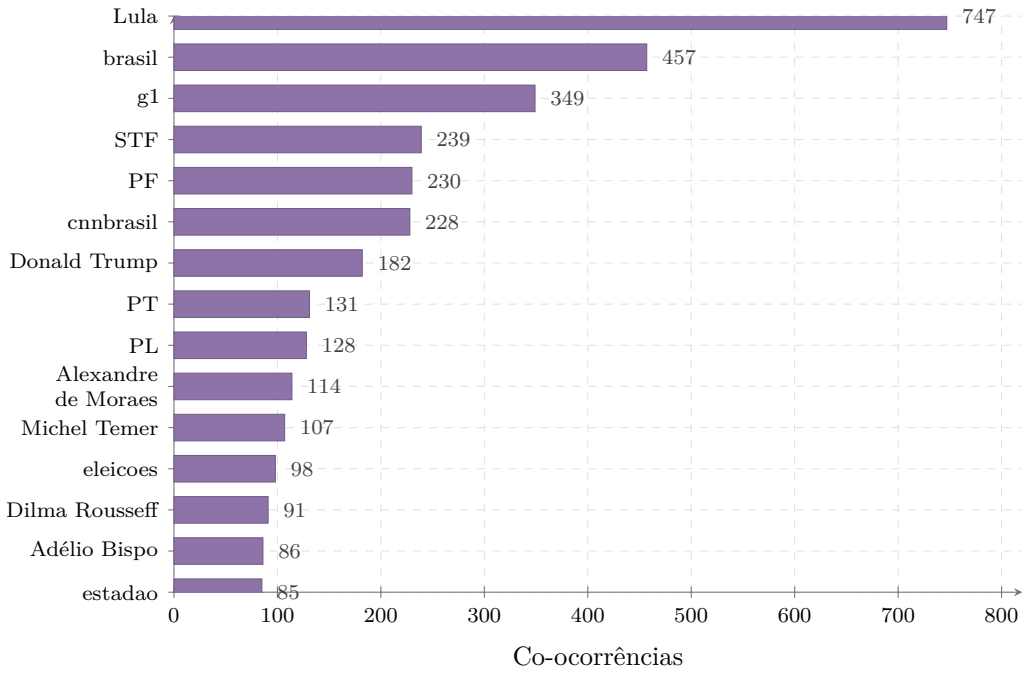

Figura 7: *Top* 15 entidades que mais co-ocorrem com Jair Bolsonaro nas notas em português, ordenadas por número de notas em que aparecem em conjunto. Lula encabeça a vizinhança discursiva (747 co-ocorrências), seguido pela menção de *Brasil* como localizador (457) e por veículos de mídia tradicional (*g1* com 349, *cnnbrasil* com 228), instituições do Estado de Direito (*STF* com 239, *PF* com 230) e outros atores políticos (Donald Trump, Alexandre de Moraes, Michel Temer, Dilma Rousseff).

a arquitetura de evidência mobilizada quando se contesta ou contextualiza alegações sobre o ex-presidente.

O padrão é replicável para qualquer entidade central; Bolsonaro é apenas a entrada de maior densidade. Combinado com a modelagem temática da Seção 6.1, o mapa de vizinhanças entrega uma leitura referencial das disputas: tópicos descrevem o eixo da controvérsia, vizinhanças descrevem o elenco que a constitui.

6.3.6 Domínios e categorização de fontes

Os 174.625 *spans* de URL foram canonicalizados a domínios e classificados em cinco categorias: mídia tradicional brasileira, mídia internacional, plataformas digitais, fontes oficiais e governamentais, e outros (residual).

A distribuição confirma o que a contagem de `VEICULO_MIDIA` já sugeria: a maior parte das URLs aponta para mídia tradicional brasileira, com participação menor e consistente de fontes oficiais e plataformas digitais. A *crowdsourced fact-checking* brasileira não produz evidência primária — referencia, em larga maioria, a infraestrutura do jornalismo institucional. Não é uma crítica; é um dado estrutural. Mas tem implicação direta para qualquer discussão normativa: a capacidade do sistema de operar independentemente dessa infraestrutura é, pelos dados, bastante limitada.

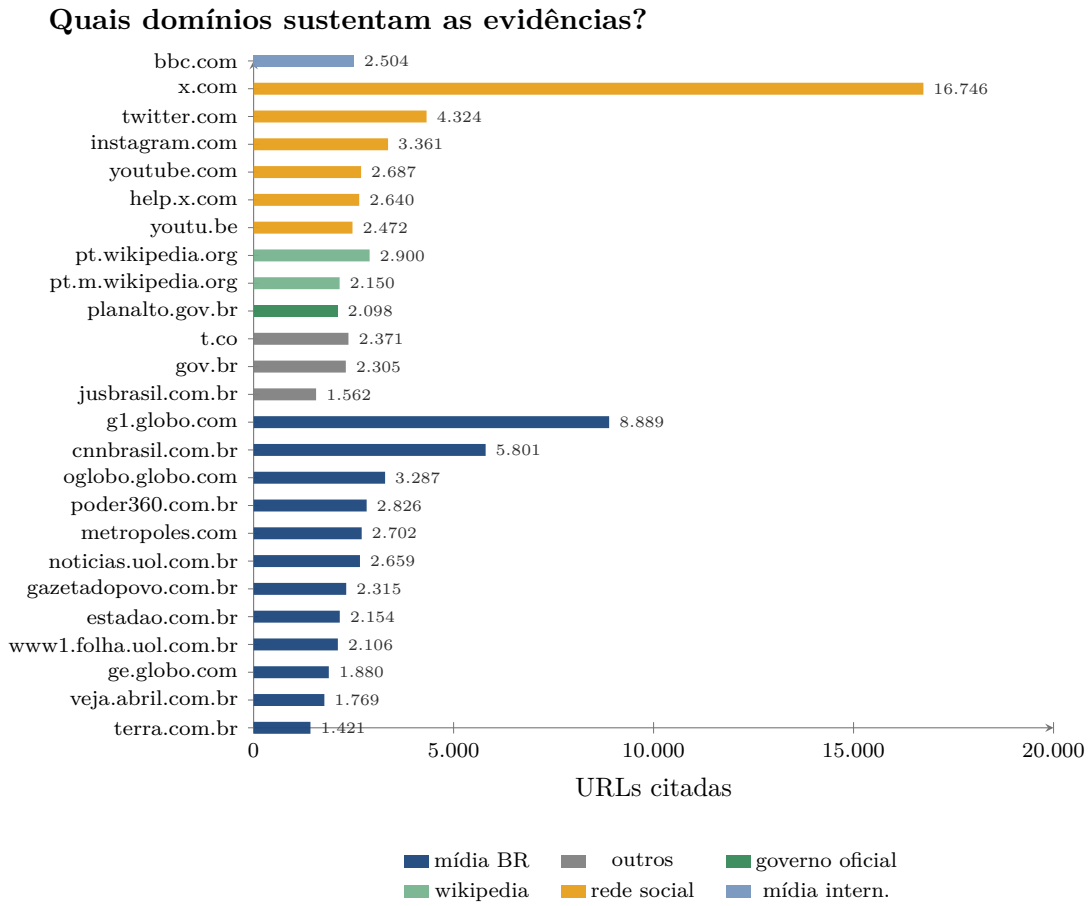

Figura 8: *Top* 25 domínios mais citados, coloridos por categoria de fonte.

6.3.7 O que essa camada acrescenta ao argumento

A análise de entidades cumpre três funções no argumento. Primeiro, dá visibilidade empírica à infraestrutura de evidência: URL é traço estrutural, jornalismo institucional ancora a maior parte da evidência externa, referência jurídica formal é exceção. Segundo, operacionaliza a camada referencial das disputas: ao mapear quem é nomeado e em que vizinhança, oferece uma leitura concreta das coalizões em torno das quais o consenso é negociado — ou não. Terceiro, fornece a granularidade que as Seções 7 e 8 vão precisar: quando o classificador de utilidade penaliza estatisticamente notas que mencionam figuras políticas, a pergunta relevante não é se penaliza, mas quais figuras, em que vizinhança, e com que assimetria — distinção que só a camada referencial torna operável.

6.4 Comparação consolidada dos classificadores de utilidade

A Tabela 6 resume os resultados das trilhas sobre o recorte supervisionado de 19.843 notas com decisão da comunidade (CRH ou CRNH), agregados sobre as cinco dobras de validação cruzada. Para A1, A2 e A3, os números vêm do *notebook* de *baseline*, executado com `GroupKFold(5)`. Para ES *frozen*, FT-Solo, GBDT-Min e GBDT-Full, vêm do *notebook* contextual, executado com `StratifiedGroupKFold(5, shuffle=True, random_state=42)`. A comparação entre os dois grupos deve ser lida como referência aproximada — os *splits* não são idênticos.

Tabela 6: Métricas dos modelos de utilidade. O valor antes do ± corresponde à métrica OOF global; o desvio corresponde à variação entre dobras. A3 contém *target leakage* deliberado e serve apenas como teto operacional. Em **negrito**, melhores resultados sem *leakage*.

Modelo	ROC-AUC OOF	MCC OOF	PR-AUC minoritária	F1
A1 — TF-IDF + LR (*baseline*)	0,8956 ± 0,0029	0,5747 ± 0,0057	0,7812 ± 0,0114	0,8997 ± 0,0012
A2 — Explicativa (+ *tweet*)	0,8957 ± 0,0030	0,5763 ± 0,0070	0,7820 ± 0,0110	0,9000
ES *frozen* — *stacking* Qwen + TF-IDF + tabular	0,9120 ± 0,0035	0,5943 ± 0,0161	0,8158 ± 0,0105	0,8936 ± 0,0030
FT-Solo — Qwen3-4B + LoRA	0,9219 ± 0,0060	**0,6380 ± 0,0375**	0,8296 ± 0,0171	**0,9100 ± 0,0182**
GBDT-Tabular-Min — XGBoost + *proba* FT-Solo	0,9194 ± 0,0065	0,6320 ± 0,0469	0,8320 ± 0,0185	0,9082 ± 0,0225
GBDT-Tabular-Full — XGBoost + 3 *probas* OOF	**0,9251 ± 0,0064**	0,6260 ± 0,0377	**0,8408 ± 0,0174**	0,9042 ± 0,0179
A3 — Operacional (`coreNoteIntercept`, *leakage*)	0,9981 ± 0,0000	0,9197 ± 0,0080	0,9936 ± 0,0020	0,9812

O texto do *tweet*-fonte não adiciona sinal discriminativo relevante ao *baseline* lexical: A1 e A2 ficam virtualmente idênticos (ROC-AUC 0,8956 vs. 0,8957). A nota já carrega a maior parte do contexto necessário para a decisão.

A passagem de TF-IDF para sinais semânticos do Qwen melhora a ordenação de modo consistente: ES *frozen* sobe para 0,9120 (+0,0164), FT-Solo para 0,9219 (+0,0263). Há sinal contextual além da superfície lexical — estilo de correção, formulação impessoal, estrutura argumentativa, marcas de evidência.

Qual é o melhor modelo depende da métrica priorizada. Para ranqueamento global: GBDT-Full, com ROC-AUC 0,9251 e PR-AUC minoritária 0,8408. Para decisão binária no limiar 0,5: FT-Solo, com MCC 0,6380 e F1 0,9100. Não há vencedor único. GBDT-Full ordena melhor; FT-Solo decide melhor no corte padrão.

A diferença GBDT-Min para GBDT-Full é real mas moderada (+0,0057 em ROC-AUC) — os sinais adicionais contribuem para ranqueamento fino, mas o Min preserva MCC e F1 ligeiramente superiores, com considerável redução de complexidade.

O teto operacional com *leakage* (ROC-AUC 0,9981) confirma que `coreNoteIntercept` é, na prática, uma variável muito próxima da decisão final do sistema. Isso valida o fluxo de dados e dimensiona o sinal que permanece inacessível quando se usa apenas texto, tema e variáveis derivadas.

Para a tarefa operacional sobre as 142.051 notas (incluindo aquelas em NEEDS_MORE_RATINGS), o *baseline* TF-IDF atinge ROC-AUC 0,8228 ± 0,004 e PR-AUC 0,3840 — acima do *baseline* trivial de proporção (0,11), mas sensivelmente abaixo dos valores do classificador central. A leitura deve ser cautelosa: o rótulo mistura rejeição da comunidade com ausência de avaliação. O modelo serve, ainda assim, para ranquear as 122.208 notas em NEEDS_MORE_RATINGS por probabilidade prevista de utilidade — uma fila de atenção heurística que o sistema não produz por si só.

6.5 Variáveis estruturais que mais pesam

Para isolar a contribuição das variáveis manuais — que tendem a ser dominadas pelo TF-IDF quando ambos coexistem —, ajustamos uma regressão logística só com as 20 variáveis tabulares e a codificação de tema, e medimos a importância de cada variável por permutação sobre o conjunto de validação. Esse modelo só-tabular atinge ROC-AUC de 0,8164 e MCC de 0,4321, indicando que uma parte substancial do desempenho do *baseline* TF-IDF é atingível sem nenhum sinal lexical fino.

Tabela 7: *Permutation importance* das 15 variáveis tabulares mais relevantes no classificador só-tabular. As importâncias representam o decréscimo médio em ROC-AUC quando os valores da variável são permutados aleatoriamente.

#	Variável	Importância média	Desvio
1	`tema_modelo`	0,1079	0,0024
2	`tem_url`	0,0466	0,0011
3	`n_chars_tweet`	0,0316	0,0013
4	`n_words_nota`	0,0166	0,0013
5	`qtd_links_nota`	0,0121	0,0021
6	`n_chars_nota`	0,0111	0,0014
7	`ratio_nota_tweet_chars`	0,0073	0,0010
8	`qtd_evidencia_v3`	0,0068	0,0024
9	`n_words_tweet`	0,0062	0,0013
10	`tem_evidencia_v3`	0,0059	0,0006
11	`menciona_figura_polit`	0,0059	0,0010
12	`mean_words_per_sentence`	0,0040	0,0010
13	`nota_desnecessaria`	0,0023	0,0006
14	`menciona_instituicao`	0,0014	0,0005
15	`qtd_numeros_nota`	0,0012	0,0005

O macrotema é, de longe, a variável mais importante — carrega cerca de duas vezes mais sinal preditivo do que a segunda colocada. Isso é coerente com o que a Seção 6.2 mostrou: apostas *online* e crise política não são apenas temas distintos, são regimes de consenso distintos. A URL é o segundo sinal: notas com fonte ostensivamente exibida são muito mais propensas a atingir utilidade, diferença que o modelo captura mesmo sem qualquer modelagem do conteúdo da fonte. Marcadores discursivos importam na terceira posição — abertura notarial, linguagem factual, evidência citada discriminam de notas que declaram a publicação desnecessária ou a tratam como mera opinião.

`menciona_figura_polit` tem taxa de utilidade inferior à das notas sem menção a políticos. A diferença bruta é moderada; o que não é moderado é sua interpretação. O classificador não introduz esse padrão: aprende o que já está no rótulo. E o rótulo reflete uma propriedade do sistema — quando a utilidade depende de aceitabilidade transversal entre avaliadores divergentes, temas que ativam clivagens partidárias pagam um pedágio sistemático que nenhuma qualidade editorial elimina.

6.6 Interpretabilidade — o que o classificador aprende, e o que isso revela sobre o sistema

A análise de interpretabilidade foi construída em três camadas, escolhidas para responderem a perguntas distintas e mutuamente irredutíveis. A primeira examina o que o GBDT-Full usa, no agregado: que variáveis pesam, em que direção e com que magnitude, ao longo das 19.843 notas do recorte supervisionado. A segunda examina a geometria do espaço aprendido pelo FT-Solo: como as 2.560 dimensões do *encoder fine-tunado* organizam as notas, e se essa organização guarda relação com o rótulo. A terceira examina quais palavras puxam a decisão local: que *tokens*, dentro de cada nota, contribuíram com mais peso para a probabilidade prevista. As três camadas conversam — e a tensão entre elas é o achado mais informativo do estudo.

6.6.1 O caso que motiva a análise

Antes da descrição metodológica, o caso que melhor sintetiza a leitura final dos resultados. A nota a seguir recebeu probabilidade 0,852 de utilidade pelo FT-Solo. Seu rótulo real é não-útil.

> *"Não existe nenhuma relação entre o encontro do Embaixador de Israel com o ex-presidente Bolsonaro e a liberação de brasileiros da Faixa de Gaza. A própria Embaixada informou que não convidou o ex-presidente Jair Bolsonaro a uma reunião com parlamentares."*

URL de veículo institucional (`g1.globo.com`). Abertura notarial: "não existe nenhuma relação". Fonte primária com voz ativa: "a própria Embaixada informou". Em todos os marcadores estruturais que o classificador aprendeu a associar à utilidade, a nota se comporta como útil. Seus seis vizinhos mais próximos no espaço de *embeddings* do FT-Solo têm rótulo positivo, todos. O modelo prevê o que o texto sugere. A comunidade decidiu de outra maneira.

Isso não é falha do modelo. É evidência de uma propriedade do rótulo: utilidade no Community Notes não é função do texto isolado da nota — é função do texto condicionada à composição ideológica dos avaliadores e ao assunto subjacente. Em casos politicamente carregados, notas factualmente sólidas podem permanecer em NEEDS_MORE_RATINGS indefinidamente. Não porque sejam ruins. Porque a aceitação cruzada entre grupos não se forma.

6.6.2 SHAP global — onde está o sinal

A primeira camada usa TreeSHAP sobre o GBDT-Tabular-Full retreinado no recorte completo das 19.843 notas. Para a leitura agregada, calculamos o valor absoluto médio do SHAP por *feature*, em todas as notas e em todas as árvores.

`proba_ft_solo` ocupa o topo com folga. Das dez *features* mais importantes: uma é a probabilidade do FT-Solo, duas são as outras probabilidades OOF (`proba_es_frozen`, `proba_e2_lexical`), uma é o *target encoding* do tema, seis são variáveis tabulares. Em magnitude, `proba_ft_solo` responde por uma fração desproporcional do sinal — o GBDT não está aprendendo a tarefa do zero; está calibrando a probabilidade do FT-Solo com ajustes marginais vindos das variáveis estruturais.

A consequência operacional é direta: a diferença entre GBDT-Full e FT-Solo isolado é de 0,0032 em ROC-AUC. O GBDT, neste estudo, serve como lente interpretativa sobre o que o FT-Solo já produz — não como aprendiz primário. Em produção, FT-Solo + cabeça linear entregaria desempenho equivalente com uma fração da complexidade.

6.6.3 Geometria do espaço — o que o FT-Solo organiza

A segunda camada projeta as 19.843 *embeddings* do FT-Solo (2.560 dimensões) em duas dimensões via UMAP, com métrica cosseno e parâmetros padrão.

As duas projeções coincidem com folga: região central do *embedding space* concentra notas de alta probabilidade prevista e alta proporção de utilidade real; regiões periféricas — notas curtas, sem fontes — concentram as não-úteis. Na inspeção dos seis vizinhos mais próximos para 60 notas amostradas, os vizinhos compartilham o rótulo do nó central na maioria dos casos. O

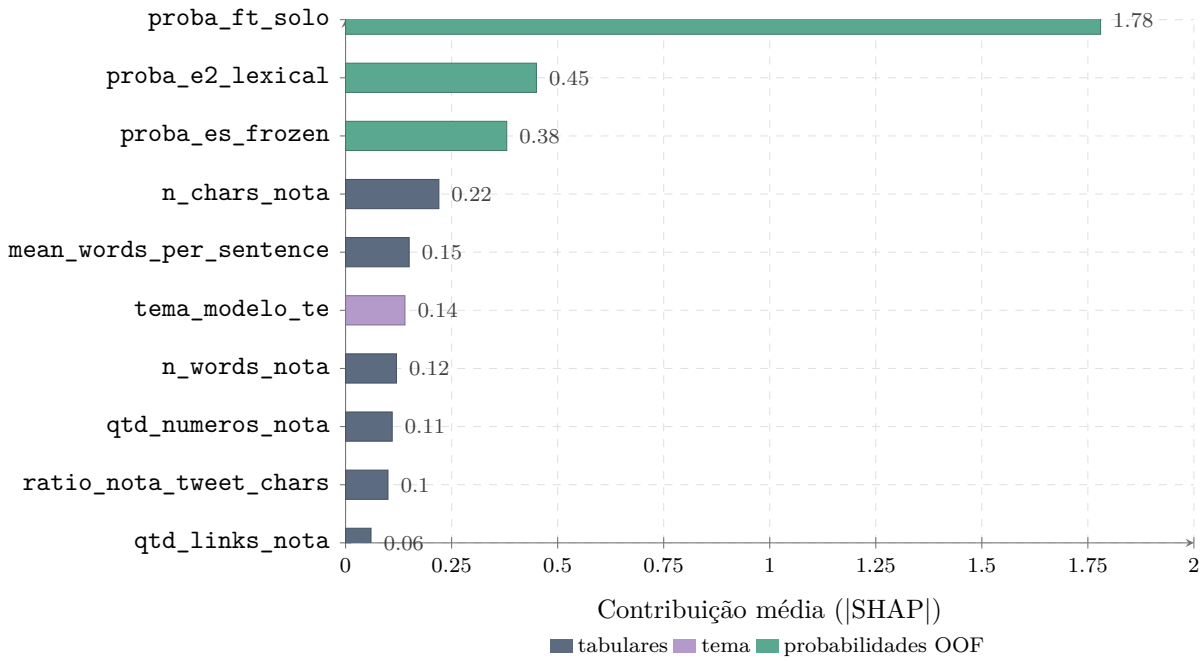

Figura 9: *Top* 10 *features* do GBDT-Full por importância global (média do valor absoluto SHAP, calculado via TreeSHAP). `proba_ft_solo` ocupa o topo do *ranking* com folga, seguida pelas outras duas probabilidades OOF (`proba_e2_lexical`, `proba_es_frozen`). As variáveis tabulares estruturais (`n_chars_nota`, `mean_words_per_sentence`, `n_words_nota`, entre outras) ocupam as posições intermediárias, e o *target encoding* do tema (`tema_modelo_te`) aparece em sexto lugar. Cores: verde para probabilidades de outros modelos, lilás para tema, cinza-azulado para tabulares.

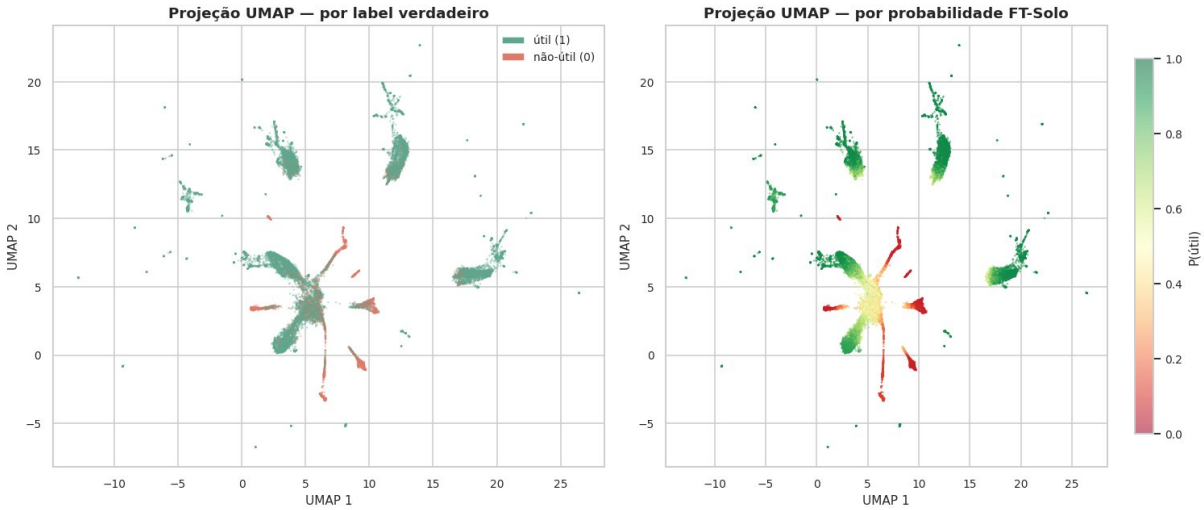

Figura 10: Projeção UMAP das 19.843 *embeddings* do FT-Solo (2.560 dimensões reduzidas a 2). À esquerda, coloração pelo *label* verdadeiro (verde = útil, vermelho = não-útil); à direita, pela probabilidade prevista pelo FT-Solo (verde escuro = $P(\text{útil}) \to 1$, vermelho = $P(\text{útil}) \to 0$). As duas projeções coincidem com folga: regiões periféricas concentram notas não-úteis com probabilidade prevista baixa, enquanto a região central agrega notas úteis com alta probabilidade. A divergência sistemática se concentra na zona central de fronteira, onde *clusters* visualmente coerentes são cortados pelo rótulo — assinatura geométrica da aceitabilidade algorítmica diferencial discutida em 6.6.

FT-Solo aprendeu uma noção operacional de *parecer útil* que tem assinatura geométrica — não correlações de superfície.

A exceção é o caso da Seção 6.6. A nota sobre o embaixador de Israel habita um *cluster* de

notas com rótulo positivo; seu próprio rótulo é negativo. O FT-Solo agrupou notas com a mesma fisionomia textual; a comunidade as separou por critério que o texto não registra. Divergência sistemática. Com assinatura espacial.

6.6.4 SHAP textual — quais palavras decidem

A terceira camada aplica o `PartitionExplainer` do SHAP sobre o FT-Solo (carregado a partir do *adapter* LoRA do *fold* 1), com tokenização por palavra e mascaramento padrão. Pelo custo do método (∼12 segundos por nota em GPU A100), amostramos 60 notas em cinco estratos balanceados: alta confiança útil (*proba* $> 0{,}88$, *label* 1), alta confiança não-útil (*proba* $< 0{,}12$, *label* 0), fronteira (*proba* entre 0,45 e 0,55), falsos positivos (*proba* $> 0{,}80$, *label* 0) e falsos negativos (*proba* $< 0{,}20$, *label* 1).

Agregados sobre o estrato de alta confiança útil, os *tokens* mais associados à classe positiva formam três grupos. Estrutural: fragmentos de URL (`https://`, `com/`, `ghtml`, `2024/`) aparecem repetidamente no topo. Discursivo: "falso", "conforme", "segundo", "informou" — o estilo notarial de correção afirmativa. Temático: nomes de instituições e termos de domínios em que a checagem opera com regularidade (anúncio, apostas).

Que `https://` seja o *token* de maior peso positivo em várias notas é, ao mesmo tempo, um achado e uma advertência. Confirma que a presença de fonte externa é sinal confiável de utilidade. Mas o modelo aprende a presença ostensiva de URL, não a qualidade do que ela contém. A triangulação com o TF-IDF (Seção 6.6.5) mostra que "falso", "errado", "conforme" têm coeficientes positivos altos mesmo na lente puramente lexical — há sinal genuíno de marcação de correção além do sinal estrutural de presença de fonte. Mas a advertência permanece: para o sistema, evidência é, em parte, fonte ostensivamente exibida; a qualidade dessa fonte fica fora do escopo do que o classificador — e provavelmente a comunidade — avalia.

Do lado negativo: marcadores de subjetividade ("opinião", "claramente", "certo", "desnecessária") e, em determinados contextos, nomes de figuras políticas. Notas que tratam a publicação-alvo como mera opinião, ou que declaram a nota desnecessária, são consistentemente rejeitadas. Independentemente do mérito do que observam.

6.6.5 Triangulação lexical e galeria curada

Para validar os padrões do SHAP textual sobre uma base maior, ajustamos uma regressão logística sobre TF-IDF (`min_df=5`, bigramas, `class_weight` balanced) no recorte inteiro, e inspecionamos os 30 termos com maior coeficiente positivo e negativo.

A leitura confirma o padrão do SHAP: termos como "é falso", "na verdade", "conforme", "segundo", e nomes de veículos de imprensa institucional dominam o lado positivo; termos como "opinião", "claramente", "desnecessária", "sátira" dominam o lado negativo. A coincidência entre as duas lentes — uma textual contextual (FT-Solo + SHAP), outra lexical exata (TF-IDF + LR) — sugere que os dois métodos estão capturando o mesmo fenômeno por caminhos distintos, o que aumenta a confiabilidade da leitura.

A galeria curada — seis casos representativos, um por estrato — acompanha o portal de resultados como conjunto de *cards* navegáveis. Cada *card* apresenta o texto da nota, o *score* do FT-Solo, a concordância entre os modelos, os principais *tokens* positivos e negativos do SHAP, e

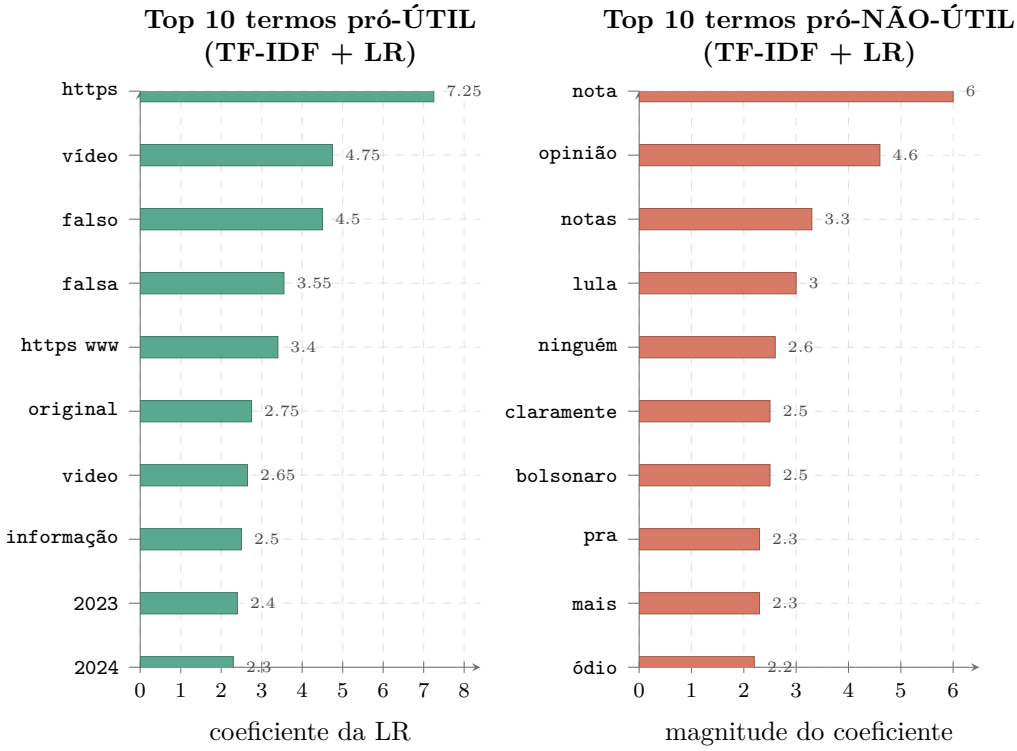

Figura 11: Coeficientes da regressão logística sobre TF-IDF (`min_df=5`, bigramas, `class_weight` balanceado), ajustada sobre o recorte supervisionado completo. À esquerda, os 10 termos com maior coeficiente positivo (pró-útil), após remoção de stopwords: dominados por marcadores estruturais de fonte (`https`, `https www`), por marcadores discursivos de correção factual (`falso`, `falsa`, `original`) e por *tokens* datais (`2023`, `2024`). À direita, os 10 termos com maior coeficiente negativo em valor absoluto (pró-não-útil): dominados por marcadores de subjetividade (`opinião`, `claramente`, `ódio`), por meta-comentários sobre a plataforma (`nota`, `notas`) e por nomes de figuras políticas em determinados contextos (`lula`, `bolsonaro`).

os seis vizinhos semânticos. A galeria não é ornamental: é a unidade interpretativa básica da interface pública de inspeção, onde qualquer usuário pode submeter o texto de uma nota e obter a probabilidade prevista, a explicação textual e a vizinhança semântica em tempo real.

7 Discussão

7.1 O que os achados sustentam — e o que não sustentam

Os três eixos empíricos apontam para a mesma direção. O estado predominante do Community Notes em português não é o consenso — é a espera por ele: 82% das notas permanecem em NEEDS_MORE_RATINGS, e essa suspensão não é neutra. Distribui-se de forma sistematicamente desigual entre tópicos. Domínios operacionais (apostas *online*, golpes financeiros, conteúdo gerado por IA) atravessam o filtro com relativa regularidade porque a estrutura da alegação admite ancoragem objetiva, reduzindo o espaço de divergência entre avaliadores politicamente distintos. *Crise política* acumula suspensão acima de 85%. E o macrotema é, de longe, o sinal preditivo mais forte sobre a utilidade — acima da presença de URL, dos marcadores discursivos, do comprimento da nota. O filtro de consenso não é apenas sensível ao conteúdo: é sensível ao campo semântico em que a disputa se insere.

A interpretabilidade torna esse padrão observável na geometria do classificador. Notas que o modelo prevê como úteis — URL institucional, abertura notarial, fonte primária com voz ativa — podem ser rejeitadas quando o assunto ativa clivagens políticas. O caso da nota sobre o embaixador de Israel (Seção 6.6) não é anedota: é a assinatura de uma propriedade estrutural do rótulo. Utilidade no Community Notes não é função do texto isolado; é função do texto condicionada à composição ideológica dos avaliadores e ao campo político ativado. É isso que o conceito de aceitabilidade algorítmica nomeia — e o que distingue esse sistema de um mecanismo de verificação factual.

O artigo descreve esse funcionamento. Não avalia se o sistema deveria existir, se deveria ser substituído por verificadores profissionais, ou se a lógica de *bridging-based ranking* é normativamente desejável em sociedades polarizadas. Qualquer leitura prescritiva extrapola o que os dados sustentam.

7.2 Limitações

Quatro limitações merecem registro. *Cobertura*: o *snapshot* público não contém todas as notas que já existiram no sistema — cerca de 12% dos registros de *ratings* e de histórico de *status* eram órfãos e foram filtrados; os achados se aplicam ao universo observável, não ao histórico completo. *Detecção de idioma*: o recorte em português depende da classificação por fastText com limiar 0,80 validado por amostragem estratificada; há erro residual documentado, e notas em variantes lusófonas periféricas podem ter sido excluídas por sub-representação no modelo. *Recorte temporal*: o *snapshot* é de 7 de abril de 2026; mudanças posteriores no algoritmo, na regulação brasileira ou na composição dos avaliadores ativos não são capturadas. *Interpretabilidade textual*: a análise via SHAP `PartitionExplainer` usou o *adapter* LoRA de uma única dobra, dado o custo computacional do método (∼12 segundos por nota); a variabilidade entre os cinco *adapters* não está quantificada nessa camada.

8 Conclusão

O Community Notes é frequentemente apresentado como alternativa à moderação centralizada: uma forma distribuída de produzir verdade pública pelo concurso de avaliadores anônimos e ideologicamente diversos. O caso brasileiro mostra que essa apresentação descreve apenas parte do mecanismo. A operação efetiva do sistema produz notas publicadas e visíveis, sim — mas produz, ao mesmo tempo, uma massa muito maior de notas em suspensão: mais de 80% do material, distribuída de forma sistematicamente desigual, concentrada nos assuntos politicamente carregados.

As 142 mil notas em português, cruzadas com a modelagem temática, o classificador interpretável e a geometria do espaço de *embeddings fine-tunados*, sustentam uma leitura específica: o que o sistema chama de utilidade é, mais precisamente, aceitabilidade algorítmica entre avaliadores divergentes. Essa aceitabilidade tem correlação parcial com qualidade editorial e acurácia factual. Não se reduz a elas. Notas factualmente sólidas, bem fundamentadas, estilisticamente consistentes podem permanecer indefinidamente em NEEDS_MORE_RATINGS quando o assunto ativa divisões partidárias profundas. A evidência disso não é apenas estatística agregada — é geométrica,

observável no espaço do classificador.

Chamar isso de *fact-checking* distribuído é impreciso. Chamar de consenso democrático, mais ainda. O que o sistema faz é governar algoritmicamente a aceitabilidade da verdade pública — e essa é, ao mesmo tempo, uma descrição técnica e uma distinção com consequências. Para a literatura empírica sobre desinformação, a análise oferece uma ancoragem brasileira para um debate até aqui dominado pelo recorte anglófono. Para a discussão pública sobre moderação, oferece uma linguagem que permite discutir o sistema pelas propriedades que de fato tem: uma infraestrutura específica, com comportamentos específicos, sobre a qual perguntas normativas ganham precisão apenas quando se sabe o que ela realmente faz.

Declaração de conflito de interesse

O autor declara não haver conflito de interesse financeiro, pessoal, institucional ou acadêmico que possa ter influenciado a concepção do estudo, a análise dos dados, a interpretação dos resultados ou a redação do manuscrito. Esta declaração segue as orientações do Committee on Publication Ethics (COPE).

Declaração de disponibilidade de dados da pesquisa

Todo o conjunto de dados de apoio aos resultados deste estudo foi disponibilizado no Hugging Face Hub, sob a licença declarada no repositório, e pode ser acessado em `histlearn/community-notes-br` (`https://huggingface.co/datasets/histlearn/community-notes-br`). O portal de resultados está disponível em `communitynotesbr.netlify.app`, e a interface pública de inspeção do classificador com explicações SHAP está disponível em `histlearn-communitynotesbr.hf.space`. O *pipeline* de processamento, os scripts de análise e as instruções de reprodução acompanham o repositório do *dataset*.

Declaração de uso de IA

Foram utilizadas ferramentas de inteligência artificial em etapas específicas e documentadas do trabalho: (i) na rotulagem semântica dos tópicos identificados por BERTopic, conforme descrito na seção de métodos; (ii) na revisão ortográfica e gramatical do texto; e (iii) como apoio à geração de hipóteses interpretativas sobre os resultados quantitativos, posteriormente validadas pelo autor contra a evidência empírica. As decisões metodológicas, a análise dos dados e a redação substantiva são de responsabilidade exclusiva do autor.

Referências

[1] ALLEN, Jennifer; MARTEL, Cameron; RAND, David G. Birds of a feather don't fact-check each other: partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In: *CHI Conference on Human Factors in Computing Systems*, 2022, New Orleans. Proceedings [...]. New York: ACM, 2022. p. 1–19. DOI: 10.1145/3491102.3502040.

[2] AYOUB, Jackie; YANG, X. Jessie; ZHOU, Feng. Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management*, v. 58, n. 4, art. 102569, 2021. DOI: 10.1016/j.ipm.2021.102569.

[3] BORENSTEIN, Nadav; WARREN, Greta; ELLIOTT, Desmond; AUGENSTEIN, Isabelle. Can Community Notes replace professional fact-checkers? In: *63rd Annual Meeting of the Association for Computational Linguistics*, 2025, Vienna. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers. Stroudsburg: Association for Computational Linguistics, 2025. p. 535–552.

[4] BOZARTH, Lia; IM, Jane; QUARLES, Christopher; BUDAK, Ceren. Wisdom of two crowds: misinformation moderation on Reddit and how to improve this process — a case study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction*, v. 7, n. CSCW1, art. 155, p. 1–33, 2023. DOI: 10.1145/3579631.

[5] CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (org.). *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*. 4. ed. São Carlos: BPLN, 2026. 1314 p. 3 v. ISBN 978-65-02-00178-3. Disponível em: http://brasileiraspln.com/livro-pln-4ed/. Acesso em: 28 abr. 2026.

[6] CHUAI, Yuwei; TIAN, Haoye; PRÖLLOCHS, Nicolas; LENZINI, Gabriele. Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, v. 8, n. CSCW2, art. 428, 2024. DOI: 10.1145/3686967.

[7] GILLESPIE, Tarleton. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press, 2018.

[8] GORWA, Robert. What is platform governance? *Information, Communication & Society*, v. 22, n. 6, p. 854–871, 2019. DOI: 10.1080/1369118X.2019.1573914.

[9] GORWA, Robert; BINNS, Reuben; KATZENBACH, Christian. Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society*, v. 7, n. 1, 2020. DOI: 10.1177/2053951719897945.

[10] GROOTENDORST, Maarten. BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint*, arXiv:2203.05794, 2022. DOI: 10.48550/arXiv.2203.05794.

[11] JIANG, Fan; ZHAO, Yu; LYU, Chenyang; SHI, Tianqi; DU, Yichao; JIANG, Feihu; WANG, Longyue; LUO, Weihua. Marco-MoE: open multilingual mixture-of-expert language models with efficient upcycling. 2026.

[12] JOULIN, Armand; GRAVE, Edouard; BOJANOWSKI, Piotr; MIKOLOV, Tomas. Bag of tricks for efficient text classification. *arXiv preprint*, arXiv:1607.01759, 2016a. DOI: 10.48550/arXiv.1607.01759.

[13] JOULIN, Armand; GRAVE, Edouard; BOJANOWSKI, Piotr; DOUZE, Matthijs; JÉGOU, Hervé; MIKOLOV, Tomas. FastText.zip: compressing text classification models. *arXiv preprint*, arXiv:1612.03651, 2016b. DOI: 10.48550/arXiv.1612.03651.

[14] KLONICK, Kate. The new governors: the people, rules, and processes governing online speech. *Harvard Law Review*, v. 131, n. 6, p. 1598–1670, 2018.

[15] MEN, Xiuping; MARIANO, Vladimir Y. Explainable fake news detection based on BERT and SHAP applied to COVID-19. *International Journal of Modern Education and Computer Science*, v. 16, n. 1, p. 11–22, 2024. DOI: 10.5815/ijmecs.2024.01.02.

[16] OVADYA, Aviv; THORBURN, Luke. *Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance*. Knight First Amendment Institute at Columbia University, 2023.

[17] PATEL, Jainee; BHATT, Chintan M.; TRIVEDI, Himani; NGUYEN, Thanh Thi. Misinformation detection using large language models with explainability. *arXiv preprint*, arXiv:2510.18918, 2025. DOI: 10.48550/arXiv.2510.18918.

[18] PICCARDI, Tiziano; SAVESKI, Martin; JIA, Chenyan; HANCOCK, Jeffrey T.; TSAI, Jeanne L.; BERNSTEIN, Michael S. Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, v. 390, n. 6776, eadu5584, 2025. DOI: 10.1126/science.adu5584.

[19] PRÖLLOCHS, Nicolas. Community-based fact-checking on Twitter's Birdwatch platform. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 16, n. 1, p. 794–805, 2022. DOI: 10.1609/icwsm.v16i1.19335.

[20] RORTY, Richard. *Contingência, ironia e solidariedade*. Tradução de Vera Ribeiro. São Paulo: Martins, 2007.

[21] SAEED, Mohammed; TRAUB, Nicolas; NICOLAS, Maelle; DEMARTINI, Gianluca; PAPOTTI, Paolo. Crowdsourced fact-checking at Twitter: how does the crowd compare with experts? In: *31st ACM International Conference on Information and Knowledge Management*, 2022, Atlanta. Proceedings [...]. New York: ACM, 2022. DOI: 10.1145/3511808.3557279.

[22] THORBURN, Luke; POLUKAROV, Maria; VENTRE, Carmine. Societal sorting as a systemic risk of recommenders. In: *18th ACM Conference on Recommender Systems*, 2024, Bari. RecSys 2024: Proceedings of the 18th ACM Conference on Recommender Systems. New York: ACM, 2024. p. 951–956. DOI: 10.1145/3640457.3688175.

[23] TÖRNBERG, Petter; VALEEVA, Diliara; UITERMARK, Justus; BAIL, Christopher. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint*, arXiv:2310.05984, 2023. DOI: 10.48550/arXiv.2310.05984.

[24] WANG, Liang; YANG, Nan; HUANG, Xiaolong; YANG, Linjun; MAJUMDER, Rangan; WEI, Furu. Multilingual E5 text embeddings: a technical report. *arXiv preprint*, arXiv:2402.05672, 2024. DOI: 10.48550/arXiv.2402.05672.

[25] WIRTSCHAFTER, Valerie; MAJUMDER, Sharanya. Future challenges for online, crowdsourced content moderation: evidence from Twitter's Community Notes. *Journal of Online Trust and Safety*, v. 2, n. 1, 2023. DOI: 10.54501/jots.v2i1.139.

[26] WOJCIK, Stefan; HILGARD, Sophie; JUDD, Nick; MOCANU, Delia; RAGAIN, Stephen; HUNZAKER, M. B. Fallin; COLEMAN, Keith; BAXTER, Jay. Birdwatch: crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint*, arXiv:2210.15723, 2022. DOI: 10.48550/arXiv.2210.15723.

[27] ZARATIANA, Urchade; TOMEH, Nadi; HOLAT, Pierre; CHARNOIS, Thierry. GLiNER: generalist model for named entity recognition using bidirectional transformer. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

2024, Mexico City. Proceedings [...]. Stroudsburg: Association for Computational Linguistics, 2024. p. 5364–5376. DOI: 10.18653/v1/2024.naacl-long.300.

[28] ZHAO, Andy; HU, Haohan. Unveiling strategic governance and user dynamics in Weibo's community-driven content moderation system. *Journal of Quantitative Description: Digital Media*, v. 5, 2025. DOI: 10.51685/jqd.2025.013.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.