Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Detecção de Toxicidade em Linguagem Codificada: um Estudo de Caso em uma Comunidade Gamer do YouTube

Israel Matias do Amaral

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Helen de Cassia Sousa da Costa Lima

COORIENTAÇÃO:
Carlos Henrique Gomes Ferreira

Julho, 2026
João Monlevade–MG

Israel Matias do Amaral

Detecção de Toxicidade em Linguagem Codificada: um Estudo de Caso em uma Comunidade Gamer do YouTube

Orientadora: Helen de Cassia Sousa da Costa Lima

Coorientador: Carlos Henrique Gomes Ferreira

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Julho de 2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Israel Matias do Amaral

Detecção de Toxicidade em Linguagem Codificada: um Estudo de Caso em uma Comunidade *Gamer* do YouTube

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação

Aprovada em 31 de julho de 2026.

Membros da banca

Dra. Helen de Cássia Sousa da Costa Lima - Orientadora - (Departamento de Computação e Sistemas - DECSI da Universidade Federal de Ouro Preto - UFOP)
Dr. Carlos Henrique Gomes Ferreira - Coorientador - (Departamento de Computação - DECOM da Universidade Federal de Ouro Preto - UFOP)
Dr. Humberto Torres Marques Neto - (Departamento de Ciência da Computação da Pontifícia Universidade Católica de Minas Gerais - PUC Minas)
Dr. Josemar Alves Caetano - (Departamento de Computação e Sistemas - DECSI da Universidade Federal de Ouro Preto - UFOP)

Helen de Cássia Sousa da Costa Lima, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 18/08/2026

Documento assinado eletronicamente por **Helen de Cassia Sousa da Costa Lima**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 18/08/2026, às 11:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1167094** e o código CRC **02D123F8**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.013575/2025-46 SEI nº 1167094

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

À minha família, pelo apoio incondicional em cada etapa desta jornada.

Agradecimentos

Agradeço primeiramente à minha família, pelo amor, paciência e suporte ao longo de toda a minha formação acadêmica.

À minha orientadora, Prof.ª Helen de Cassia Sousa da Costa Lima, pela orientação dedicada, pelos ensinamentos e pela confiança depositada neste trabalho.

Aos professores do Departamento de Computação e Sistemas da Universidade Federal de Ouro Preto, que contribuíram para a minha formação e ampliaram minha visão sobre a ciência e a tecnologia.

Aos meus amigos e colegas de curso, pelas trocas de experiências, pelo companheirismo e por tornarem essa trajetória mais leve e significativa.

E a todos que, de alguma forma, contribuíram para a realização deste trabalho.

"If I have seen further it is by standing on the shoulders of Giants."

— Isaac Newton (1642 – 1727)

Resumo

A detecção automática de toxicidade já lida bem com as formas explícitas, mas comunidades podem transmitir conteúdo nocivo por uma linguagem codificada própria, feita de palavras e siglas de aparência cotidiana que ganham sentido nocivo por acordo interno do grupo. Este trabalho avalia quanto os detectores de estado da arte enxergam desse fenômeno em um estudo de caso sobre uma comunidade de jogos do YouTube brasileiro, notoriamente tóxica, cujos chats ao vivo foram coletados em sete canais ao longo de dois meses, num total de 1,74 milhão de mensagens. Foram construídos dois instrumentos. O primeiro é um catálogo dos termos codificados da comunidade, que é o registro escrito de cada termo com o significado que ele tem no uso interno, suas variantes e exemplos reais, levantado a partir da vivência do pesquisador nesses chats e confirmado pela análise estatística das palavras que acompanham cada termo no corpus completo. O segundo é um conjunto de referência de 3.500 mensagens rotuladas à mão, em que cada mensagem tóxica é ainda classificada segundo o modo como a toxicidade chega ao leitor: de forma explícita ou codificada. Sobre essa mesma amostra foram avaliadas três famílias de detectores: um classificador multilíngue de prateleira, modelos ajustados sobre bases de toxicidade em português brasileiro e modelos de linguagem de grande porte instruídos com quantidade crescente de contexto cultural. Sem acesso ao código da comunidade, os detectores encontram de 1,9% a 9,4% das mensagens de toxicidade codificada, contra 48% a 90% das explícitas, e o resultado se repete em arquiteturas e escalas diferentes. Fornecido o catálogo junto com a instrução, os três modelos de linguagem passam a encontrar de 96% a 98% das mensagens codificadas. É a documentação do código que recupera a detecção, e não o modelo, ao passo que acrescentar exemplos genéricos chega a piorar o desempenho. A precisão permanece baixa em todas as configurações testadas, entre 6,9% e 24,6%, de modo que esses modelos erram na maior parte das mensagens que classificam como tóxicas. A moderação eficaz de comunidades com repertório codificado depende de conhecimento cultural documentado, sendo o catálogo de termos codificados uma intervenção viável para melhoria de sistemas de moderação e uma técnica replicável em outros contextos, como o de comunidades que desenvolvam uma linguagem própria.

Palavras-chave: detecção de toxicidade; moderação de conteúdo; linguagem codificada; chat ao vivo; modelos de linguagem de grande porte.

Abstract

Automatic toxicity detection already handles explicit forms well, but communities can convey harmful content through a coded language of their own, made of everyday-looking words and acronyms that acquire harmful meaning by internal agreement of the group. This work assesses how much of that phenomenon state-of-the-art detectors actually see, in a case study of a notoriously toxic Brazilian gaming community on YouTube, whose live chats were collected from seven channels over two months, totaling 1.74 million messages. Two instruments were built. The first is a catalog of the community's coded terms, which is the written record of each term with the meaning it has in internal usage, its variants and real examples, drawn from the researcher's own experience in these chats and confirmed by the statistical analysis of the words that accompany each term in the full corpus. The second is a reference set of 3,500 hand-labeled messages, in which each toxic message is further classified by the way its toxicity reaches the reader: explicit or coded. Three families of detectors were evaluated on this same sample: an off-the-shelf multilingual classifier, models fine-tuned on Brazilian Portuguese toxicity datasets, and large language models prompted with increasing amounts of cultural context. Without access to the community's code, the detectors find 1.9% to 9.4% of the messages carrying coded toxicity, against 48% to 90% of the explicit ones, and the result recurs across different architectures and scales. Given the catalog along with the instruction, the three language models come to find 96% to 98% of the coded messages. It is the documentation of the code that restores detection, and not the model, whereas adding generic examples can even worsen performance. Precision remains low in every configuration tested, between 6.9% and 24.6%, so that these models are wrong about most of the messages they classify as toxic. Effective moderation of communities with a coded repertoire depends on documented cultural knowledge, making the catalog of coded terms a viable intervention for improving moderation systems and a technique replicable in other contexts, such as that of communities that develop a language of their own.

Keywords: toxicity detection; content moderation; coded language; live chat; large language models.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API *Application Programming Interface* (interface de programação de aplicações)

BERT *Bidirectional Encoder Representations from Transformers*

FP Falso positivo

LLM *Large Language Model* (modelo de linguagem de grande porte)

PLN Processamento de Linguagem Natural

PMI *Pointwise Mutual Information* (informação mútua pontual)

SOTA *State of the Art* (estado da arte)

TP *True positive* (verdadeiro positivo)

XLM-R *Cross-lingual Language Model* – RoBERTa (modelo multilíngue)

Sumário

1 Introdução

A moderação de conteúdo em plataformas de vídeo ao vivo opera numa escala que inviabiliza a revisão humana mensagem a mensagem. O chat de uma única transmissão popular produz milhares de mensagens por hora, e cada uma permanece visível por segundos. A tarefa recai, na prática, sobre a detecção automática de toxicidade, área que amadureceu rapidamente na última década, das listas de palavras aos modelos Transformer pré-treinados (FORTUNA; NUNES, 2018; RAMOS et al., 2024). Hoje a detecção da toxicidade explícita é um problema maduro, com desempenho alto mesmo em chat ao vivo e em idiomas com poucos recursos (POOKPANICH; SIRIBORVORNRATANAKUL, 2024; RAMOS et al., 2024).

Esse amadurecimento, porém, provoca uma reação adaptativa: se é a palavra explícita que dispara a moderação, basta substituí-la por outra que o detector não reconheça, e o vocabulário muda enquanto o conteúdo tóxico permanece. A literatura documenta as formas dessa evasão, o discurso de ódio implícito, transmitido por sarcasmo e formas indiretas (ELSHERIEF et al., 2021; OCAMPO et al., 2023), o disfarce humorístico (SCHMID, 2025) e o repertório visual de emojis e símbolos (KIRK et al., 2022b; ZHOU et al., 2025). Há, contudo, uma forma mais radical, a linguagem codificada da comunidade, feita de palavras e siglas de aparência cotidiana que carregam significados nocivos por acordo interno de um grupo. Mendelsohn et al. (2023) chamam essas expressões de *dogwhistles*, as que transmitem um sentido ao público amplo e um segundo sentido, com frequência hostil, ao grupo que conhece o código. Esse recurso já foi documentado no discurso político dos Estados Unidos, onde a troca do termo que nomeia o grupo pelo código reduz a pontuação de toxicidade atribuída por um classificador comercial. Esse levantamento, porém, foi feito em inglês e a partir de fontes externas às comunidades que usam os termos, de modo que o repertório que uma comunidade cria para si mesma, em outro idioma, fica fora do seu alcance. Nesse regime, a toxicidade não está na forma como a mensagem é dita, que um leitor atento poderia desfazer, mas num dicionário compartilhado que o leitor de fora não possui.

Este trabalho investiga esse fenômeno num caso concreto, uma comunidade de jogos do YouTube brasileiro, cujos chats ao vivo foram monitorados em sete canais ao longo de dois meses, somando cerca de 1,74 milhão de mensagens. A comunidade é notoriamente tóxica, a ponto de atrair atenção recorrente de canais de opinião e reação dedicados a catalogar publicamente o seu comportamento, e é essa notoriedade que a qualifica como caso incomum (YIN, 2018), escolhido não por representar a média das comunidades, mas por exibir de forma extrema e visível o fenômeno de interesse, que aqui não é o comportamento tóxico responsável pela sua fama, e sim o vocabulário codificado que

circula no chat.

Nesse repertório, uma sigla de dois caracteres designa pornografia infantil, um verbo escolar comum serve de eufemismo para estupro, saudações nazistas circulam como sequências de emojis de aparência lúdica e siglas de facções criminosas funcionam como saudação de filiação. Não se trata de gíria regional que os detectores ainda não aprenderam, mas de um repertório em que cada termo circula acompanhado de variantes e disfarces que permitem a quem o usa negar a intenção.

Dois termos são usados em todo o restante do texto. O **código** da comunidade é o conjunto desses termos, com o significado que o grupo dá a eles. **Documentar** esse código é registrar cada termo por escrito, com o sentido que ele tem no uso interno, as variantes e os disfarces com que aparece e exemplos reais de uso, num registro que este trabalho chama de **catálogo de termos codificados**. Um registro assim pode ser entregue a um detector junto com a mensagem a classificar.

Os conjuntos anotados em português cobrem domínios genéricos e não separam a toxicidade codificada da explícita (LEITE et al., 2020; VARGAS et al., 2022), os conjuntos de avaliação de toxicidade implícita tratam do sentido que fica subentendido, quase sempre em inglês (ELSHERIEF et al., 2021; OCAMPO et al., 2023), e as medições em escala de chats ao vivo brasileiros se apoiam em classificadores genéricos, sem avaliação específica do repertório local (LOBO; BARBOSA, 2025). A pergunta sobre o que a moderação automática deixa de ver nessa comunidade permanece, portanto, sem instrumento de medição.

1.1 Questão de Pesquisa

O trabalho é orientado por uma questão central:

> **QP1.** *Modelos de detecção automática de toxicidade de estado da arte detectam a toxicidade veiculada pela linguagem codificada de uma comunidade?*

A resposta exige mais do que um número único para a amostra inteira. A questão é respondida separando os erros cometidos nas mensagens de toxicidade explícita dos cometidos nas de toxicidade codificada, conforme a Seção 3.7.3. Uma questão secundária, condicionada à primeira, avalia o catálogo como intervenção:

> **QP2.** *Se a detecção não alcança essa toxicidade, fornecer o código documentado da comunidade a um modelo de linguagem recupera a detecção?*

1.2 Objetivos

O objetivo geral é medir quanto os modelos de detecção automática de estado da arte alcançam da toxicidade codificada de uma comunidade de jogos do YouTube brasileiro, usando os detectores existentes como instrumento de medição. Desdobra-se nos objetivos específicos:

1. construir e validar um **catálogo de termos codificados** da comunidade;
2. produzir um **conjunto de referência rotulado** em que cada mensagem tóxica é classificada pelo mecanismo da sua toxicidade: explícita, codificada ou mista;
3. medir quanto os detectores de estado da arte alcançam da toxicidade explícita e quanto alcançam da codificada;
4. medir quanto a detecção da toxicidade codificada muda quando o catálogo é fornecido a modelos de linguagem de grande porte.

1.3 Contribuições

O trabalho oferece quatro contribuições:

1. o **catálogo de termos codificados** da comunidade, registro documentado do repertório, com o significado de cada entrada no uso comunitário, suas variantes e disfarces, a política de moderação correspondente e exemplos reais de uso;
2. um **protocolo replicável** de construção e validação do catálogo, em que a vivência na comunidade levanta as hipóteses e a análise do corpus completo as verifica, aplicável a outras comunidades com repertório codificado;
3. um **conjunto de referência rotulado** com a toxicidade separada pelo mecanismo, o que permite medir em separado quanto cada modelo detecta das mensagens explícitas e das codificadas;
4. **evidência empírica** da comparação de três famílias de detectores diante da linguagem codificada de uma comunidade, e da melhoria alcançada quando a documentação desse código lhes é fornecida como intervenção.

1.4 Organização do Trabalho

O restante do texto está organizado em quatro capítulos. O Capítulo 2 revisa a literatura de detecção de toxicidade e situa a lacuna que o trabalho ocupa. O Capítulo 3

descreve o desenvolvimento do trabalho, que inclui a coleta, a amostragem, a construção e a validação do catálogo, a rotulagem, o desenho dos experimentos e as considerações éticas. O Capítulo 4 apresenta os resultados e responde às duas questões de pesquisa. O Capítulo 5 conclui, com limitações e trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo situa o trabalho na literatura de detecção automática de conteúdo nocivo em plataformas digitais. A Seção 2.1 define os conceitos que o texto utiliza, dos tipos de toxicidade aos paradigmas de anotação. A Seção 2.2 descreve a evolução técnica dos detectores até o estado da arte. A Seção 2.3 discute os trabalhos relacionados, agrupados por tema. A Seção 2.4 sintetiza a literatura e delimita a lacuna que o estudo ocupa.

2.1 Conceitos fundamentais

2.1.1 Discurso de ódio, linguagem ofensiva e toxicidade

Não há uma definição única aceita por toda a literatura para o fenômeno investigado. Vários termos convivem e se sobrepõem, como discurso de ódio, linguagem ofensiva e toxicidade, cada um com um ângulo diferente (FORTUNA; NUNES, 2018). Fortuna e Nunes (2018) definem discurso de ódio como a linguagem que ataca ou diminui um grupo por características como aparência física, religião, origem étnica ou nacional, orientação sexual ou identidade de gênero. Para os autores, mesmo a discriminação disfarçada de piada é discurso de ódio.

Davidson et al. (2017) tratam a linguagem ofensiva como uma categoria distinta do discurso de ódio, definida como mensagens que empregam termos vulgares ou insultuosos sem atacar um grupo protegido.

O termo toxicidade entrou no vocabulário da área pelo uso prático da Perspective API (Jigsaw; Google, 2017), que o define como um comentário rude, desrespeitoso ou capaz de levar alguém a abandonar uma discussão. Por ser mais amplo que discurso de ódio, funciona como um termo guarda-chuva. Este trabalho adota toxicidade nesse sentido amplo, porque o repertório velado da comunidade estudada corresponde a políticas de moderação de plataforma de tipos diferentes, como segurança infantil, extremismo violento, discurso de ódio e assédio, das quais apenas uma coincide com o discurso de ódio em sentido estrito.

2.1.2 Toxicidade explícita, implícita e codificada

O conteúdo nocivo costuma ser dividido em duas formas, a explícita e a implícita (ELSHERIEF et al., 2021; OCAMPO et al., 2023).

A toxicidade explícita é aquela em que a própria palavra usada carrega a ofensa, como um xingamento ou um termo de ódio direto (ELSHERIEF et al., 2021; OCAMPO

et al., 2023). É a forma que os detectores atuais identificam com mais facilidade.

A toxicidade implícita é aquela que ataca uma pessoa ou um grupo de maneira indireta, sem empregar uma palavra abertamente ofensiva. ElSherief et al. (2021) a definem como o ódio transmitido por sarcasmo, metáfora ou insinuação, sem os sinais visíveis da forma explícita. Como não há uma palavra ofensiva evidente na frase, quem escreve preserva a chamada negabilidade plausível, ou seja, pode alegar que não quis dizer aquilo, e a mensagem escapa dos detectores que se baseiam em palavras-chave ofensivas (ELSHERIEF et al., 2021). A ironia e a metáfora, ainda assim, deixam algum rastro no texto, e um leitor atento costuma recuperar o sentido.

Este trabalho investiga uma terceira forma, a toxicidade codificada. Ela ocorre quando palavras e siglas de aparência comum passam a ter um significado nocivo por acordo interno de uma comunidade. Um exemplo é uma sigla de dois caracteres que, no uso do grupo, designa pornografia infantil. Esse vocabulário é o que este trabalho chama de código da comunidade, ou linguagem codificada. A literatura de toxicidade implícita já registra a estratégia, e ElSherief et al. (2021) apontam a linguagem codificada como um recurso usado para escapar de detectores por palavra-chave.

O recurso também tem nome próprio na área. Mendelsohn et al. (2023) definem *dogwhistle* como a expressão que transmite um sentido ao público amplo e um segundo sentido, com frequência hostil, ao grupo que conhece o código, empregada para escapar tanto da repercussão pública quanto da moderação automática. A estrutura é a mesma examinada aqui. O termo, porém, cobre um conjunto maior, e os próprios autores observam que nem todo dogwhistle é hostil, já que alguns apenas sinalizam a identidade religiosa ou política de quem fala. Este trabalho trata do subconjunto que carrega conteúdo nocivo, e é a esse recorte que a expressão toxicidade codificada se refere.

A diferença em relação à toxicidade implícita está em onde fica o sinal ofensivo. Na forma implícita, o sentido ofensivo ainda está no texto, na ironia ou na metáfora. Na forma codificada, a aparência da mensagem não revela a ofensa, porque o significado real depende de um vocabulário próprio da comunidade, que o leitor de fora não domina.

2.1.3 Paradigmas de anotação humana

Um detector treinado herda a qualidade e os vieses das anotações humanas sobre as quais foi ajustado, e anotar toxicidade é reconhecidamente difícil, sobretudo nas formas não explícitas (DAVIDSON et al., 2017; ELSHERIEF et al., 2021; OCAMPO et al., 2023). Röttger et al. (2022) distinguem dois paradigmas de anotação para tarefas subjetivas. No paradigma descritivo, busca-se capturar o julgamento espontâneo de cada anotador, e a divergência entre eles é tratada como informação legítima. No paradigma prescritivo, um conjunto de diretrizes fixa uma definição específica do que se quer medir, e a divergência

passa a ser erro de aplicação da regra. Nenhum é superior em tese, mas a distinção decide o desenho da rotulagem aqui. Um rótulo que depende de conhecimento específico da cultura de uma comunidade não pode ser obtido pela soma de opiniões espontâneas de anotadores leigos, que tenderiam a marcar como inofensivas justamente as mensagens em disputa. Por isso a rotulagem deste trabalho adota o paradigma prescritivo, com um guia de rotulagem e regras explícitas, apresentadas no Capítulo 3.

2.2 Detecção automática de toxicidade

A detecção automática de conteúdo nocivo evoluiu por três gerações de técnicas e hoje conta ainda com uma quarta via, a dos modelos de linguagem de grande porte (LLMs) (FORTUNA; NUNES, 2018; RAMOS et al., 2024).

A primeira geração usava listas de palavras e regras. A presença, na mensagem, de um termo previamente listado como ofensivo já disparava o alarme. Essa abordagem trata a palavra como prova e ignora o contexto em que ela aparece (DAVIDSON et al., 2017).

A segunda geração trouxe o aprendizado de máquina clássico, em que o modelo aprende a classificar a partir de exemplos já rotulados, com algoritmos como máquinas de vetores de suporte, regressão logística e Naive Bayes. Em seguida vieram as redes neurais profundas, capazes de captar relações no texto sem que o pesquisador precise descrever manualmente cada característica relevante (FORTUNA; NUNES, 2018; RAMOS et al., 2024).

A terceira geração é a dos modelos Transformer (VASWANI et al., 2017) e é a que alcança o melhor desempenho nos levantamentos recentes da área (RAMOS et al., 2024). O BERT (DEVLIN et al., 2019) é um dos seus exemplos mais difundidos. Ele é primeiro pré-treinado em um grande volume de texto comum, para aprender o funcionamento da língua, e depois ajustado a uma tarefa específica com um conjunto menor de exemplos, processo conhecido como *fine-tuning*. Existem versões que atendem a muitos idiomas ao mesmo tempo, como o XLM-R (CONNEAU et al., 2020), e uma versão treinada especificamente para o português, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), ponto de partida comum para tarefas de classificação de texto na língua.

A quarta via é a dos LLMs usados por instrução. Brown et al. (2020) mostraram que modelos suficientemente grandes executam tarefas novas sem treino adicional, apenas a partir de uma descrição da tarefa escrita em linguagem comum. Essa descrição pode vir sozinha, situação chamada de *zero-shot*, ou acompanhada de alguns exemplos já resolvidos, o *few-shot*. Em vez de treinar um classificador, passa-se a instruir um modelo de uso geral, e tudo o que ele precisa saber sobre a tarefa precisa estar no próprio texto da instrução, chamado de *prompt*.

A revisão de Ramos et al. (2024), que reúne mais de cem estudos, confirma que os modelos Transformer alcançam o melhor desempenho, ao custo de maior exigência de processamento, e observa que a grande maioria dos trabalhos é em inglês. O mesmo levantamento ressalva que nem esses modelos resolvem o problema por completo. A linguagem nociva muda continuamente, o contexto, o sarcasmo e a ironia continuam difíceis, e os dados usados no treino raramente contemplam as particularidades de comunidades específicas (GEETANJALI; KUMAR, 2025; RAMOS et al., 2024). Os detectores avaliados neste trabalho, um modelo multilíngue ajustado para toxicidade (HANU; Unitary team, 2020) e modelos BERTimbau ajustados sobre dados brasileiros (SOUZA; NOGUEIRA; LOTUFO, 2020), pertencem a essa terceira geração. A pergunta que o trabalho investiga é se ela resolve, ou apenas adia, esses problemas quando a linguagem nociva é o código de uma comunidade.

2.3 Trabalhos relacionados

Esta seção discute os trabalhos empíricos mais próximos deste estudo, agrupados por tema, em dois níveis de detalhe. Três deles são descritos pelo problema que se propuseram a resolver, pelo modo como o resolveram e pela conclusão a que chegaram. São os três de que este trabalho depende, uns porque o seu método reaparece no desenho dos experimentos, outros porque a contribuição defendida aqui se define por contraste com eles. Os demais entram pela afirmação que sustentam. A Seção 2.4 reúne todos em uma tabela comparativa.

2.3.1 Toxicidade implícita e formas de disfarce

A toxicidade implícita concentrou o esforço recente de construção de recursos e continua sem solução. O primeiro conjunto anotado especificamente para essa forma, organizado por uma taxonomia de categorias de ódio implícito, expôs o baixo desempenho das ferramentas de uso geral (ELSHERIEF et al., 2021), e a avaliação de sete conjuntos já consolidados chegou ao mesmo diagnóstico, tanto no implícito, que não traz nenhuma palavra ofensiva, quanto no sutil, que depende de ironia (OCAMPO et al., 2023). É esse diagnóstico que sustenta a expectativa de que os detectores também falhariam diante do código da comunidade estudada.

Duas outras formas de disfarce foram estudadas por conta própria. Diante do emoji os detectores falham de forma sistemática, porque esse símbolo é pouco processado e pouco representado nos dados de treino (KIRK et al., 2022b), e a sua carga tóxica não vem de um significado ofensivo próprio, e sim da associação com outros símbolos e do uso em contexto (ZHOU et al., 2025). No humor, a moldura cômica exige do receptor um passo a mais de interpretação e torna posições hostis mais aceitáveis sem reduzir o dano (SCHMID, 2025), mecanismo idêntico ao escudo “é só pra rir” recorrente na comunidade estudada.

2.3.2 Toxicidade codificada

Mendelsohn et al. (2023) trataram da expressão codificada que carrega um sentido para o público amplo e outro para o grupo que conhece o código, até então estudada sobretudo fora da computação. Reuniram um glossário de 340 termos em inglês a partir de literatura acadêmica, cobertura jornalística e enciclopédias colaborativas da internet, com o significado de cada entrada atribuído à fonte que o havia descrito, e o aplicaram a discursos do Congresso dos Estados Unidos e a um teste em que frases hostis tinham o termo que nomeia o grupo trocado por um desses códigos. Concluíram que as frases assim codificadas recebem pontuação de toxicidade mais baixa de um classificador comercial, e que um modelo de linguagem de grande porte questionado sobre esses termos recupera menos da metade do glossário.

Kruk et al. (2024) atacaram o problema deixado em aberto pelo trabalho anterior, o de decidir, diante de uma ocorrência concreta, se o termo está no sentido codificado ou no sentido comum. Partiram desse mesmo glossário, procuraram os seus termos em comentários do Reddit e em registros do Congresso dos Estados Unidos, e entregaram a um modelo de linguagem o termo junto com o significado que o glossário documenta, pedindo a classificação de cada ocorrência. Concluíram que os modelos são pouco confiáveis para reconhecer um termo codificado por conta própria, mas separam bem o uso codificado do uso comum quando recebem o significado documentado.

2.3.3 Detecção de toxicidade em português brasileiro

Dois conjuntos anotados servem de referência para o português brasileiro, ambos sobre domínios genéricos, e são a base dos detectores ajustados avaliados aqui. O ToLD-BR reúne 21 mil mensagens do Twitter anotadas por pessoas de perfis variados, com rótulos para categorias como insulto, misoginia, racismo, homofobia e obscenidade, e mostrou que a toxicidade não passa facilmente de um idioma para outro, com 76% de macro-F1 no modelo treinado diretamente em português contra 56% no aproveitado do inglês (LEITE et al., 2020). O HateBR reúne 7 mil comentários de Instagram publicados em contas de políticos brasileiros, anotados por especialistas em três níveis, da simples presença de ofensa até o alvo e o grau de agressividade (VARGAS et al., 2022). Sobre o primeiro, um modelo de linguagem sem nenhum exemplo prévio alcançou 73% de F1, resultado próximo ao dos modelos treinados, porém muito sensível à forma como a instrução é escrita (OLIVEIRA et al., 2023).

2.3.4 Toxicidade em chats ao vivo e comunidades de jogos

O chat ao vivo tem características próprias, diferentes dos comentários estáticos que dominam os dados da área. Em chats da Twitch em português, a toxicidade se concentra

e se espalha em cascatas, isto é, em sequências encadeadas de mensagens tóxicas em que uma parece puxar a seguinte, e há canais que concentram essa cultura sem moderação à altura (LOBO; BARBOSA, 2025). A medição se apoia em ferramentas de uso geral, como a Perspective API, justamente a classe de ferramenta cujo alcance sobre o código de uma comunidade este trabalho mede.

Martins et al. (2025) realizaram o estudo de caso mais próximo deste trabalho, sobre 28 canais da Manosfera *Red Pill* brasileira acompanhados ao longo de dezoito meses, descrevendo como evoluíram no tempo, como atraíam audiência e sobre o que falavam. A diferença de recorte define a contribuição aqui. Martins et al. descrevem o conteúdo produzido pelos canais, em vídeos e falas, em uma comunidade de toxicidade ideológica e, em boa parte, legível. Este trabalho olha para a linguagem da audiência, no chat ao vivo, de uma comunidade cuja toxicidade é deliberadamente codificada, e pergunta não o que a comunidade diz, mas se os detectores automáticos conseguem ler o que ela diz.

No chat de futebol tailandês, cinco modelos da família BERT alcançaram F1 acima de 0,95 (POOKPANICH; SIRIBORVORNRATANAKUL, 2024), o que indica que essa geração vai bem mesmo em idioma de poucos recursos quando o alvo é o insulto aberto, de vocabulário claro. O problema, portanto, não está em ser um chat ao vivo, mas no que a comunidade faz com a linguagem dentro dele.

2.3.5 Modelos de linguagem de grande porte como detectores e anotadores

Os modelos de linguagem de grande porte já foram avaliados como detectores e como anotadores, com dois resultados que interessam ao experimento deste trabalho. Modelos instruídos, sem treino específico para a tarefa, classificam discurso de ódio com acurácia e F1 entre 80% e 90% numa suíte de testes em inglês, e nessa mesma avaliação as instruções diretas e curtas superaram as instruções longas com contexto adicional, resultado que contrariou a expectativa dos próprios autores e indica que acrescentar contexto genérico não ajuda a detecção e pode até atrapalhar (KUMARAGE; BHATTACHARJEE; GARLAND, 2024). Na anotação, o ChatGPT concordou com anotadores humanos em 80% dos casos de ódio implícito em inglês, o que indica que nessa forma o modelo já rivaliza com pessoas (HUANG; KWAK; AN, 2023).

Juntos, esses resultados (KUMARAGE; BHATTACHARJEE; GARLAND, 2024; HUANG; KWAK; AN, 2023) deixam em aberto a pergunta que o experimento deste trabalho enfrenta. Se os modelos já lidam bem com a toxicidade implícita e acrescentar contexto genérico não ajuda, o que acontece quando o que falta ao modelo não é a capacidade geral, mas o vocabulário de uma comunidade específica? A hipótese natural é que o complemento útil não seja mais um exemplo genérico, e sim o próprio código, ou seja, o catálogo de termos.

2.4 Síntese e lacuna de pesquisa

A Tabela 1 resume os trabalhos revisados que definem mais diretamente a posição deste estudo. Ela os organiza pelos mesmos eixos usados no capítulo, a saber, o ambiente e o idioma estudados, o tipo de toxicidade em foco e a abordagem adotada, e os apresenta na ordem em que foram discutidos. A única exceção é Martins et al. (2025), deslocado para a última posição por ser o estudo mais próximo deste, o que permite a comparação direta com a linha final.

Lidos em conjunto, os trabalhos revisados concordam em três pontos que este estudo toma como base. Primeiro, detectar a toxicidade explícita já é um problema bem resolvido, com bom desempenho até em chat ao vivo de idiomas com poucos recursos (POOKPANICH; SIRIBORVORNRATANAKUL, 2024; RAMOS et al., 2024). Segundo, a toxicidade não explícita continua em aberto, com falhas que se repetem diante do implícito (ELSHERIEF et al., 2021; OCAMPO et al., 2023), do código (MENDELSOHN et al., 2023), do humor (SCHMID, 2025) e do emoji (KIRK et al., 2022b; ZHOU et al., 2025), e sem boa passagem de um idioma para outro (LEITE et al., 2020). Terceiro, o uso de LLMs muda o problema, mas não o resolve, pois esses modelos já lidam bem com o implícito em inglês, sem treino (HUANG; KWAK; AN, 2023), mas dependem muito da instrução e não melhoram com contexto genérico (KUMARAGE; BHATTACHARJEE; GARLAND, 2024; OLIVEIRA et al., 2023).

A lacuna fica clara ao cruzar as colunas da Tabela 1. O fenômeno estudado aqui, a linguagem codificada de uma comunidade, não é o implícito dos conjuntos em inglês, e os conjuntos de dados disponíveis em português não o documentam, porque cobrem a toxicidade explícita de tuítes e de comentários políticos (LEITE et al., 2020; VARGAS et al., 2022). Os trabalhos que documentam o mesmo fenômeno o fazem em inglês, no debate político e em redes de texto público (MENDELSOHN et al., 2023; KRUK et al., 2024), e os dois partem de um repertório levantado de fora das comunidades que estudam. Mendelsohn et al. (2023) registram esse ponto como limite do próprio trabalho, ao afirmar que se descrevem como observadores de fora em quase todas as entradas do glossário e que a validação por membros dos grupos estudados seria desejável, embora improvável de obter, porque revelar o código destruiria a sua utilidade. O trabalho seguinte herda a mesma condição, pois toma aquele glossário como ponto de partida. É essa a lacuna que o catálogo construído aqui ocupa, porque documenta de dentro, e sobre o corpus da própria comunidade, um repertório que nenhum desses recursos alcança. O ambiente, o chat ao vivo em português, já foi medido em larga escala, mas com ferramentas de uso geral e sem avaliação específica do repertório local (LOBO; BARBOSA, 2025). E o estudo de caso brasileiro mais próximo descreve o conteúdo produzido pelos canais, não a linguagem codificada da audiência (MARTINS et al., 2025). Por fim, nenhum dos trabalhos da Tabela 1 reúne o desenho completo deste estudo, que combina a documentação formal

do código, o catálogo, com a avaliação de três famílias de detectores em uso real.

É nessa lacuna que o trabalho se posiciona. Ele não propõe um detector novo, mas usa os detectores existentes como instrumento para medir o quanto a moderação automática deixa de detectar quando a toxicidade circula pela linguagem codificada de uma comunidade.

Tabela 1 – Síntese comparativa dos trabalhos relacionados

Trabalho	**Ambiente / idioma**	**Toxicidade estudada**	**Abordagem**
ElSherief et al. (2021)	Twitter / inglês	Implícita (retórica)	Taxonomia + *benchmark* + modelos-base BERT
Ocampo et al. (2023)	7 *benchmarks* / inglês	Implícita e sutil	Análise sistemática + avaliação de modelos de estado da arte
Mendelsohn et al. (2023)	Discurso político EUA / inglês	Codificada (*dogwhistle*)	Glossário curado de fontes secundárias + GPT-3 + Perspective API
Kruk et al. (2024)	Reddit e Congresso EUA / inglês	Codificada (*dogwhistle*)	Desambiguação por LLM + conjunto de 16.550 ocorrências
Kirk et al. (2022b)	Sintético-adversarial / inglês	Transmitida por emoji	Suíte de testes + avaliação de modelos de estado da arte
Schmid (2025)	Memes em redes sociais / alemão	Humorística (ódio com moldura de piada)	Estudo de recepção com usuários, métodos mistos
Zhou et al. (2025)	Twitter / inglês	Transmitida por emoji	Análise de distribuição + moderação via LLM
Leite et al. (2020)	Twitter / PT-BR	Explícita (geral)	Conjunto (21 mil) + BERT + análise multilíngue
Vargas et al. (2022)	Instagram político / PT-BR	Explícita (ofensa)	Conjunto especialista (7 mil) + modelos-base
Oliveira et al. (2023)	Twitter / PT-BR	Explícita (ódio)	ChatGPT *zero-shot* vs. modelos ajustados
Lobo e Barbosa (2025)	Chat ao vivo Twitch / PT-BR	Explícita (via Perspective API)	Medição em escala + cascatas de toxicidade
Pookpanich e Siriborvornratanakul (2024)	Chat ao vivo YouTube / tailandês	Explícita (ofensa)	Cinco modelos BERT para moderação ao vivo
Kumarage, Bhattacharjee e Garland (2024)	HateCheck / inglês	Nociva (geral)	Revisão + testes de LLMs por instrução
Huang, Kwak e An (2023)	Twitter / inglês	Implícita (retórica)	ChatGPT vs. anotadores humanos
Martins et al. (2025)	Vídeos YouTube / PT-BR	Ideológica (misoginia)	Estudo de caso de comunidade (caracterização)
Este trabalho	**Chat ao vivo YouTube, comunidade de jogos / PT-BR**	**Codificada (código da comunidade)**	**Catálogo documentado + avaliação de três famílias de detectores**

Fonte: elaboração própria.

3 Desenvolvimento

Este capítulo descreve o desenvolvimento do trabalho. A Seção 3.1 apresenta uma visão geral da metodologia e o encadeamento entre suas etapas. A Seção 3.2 descreve a coleta de mensagens em escala. A Seção 3.3 descreve o pré-processamento dos dados e a Seção 3.4 a amostragem estratificada. A Seção 3.5 detalha a construção e a validação do catálogo de termos codificados. A Seção 3.6 apresenta a rotulagem da amostra. A Seção 3.7 descreve o desenho dos experimentos de análise de erro. As considerações éticas encerram o capítulo na Seção 3.8.

3.1 Visão Geral da Metodologia

A metodologia divide-se em etapas encadeadas, da coleta das mensagens à avaliação dos detectores. O instrumento que as articula é o **catálogo de termos codificados** da comunidade, um inventário dos termos que transmitem conteúdo tóxico sob aparência inocente. O catálogo permite distinguir, na amostra, a toxicidade explícita da toxicidade codificada, cujo reconhecimento depende do repertório da comunidade.

O trabalho se organiza em dois momentos. No primeiro, a coleta e o pré-processamento preparam o corpus, a construção do catálogo documenta o repertório codificado e a rotulagem produz um conjunto de referência de 3.500 mensagens rotuladas manualmente, usado para medir os detectores. Nesse conjunto, cada mensagem tóxica recebe também o mecanismo de sua toxicidade, que pode ser explícita, codificada ou mista. No segundo momento, faz-se a análise de erro. Classificadores de estado da arte e modelos de linguagem de grande porte rodam sobre essa mesma amostra, e seus erros são decompostos pelo eixo explícito/codificado. É essa decomposição, e não a métrica agregada, que responde às questões de pesquisa formuladas no Capítulo 1.

A separação entre os dois momentos tem consequência para a validade do desenho. O catálogo é construído e validado **antes** e **independentemente** da execução dos classificadores, de modo que a definição de "toxicidade codificada" não pode ser ajustada retroativamente aos erros observados.

A Figura 1 resume esse encadeamento, da coleta à análise de erro.

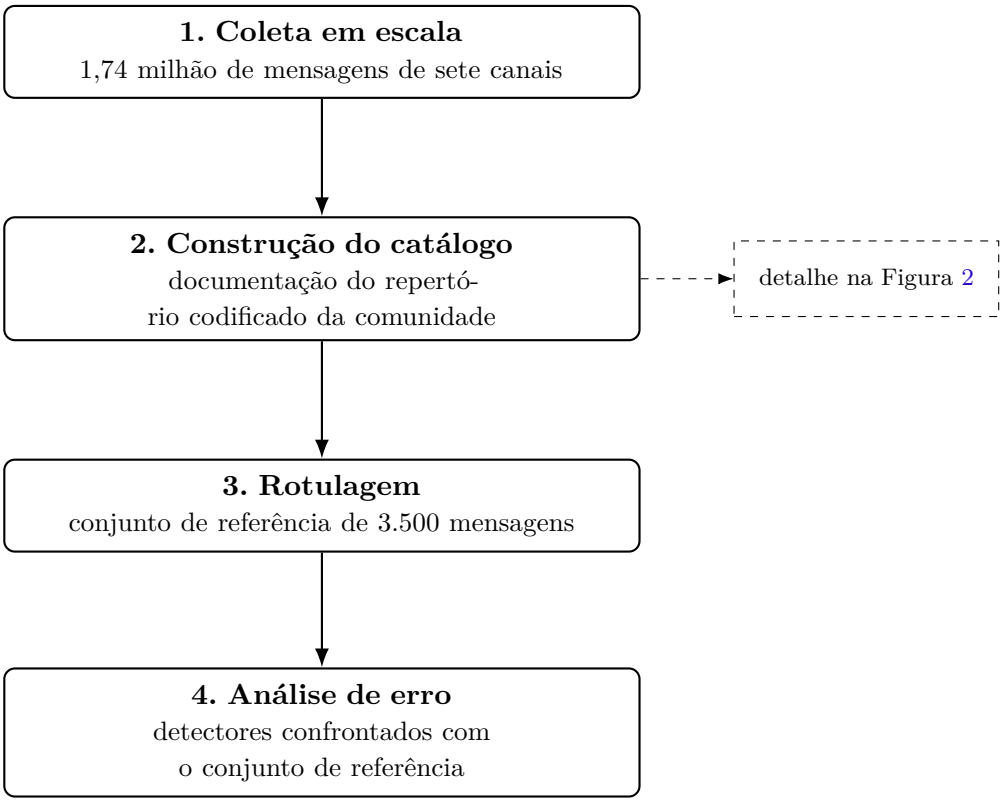

Figura 1 – Visão geral da metodologia e o encadeamento entre suas etapas

Fonte: elaborado pelo autor.

3.2 Coleta de Dados

A coleta de mensagens foi realizada por um sistema próprio de monitoramento de transmissões ao vivo, desenvolvido sobre a YouTube Data API v3 (Google Developers, s.d.). O corpus é composto por mensagens publicadas em chats públicos, obtidas pela API oficial da plataforma, dentro dos limites de quota e sem raspagem não autorizada. Os dados são tratados apenas para fins acadêmicos, e as demais decisões éticas estão reunidas na Seção 3.8.

Os sete canais monitorados pertencem a uma mesma subcomunidade gamer brasileira do YouTube, voltada à transmissão ao vivo de jogos, com audiências grandes e chats de alto volume. A comunidade foi escolhida pela notoriedade da sua toxicidade, atestada por canais externos que produzem conteúdo sobre o seu comportamento, e não por representar a média das comunidades de jogos, conforme o Capítulo 1. Os canais são referidos pelos pseudônimos Canal_A a Canal_G, conforme a política de anonimização da Seção 3.8. A coleta consolidada cobre de 15 de junho a 14 de agosto de 2025, dois meses, e totaliza **1.740.527 mensagens** de **231 transmissões ao vivo**, cerca de 7,5 mil por transmissão. Estão excluídas as 177 mensagens publicadas pelo próprio pesquisador

como membro da comunidade, com a autoexclusão detalhada na Seção 3.8. A Tabela 2 apresenta a distribuição por canal.

Tabela 2 – Distribuição de mensagens por canal (coleta consolidada)

Canal	**Mensagens**	**Proporção**
Canal_F	984.782	56,6%
Canal_E	532.134	30,6%
Canal_C	98.156	5,6%
Canal_A	61.294	3,5%
Canal_D	39.784	2,3%
Canal_G	16.827	1,0%
Canal_B	7.550	0,4%
Total	**1.740.527**	**100%**

Fonte: dados da pesquisa.

A distribuição concentra-se em dois canais, que respondem por 87,2% do volume, uma das razões da amostragem uniforme por canal (Seção 3.4).

3.3 Pré-processamento

Antes da amostragem, o corpus completo passou por uma etapa de pré-processamento. A limpeza foi mínima e deliberada, porque parte do fenômeno em estudo está nas mensagens curtas e repetidas que um tratamento mais agressivo eliminaria.

Duas classes de mensagens foram removidas do corpus completo por expressões regulares. As mensagens com URLs saíram porque a divulgação externa e o spam de links estão fora do fenômeno estudado, num total de 9.796, ou 0,6% do corpus. As mensagens de riso puro, compostas apenas por sequências da letra “k”, saíram por não conterem conteúdo linguístico analisável, num total de 229.569, ou 13,2% do corpus. Após a limpeza, o corpus passou de 1.740.527 para 1.501.162 mensagens.

Dois procedimentos comuns de pré-processamento foram deliberadamente evitados, a desduplicação global e o filtro de tamanho mínimo. Os termos codificados da comunidade são tipicamente curtos, e seu uso frequente envolve repetição intensa, em ondas coordenadas da mesma mensagem. Desduplicar ou exigir um tamanho mínimo removeria parte do fenômeno que o trabalho estuda. A amostra final ficou com 8,1% de mensagens duplicadas.

O chat do YouTube tem emojis próprios da plataforma, guardados no texto como códigos curtos e exibidos como pequenas imagens, como  . A amostra contém 41 deles, com 1.835 ocorrências em 10,7% das mensagens. Parte do repertório codificado é transmitida por combinações desses emojis, então a camada visual foi preservada por inteiro. Um

mapeamento público converte cada código na imagem correspondente, e os emojis foram exibidos na forma gráfica, dentro das mensagens, no instrumento de rotulagem.

3.4 Amostragem Estratificada

Do corpus já pré-processado, extraiu-se uma amostra de **3.500 mensagens** para a rotulagem manual, com **500 mensagens por canal** sorteadas ao acaso. O sorteio usou semente fixa para permitir a reprodução exata da amostra. A alocação uniforme foi escolhida para que nenhum canal ficasse sub-representado. Sob a alocação proporcional ao volume, os cinco canais menores somariam cerca de 13% da amostra, e o menor entraria com cerca de 15 mensagens, o que inviabilizaria a leitura por canal. Com o mesmo número de mensagens em cada um, todos reúnem material suficiente para serem analisados e comparados entre si.

A Tabela 3 resume o percurso dos dados, do corpus completo à amostra rotulada.

Tabela 3 – Pré-processamento e amostragem

Nível	**Descrição**	**Mensagens**
Corpus completo	Coleta consolidada dos 7 canais (15/06–14/08/2025), após autoexclusão do pesquisador (Seção 3.8)	1.740.527
Corpus filtrado	Após remoção de URLs (−9.796) e riso puro (−229.569)	1.501.162
Amostra rotulada	Sorteio uniforme de 500 mensagens por canal	3.500

Fonte: dados da pesquisa.

3.5 Construção do Catálogo de Termos Codificados

O principal instrumento desta pesquisa é um **catálogo de termos codificados**, um inventário dos termos que a comunidade usa para transmitir conteúdo tóxico sob aparência inocente. Cada entrada apresenta o significado do termo no uso comunitário, suas variantes e disfarces, a política de moderação correspondente e exemplos reais de uso. O catálogo cumpre um papel duplo. É ao mesmo tempo o **produto** que documenta o repertório da comunidade, apresentado no Capítulo 4, e o **instrumento de medição** que separa os erros dos classificadores entre toxicidade explícita e codificada.

A construção seguiu um protocolo de três passos, resumido na Figura 2. No primeiro, a **imersão** do pesquisador na comunidade levanta os termos-semente, os candidatos a termo codificado. No segundo, uma **mineração de vizinhança** percorre o corpus completo

e recupera, para cada semente, as palavras que a acompanham no uso real. No terceiro, a **validação convergente** admite o termo apenas quando o significado sugerido pela imersão coincide com o que essas palavras revelam no corpus.

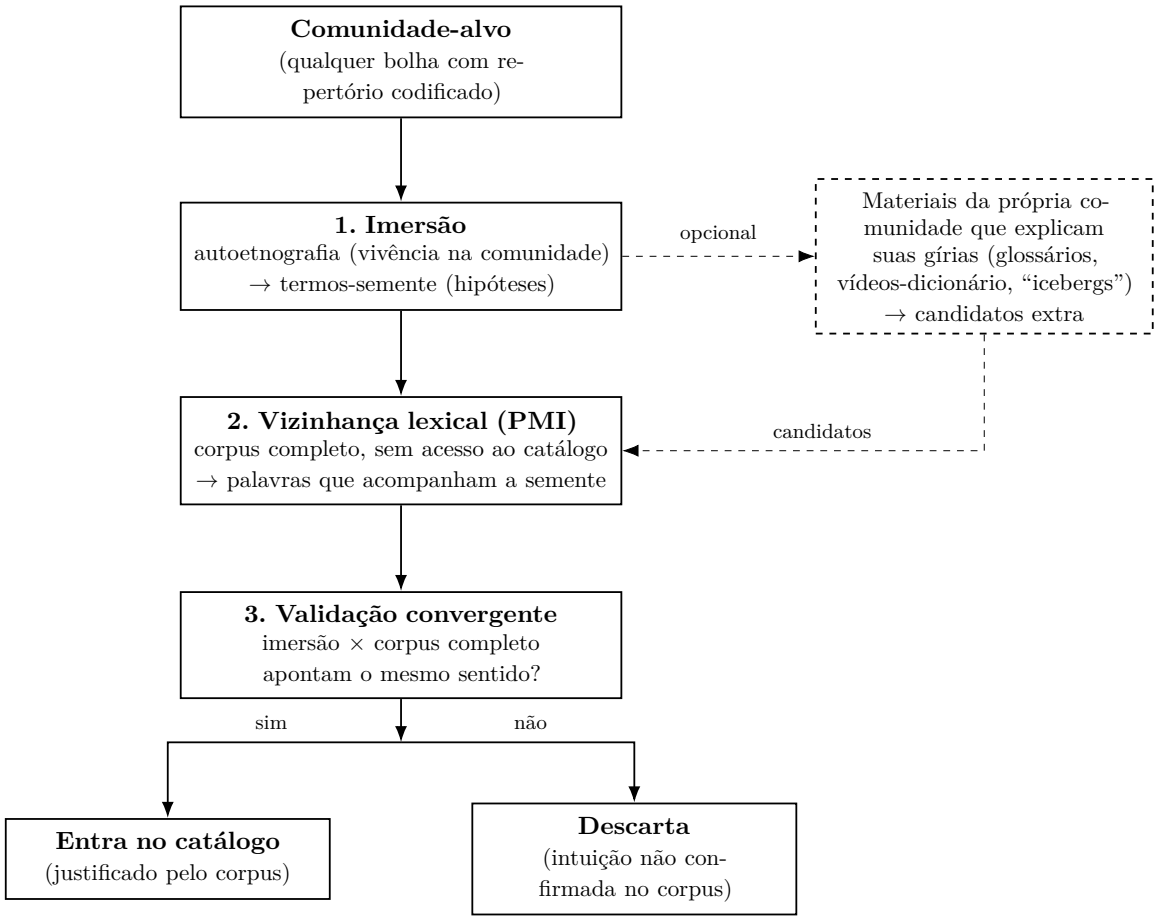

Figura 2 – Fluxo metodológico de construção e validação do catálogo de termos codificados

Fonte: elaborado pelo autor.

3.5.1 Imersão Autoetnográfica como Fonte de Descoberta

Os termos-semente não vieram de uma observação formal com notas de campo, e sim da experiência do pesquisador como membro da comunidade. Usar a própria experiência vivida na comunidade como dado de pesquisa é o que se chama de **autoetnografia** (ELLIS; ADAMS; BOCHNER, 2011). Como a comunidade é online, o trabalho se situa também na tradição da **netnografia**, a etnografia adaptada a comunidades online (KOZINETS, 2002; KOZINETS, 2020). Entre as duas vertentes da autoetnografia, adota-se a que exige apoio em dados verificáveis, e não a que se apoia apenas no relato pessoal (ANDERSON, 2006).

Essa experiência serve apenas para **levantar hipóteses**, e não para comprová-las, uma separação clássica entre descoberta e prova na filosofia da ciência (REICHENBACH, 1938). Ela sugere quais termos investigar e o que eles podem significar, mas não é a

evidência que sustenta o catálogo. Essa evidência vem de uma fonte independente da memória do pesquisador, a forma como as palavras se combinam no corpus completo, descrita na Seção 3.5.2. Um termo que o pesquisador suspeitava, mas que não aparece dessa forma no corpus, não entra. O registro dos termos descartados, na Seção 3.5.3, mostra que isso de fato acontece, o que responde à crítica de que a memória do pesquisador não poderia ser verificada. A experiência continua pessoal e de um único observador, e esse limite é controlado ao usá-la só para levantar hipóteses, nunca como evidência. O que se aproveita em outras comunidades é o método, não a experiência em si.

A formalização dos critérios seguiu a lógica da *grounded theory* (GLASER; STRAUSS, 1967; CHARMAZ, 2006). Em vez de definir os critérios de antemão, os termos foram listados primeiro, e os critérios surgiram do que as entradas confirmadas tinham em comum, a aparência inocente que esconde uma carga tóxica dependente do contexto, a regra da sigla isolada e o limiar de repetição de emojis, todos detalhados na Seção 3.6.2.

3.5.2 Validação Convergente por Mineração de Vizinhança Lexical

A evidência que sustenta cada entrada do catálogo vem da **mineração de vizinhança lexical**, aplicada ao corpus completo de 1.740.527 mensagens. Para cada termo-semente, o procedimento recolhe as mensagens publicadas até **60 segundos** antes e depois de cada ocorrência, dentro da mesma transmissão. A vizinhança é delimitada por tempo, e não por distância em palavras como se faz em texto corrido. Em chats de transmissão ao vivo as mensagens são curtas, repetem tokens e não formam frases completas (RINGER; NICOLAOU; WALKER, 2020), de modo que a proximidade no tempo é o que indica relação entre os termos.

A duração da janela é um parâmetro da mineração de vizinhança, e sua influência sobre o resultado foi verificada repetindo o cálculo com janelas de 15, 30, 120 e 180 segundos. Nenhuma dessas variações trouxe vizinhos que apontassem para outro referente. O que muda é a força da ligação entre as palavras, que se reforça quando a janela encolhe e se dilui quando ela cresce, a ponto de dois dos vizinhos citados saírem do corte nas janelas mais largas. É o efeito de escala já descrito por Church e Hanks (1990), para quem janelas menores destacam relações de curto alcance e janelas maiores destacam relações temáticas mais amplas. A leitura da vizinhança, portanto, não depende do valor exato escolhido. Os números de cada janela estão na Seção A.3.

As palavras dessa vizinhança são então ordenadas pela informação mútua pontual (PMI), medida de associação lexical proposta por Church e Hanks (1990) e calculada na Equação 3.1.

$$\mathrm{PMI}(s, w) = \log_2 \frac{P(s, w)}{P(s)\,P(w)} \tag{3.1}$$

Nessa fórmula, s é o termo-semente e w é uma palavra encontrada na vizinhança dele. A letra P indica proporção, ou seja, o quanto algo aparece em relação ao total. Assim, $P(w)$ é a proporção com que a palavra aparece no corpus, $P(s)$ é a proporção com que a semente aparece, e $P(s, w)$ é a proporção com que as duas aparecem juntas na mesma janela. Multiplicar $P(s)$ por $P(w)$ dá a proporção com que elas apareceriam juntas se não houvesse relação alguma entre as duas. A divisão compara, portanto, o que de fato se observa com o que o acaso previa, e o resultado é quantas vezes o observado supera o esperado.

O logaritmo apenas converte essa razão em uma escala mais fácil de ler. Uma razão de um, que é o caso em que observado e esperado coincidem, vira zero. Razões maiores viram valores positivos e razões menores viram valores negativos. Como o logaritmo é tomado na base dois, cada unidade a mais equivale a dobrar a razão, de modo que o valor um significa o dobro do esperado, o valor dois significa quatro vezes e o valor três significa oito vezes. Os vizinhos apresentados na Tabela 4 ficaram entre 0,6 e 2,9, ou seja, aparecem de uma vez e meia a quase oito vezes mais do que o acaso previa.

O PMI sozinho favorece palavras raras, que alcançam valores altos só por aparecerem poucas vezes e quase sempre ao lado da semente. Essa é uma limitação conhecida da medida, e é o que motiva suas versões normalizadas (BOUMA, 2009). Para conter o efeito, a ordenação pondera o valor de PMI pela frequência da palavra na vizinhança, e entram na lista apenas palavras que aparecem ao menos 30 vezes nessas janelas, 50 vezes no corpus inteiro e vindas de pelo menos cinco autores diferentes, o que descarta bordão repetido por um único usuário. De cada semente foram examinadas as sessenta palavras mais bem colocadas, sem valor de corte fixo.

Duas propriedades mantêm esse cálculo independente do pesquisador. Ele é **cego ao catálogo**, porque recebe apenas o termo escrito, e não o significado que o pesquisador atribui a ele, e roda sobre o corpus inteiro, e não sobre a amostra que seria rotulada depois. A validação de cada entrada é, portanto, uma **triangulação entre métodos** (DENZIN, 1978; PATTON, 1999), apoiada em dois eixos que não dependem um do outro. O primeiro é a experiência do pesquisador na comunidade, descrita na Seção 3.5.1, que serve apenas para levantar as hipóteses. O segundo é a forma como as palavras se combinam no corpus, que independe do que o pesquisador lembra ou acredita. Os materiais que a própria comunidade produz para explicar suas gírias, como vídeos que funcionam de glossário, entram apenas como apoio opcional, fora desses dois eixos.

A Tabela 4 mostra essa convergência para os três primeiros termos-semente. As palavras que a mineração devolveu, sem acesso ao significado registrado pelo pesquisador, apontam para aquilo que esse significado descreve.[1]

[1] Os nomes e as siglas das duas organizações criminosas presentes no catálogo são mascarados em todo o documento pelos marcadores `FA` e `FB` (Facção A e Facção B), que substituem a sigla real em

Tabela 4 – Convergência entre significado documentado e vizinhança estatística (PMI, corpus completo)

Semente	**Significado documentado**	**Vizinhos de PMI alto (procedimento cego ao catálogo)**
`CP`	pornografia infantil (*child porn*)	`criança`, `trade`, `bio` (*link in bio*), referência a denunciante público de pedofilia
`Estudar`	eufemismo para "estuprar"	`estupro`, `estrupo` (grafia popular), `lei` (bordão associado), `escola` (contexto desambiguador)
`FA`	Facção A (facção criminosa; nome mascarado)	🏴, emoji de bandeira vermelha, 🫳, 🫵, `FB` (sigla da facção rival, mascarada)

Fonte: dados da pesquisa.

A mineração foi executada em rodadas sucessivas, e cada rodada atacou uma dimensão diferente do fenômeno. A Tabela 5 lista os métodos usados e o que cada um deles rendeu. Lidos em conjunto, eles mostram que, neste caso, nenhum método sozinho recuperou o repertório por inteiro.

qualquer grafia, inclusive nos exemplos citados do corpus. A política de anonimização está detalhada na Seção 3.8.

Tabela 5 – Inventário de métodos de descoberta e validação do catálogo

Método	**Dimensão atacada**	**Rendimento**
Experiência do pesquisador na comunidade	conhecimento prévio da comunidade	`CP`, `Estudar` e `FA` (sementes textuais), marcadores visuais de `Hitler/Hail`
Vizinhança PMI (sementes textuais)	co-ocorrência textual	`FB`, `Trade`, `Bio`
Vizinhança PMI (sementes visuais)	co-ocorrência visual	marcadores textuais de `Hitler/Hail`, validação bidirecional
Análise dos autores que postaram a semente ao menos três vezes	autoria	variante `fechadão`, sobreposições entre repertórios
Expressões de duas ou mais palavras	combinação de palavras	desambiguação de `Bio`, estruturas `hail X`, `fechadão com FA`
Rajadas de mensagens concentradas no tempo	temporal	taxonomia de ondas coordenadas, nenhum termo novo
Co-ocorrência com cada termo como semente própria	co-ocorrência, todos os termos	confirmação, nenhum termo novo
Busca por grafias alteradas	letra a letra	nenhum termo novo, vetor de evasão distinto

Fonte: dados da pesquisa.

Para a entrada associada a simbologia nazista (`Hitler/Hail`), a mineração rodou nos dois sentidos. Partindo dos marcadores visuais, como o emoji de pinguim e o emoji da bandeira da Alemanha, a vizinhança devolveu os marcadores escritos `hail` e `hitler`. Partindo do texto, devolveu o mesmo conjunto de emojis. A convergência se confirma, assim, nas duas direções.

A partir da quarta rodada, os métodos novos deixaram de produzir entradas e passaram apenas a confirmar ou a detalhar as sete já existentes, que é o sinal de **saturação teórica** descrito pela *grounded theory* (GLASER; STRAUSS, 1967). Isso encerrou a expansão do catálogo, cujo objetivo é sustentar a análise de erro, e não esgotar o vocabulário da comunidade.

Para aplicar esse **protocolo** a outras comunidades e construir um catálogo equivalente, três condições precisam estar presentes. É preciso um corpus em que cada termo apareça vezes suficientes para atender aos mínimos de frequência adotados acima, já

que o PMI é calculado a partir de contagens e oscila bastante quando elas são baixas. É preciso alguém com convivência na comunidade estudada, seja o próprio pesquisador ou um colaborador, e é preciso que os termos procurados tenham o perfil de aparência inocente com carga tóxica dada pelo contexto. Esse conjunto de condições é a contribuição de reprodutibilidade do trabalho.

3.5.3 Critério Editorial e Esquema de Categorias

Um termo só entra no catálogo quando reúne duas condições ao mesmo tempo. A primeira é a **inocência lexical**, ou seja, a palavra tem uma leitura cotidiana comum, como acontece com `estudar` e `bio`. A segunda é a **carga tóxica dada pelo contexto**, ou seja, no uso da comunidade o termo passa a apontar para algo potencialmente nocivo.

Cada entrada recebe depois uma categoria. A categoria é a política das Diretrizes da Comunidade do YouTube (YouTube, 2026b) mais próxima do uso que a comunidade faz do termo, com os nomes oficiais da plataforma mantidos. *Child Safety* recebe `CP`, `Trade` e `Bio`, *Harassment and Cyberbullying* recebe `Estudar`, *Violent Extremist or Criminal Organizations* (YouTube, 2026a) recebe `FA` e `FB`, e *Hate Speech* recebe `Hitler/Hail`. O argumento deixa de depender do que o pesquisador considera tóxico. Passa a ser que o termo corresponde ao tipo de conteúdo que a política define e que a plataforma se compromete a moderar.

Essa correspondência é uma ferramenta de análise e não um veredito. O trabalho não constata que uma infração ocorreu nem trata mensagem ou autor como responsável por crime, o que cabe à plataforma e às autoridades competentes. As categorias das Diretrizes são um conjunto de regras públicas, escritas por quem se propõe a moderar o conteúdo. Elas apoiam a gravidade de cada entrada em algo que não foi definido pelo pesquisador. A evidência de que o uso comunitário do termo se encaixa nessa classificação vem da imersão e da mineração de vizinhança descritas nas Seções 3.5.1 e 3.5.2, e não de qualquer decisão do YouTube sobre aquelas mensagens.

O catálogo final tem **sete entradas confirmadas**, e o caminho até elas foi feito de descartes. Uma lista inicial de 100 candidatos, montada com auxílio de um modelo de linguagem sobre as transcrições e os comentários dos vídeos que a própria comunidade produziu, foi abandonada por inteiro, porque os termos ou apareciam pouco no corpus ou não podiam ser verificados nele. Já com o catálogo em uso, duas entradas foram **removidas** depois de registrarem uma única ocorrência tóxica cada. O ramo de descarte do protocolo apresentado na Figura 2 funciona, portanto. A experiência do pesquisador propõe, mas é o corpus que decide.

O protocolo tem limites conhecidos. Ele só alcança termos que aparecem no corpus vezes suficientes para atender aos mínimos de frequência da Seção 3.5.2, e um termo

raro ou recém-criado fica de fora, como aconteceu com a lista inicial de candidatos. A mineração parte sempre de um termo já suspeito, de modo que um termo de que ninguém suspeitou não tem por onde ser alcançado, e a suspeita vem da vivência de um único pesquisador. A medida usada capta as palavras que aparecem perto da semente, e não as que ocupam o lugar dela, o que deixa de fora eufemismos que não convivem com o termo que substituem. O catálogo é, portanto, um registro parcial do repertório, suficiente para separar os estratos da análise de erro e não um inventário fechado do vocabulário da comunidade. O que esses limites significam para a leitura dos resultados está na Seção 5.3.

3.6 Rotulagem

A rotulagem manual da amostra de 3.500 mensagens produziu o conjunto de referência contra o qual os classificadores do Capítulo 4 são avaliados. Cada mensagem recebeu uma classificação binária, tóxica ou não-tóxica, e cada mensagem tóxica recebeu ainda uma segunda classificação, que registra por qual mecanismo a toxicidade é transmitida, se explícito, codificado ou misto. É essa segunda classificação que permite separar o desempenho dos modelos por tipo de toxicidade no Capítulo 4.

A rotulagem foi precedida por uma passagem de calibração sobre a amostra, descrita na Seção 3.6.1. Ela foi feita numa interface construída para este trabalho, que apresenta uma mensagem por vez, com os emojis exibidos como imagem e não como o código de texto em que são armazenados, e mantém o catálogo à vista. As decisões foram orientadas por um guia de rotulagem deliberadamente cético, segundo o qual só se marca como tóxico o que tem evidência concreta na própria mensagem, e toda dúvida se resolve a favor de não-tóxico. Essa assimetria é uma escolha de desenho. Como o trabalho argumenta que os classificadores erram na toxicidade codificada, marcar como tóxicos os casos apenas suspeitos aumentaria o número de erros atribuídos aos modelos e favoreceria a hipótese da pesquisa. O viés que se admite é o oposto, contra ela.

3.6.1 Calibração dos Critérios e Rotulagem

Antes da rotulagem definitiva, uma passagem de calibração sobre a amostra serviu para fixar os critérios de decisão. Foi nela que se calibrou o limiar de repetição de emoji apresentado na Seção 3.6.2 e que se anotaram os tipos de mensagem cuja classificação ficava em dúvida, o que orientou a reescrita do guia. Nenhuma entrada do catálogo final veio dessa passagem.

Com os critérios consolidados, o guia foi reescrito antes da rotulagem definitiva, em torno da regra de que toda dúvida se resolve como não-tóxico, sem exceções. Ele fixa o que conta como evidência concreta na mensagem e como decidir os casos ambíguos que a calibração havia mapeado, boa parte deles situações em que a superfície agressiva não

corresponde a toxicidade no uso local da comunidade. Um termo do catálogo só sustenta o rótulo tóxico quando o contexto ao redor sustenta essa leitura.

Com o guia estabilizado, as 3.500 mensagens da amostra foram rotuladas segundo esses critérios. O resultado é o conjunto de referência do trabalho, com **157 mensagens tóxicas**, ou 4,49% do total, e 3.343 não-tóxicas. Quando as duas entradas descartadas na Seção 3.5.3 saíram do catálogo, os rótulos que se apoiavam apenas nelas foram revistos, o que mantém catálogo e rótulos coerentes entre si. Esse conjunto é o que o texto chama adiante de *gold standard*.

Os rótulos foram produzidos por um único rotulador, o próprio pesquisador, seguindo o guia de rotulagem e as regras da Seção 3.6.2. A anotação por vários rotuladores independentes, que permitiria medir a concordância entre eles, fica como etapa posterior de consolidação do conjunto.

3.6.2 Regras de Decisão

Duas decisões se repetiam ao longo da rotulagem e foram fixadas em regras explícitas. As duas vieram da calibração e foram verificadas depois contra o corpus completo.

A primeira regra trata da sigla isolada. Uma mensagem cujo conteúdo é apenas a sigla `CP` ou `FA`, com ou sem pontuação e inclusive na forma interrogativa `FA?`, é classificada como tóxica, sem exceção. Três ordens de argumento sustentam essa decisão.

O primeiro argumento é empírico. Todas as ocorrências isoladas presentes na amostra rotulada, dez de `CP` e cinco de `FA`, receberam classificação tóxica na inspeção contextual, sem contraexemplo.

O segundo argumento é linguístico. A sigla isolada, sem nenhuma palavra ao redor, não funciona como informação, e sim como sinal. Em chats de transmissão ao vivo, esse tipo de mensagem curta e repetida cumpre sobretudo a função fática da linguagem (JAKOBSON, 1960), a de marcar presença e pertencimento, e não a de transmitir conteúdo. É o registro participativo típico desses chats de grande volume (FORD et al., 2017), e não o registro conversacional. A distinção aparece na própria sintaxe. O uso cotidiano dessas siglas vem acompanhado de complemento, um numeral, um verbo ou um objeto, como em "comprei 1000 CP no jogo". A sigla nua não tem complemento, e não sobra informação a comunicar. O que resta é a sinalização de filiação.

O terceiro argumento é de evasão. A forma isolada é justamente aquela em que o classificador não tem contexto sintático para desambiguar, restando-lhe apenas o léxico, no qual essas siglas não figuram como tóxicas.

O apoio da amostra é pequeno, quinze ocorrências ao todo, e por isso a regra foi verificada contra o corpus completo. Foram examinadas as **7.087** ocorrências de `CP` isolado

e as **1.741** de `FA` isolado em 1,74 milhão de mensagens, em busca de contraexemplos. No caso de `CP`, apenas três mensagens, ou 0,04% do total, ocorreram em transmissões de um jogo cuja moeda virtual tem a mesma sigla, e a inspeção qualitativa mostrou que nenhuma delas era uso informacional. No caso de `FA`, as 47 mensagens cuja vizinhança admitia a leitura cotidiana da sigla se reduziram, depois da inspeção, a menções sem relação com essa leitura. Nenhum contraexemplo se confirmou.

A regra também responde à objeção de que as siglas isoladas seriam spam de poucas contas, e não comportamento de comunidade. A distribuição por autor, calculada sobre o corpus já limpo na Seção 3.3, aponta o contrário. As 7.087 ocorrências de `CP` isolado vêm de 1.859 autores diferentes, em 140 vídeos de seis canais. Desses autores, 52,8% aparecem uma única vez, o mais ativo responde por 2,6% do total e são necessários 155 autores para cobrir metade das ocorrências. Muitos autores com contribuição pequena é o oposto do que se espera de contas automatizadas ou de poucos usuários intensos. O mesmo perfil se repete em `FA`, com 1.741 ocorrências distribuídas por 544 autores, 57,0% deles com uma única ocorrência.

A segunda regra trata da repetição de emoji. Para os sinais visuais do catálogo, o emoji conta como elemento dominante da mensagem, e o rótulo tóxico se torna aplicável quando aparece repetido quatro vezes ou mais, ou quando não há texto competindo com ele. A literatura revisada não oferece um limiar pronto para esse caso. A detecção de spam opera em outros níveis, como a proporção de palavras repetidas e a taxa de mensagens por usuário (SAHAMI et al., 1998), e os estudos de alongamento lexical tratam da repetição de caracteres dentro da palavra, não da repetição do sinal inteiro (BRODY; DIAKOPOULOS, 2011). O limiar foi então calibrado sobre a amostra e submetido a um diagnóstico de sensibilidade, apresentado na Tabela 6. Cada linha refaz a regra com uma exigência diferente, de duas a sete repetições, e mostra quantas mensagens da amostra essa versão atingiria e quantas delas haviam sido marcadas como tóxicas na rotulagem.

Tabela 6 – Sensibilidade do limiar de repetição de emoji, com a regra simulada contra os rótulos reais

Limiar N	Mensagens atingidas	Rotuladas tóxicas	Acordo com a regra
2	16	13	81,2%
3	**14**	**13**	**92,9%**
4	**14**	**13**	**92,9%**
5	**14**	**13**	**92,9%**
6	13	12	92,3%
7	10	10	100,0%

Fonte: dados da pesquisa.

Exigir três, quatro ou cinco repetições dá exatamente o mesmo resultado, e a razão

está na amostra. Nenhuma mensagem repete o emoji três ou quatro vezes. As repetições observadas são de uma ou duas vezes, ou então de cinco para cima. Como esse intervalo é vazio, qualquer exigência dentro dele atinge o mesmo conjunto de catorze mensagens. A escolha de quatro é, portanto, indistinguível de suas vizinhas, e o que sustenta a regra é esse vazio na distribuição, não o número escolhido.

3.6.3 Classificação por Mecanismo: Explícito, Codificado e Misto

Concluída a rotulagem binária, cada uma das 157 mensagens tóxicas foi classificada pelo mecanismo da sua toxicidade. Uma mensagem é explícita quando a toxicidade está na superfície do texto, como no insulto direto, na ameaça ou no termo chulo dirigido a alguém, e a leitura independe do repertório da comunidade. É codificada quando a toxicidade é transmitida apenas por termo, estrutura ou sinal visual do catálogo, sem marca tóxica na superfície. É mista quando os dois sinais aparecem na mesma mensagem.

A classificação não foi atribuída de forma automática. Uma busca por expressões regulares, que procura os termos do catálogo e os emojis, serviu de triagem inicial, e as 157 mensagens foram revistas uma a uma pelo pesquisador. A revisão corrigiu erros nas duas direções, tanto mensagens postas no codificado sem que estivessem quanto mensagens codificadas que a busca não havia alcançado, e no saldo reduziu o número de codificadas em relação ao que a triagem automática indicava, o que caminha contra a hipótese da pesquisa, não a favor dela. A distribuição final, **102 explícitas, 53 codificadas e 2 mistas**, define os estratos sobre os quais a análise de erro é reportada.

3.7 Desenho dos Experimentos de Análise de Erro

Com o catálogo construído e o *gold standard* estabelecido, a segunda frente do trabalho executa duas famílias de experimentos sob o mesmo protocolo. A primeira compara três classificadores de estado da arte, na Seção 3.7.1, e testa se a falha na toxicidade codificada aparece em modelos de naturezas diferentes ou se decorre de uma escolha ruim de modelo. A segunda, na Seção 3.7.2, fornece o catálogo a modelos de linguagem de grande porte e testa se isso corrige a falha. As duas rodam sobre as mesmas 3.500 mensagens, contra o mesmo *gold standard* e com a mesma separação por mecanismo, o que torna os resultados comparáveis entre si.

3.7.1 Seleção dos Classificadores de Estado da Arte

A seleção não buscou os três melhores modelos de uma lista de desempenho. Cada modelo escolhido tem uma vantagem diferente sobre o problema, e cada vantagem corresponde a uma explicação possível para a falha, como mostra a Tabela 7. Ao falhar, o modelo derruba a explicação que representa. Se três modelos com vantagens diferentes

falham no mesmo estrato, a causa não está na vantagem particular de cada um, e resta o que é comum aos três, o desconhecimento do código local da comunidade.

Tabela 7 – Os três classificadores e a explicação concorrente que cada um testa

Modelo	**Vantagem que deveria bastar**	**Explicação concorrente que derruba**
Detoxify *multilingual* (XLM-RoBERTa / Jigsaw) (HANU; Unitary team, 2020)	Classificador genérico multilíngue, com português	"Um moderador automático genérico resolveria"
BERTimbau *fine-tuned* em HateBR (SOUZA; NOGUEIRA; LOTUFO, 2020; VARGAS et al., 2022)	PT-nativo, treinado em ofensa brasileira	"Falhou porque não é português / não viu ofensa BR"
BERTweet.BR *fine-tuned* em ToLD-BR (LEITE et al., 2020)	Registro de rede social BR ponta a ponta (pré-treino e ajuste em tweets brasileiros)	"Falhou porque não viu gíria informal de internet"

Fonte: elaborado pelo autor.

Os três modelos rodam sob a mesma configuração. O texto entra cru, com os códigos de emoji preservados, que é a forma como um sistema de moderação da própria plataforma receberia a mensagem. Cada modelo devolve um valor contínuo, convertido em rótulo tóxico ou não-tóxico no ponto de corte padrão de 0,5. Esses valores ficam salvos mensagem a mensagem, o que permite a varredura de pontos de corte descrita na Seção 3.7.3.

3.7.2 Experimento com LLMs: o Catálogo como Intervenção

A comparação anterior avalia os classificadores como eles são entregues. O segundo experimento pergunta o que acontece quando o código da comunidade é fornecido de forma explícita ao modelo, ou seja, se dar o catálogo a um modelo de linguagem altera a detecção da toxicidade codificada. O desenho informa nos dois desfechos possíveis. Se altera, o catálogo passa a valer também como intervenção de moderação, e não apenas como instrumento de medida. Se não altera, o que o primeiro experimento tiver medido não vem apenas da ausência de documentação, já que nem a instrução explícita recuperaria o código.

Cada modelo processa as 3.500 mensagens em três condições, apresentadas na Tabela 8. O que varia entre elas é o que acompanha a mensagem no pedido enviado ao

modelo, nos regimes *zero-shot* e *few-shot*. As condições são comparadas duas a duas. De c0 para c1 acrescentam-se exemplos já resolvidos, e de c1 para c2 acrescenta-se o catálogo. Cada comparação isola assim o efeito de um único elemento. Os contrastes foram fixados antes da execução, e não escolhidos depois de vistos os resultados.

Tabela 8 – Condições experimentais do experimento com LLMs

#	**Condição**	**O prompt contém**
c0	*Zero-shot*	Definição geral de toxicidade + a mensagem
c1	*Few-shot* genérico	c0 + 5 exemplos tóxicos explícitos + 5 não-tóxicos
c2	*Few-shot* + catálogo	c1 + o catálogo de termos (*codebook*)

Fonte: elaborado pelo autor.

Dois cuidados protegem a validade do experimento. Os exemplos resolvidos da condição c1 foram escolhidos no corpus completo, fora da amostra avaliada, e uma verificação automática confirma que nenhum deles está entre as 3.500 mensagens. O catálogo entregue na condição c2 é gerado a partir da planilha, com as entradas transcritas sem alteração. A instrução enviada aos modelos nas três condições está reproduzida na Seção A.4.

O desenho separa dois papéis entre os modelos. O primeiro é o de teto comercial, ocupado pela família Gemini 2.5, com a versão *pro* como teto propriamente dito e a *flash* como segundo ponto, o que verifica se o resultado se mantém dentro de um mesmo fornecedor. O acesso se deu por via institucional, sem custo. O trabalho os descreve como modelos comerciais de fronteira, e não como os melhores modelos existentes. O segundo papel é o dos modelos de pesos abertos, Qwen3.5-9B (Qwen Team, 2026) e Llama-3.1-8B-Instruct (Llama Team, AI @ Meta, 2024), que podem ser reexecutados por terceiros sem custo e representam o caminho viável para aplicar o processo ao corpus completo. Duas arquiteturas que convergem no mesmo resultado afastam a leitura de que o achado seja particularidade de um modelo.

Todos os modelos seguem o mesmo protocolo de execução, para que a comparação entre eles seja justa. A instrução e a mensagem a ser classificada seguem juntas num único pedido, sem uso do campo de instrução de sistema que algumas interfaces oferecem à parte. A interface também não obriga o modelo a devolver a resposta num formato fixo, recurso que não estava disponível no acesso comercial usado e foi desligado também nos modelos abertos, para manter a uniformidade. O formato JSON é pedido no próprio texto da instrução, e não garantido pela ferramenta, de modo que a leitura da resposta tolera

pequenas variações na formatação. A temperatura é fixada em 0, o que reduz a variação entre execuções. A execução é retomável, já que cada mensagem processada é salva no momento, e uma interrupção não obriga a reprocessar tudo desde o início. Os modelos abertos rodam em modo direto, sem produzir uma cadeia de raciocínio explícita antes da resposta, pela eficiência que o processo exige quando for aplicado ao corpus completo.

3.7.3 Métricas e Protocolo de Avaliação

A análise de erro se apoia no *recall*, a proporção das mensagens tóxicas que o modelo de fato detecta. Ele é calculado duas vezes, uma sobre as 102 mensagens explícitas e outra sobre as 53 codificadas, seguindo os estratos da Seção 3.6.3. Um único valor, calculado sobre as 157 mensagens tóxicas de uma vez, seria dominado pelas explícitas, que são a maioria, e um modelo cego ao código ainda assim apareceria com um número aceitável.

As demais métricas seguem o padrão adotado no trabalho. A precisão é a proporção das mensagens marcadas como tóxicas que de fato são tóxicas, e é reportada de forma explícita, porque em moderação o falso positivo tem custo prático, já que cada um é uma mensagem legítima sinalizada sem motivo. O F1 da classe tóxica resume precisão e *recall* numa medida só, que fica alta apenas quando as duas são altas ao mesmo tempo, e é a métrica principal de comparação entre modelos. A acurácia, que é a proporção de acertos sobre todas as mensagens, não é usada, porque a amostra tem 4,49% de mensagens tóxicas e um modelo que marcasse todas como não-tóxicas já acertaria 95,5%.

Para os modelos de saída contínua, os pontos de corte são varridos por completo a partir dos valores brutos salvos, o que antecipa a objeção de que bastaria calibrar o corte. Os modelos de linguagem devolvem rótulo já binário, para o qual a varredura se reduz a um ponto único e não é reportada, e a comparação com eles se faz pelas métricas pontuais comuns aos dois grupos.

Uma análise de sensibilidade acompanha o desenho. Uma variante da entrada retira a camada visual, tanto os códigos de emoji quanto os emojis do padrão Unicode, mantendo os mesmos rótulos. Ela responde à objeção de que os códigos de emoji no meio do texto atrapalhariam a leitura do modelo. Se o resultado se mantém com e sem essa camada, a falha está no sinal que o modelo não reconhece, e não na forma como o texto chega até ele.

Os três classificadores da primeira família rodaram na máquina do pesquisador, em hardware de consumo, uma placa de vídeo GTX 1060 de 6 GB. A escolha mantém os dados sensíveis fora de serviços de nuvem, conforme a Seção 3.8. Os modelos de pesos abertos rodaram no Colab, com placa T4, e os comerciais pela via institucional já descrita. Em todos os casos os valores brutos de cada mensagem ficam salvos, o que permite recalcular as métricas e avaliar outros pontos de corte sem executar os modelos de novo, com as

versões das bibliotecas fixadas em arquivo de requisitos.

3.8 Considerações Éticas

A natureza pública do material e os limites observados na coleta estão descritos na Seção 3.2. Esta seção reúne as demais decisões éticas do trabalho.

A vivência do pesquisador como membro da comunidade, apresentada na Seção 3.5.1, é fonte de descoberta e não procedimento de geração de dados. Ainda assim, mensagens publicadas por ele na janela de coleta aparecem no material. Para evitar circularidade entre observador e objeto, essas mensagens foram identificadas e retiradas de todas as análises, num total de 177, ou 0,010% do material, removidas já na origem do corpus, como registra a Tabela 3. Nenhuma delas integra a amostra rotulada. As estatísticas de concentração por autor e de co-ocorrência da Seção 3.6.2 foram recalculadas depois desse filtro, de modo que os resultados apresentados já consideram a remoção.

O objeto da análise são as mensagens e o repertório da comunidade, e não os indivíduos. Os canais são referidos por pseudônimos, de Canal_A a Canal_G, com o mapeamento real mantido fora do repositório público. Nenhum autor é identificado, os exemplos são citados sem autoria e os identificadores de autor e de vídeo são substituídos, nos arquivos derivados, por códigos que não permitem voltar ao original. Quando o texto de uma mensagem reproduzida neste documento nomeia uma pessoa, seja um usuário do chat ou o dono do canal, o nome é substituído por um marcador entre colchetes. Os dados brutos não são publicados, apenas material agregado ou anonimizado.

Pela mesma razão, as duas organizações criminosas cujas siglas integram o catálogo não são nomeadas. O nome e a sigla de cada uma são substituídos, em todo o documento, pelos marcadores Facção A, ou `FA`, e Facção B, ou `FB`. A análise não depende de qual organização é qual, e sim do uso da sigla como sinal de filiação. O mascaramento é uma medida de segurança dos pesquisadores, que evita associar o documento e seus autores aos nomes reais dessas organizações.

Este documento reproduz exemplos reais da comunidade, incluindo referências veladas a conteúdo que corresponde às categorias *Child Safety* e *Hate Speech* das Diretrizes da Comunidade do YouTube, apenas quando são indispensáveis ao argumento, seguindo a literatura sobre tratamento de texto nocivo em pesquisa de PLN (KIRK et al., 2022a). A severidade do material não é acidental, e sim a razão de documentá-lo, já que registrar um repertório codificado sob superfície inocente é condição para que plataformas e moderadores possam detectá-lo.

A execução dos classificadores da primeira família foi inteiramente local, como descreve a Seção 3.7.3, e nessa etapa as mensagens não passaram por serviços de terceiros.

No experimento com modelos comerciais, apenas o texto das mensagens, já sem identificação de autores, compõe os pedidos enviados.

4 Resultados

A Seção 4.1 apresenta o catálogo de termos codificados, que é ao mesmo tempo um produto do trabalho e o critério que separa as mensagens de toxicidade codificada das demais. As Seções 4.2 e 4.3 medem o desempenho sobre essas mensagens em duas etapas: primeiro três classificadores de estado da arte, tal como são entregues, depois modelos de linguagem de grande porte, aos quais o catálogo é fornecido como intervenção. A Seção 4.4 reúne os detectores e responde à questão de pesquisa. Exceto quando indicado, os números vêm da amostra de 3.500 mensagens de chat da Seção 3.4 e comparam a resposta de cada detector aos rótulos que o pesquisador atribuiu a essas mesmas mensagens. Dessas 3.500, são 157 as tóxicas, 102 por toxicidade explícita, 53 por toxicidade codificada e 2 pelas duas ao mesmo tempo, conforme a Seção 3.6.3. As métricas e os limiares seguem o protocolo da Seção 3.7.3.

4.1 O catálogo de termos codificados

O primeiro resultado é o próprio instrumento, o catálogo de termos codificados da comunidade, construído e validado na Seção 3.5. A versão final tem sete entradas confirmadas, que passaram pela validação por mineração de vizinhança e pelo critério editorial de inocência lexical. A Tabela 9 apresenta cada entrada com o significado que tem no uso da comunidade, a política de moderação correspondente e o seu alcance no corpus completo, de 1,74 milhão de mensagens e sem as mensagens do próprio pesquisador, conforme a Seção 3.8. A versão íntegra, com as variantes, os disfarces, os exemplos de uso e as observações de rotulagem de cada entrada, está no Apêndice A.

Tabela 9 – O catálogo de termos codificados e o seu alcance no corpus completo (versão sintética, com a íntegra no Apêndice A)

Termo	**Significado no uso comunitário**	**Política correspondente (YouTube)**	**Mens.**	**Autores**
`CP`	Referência a pornografia infantil	*Child Safety*	14.915	3.891
`Estudar`	Eufemismo para estuprar	*Harassment and Cyberbullying*	6.275	2.626
`FA`	Sigla de facção criminosa (Facção A, nome mascarado)	*Violent Extremist or Criminal Organizations*	2.669	844
`Hitler / Hail`	Saudação nazista e simbologia associada	*Hate Speech*	1.725	882
`Trade`	Eufemismo para comercialização de pornografia infantil	*Child Safety*	481	193
`Bio / link in bio`	Ponteiro a canal de distribuição	*Child Safety*	303	166
`FB`	Sigla de facção criminosa rival (Facção B, nome mascarado)	*Violent Extremist or Criminal Organizations*	197	117

Fonte: dados da pesquisa. As contagens medem a ocorrência do padrão textual, que inclui usos inocentes. Nomes e siglas de facções mascarados (`FA`, `FB`), conforme a política de anonimização da Seção 3.8.

A tabela permite três leituras. A primeira é a do alcance das políticas, já que as sete entradas correspondem a quatro delas, de modo que o repertório não se concentra num único tipo de conteúdo. A segunda é a da complementaridade dentro de *Child Safety*, a política de proteção à criança, em que `CP`, `Trade` e `Bio / link in bio` cobrem lados diferentes do mesmo referente: o consumo, a transação e o caminho de acesso. A terceira é a de que uma mesma entrada costuma reunir sinais de naturezas diferentes: textuais e visuais, como registram as entradas de `FA` e de `Hitler / Hail` no Apêndice A.

Duas qualificações acompanham as contagens. A primeira retoma o aviso da tabela, o de que a contagem é de superfície, e a entrada `Bio` mostra por que isso exige cuidado. A palavra `bio` sozinha tem uso comum e inocente, e quem carrega o sentido codificado é a expressão inteira, `link in bio` ou `link na bio`. Das 303 ocorrências contadas, 93% já são a expressão inteira, e apenas 22 são a palavra sozinha. A segunda diz respeito à escala e à dispersão. `CP` lidera o repertório tanto em mensagens quanto em autores distintos, e a análise de concentração da Seção 3.6.2 mostrou que o uso se espalha por muitas contas em vez de se concentrar em poucas, o que afasta a leitura de que seja spam de poucas contas.

Cada termo do catálogo tem, portanto, centenas ou milhares de autores distintos.

Foi o catálogo que decidiu quais mensagens da amostra contam como de toxicidade codificada, as 53 que formam o estrato codificado, conforme a Seção 3.6.3.

4.2 Os classificadores de estado da arte

Esta seção reporta os três classificadores selecionados na Seção 3.7.1, na ordem da Tabela 7. Cada um deles testa uma explicação concorrente para a falha na toxicidade codificada. Os números seguem o protocolo da Seção 3.7.3, com as mesmas 3.500 mensagens, o mesmo conjunto de referência, a mesma separação por mecanismo e o limiar padrão de 0,5, acompanhado da varredura de limiares feita a partir das pontuações salvas.

4.2.1 Detoxify e a cegueira ao código

O classificador genérico multilíngue produziu F1 de 0,171 para a classe tóxica, e esse valor não muda quando o limiar baixa de 0,5, o padrão, para 0,35, o melhor ponto da varredura. A métrica global é o ponto de partida, e a estrutura do erro aparece na separação por mecanismo da Tabela 10.

Tabela 10 – Recall do Detoxify por mecanismo de toxicidade

Estrato de referência	**n**	**Recall @0,5**	**Recall @ótimo (0,35)**
Explícito	102	49,0%	58,8%
Misto	2	100%	100%
Codificado	53	**1,9%**	**3,8%**

Fonte: dados da pesquisa.

O achado tem três leituras. A primeira é a diferença entre os estratos, já que o modelo detecta cerca de metade das mensagens explícitas e apenas 1 das 53 codificadas. A segunda responde à objeção mais imediata, a de que o limiar estaria mal escolhido. Baixar o corte para o melhor ponto da varredura recupera apenas mais uma mensagem codificada, de 1 para 2, enquanto o estrato explícito sobe quase dez pontos, de modo que nenhum limiar testado reduz a diferença entre os dois estratos. A terceira vem do estrato misto, em que as duas mensagens que combinam código e material explícito foram detectadas, o que sugere que o modelo responde à parte explícita da mensagem.

4.2.2 BERTimbau/HateBR e o colapso de calibração

O segundo modelo, o BERTimbau ajustado em ofensa brasileira, produziu o pior número global dos três, com F1 de 0,126, abaixo do genérico, mas por um modo de falha

diferente. A taxa de positivos ficou em 47,9%, ou seja, quase metade das mensagens da amostra foi marcada como tóxica, o que caracteriza um colapso de calibração quando o modelo é aplicado fora do domínio em que foi treinado.

Esse colapso cria uma leitura enganosa. O recall do HateBR no estrato codificado é de 71,7%, e lido isoladamente sugeriria que o modelo reconheceu o código que o genérico não viu. Três números indicam que não foi isso que aconteceu. Um classificador que marcasse 47,9% das mensagens ao acaso já alcançaria cerca de 47,9% de recall nos dois estratos, de modo que a taxa observada fica pouco acima do que a marcação indiscriminada produziria. O recall codificado, de 71,7%, é praticamente igual ao explícito, de 74,5%, quando um modelo que reconhecesse o código deveria separar os dois estratos. E a varredura de limiar não resgata o desempenho, pois o melhor corte eleva o F1 apenas a 0,131.

O modelo cumpre o papel que tinha no desenho. A explicação de que a falha viria do idioma ou da ausência de ofensa brasileira nos dados de treino não se sustenta, já que o modelo treinado em português também não recupera a toxicidade codificada. Os dois modos de falha observados até aqui, a cegueira do Detoxify e o disparo excessivo do HateBR, mostram que o recall e a precisão falham de forma independente.

4.2.3 BERTweet.BR/ToLD-BR e o limite da exposição ao registro

O terceiro modelo, pré-treinado e ajustado em toxicidade de tuítes brasileiros, é o melhor e o mais bem calibrado dos três, com F1 de 0,176 no limiar padrão, 0,204 no melhor corte, que fica em 0,54, e taxa de positivos de 12,4%, sem o colapso observado no modelo anterior. É esse desempenho que dá peso ao seu resultado no estrato codificado, em que o recall fica em 3,8%, ou 2 mensagens em 53, praticamente igual ao 1,9% do Detoxify genérico. O modo de falha também é o mesmo, com 48,0% de recall no estrato explícito contra os 3,8% do codificado.

Este era, entre os três, o modelo com maior exposição ao registro da amostra, com linguagem informal, abreviações e gíria de rede social brasileira, e essa exposição não melhorou o desempenho no estrato codificado. A explicação de que a falha viria do desconhecimento da gíria de internet também não se sustenta, o que reduz as explicações disponíveis ao conhecimento do repertório da própria comunidade.

4.2.4 Análise de sensibilidade à camada visual

Resta a objeção prevista na Seção 3.7.3, a de que os códigos de emoji preservados no texto cru poderiam atrapalhar a leitura do modelo, de modo que a falha seria de formato da entrada e não de desconhecimento do código. A variante de controle executa o Detoxify outra vez com a camada visual removida da entrada, e a Tabela 11 compara as duas execuções.

Tabela 11 – Análise de sensibilidade do Detoxify à camada visual

Métrica @0,5	**Com camada visual**	**Sem camada visual**
F1 (classe tóxica)	0,171	0,173
Recall explícito	49,0%	49,0%
Recall codificado	**1,9%**	**1,9%**

Fonte: dados da pesquisa.

O resultado praticamente não muda. Remover a camada visual não recupera nenhuma mensagem codificada, o que indica que falta sinal no texto e não que o formato da entrada atrapalhe. Em 16 das 53 mensagens codificadas o texto fica vazio depois da remoção, com o conteúdo inteiro apoiado nos emojis, e nesses casos não há o que o pré-processamento corrija.

4.2.5 Síntese dos três classificadores

A Tabela 12 reúne os números dos três modelos e o modo de falha de cada um.

Tabela 12 – Síntese dos três classificadores de estado da arte

Modelo	**F1 @0,5**	**F1 @melhor**	**Tx. positivos**	**Recall codif.**	**Modo de falha**
Detoxify *multilingual*	0,171	0,171	13,2%	**1,9%**	cegueira
BERTimbau/ HateBR	0,126	0,131	47,9%	71,7%*	colapso
BERTweet.BR/ ToLD-BR	0,176	0,204	12,4%	**3,8%**	cegueira

* Não é detecção do código. O modelo marca 47,9% da amostra, e marcar essa proporção ao acaso já produziria recall próximo desse valor, sem reconhecer nada. Ver a Seção 4.2.2.

Fonte: dados da pesquisa.

Três modelos, com três vantagens diferentes, falham no estrato codificado. Dois por cegueira, com no máximo 3,8% de recall, e um por colapso que não separa os estratos. Cada um afasta a explicação concorrente que carregava, já que o Detoxify é um moderador automático genérico e multilíngue, aplicado sem adaptação ao domínio, o BERTimbau foi treinado em português e em ofensa brasileira e o BERTweet.BR foi pré-treinado e ajustado em tuítes brasileiros. Afastadas essas explicações, sobra a que os três compartilham, a de que o repertório codificado da comunidade não está representado nos dados com que esses

modelos foram treinados. O que falta a eles não é capacidade, e sim o conhecimento que o catálogo documenta.

4.3 O catálogo como intervenção no experimento com LLMs

Os três classificadores da seção anterior estabeleceram o diagnóstico, e este experimento testa a intervenção. A pergunta, formulada na Seção 3.7.2, é se o catálogo corrige a cegueira à toxicidade codificada. Nas três condições da Tabela 8, a passagem de c0 para c1 mede o efeito dos exemplos genéricos e a de c1 para c2 isola a contribuição do catálogo. Os resultados vêm primeiro para os modelos de pesos abertos e depois para o modelo comercial.

4.3.1 Modelos de pesos abertos

A Tabela 13 consolida as três condições para os dois modelos de pesos abertos.

Tabela 13 – Recall por eixo e precisão dos modelos de pesos abertos (Qwen3.5-9B e Llama-3.1-8B) nas três condições

Condição	**Codificado (n=53)**		**Explícito (n=102)**		**Precisão**	
	Qwen	Llama	Qwen	Llama	Qwen	Llama
c0 (*zero-shot*)	9,4%	30%	80%	74%	12,3%	12,8%
c1 (*few-shot* genérico)	19%	49%	87%	83%	11,8%	10,8%
c2 (*few-shot* + catálogo)	**96%**	**96%**	80%	72%	16,2%	12,2%

Fonte: dados da pesquisa.

A linha de partida mantém a continuidade com a seção anterior. Em c0, em que o pedido traz apenas a definição geral de toxicidade e a mensagem, os dois abertos ficam no estrato codificado bem abaixo do que alcançam no explícito, com 9,4% no Qwen e 30% no Llama. O primeiro fica abaixo de 10%, como os classificadores dedicados, que ficaram entre 1,9% e 3,8%, e o segundo, ainda que acima deles, perde 7 de cada 10 mensagens codificadas. São dois tipos de detector diferentes com o mesmo ponto de falha.

Os contrastes contam o resto. Os exemplos genéricos, de c0 para c1, elevam nos dois modelos o recall no estrato explícito, que é o que eles cobrem, e elevam também o codificado, de 9,4% para 19% no Qwen e de 30% para 49% no Llama, mas sem que ele chegue à metade das 53 mensagens. A entrada do catálogo, de c1 para c2, tem outra ordem de efeito, com os dois modelos indo a 96%, o que corresponde a mais 41 detecções no Qwen e mais 25 no Llama.

A convergência entre as arquiteturas qualifica o achado. Em c2, Qwen e Llama não apenas alcançam o mesmo recall como recuperam praticamente o mesmo conjunto de

mensagens, com 50 das 53 detectadas pelos dois e uma detecção exclusiva de cada. São duas arquiteturas distintas, treinadas por organizações diferentes, com o mesmo resultado sob a mesma intervenção, o que indica que o efeito vem do catálogo e não de uma particularidade de um dos modelos.

Há ainda uma assimetria a registrar. No Qwen, c2 é o melhor ponto do experimento, com F1 de 0,273, o único acima de 0,21 em todo o experimento, e o catálogo elevou tanto o recall quanto a precisão. No Llama, o ganho ficou restrito ao recall, e mesmo nesse ponto a precisão fica em 16,2%.

4.3.2 Gemini 2.5 e a dependência da documentação

A execução comercial usou o modelo intermediário da família Gemini 2.5, o *flash*, segundo ponto previsto na Seção 3.7.2. A Tabela 14 consolida as três condições.

Tabela 14 – As três condições no Gemini 2.5 *flash*

Condição	**Recall codif. (n=53)**	**Recall explíc. (n=102)**	**Precisão**	**F1**	**TP / FP**
c0 (*zero-shot*)	43%	90%	24,6%	0,370	117 / 358
c1 (*few-shot* genérico)	**2%**	94%	15,8%	0,253	99 / 527
c2 (*few-shot* + catálogo)	**98%**	89%	18,5%	0,309	145 / 637

Fonte: dados da pesquisa.

A leitura tem três partes. A primeira é o conhecimento que o modelo já traz. Em c0, o comercial detecta 43% do estrato codificado, acima dos abertos, que ficaram em 30% e 9,4%, o que indica que ele reconhece parte do código sem que o pedido o descreva. Ainda assim perde 57% do estrato, de modo que esse conhecimento prévio não substitui a documentação.

A segunda é o comportamento de c1. Os exemplos genéricos, que nos modelos abertos ajudavam, aqui reduzem o recall codificado de 43% para 2%, enquanto o explícito sobe. A explicação mais provável é que os cinco exemplos de toxicidade explícita concentrem a atenção do modelo num padrão de superfície e afastem o conhecimento prévio sobre o código. O resultado contraria a expectativa que acompanha o *few-shot* desde Brown et al. (2020), a de que exemplos resolvidos melhoram a resposta do modelo, já que para a toxicidade codificada eles não foram apenas pouco eficazes, mas prejudicaram o desempenho deste modelo.

A terceira é o efeito do catálogo, que se repete aqui. De c1 para c2, o recall codificado sobe de 2% para 98%, contra os 96% que Qwen e Llama alcançaram na mesma condição.

4.3.3 Convergência entre arquiteturas e o resto irredutível

Reunindo os três modelos em c2, a convergência observada entre os abertos se estende ao modelo comercial. A interseção das detecções codificadas de Gemini, Qwen e Llama é de 50 das 53 mensagens, e a união, de 52. Do modelo aberto de 9 bilhões de parâmetros ao comercial de ponta, a mesma intervenção recupera praticamente o mesmo conjunto, o que indica que o resultado varia pouco com a arquitetura e com a escala.

A única mensagem codificada que escapa aos três modelos, mesmo com o catálogo disponível, é *"[streamer] eu gosto de CPF sem F oque você acha"*, em que a sigla `CP` é reconstruída pela subtração de uma letra de outra sigla, a do documento de identificação nacional, que é inocente e de uso corrente. Nenhuma palavra da mensagem está no catálogo, e o referente só aparece a partir de uma operação sobre elas. É uma mensagem apenas, mas ela delimita o alcance da intervenção, que cobre as formas registradas no catálogo e não as que se produzem por operação sobre palavras de fora dele, ainda que o resultado seja um termo documentado.

4.3.4 O eixo da precisão e o custo operacional

Tudo o que foi exposto até aqui diz respeito ao recall. No outro eixo o quadro é diferente, já que a precisão fica baixa em todas as condições e em todos os modelos testados, entre 10,8% e 16,2% nos abertos e entre 15,8% e 24,6% no comercial. Na melhor condição de precisão do experimento, que é o Gemini em c0, com 24,6% e F1 de 0,370, 3 de cada 4 mensagens sinalizadas são falsos positivos, e nas demais condições a proporção chega a 8 ou 9 em cada 10. O catálogo não responde por esse excesso, que já aparece em c0 e c1, condições em que ele não é fornecido.

O experimento mostra, portanto, uma falha em duas direções independentes, estrutura que os classificadores da Seção 4.2 já haviam apresentado na forma da cegueira e do colapso. A do recall tem conserto, e o conserto é o catálogo, que recupera de 96% a 98% do estrato codificado nas três arquiteturas testadas. A da precisão permanece sem conserto, já que em nenhuma condição testada, nem mesmo no modelo comercial, mais de um quarto das mensagens sinalizadas é de fato tóxica, e cada mensagem legítima sinalizada tem um custo para quem modera.

4.4 Síntese e resposta à questão de pesquisa

A Tabela 15 reúne todos os detectores do capítulo sob as três métricas que estruturaram a análise. Para os modelos de linguagem, são mostradas as duas condições que respondem à questão de pesquisa, a que não fornece o catálogo, c0, e a que o fornece, c2. As condições intermediárias estão na Seção 4.3.

Tabela 15 – Síntese geral dos detectores sobre as mesmas 3.500 mensagens

Detector	**Acesso ao código**	**Recall codif.**	**Recall explíc.**	**Precisão**
Detoxify *multilingual*	nenhum	1,9%	49,0%	11,5%
BERTimbau/HateBR	nenhum	71,7%*	74,5%	6,9%
BERTweet.BR/ToLD-BR	nenhum	3,8%	48,0%	12,0%
Qwen3.5-9B	nenhum (c0)	9,4%	80%	12,3%
Qwen3.5-9B	**catálogo (c2)**	**96%**	80%	16,2%
Llama-3.1-8B	nenhum (c0)	30%	74%	12,8%
Llama-3.1-8B	**catálogo (c2)**	**96%**	72%	12,2%
Gemini 2.5 *flash*	nenhum (c0)	43%	90%	24,6%
Gemini 2.5 *flash*	**catálogo (c2)**	**98%**	89%	18,5%

* Efeito do colapso de calibração, já que o modelo marca 47,9% da amostra. Ver a Seção 4.2.2.

Fonte: dados da pesquisa.

A primeira leitura é a de que a falha é estrutural. Na coluna do recall codificado, entre as linhas sem catálogo, o classificador genérico multilíngue fica em 1,9%, o modelo de registro brasileiro de rede social em 3,8% e os modelos de linguagem entre 9,4% e 43%, enquanto os mesmos detectores ficam entre 48% e 90% no estrato explícito. Detectores de naturezas diferentes, classificadores dedicados de duas linhagens e modelos de linguagem de grande porte, apresentam o mesmo padrão, com desempenho muito menor no codificado do que no explícito. Uma diferença que se repete assim entre arquiteturas e escalas dificilmente vem da limitação de um detector em particular. A explicação mais provável está no próprio código, cujo reconhecimento depende do repertório da comunidade. As duas exceções aparentes se resolvem quando examinadas, já que o recall alto do HateBR vem do colapso de calibração e não da detecção, e o conhecimento prévio do modelo comercial, de 43%, ainda deixa passar a maioria das mensagens codificadas.

A segunda leitura é a resposta à questão de pesquisa, que se dá em dois eixos. No eixo do recall, a falha tem conserto, e o experimento identifica qual é, já que nas linhas com catálogo os três modelos de linguagem chegam a valores entre 96% e 98%, recuperando praticamente o mesmo conjunto de mensagens. É a documentação que recupera, e não o modelo. Sem ela, o modelo comercial perde 57% do estrato e os exemplos genéricos chegam a prejudicar o desempenho, e com ela a arquitetura e a escala deixam de fazer diferença entre os modelos testados. No eixo da precisão a resposta é negativa em todas as linhas da tabela, com valores entre 6,9% e 24,6%, e o maior deles vem de uma condição sem catálogo.

A terceira leitura reúne os dois eixos. A precisão baixa é uniforme, atravessa classificadores dedicados e modelos de linguagem, com e sem catálogo, numa amostra em

que 4,49% das mensagens são de fato tóxicas. A leitura que este trabalho propõe se apoia no registro comunitário mapeado durante a rotulagem, cujos protocolos de desambiguação foram criados porque muita mensagem de aparência agressiva não é tóxica no uso da comunidade, conforme a Seção 3.6.1. Nesta comunidade, o código torna tóxico o que parece inocente, o que derruba o recall de quem não o conhece, e o registro comunitário torna inofensivo muito do que parece tóxico, o que derruba a precisão de quem julga pela superfície. Os falsos negativos e os falsos positivos seriam, então, dois efeitos da mesma ausência, a do contexto cultural local, e é por isso que o catálogo corrige um eixo e não o outro, pois documenta o código e não o registro comunitário. Moderar esta comunidade exigiria as duas competências, e a segunda permanece sem instrumento.

Os resultados dialogam com a revisão do Capítulo 2 em quatro pontos. A literatura de toxicidade implícita já havia estabelecido que as ferramentas genéricas falham onde não há palavra ofensiva na superfície (ELSHERIEF et al., 2021; OCAMPO et al., 2023), e os recalls codificados deste trabalho, de 1,9% e 3,8% nos dois classificadores dedicados que não colapsaram, situam o código de comunidade num regime mais severo do que o do implícito desses conjuntos de avaliação. Isso é coerente com a distinção feita na Seção 2.1.2, já que a ironia e o estereótipo deixam algum sinal no texto, e o código não deixa porque a sua superfície é a de uma palavra comum.

Mendelsohn et al. (2023) estudaram o mesmo fenômeno em inglês, e o capítulo se compara a eles em três frentes. Eles mediram a queda na pontuação de toxicidade de um classificador comercial em frases hostis montadas para o teste, com o termo que nomeia o grupo trocado pelo código, e aqui a perda aparece em mensagens que a própria comunidade escreveu, com recall de 1,9% e 3,8% nos dois classificadores dedicados que não colapsaram. Eles também pediram a um modelo de linguagem que produzisse termos codificados, e ele devolveu 153 dos 340 do glossário. Essa medida é sobre o que o modelo consegue listar, e não sobre a detecção em mensagens, mas o mesmo conhecimento parcial reaparece aqui, já que os modelos sem catálogo ficam entre 9,4% e 43%. E registraram como limite do próprio trabalho o fato de terem descrito o repertório de fora, apontando a validação por membros dos grupos estudados como desejável, que é o caminho seguido aqui, com o catálogo levantado por quem acompanha a comunidade e verificado no corpus completo, conforme a Seção 3.5.

Kruk et al. (2024) entregaram a modelos de linguagem o termo codificado junto com o significado documentado no glossário e pediram que apontassem, entre exemplos de uso, quais estavam no sentido codificado. Concluíram que os modelos são pouco confiáveis para reconhecer o termo por conta própria e ficam bons nessa separação quando recebem o significado. A condição c2 usa o mesmo recurso em português, com duas diferenças. O significado entregue não vem de um glossário montado a partir de fontes externas, e sim do catálogo levantado na própria comunidade, e a tarefa aqui é mais exigente, porque o

modelo percorre as 3.500 mensagens em vez de julgar ocorrências já localizadas de um termo conhecido. Ainda assim, os três modelos testados chegam a valores entre 96% e 98% no estrato codificado.

Por fim, o experimento com modelos de linguagem responde à pergunta que a Seção 2.3.5 deixou aberta. Kumarage, Bhattacharjee e Garland (2024) verificaram que enriquecer o pedido com contexto genérico não melhora a classificação, e a condição c1 deste trabalho vai na mesma direção, com os exemplos genéricos chegando a prejudicar o desempenho. O salto da condição c2 mostra onde essa generalização termina, pois quando a falha é de vocabulário cultural, e não de capacidade, o acréscimo do dicionário local muda o patamar de desempenho.

Em síntese, a questão de pesquisa recebe uma resposta em três partes. Os modelos de detecção automática de estado da arte falham de forma sistemática na toxicidade codificada desta comunidade, e a falha não se explica pelo idioma, pelo domínio, pelo registro das redes sociais ou pela capacidade do modelo, já que atravessa as arquiteturas e as escalas testadas. Essa falha tem conserto no eixo da detecção, e o conserto é o conhecimento cultural documentado, o catálogo, e não modelos de maior escala. E permanece sem conserto no eixo do custo operacional, já que nenhuma configuração testada modera esta comunidade com precisão utilizável, porque a competência que falta, a de reconhecer quando uma mensagem de aparência agressiva não é tóxica no uso da comunidade, não está documentada em nenhum instrumento que este trabalho examinou.

5 Conclusão

Este trabalho perguntou se os modelos de detecção automática de toxicidade de estado da arte enxergam a toxicidade veiculada pela linguagem codificada de uma comunidade e, quando não enxergam, se fornecer o código documentado corrige a falha. Para responder, construiu dois artefatos sobre um corpus de 1,74 milhão de mensagens de chat ao vivo de uma comunidade de jogos do YouTube brasileiro: o catálogo de termos codificados e um conjunto de referência de 3.500 mensagens, com cada mensagem tóxica classificada como explícita, codificada ou mista, e os usou para separar os erros de três famílias de detectores entre as mensagens explícitas e as codificadas. Este capítulo sintetiza os achados, revisita as contribuições, assume as limitações e aponta os desdobramentos.

5.1 Síntese dos Achados

A questão de pesquisa recebeu resposta em três partes, apresentadas na Seção 4.4. A primeira é que os modelos de detecção automática de estado da arte falham de forma sistemática na toxicidade codificada da comunidade. Sem acesso ao código, detectores de naturezas diferentes ficam entre 1,9% e 43% de recall no estrato codificado, nos modelos que não colapsaram, contra 48% a 90% no estrato explícito, o que desloca a explicação do detector para o próprio código, cujo reconhecimento depende do repertório da comunidade. A segunda é que, no eixo da detecção, a falha tem conserto, e o conserto é a documentação, pois com o catálogo fornecido na instrução os três modelos de linguagem sobem para valores entre 96% e 98% de recall codificado, sem que a arquitetura e a escala façam diferença, enquanto o enriquecimento convencional, feito de exemplos genéricos, chega a prejudicar o desempenho. A terceira é que, no eixo do custo operacional, nenhuma configuração testada corrige a precisão, que fica entre 6,9% e 24,6%, porque a competência que falta, a de reconhecer quando uma mensagem de aparência agressiva não é tóxica no uso da comunidade, não está documentada em nenhum instrumento que este trabalho examinou.

As duas faces são a mesma ausência, a do contexto cultural local. O código torna tóxico o que parece inocente, o que derruba o recall de quem não o conhece, e o registro comunitário torna inofensivo muito do que parece tóxico, o que derruba a precisão de quem julga pela superfície. O catálogo documenta a primeira competência e por isso conserta um eixo só.

5.2 Contribuições

As quatro contribuições anunciadas na Seção 1.3 saem do trabalho qualificadas pelos resultados. O catálogo de termos codificados entrou como instrumento de medição e saiu também validado como intervenção, já que é o único fator testado que recupera a detecção do código. O protocolo de construção e validação, que usa a imersão como fonte de hipóteses, a mineração de vizinhança lexical no corpus completo como verificação e a convergência entre as duas como critério, não depende da comunidade estudada e pode ser repetido onde houver repertório codificado. O conjunto de referência rotulado por mecanismo viabilizou a métrica que sustenta o argumento, porque sem a separação entre o estrato explícito e o codificado a falha desapareceria dentro de médias que pareceriam aceitáveis. E a evidência empírica delimita o que os modelos maiores não resolvem, já que a falha não é de capacidade e sim de vocabulário cultural, e por isso se corrige com documentação e não com escala.

5.3 Limitações

O desenho é um estudo de caso único (YIN, 2018), com sete canais de uma comunidade e dois meses de coleta. Os números valem para esta comunidade e para este período. A coleta não tem uma janela de comparação em outra época do ano, e um acontecimento externo dentro do período também pode alterar o que se fala no chat, efeitos que não foram medidos. O contraste entre os estratos depende de os termos serem usados na comunidade, e não da taxa com que aparecem numa janela. O que se pode repetir em outra comunidade é o achado geral, o de que a linguagem codificada escapa da detecção e a documentação a recupera.

O conjunto de referência foi produzido por um único rotulador, o próprio pesquisador, seguindo o guia de rotulagem e as regras da Seção 3.6.

A rotulagem adota por desenho a regra de classificar como não-tóxica a mensagem em dúvida, conforme a Seção 3.6, o que facilita a leitura de que a precisão baixa dos detectores, medida na Seção 4.3.4, acontece porque muita mensagem de aparência agressiva não é tóxica no uso da comunidade. As duas explicações não foram separadas aqui, porque um falso positivo tanto pode ser uma mensagem que a comunidade de fato lê como inofensiva quanto uma mensagem duvidosa que a regra mandou classificar como não-tóxica. Separar as duas exigiria uma inspeção qualitativa dos falsos positivos, que este trabalho não fez.

As mensagens de toxicidade codificada são apenas 53, mais 2 mistas, e sobre uma base tão pequena cada mensagem a mais ou a menos desloca bastante os percentuais. As regras de rotulagem foram validadas no corpus completo, conforme a Seção 3.6.2, mas os

valores de recall por estrato continuam apoiados nessas 53.

O catálogo registra os termos que o protocolo alcançou, dentro dos limites da Seção 3.5. Um termo de que ninguém suspeitou não entra no catálogo e também não é reconhecido na rotulagem, e a mensagem que o carrega é contada como não-tóxica. A toxicidade codificada medida aqui é, portanto, um piso, e não o total.

O repertório se transforma, já que a comunidade autocensura e desloca termos, prática documentada em outras plataformas (STEEN; YURECHKO; KLUG, 2023), de modo que o catálogo tem prazo de validade e exige manutenção. A intervenção validada aqui recupera o código documentado, e não o que venha a substituí-lo.

O experimento usou um único modelo comercial de fronteira, de modo que a conclusão de que um modelo maior não conserta a falha se apoia na convergência entre as arquiteturas e escalas testadas, e não numa varredura da oferta comercial.

5.4 Trabalhos Futuros

A anotação da amostra por mais de um rotulador é o passo seguinte de consolidação do conjunto de referência. Antes dela é preciso definir o que o reconhecimento da toxicidade codificada exige de quem anota, já que ele depende do repertório da comunidade.

A aplicação ao corpus completo consiste em executar a melhor configuração sobre as 1,74 milhão de mensagens, o que permitiria estimar o quanto o código aparece fora da amostra rotulada.

O uso de modelos comerciais de maior capacidade verificaria se o padrão observado aqui, o salto de detecção com o catálogo e a precisão que não sobe, se mantém no topo da oferta.

A expansão do catálogo pede métodos de descoberta que não dependam da vivência do pesquisador, capazes de apontar candidatos a partir do próprio corpus e de acompanhar a mudança do repertório ao longo do tempo.

Fica em aberto se o termo codificado inicia o trecho tóxico da conversa ou se entra num trecho que já estava em curso. A mineração de vizinhança não responde a isso, porque registra que as palavras ocorrem na mesma janela e não a ordem em que aparecem. Cada mensagem do corpus tem o instante em que foi enviada, o que permite examinar essa ordem e separar as duas leituras.

A aplicação do protocolo a outras comunidades com repertório codificado testaria se o mecanismo descrito aqui se repete fora deste caso.

O resultado central do trabalho permanece como a sua melhor síntese. Diante de uma comunidade que decide não ser vista, a moderação automática eficaz não passa por

mais modelo, passa por documentar cultura.

Referências

ANDERSON, L. Analytic autoethnography. *Journal of Contemporary Ethnography*, v. 35, n. 4, 2006. Citado na página 32.

BOUMA, G. Normalized (pointwise) mutual information in collocation extraction. In: *Proceedings of the Biennial GSCL Conference*. [S.l.: s.n.], 2009. p. 31–40. Citado na página 34.

BRODY, S.; DIAKOPOULOS, N. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2011. Citado na página 40.

BROWN, T. B. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. [S.l.: s.n.], 2020. Citado 2 vezes nas páginas 21 e 53.

CHARMAZ, K. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. [S.l.]: SAGE Publications, 2006. Citado na página 33.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, v. 16, n. 1, p. 22–29, 1990. Citado na página 33.

CONNEAU, A. et al. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. [S.l.: s.n.], 2020. Citado na página 21.

DAVIDSON, T. et al. Automated hate speech detection and the problem of offensive language. In: *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*. [S.l.: s.n.], 2017. Citado 3 vezes nas páginas 19, 20 e 21.

DENZIN, N. K. *The Research Act: A Theoretical Introduction to Sociological Methods*. [S.l.]: McGraw-Hill, 1978. Citado na página 34.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. [S.l.: s.n.], 2019. Citado na página 21.

ELLIS, C.; ADAMS, T. E.; BOCHNER, A. P. Autoethnography: An overview. *Historical Social Research / Historische Sozialforschung*, v. 36, n. 4, 2011. Citado na página 32.

ELSHERIEF, M. et al. Latent hatred: A benchmark for understanding implicit hate speech. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2021. Citado 8 vezes nas páginas 15, 16, 19, 20, 22, 25, 27 e 56.

FORD, C. et al. Chat speed op pogchamp: Practices of coherence in massive twitch chat. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. [S.l.]: ACM, 2017. p. 858–871. Citado na página 39.

FORTUNA, P.; NUNES, S. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, v. 51, n. 4, 2018. Citado 3 vezes nas páginas 15, 19 e 21.

GEETANJALI; KUMAR, M. Exploring hate speech detection: Challenges, resources, current research and future directions. *Multimedia Tools and Applications*, 2025. Citado na página 22.

GLASER, B. G.; STRAUSS, A. L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. [S.l.]: Aldine, 1967. Citado 2 vezes nas páginas 33 e 36.

Google Developers. *YouTube Data API v3*. s.d. <https://developers.google.com/youtube/v3>. Acesso em: 17 jul. 2026. Citado na página 29.

HANU, L.; Unitary team. *Detoxify*. 2020. <https://github.com/unitaryai/detoxify>. Citado 2 vezes nas páginas 22 e 42.

HUANG, F.; KWAK, H.; AN, J. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In: *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*. [S.l.: s.n.], 2023. Citado 3 vezes nas páginas 24, 25 e 27.

JAKOBSON, R. Closing statement: Linguistics and poetics. In: SEBEOK, T. A. (Ed.). *Style in Language*. Cambridge, MA: MIT Press, 1960. p. 350–377. Citado na página 39.

Jigsaw; Google. *Perspective API*. 2017. <https://www.perspectiveapi.com/>. Acesso em: 17 jul. 2026. Citado na página 19.

KIRK, H. R. et al. Handling and presenting harmful text in NLP research. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. [S.l.: s.n.], 2022. p. 497–510. Citado na página 45.

KIRK, H. R. et al. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2022. Citado 4 vezes nas páginas 15, 22, 25 e 27.

KOZINETS, R. V. The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, v. 39, n. 1, 2002. Citado na página 32.

KOZINETS, R. V. *Netnography: The Essential Guide to Qualitative Social Media Research*. 3. ed. [S.l.]: SAGE Publications, 2020. Citado na página 32.

KRUK, J. et al. Silent signals, loud impact: LLMs for word-sense disambiguation of coded dog whistles. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2024. Citado 4 vezes nas páginas 23, 25, 27 e 56.

KUMARAGE, T.; BHATTACHARJEE, A.; GARLAND, J. *Harnessing Artificial Intelligence to Combat Online Hate: Exploring the Challenges and Opportunities of Large Language Models in Hate Speech Detection*. 2024. ArXiv preprint arXiv:2403.08035. Citado 4 vezes nas páginas 24, 25, 27 e 57.

LEITE, J. A. et al. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL-IJCNLP)*. [S.l.: s.n.], 2020. p. 914–924. Citado 5 vezes nas páginas 16, 23, 25, 27 e 42.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. Citado na página 43.

LOBO, J. V. C.; BARBOSA, D. M. Investigando a dinâmica da propagação do ódio em cascatas de toxicidade nos chats ao vivo da Twitch. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia 2025)*. [S.l.]: SBC, 2025. Citado 4 vezes nas páginas 16, 24, 25 e 27.

MARTINS, V. et al. A misoginia no YouTube brasileiro: Um estudo de caso sobre o conteúdo produzido pela comunidade Red Pill. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia 2025)*. [S.l.]: SBC, 2025. Citado 3 vezes nas páginas 24, 25 e 27.

MENDELSOHN, J. et al. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: [s.n.], 2023. p. 15162–15180. Citado 6 vezes nas páginas 15, 20, 23, 25, 27 e 56.

OCAMPO, N. B. et al. An in-depth analysis of implicit and subtle hate speech messages. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. [S.l.: s.n.], 2023. p. 1997–2013. Citado 8 vezes nas páginas 15, 16, 19, 20, 22, 25, 27 e 56.

OLIVEIRA, A. S. et al. How good is ChatGPT for detecting hate speech in Portuguese? In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. [S.l.]: SBC, 2023. Citado 3 vezes nas páginas 23, 25 e 27.

PATTON, M. Q. Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, v. 34, n. 5 Pt 2, 1999. Citado na página 34.

POOKPANICH, P.; SIRIBORVORNRATANAKUL, T. Offensive language and hate speech detection using deep learning in football news live streaming chat on YouTube in Thailand. *Social Network Analysis and Mining*, v. 14, n. 18, 2024. Citado 4 vezes nas páginas 15, 24, 25 e 27.

Qwen Team. *Qwen3.5 Technical Report*. 2026. TODO conferir. Citado na página 43.

RAMOS, G. et al. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, v. 14, n. 204, 2024. Citado 4 vezes nas páginas 15, 21, 22 e 25.

REICHENBACH, H. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. [S.l.]: University of Chicago Press, 1938. Citado na página 32.

RINGER, C.; NICOLAOU, M. A.; WALKER, J. A. TwitchChat: A dataset for exploring livestream chat. In: *Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-20)*. [S.l.: s.n.], 2020. Citado na página 33.

RÖTTGER, P. et al. Two contrasting data annotation paradigms for subjective NLP tasks. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. [S.l.: s.n.], 2022. p. 175–190. Citado na página 20.

SAHAMI, M. et al. A Bayesian approach to filtering junk e-mail. In: *AAAI Workshop on Learning for Text Categorization*. [S.l.: s.n.], 1998. Citado na página 40.

SCHMID, U. K. Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society*, v. 27, n. 3, p. 1588–1606, 2025. Citado 4 vezes nas páginas 15, 22, 25 e 27.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In: *Intelligent Systems (BRACIS 2020)*. [S.l.]: Springer, 2020. (Lecture Notes in Computer Science). Citado 3 vezes nas páginas 21, 22 e 42.

STEEN, E.; YURECHKO, K.; KLUG, D. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media + Society*, v. 9, n. 3, p. 1–17, 2023. Citado na página 60.

VARGAS, F. et al. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*. [S.l.: s.n.], 2022. Citado 5 vezes nas páginas 16, 23, 25, 27 e 42.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. [S.l.: s.n.], 2017. Citado na página 21.

YIN, R. K. *Case Study Research and Applications: Design and Methods*. 6. ed. [S.l.]: SAGE Publications, 2018. Citado 2 vezes nas páginas 15 e 59.

YouTube. *Violent extremist or criminal organizations policy*. 2026. YouTube Help. Acesso em: 25 jul. 2026. Disponível em: <https://support.google.com/youtube/answer/9229472>. Citado na página 37.

YouTube. *YouTube's Community Guidelines*. 2026. YouTube Help. Acesso em: 25 jul. 2026. Disponível em: <https://support.google.com/youtube/answer/9288567>. Citado na página 37.

ZHOU, Y. et al. *The Hidden Language of Harm: Examining the Role of Emojis in Harmful Online Communication and Content Moderation*. 2025. ArXiv preprint arXiv:2506.00583. Citado 4 vezes nas páginas 15, 22, 25 e 27.

Apêndices

APÊNDICE A – Materiais Elaborados pelo Autor

A.1 Catálogo de Termos Codificados em Versão Íntegra

Esta seção apresenta a versão íntegra das sete entradas confirmadas do catálogo de termos codificados, cuja versão sintética está na Tabela 9. Cada entrada registra o significado que o termo tem no uso da comunidade, o método pelo qual foi descoberto, as variantes e os disfarces observados, um exemplo real de uso no corpus e as observações que sustentaram a rotulagem. As entradas seguem a mesma ordem da tabela, por alcance no corpus.

Os exemplos são reproduzidos apenas quando indispensáveis à validade do catálogo, pela razão dada na Seção 3.8, a de que documentar um repertório de superfície inocente é condição para que ele possa ser detectado. Os nomes e as siglas das duas facções criminosas permanecem mascarados pelos marcadores `FA` e `FB`, como no restante do documento, e os nomes de pessoas que apareciam no texto das mensagens citadas foram substituídos por marcadores entre colchetes.

A.1.1 CP

Significado Referência a pornografia infantil, do inglês *child porn*.

Política correspondente *Child Safety*.

Origem Experiência do pesquisador na comunidade, como uma das três sementes textuais que iniciaram o catálogo.

Variantes e disfarces A forma em minúsculas `cp`. Como disfarces, “Café Preto” e “Cod Points”, que dão ao termo uma leitura inocente e sustentam a negabilidade plausível. Nenhum dos dois teve uso informacional confirmado no corpus.

Exemplo de uso *“o que esse CP?”* e *“Child p0rn”*.

Observações de rotulagem A entrada recebe a regra da sigla isolada, pela qual uma mensagem cujo conteúdo é apenas `CP` ou `cp` é classificada como tóxica, sem exceção. Na amostra rotulada, as 10 ocorrências isoladas eram tóxicas. No corpus completo há 7.087 ocorrências isoladas, vindas de 1.859 autores distintos, e apenas 3 delas, ou 0,04%, apareceram em transmissões de um jogo cuja moeda virtual usa a mesma

sigla, único contexto em que caberia a leitura inocente. A inspeção qualitativa não confirmou nenhuma como uso comum. O argumento linguístico completa o empírico. Uma sigla sozinha num chat ao vivo serve para marcar presença e pertencimento, e não para transmitir informação, e o uso comum da sigla vem acompanhado de complemento, como em "comprei 1000 CP no jogo". A sigla nua não tem esse complemento. Os números e o método da verificação estão na Seção 3.6.2.

A.1.2 Estudar

Significado Eufemismo para "estuprar".

Política correspondente *Harassment and Cyberbullying*.

Origem Experiência do pesquisador na comunidade, como uma das três sementes textuais que iniciaram o catálogo.

Variantes e disfarces O termo se conjuga como qualquer verbo e aparece como `estudo`, `estudar` e `estudei`, além de construções como "pode estudar" e "estudo coletivo". Diferente das siglas do catálogo, não tem forma fixa.

Exemplo de uso *"estudei uma sub14"* e *"[apelido] foi estudado pelo meno"*.

Observações de rotulagem O termo foi confirmado na rotulagem manual e aparece em chats de canais diferentes daqueles em que foi observado primeiro, o que indica que não é gíria restrita a um só grupo. Por funcionar como verbo comum, não recebe a regra da sigla isolada. A separação entre o uso codificado e o literal, como em "estudar para a prova", depende de ler o sujeito, o objeto e o tempo verbal da frase.

A.1.3 FA (sigla mascarada)

Significado Sigla de uma facção criminosa, cujo nome é mascarado. No uso da comunidade funciona como saudação de filiação.

Política correspondente *Violent Extremist or Criminal Organizations*.

Origem Experiência do pesquisador na comunidade, como uma das três sementes textuais que iniciaram o catálogo.

Variantes e disfarces A sigla sozinha e a sua forma interrogativa, a sigla escrita por extenso conforme a pronúncia, as grafias com pontuação separando as letras e uma saudação de duas palavras associada à facção, também mascarada. A gíria "fechadão" entra como variante quando vem acompanhada de complemento que nomeie a facção, no sentido de estar aliado a ela. Há ainda as codificações visuais, com o emoji de

cartão vermelho, o de bandeira vermelha e o de apito vermelho repetidos em padrão de spam, e a combinação do emoji de dedo indicador com o de cartão vermelho.

Exemplo de uso Repetição da sigla em texto, repetição e combinação dos emojis vermelhos listados acima e repetição do emoji que codifica visualmente a saudação mascarada.

Observações de rotulagem A entrada recebe a regra da sigla isolada, com o mesmo embasamento de `CP`. Na amostra rotulada, as 5 ocorrências isoladas eram tóxicas. No corpus completo há 1.741 ocorrências isoladas, vindas de 544 autores distintos, das quais 47 tinham na vizinhança temporal alguma menção a emprego ou a recrutador, que é o contexto em que caberia a leitura cotidiana da sigla. Nenhuma delas, depois da inspeção, era de fato desse tipo. Para os sinais visuais vale o limiar de repetição, pelo qual o emoji conta como elemento dominante quando aparece quatro vezes ou mais, ou quando não há texto competindo com ele. A literatura revisada não oferece limiar pronto para esse caso, de modo que ele foi calibrado sobre a amostra e submetido ao diagnóstico de sensibilidade da Seção 3.6.2. A gíria "fechadão" sozinha, sem o complemento que nomeie a facção, fica fora da regra da sigla isolada, porque o termo tem uso cômico corrente.

A.1.4 Hitler / Hail

Significado Menção à ideologia nazista e à supremacia branca, tanto por referência textual direta, com "hail" na forma abreviada da saudação nazista e a menção literal a "Hitler", quanto por codificações visuais do mesmo referente.

Política correspondente *Hate Speech.*

Origem Os marcadores visuais vieram da experiência do pesquisador na comunidade, e os textuais foram recuperados depois pela mineração de vizinhança a partir das sementes visuais, de modo que a validação percorreu os dois sentidos.

Variantes e disfarces Um emoji de pinguim específico, que dispensa qualquer acompanhamento porque o gesto de saudação está embutido na própria imagem. O emoji da bandeira alemã quando aparece junto de um marcador ariano ou de alguma variante do emoji de mão levantada, que representa o gesto. E um emoji de figura humana com balão de fala, que funciona como outra forma visual da mesma saudação.

Exemplo de uso O emoji de pinguim isolado ou repetido, a bandeira alemã junto de um dos marcadores acima, *"HAIL [streamer]"* seguido do emoji de pinguim e menções diretas como *"HITLER"* ou *"METEU UM HITLER"*.

Observações de rotulagem Vale o mesmo limiar de repetição da entrada `FA`, de quatro repetições ou ausência de texto competindo, de modo que 2 ou 3 ocorrências isoladas do emoji de pinguim não bastam. A bandeira alemã sozinha pode ser inócua, em contexto de futebol, viagem ou idioma, e o sinal tóxico só aparece na combinação com um dos marcadores. A mineração de vizinhança, que roda sem acesso ao catálogo, recuperou "hail" e "hitler" como as palavras de maior associação estatística com o emoji de pinguim, o que confirma por um método independente que os três apontam para o mesmo referente. Foi por isso que a entrada passou a ser organizada pelo referente, o conteúdo nazista, e não pelo sinal que a descobriu primeiro.

A.1.5 Trade

Significado Eufemismo para a comercialização ou a troca de pornografia infantil.

Política correspondente *Child Safety*.

Origem Mineração de vizinhança lexical a partir da semente `CP`.

Variantes e disfarces A palavra `trade` em qualquer combinação de maiúsculas e minúsculas, e a estrutura "you trade my X for your X", em forma de pergunta ou de condição.

Exemplo de uso *"Trade"*, *"TRADE"* e *"you trade my wife for your wife?"*.

Observações de rotulagem O termo cobre um lado diferente do mesmo referente de `CP`, já que não aponta para o consumo e sim para a transação. É palavra inocente em contexto comercial comum, como a troca de itens colecionáveis, e a carga tóxica vem do uso em conjunto com `CP` e `Bio`, confirmada por uma mensagem do corpus em que os três aparecem juntos, citada na entrada `Bio`.

A.1.6 Bio (link in bio)

Significado Eufemismo para o ponteiro a um canal externo de distribuição do conteúdo, o link na biografia de um perfil de rede social.

Política correspondente *Child Safety*.

Origem Mineração de vizinhança lexical a partir da semente `CP`, com a desambiguação feita depois pela análise de expressões de duas ou mais palavras.

Variantes e disfarces A forma composta `link in bio` e `link na bio`, em inglês e em português, além de `link em bio` e `legenda link na bio`. A palavra `bio` sozinha aparece como forma reduzida.

Exemplo de uso *"LINK IN BIO"*, *"link na bio"* e *"como eu durmo depois de fazer 5K em vendas de CP com o método da legenda link na bio"*.

Observações de rotulagem A forma que carrega o sentido codificado é a composta. Das 303 ocorrências genuínas da palavra no corpus, 93% fazem parte da expressão de duas ou três palavras, e apenas 22 são a palavra sozinha, base pequena demais para uma regra de termo isolado. A entrada é complementar a `Trade`, que nomeia a transação, enquanto `Bio` indica o caminho de entrega.

A.1.7 FB (sigla mascarada)

Significado Sigla da facção criminosa rival da Facção A, cujo nome também é mascarado. Cumpre o mesmo papel de sinalizar filiação, para outra organização.

Política correspondente *Violent Extremist or Criminal Organizations*.

Origem Mineração de vizinhança lexical a partir da semente `FA`, entre as palavras de associação mais forte.

Variantes e disfarces A sigla sozinha e a sua forma interrogativa. Como disfarce, a leitura da mesma sigla como a de um partido político estrangeiro.

Exemplo de uso *"FB?"*, *"fb"* e a sigla seguida do nome do partido político estrangeiro que serve de disfarce.

Observações de rotulagem A entrada tem a mesma estrutura de `FA`, em que a sigla isolada ou interrogativa comunica filiação e não informação, e por isso recebe a regra da sigla isolada pelo mesmo embasamento. O fato de o mecanismo se repetir numa segunda sigla indica que ele não é peculiaridade de uma única organização.

A.2 Vizinhos de Maior Associação por Termo-Semente

A mineração de vizinhança lexical da Seção 3.5.2 devolve, para cada termo-semente, uma lista ordenada das palavras que mais aparecem perto dele no corpus completo. As Tabelas 16, 17 e 18 trazem as vinte primeiras posições das listas das três sementes textuais que iniciaram o catálogo, as mesmas da Tabela 4, das sessenta posições examinadas em cada uma. A lista é ordenada por um índice que combina o PMI e a frequência na vizinhança, e não pelo PMI sozinho. São essas as listas em que foram encontrados `Trade`, `Bio` e `FB`, que entraram no catálogo como resultado da mineração, conforme a Tabela 5.

Os nomes de canais, de streamers, de usuários do chat e de terceiros citados nas mensagens aparecem substituídos por marcadores entre colchetes, conforme a Seção 3.8. Os

emojis próprios da plataforma são exibidos como imagem e os emojis comuns são descritos em palavras, como na Tabela 4.

Tabela 16 – Vizinhos de maior associação da semente `CP`, vinte primeiras posições

Pos.	Palavra	PMI	Ocorrências	Autores
1		0,49	12.161	301
2	[denunciante público de pedofilia]	0,78	2.991	444
3	`koe`	0,52	5.386	1.789
4	`god`	0,81	2.125	531
5	`tekken`	0,86	1.807	411
6	[canal]	0,69	2.681	420
7	emoji de bandeira vermelha	0,45	4.933	165
8	`uga`	0,91	1.067	599
9	`buga`	0,95	904	537
10	[streamer]	0,34	7.068	2.377
11	`trade`	1,28	460	179
12	`koeee`	0,69	1.524	614
13	`mad`	1,19	493	13
14	`koeeee`	0,70	1.222	563
15	`criança`	0,61	1.580	976
16		0,32	5.771	342
17		0,81	839	52
18	`freediddy`	1,08	461	13
19	`padre`	1,07	458	238
20	`points`	1,33	268	193

Fonte: dados da pesquisa. “Ocorrências” é o número de vezes que a palavra aparece nas janelas ao redor da semente, e “Autores” é o número de autores distintos que a usaram nessas janelas.

Tabela 17 – Vizinhos de maior associação da semente `Estudar`, vinte primeiras posições

Pos.	**Palavra**	**PMI**	**Ocorrências**	**Autores**
1	`tiktok`	1,59	603	39
2	`plataforma`	1,54	622	236
3	`lei`	1,08	843	616
4	[streamer]	0,26	7.656	3.155
5	`escola`	1,03	362	266
6	`estupro`	1,25	243	172
7	[usuário do chat]	0,88	267	128
8	`rage`	0,69	404	237
9	`estrupo`	1,40	93	72
10	`kick`	0,35	1.362	197
11	[apelido de streamer]	0,49	656	435
12	`aplicou`	1,32	86	78
13	`uh`	1,23	99	20
14	`venha`	0,67	332	17
15	`quiz`	1,56	58	39
16	`around`	0,85	186	9
17	[apelido de streamer]	0,65	309	61
18	`bait`	0,60	348	219
19	`ocultando`	1,33	70	5
20	`mexe`	1,15	90	51

Fonte: dados da pesquisa. As duas posições marcadas como apelido de streamer correspondem a formas distintas do mesmo apelido.

Tabela 18 – Vizinhos de maior associação da semente `FA`, vinte primeiras posições

Pos.	Palavra	PMI	Ocorrências	Autores
1	🧨	3,16	21.396	391
2	emoji de bandeira vermelha	3,16	8.931	234
3	🫳	3,02	1.318	46
4	🫰	2,76	1.425	52
5	emoji de sinal de paz	1,73	2.571	111
6	▶️	2,82	939	44
7	emoji de cem pontos	3,05	544	11
8	❤️‍🩹	2,24	472	24
9	🏠	2,77	285	23
10	`mad`	2,26	286	5
11	`FB`	2,92	154	86
12	`antares`	2,92	150	58
13	emoji da bandeira da Alemanha	0,76	2.108	150
14	[canal]	1,65	363	21
15	`borat`	2,69	136	19
16	[usuário do chat]	1,73	304	20
17	`fechadão`	3,04	96	29
18	`tabacudo`	3,15	76	33
19	[denunciante público de pedofilia]	0,91	908	117
20	🛑	1,96	182	22

Fonte: dados da pesquisa.

As listas trazem, ao lado dos vizinhos que sustentam cada entrada do catálogo, muita palavra sem relação com ela, como nomes de jogos e saudações do chat. É o resultado esperado de um procedimento que recebe apenas o termo escrito e ordena o que encontra ao redor dele. A leitura que levou às entradas do catálogo, resumida na Tabela 4, foi feita sobre listas nesta forma.

A.3 Sensibilidade da Janela Temporal na Mineração de Vizinhança

A mineração de vizinhança lexical descrita na Seção 3.5.2 recolhe as mensagens publicadas até 60 segundos antes e depois de cada ocorrência da semente. Esta seção traz os números da verificação que repetiu o cálculo com janelas de 15, 30, 120 e 180 segundos, mantendo tudo o mais igual. A Tabela 19 mostra quantas mensagens cada janela recolhe e quanto a sua lista das vinte palavras mais bem colocadas coincide com a da janela de 60 segundos, que serve de referência e por isso aparece com 100%.

Tabela 19 – Tamanho da vizinhança e coincidência do top-20 com a janela de referência

Semente	**Janela**	**Mensagens na vizinhança**	**Coincidência do top-20**
`CP`	15 s	280.326	65,0%
`CP`	30 s	427.076	75,0%
`CP`	60 s	642.620	100,0%
`CP`	120 s	925.013	70,0%
`CP`	180 s	1.107.645	65,0%
`Estudar`	15 s	163.173	55,0%
`Estudar`	30 s	271.057	65,0%
`Estudar`	60 s	446.480	100,0%
`Estudar`	120 s	712.216	65,0%
`Estudar`	180 s	902.417	55,0%
`FA`	15 s	51.832	65,0%
`FA`	30 s	80.024	80,0%
`FA`	60 s	128.667	100,0%
`FA`	120 s	208.347	85,0%
`FA`	180 s	273.568	70,0%

Fonte: dados da pesquisa.

A Tabela 20 acompanha, janela a janela, os vizinhos que a Tabela 4 apresenta como evidência de cada entrada. A marca “fora” indica que a palavra deixou de figurar entre as sessenta mais bem colocadas naquela janela.

Tabela 20 – Valor de PMI dos vizinhos citados na Tabela 4, por duração da janela

Semente	**Vizinho**	**15 s**	**30 s**	**60 s**	**120 s**	**180 s**
`CP`	`criança`	1,04	0,86	0,61	0,40	0,28
`CP`	`trade`	2,02	1,68	1,28	0,81	0,61
`CP`	`bio`	1,81	1,50	1,12	fora	fora
`Estudar`	`estupro`	2,02	1,68	1,25	0,84	0,62
`Estudar`	`estrupo`	2,14	1,83	1,40	0,94	fora
`Estudar`	`lei`	1,68	1,43	1,08	0,74	0,58
`Estudar`	`escola`	1,52	1,34	1,03	0,67	0,50
`FA`	`FB`	3,42	3,23	2,92	2,49	2,18

Fonte: dados da pesquisa.

As duas tabelas mostram o mesmo comportamento. A janela larga recolhe mais mensagens e dilui a associação, e a janela estreita a concentra, sem que nenhum vizinho passe a apontar para outro referente. Nas duas janelas mais largas, `bio` sai da lista de `CP`

e `estrupo` sai da lista de `Estudar`, o que é efeito da mesma diluição, já que ambos são os vizinhos de menor frequência entre os citados.

A.4 Instrução Fornecida aos Modelos de Linguagem

Esta seção reproduz a instrução enviada aos modelos de linguagem no experimento da Seção 3.7.2, nas três condições da Tabela 8. O pedido é montado por blocos. A definição geral de toxicidade e o pedido de formato entram nas três condições, os exemplos resolvidos entram em c1 e c2 e o catálogo entra apenas em c2. Os blocos seguem juntos num único pedido, e a mensagem a classificar vem depois deles, conforme a Seção 3.7.2.

O texto aparece como foi enviado, inclusive na ausência de acentos das mensagens de exemplo e das justificativas que as acompanham, com as linhas longas quebradas para caber na largura da página. O bloco do catálogo reproduz as entradas da planilha de termos, cuja versão íntegra está na Seção A.1, e por isso aqui figura apenas pela linha que o introduz. No pedido enviado aos modelos as siglas das facções aparecem como a comunidade as escreve, sem o mascaramento que este documento adota.

A resposta pedida tem três campos, e apenas o primeiro deles, o rótulo binário, entra nas métricas do Capítulo 4. A separação por mecanismo usada ali vem da classificação feita pelo pesquisador, descrita na Seção 3.6.3, e não do campo que o modelo devolve.

A definição geral de toxicidade abre o pedido nas três condições.

```
Você é um moderador de conteúdo. Classifique a MENSAGEM de um chat
ao vivo do YouTube como tóxica ou não-tóxica.
É TÓXICO: linguagem que ataca, humilha, ameaça ou expressa
ódio/assédio contra pessoas ou grupos; conteúdo sexual envolvendo
menores; apologia a crime ou violência; discurso de ódio (racismo,
nazismo etc.).
Mensagens podem conter emojis no formato :shortcode:
(ex.: :card-red-penalty:).
```

Na condição c2 entra em seguida o bloco do catálogo, apresentado pela linha que o introduz.

```
Use também o CATÁLOGO DE TERMOS CODIFICADOS desta comunidade -
termos de aparência inocente mas tóxicos no contexto local:

[entradas do catálogo, conforme a Seção A.1]
```

Os exemplos resolvidos entram em c1 e em c2, alternando uma mensagem tóxica e uma não-tóxica. Os cinco exemplos tóxicos são de toxicidade explícita, e nenhum deles usa termo do catálogo. Todos vieram do corpus completo, fora das 3.500 mensagens avaliadas.

```
EXEMPLOS:

Mensagem: fala de futebol seu lixo, unica coisa q sabe
Resposta: {"toxico": true, "mecanismo": "explicito",
"justificativa": "ataque pessoal dirigido (vocativo 'seu lixo')"}

Mensagem: warzone paizao
Resposta: {"toxico": false, "mecanismo": "nenhum",
"justificativa": "giria de jogo, sem ataque"}

Mensagem: para de estourar o microfone seu macaco
Resposta: {"toxico": true, "mecanismo": "explicito",
"justificativa": "racismo dirigido (slur 'macaco' em vocativo de
ataque)"}

Mensagem: AMONG PRO CHAT JOGAR JUNTO
Resposta: {"toxico": false, "mecanismo": "nenhum",
"justificativa": "convite pra jogar"}

Mensagem: mod chato pra krl, fds seu arrombado
Resposta: {"toxico": true, "mecanismo": "explicito",
"justificativa": "insulto dirigido a moderador"}

Mensagem: chama pra jogar
Resposta: {"toxico": false, "mecanismo": "nenhum",
"justificativa": "pedido casual de interacao"}

Mensagem: ngm gosta de vc nao , ngm liga smt lixo
Resposta: {"toxico": true, "mecanismo": "explicito",
"justificativa": "ataque de exclusao/humilhacao dirigido"}

Mensagem: jogam tudo errado kkkkkk
Resposta: {"toxico": false, "mecanismo": "nenhum",
"justificativa": "critica leve ao gameplay, sem alvo pessoal"}

Mensagem: ele falou que removeu, tu e imbecil krl
Resposta: {"toxico": true, "mecanismo": "explicito",
"justificativa": "insulto dirigido ('tu e imbecil')"}

Mensagem: Fui banido atoa
Resposta: {"toxico": false, "mecanismo": "nenhum",
"justificativa": "reclamacao trivial sobre ban, sem ataque"}
```

O pedido de formato fecha a instrução nas três condições.

```
Responda SOMENTE com um objeto JSON válido com as chaves:
  "toxico": true ou false
  "mecanismo": um de [lexico, sigla, visual, composto, explicito,
  nenhum] - use "nenhum" se não-tóxico; "explicito" se a toxicidade
  for aberta/literal
  "justificativa": uma frase curta explicando a decisão
```

A mensagem a ser classificada vem depois de todos esses blocos, no mesmo formato dos exemplos.

```
Mensagem: [texto da mensagem a classificar]
Resposta:
```