“Análise de toxicidade na comunidade gamer do YouTube com o uso de modelos de linguagem”

Aluno: Israel Matias do Amaral
Orientadora: Helen

UFOP

icea

DECSI
DEPARTAMENTO DE COMPUTAÇÃO E SISTEMAS

Índice

1. Introdução

1.1 Definição de toxicidade

- "Comentários tóxicos são mensagens rudes, desrespeitosas ou irracionais que podem fazer com que uma pessoa abandone uma discussão” **(JIGSAW, 2017 apud FORTUNA; NUNES, 2018)**
- É o termo mais adequado para uma abordagem abrangente e pode servir como categoria superior para "linguagem ofensiva" e **"discurso de ódio"**.
- **Fortuna e Nunes (2018)** destacam que o discurso de ódio pode ocorrer com "diferentes estilos linguísticos, mesmo em formas sutis ou quando o **humor** é usado",

1.2 O debate sobre os limites do humor

Condenação de Léo Lins reacende debate sobre limites do humor e da liberdade de expressão; veja o que dizem juristas

Justiça condenou humorista a 8 anos e 3 meses de prisão, além do pagamento de multa e indenização por danos morais. Defesa alega que não houve intenção de ofender ninguém.

Por Redação GloboNews e g1 SP

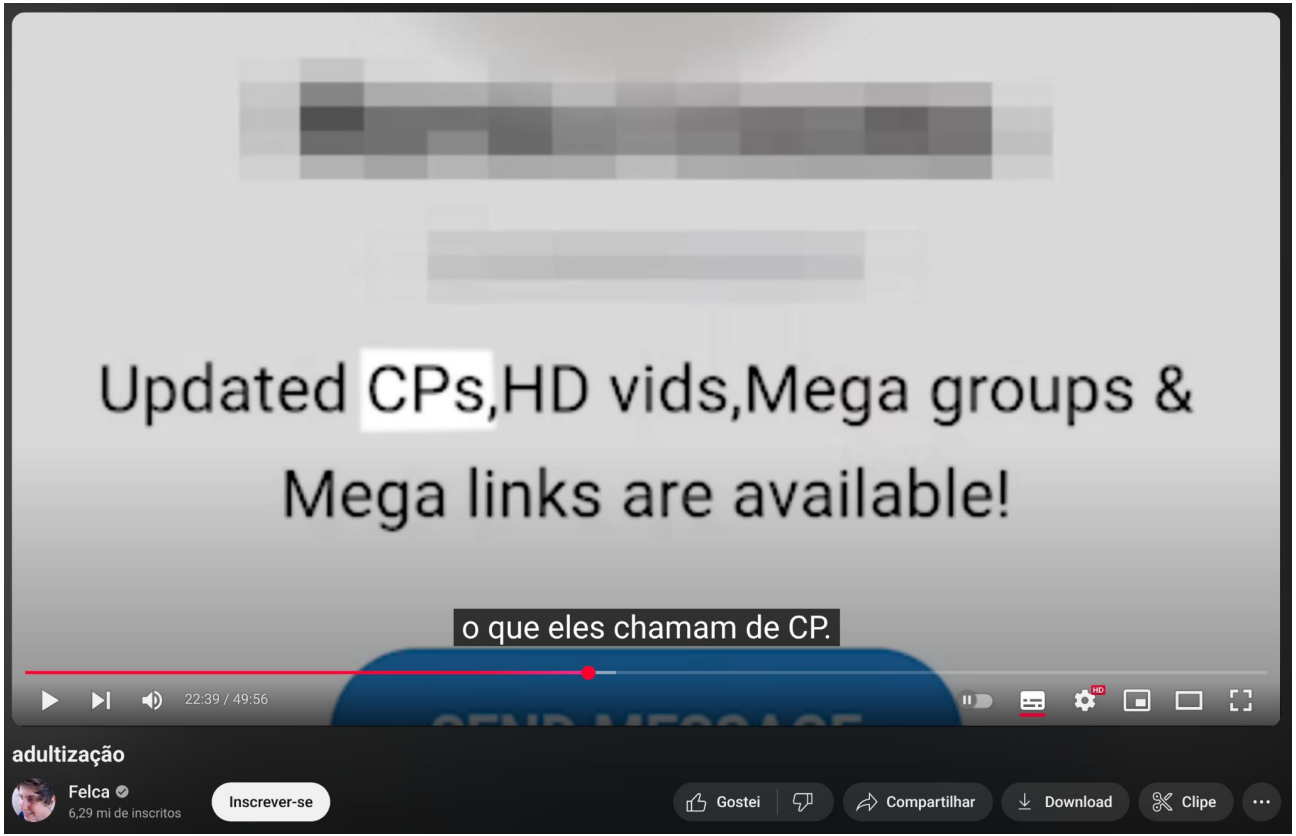
04/06/2025 15h29 · Atualizado há uma semana

1.3 Cenário gamer no YouTube

- Com o crescimento da cultura gamer e das transmissões ao vivo no YouTube, cresce também a presença de comentários tóxicos.
- Muitos desses comentários são expressos por meio de ironias e **expressões codificadas**, o que dificulta sua detecção automática. **(OLIVEIRA et al., 2023)**

1.4 Expressões Codificadas

1.5 Expressões codificadas na mídia



1.6 O foco deste trabalho

- Este trabalho foca em transmissões ao vivo realizadas por uma **sub comunidade gamer do YouTube**, composta por streamers como *luangameplay* e *renanplay*, conhecidos por adotar um estilo de humor mais ácido, politicamente incorreto e usar **linguagem codificada.**
- Essa comunidade apresenta alto engajamento — alguns canais chegam a mais de **1,4 milhões** de inscritos e registram **mais de mil** espectadores por transmissão.

1.7 Justificativa da escolha

- A escolha dessa sub comunidade se justifica por combinar 2 elementos críticos para o estudo:
  1. **volume elevado de interações** (média ~9 mil[1] mensagens por live)
  2. **alta incidência de linguagem ambígua ou ofensiva**, o que o torna um ambiente ideal para avaliar a eficácia de **modelos** de detecção de discurso tóxico.

[1]Média = 8.826,30, Mediana = 3.332,50.

2. Trabalhos Relacionados

2.1 O desempenho de modelos BERT

Pookpanich e Siriborvornratanakul (2024) - "Offensive language and hate speech detection using deep learning in football news live streaming chat on YouTube in Thailand"

- **Objetivo:** Detecção de linguagem ofensiva e discurso de ódio para o tailandês em chats de transmissão ao vivo do YouTube, utilizando modelos de linguagem baseados em Transformer. O estudo visou avaliar o desempenho de vários modelos BERT.
- **Principais Resultados:**
  - O XLM-RoBERTa apresentou o melhor desempenho em termos de recall e F1-score (média de recall de 0.9669 e F1-score de 0.9530).
  - No entanto, não houve diferenças estatisticamente significativas no desempenho entre os cinco modelos avaliados.

2.2 O efeito do humor no discurso tóxico

Schmid (2025) - "Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and proces."

- **Objetivo:** Investigar como o público percebe e processa a combinação de humor (ex: ironia) e discurso de ódio (ex: desumanização) em memes, e de que forma o discurso de ódio humorístico contribui para a normalização de ideologias hostis nas redes sociais.
- **Principais Resultados:**
    - O discurso de ódio humorístico implícito é mais tolerado pelo público do que o discurso de ódio explicitamente expresso.
    - Participantes mais jovens, familiarizados com a cultura dos memes, tendiam a achar o discurso de ódio humorístico mais divertido, mas também a rejeitar o discurso de ódio, embora essa rejeição fosse frequentemente atrasada e precedida por um prazer inicial.

3. Oportunidades

3.1 Lacunas

1. Desafios na Detecção de Linguagem Ambígua e Humorística
2. Escassez de Pesquisa em Idiomas de Baixo Recurso
3. Dificuldade em Capturar Nuances Culturais e Individuais

3.2 Oportunidades

1. Estudo Aprofundado de Linguagem Ambígua e Humor no Contexto Gamer
2. Contribuição Direta para o Português do Brasil
3. Investigação de Nuances Locais e Culturais

3.4 Questões de pesquisa

1. Como classificadores de toxicidade percebem e classificam a linguagem ambígua, codificada e humorística, e quais são os desafios específicos impostos por essas nuances linguísticas e culturais?
2. Em que medida o pré-processamento textual influencia a eficácia dos modelos de linguagem na detecção de toxicidade?
3. Quais vieses algorítmicos ou desafios éticos (ex: falsos positivos para humor, falsos negativos para hate speech sutil) são evidentes, e quais são as implicações para a moderação de conteúdo e a salvaguarda da liberdade de expressão?

3.5 Objetivo Geral

- Aplicar modelos de linguagem para detectar e analisar o discurso tóxico em português, especificamente em comentários extraídos de chats ao vivo de uma sub comunidade gamer brasileira do YouTube, considerando suas nuances linguísticas e culturais.

3.6 Objetivos Específicos

1. **Avaliar como classificadores de toxicidade percebem e classificam a linguagem** ambígua, codificada, irônica e humorística presente nos chats ao vivo da comunidade gamer estudada.
2. **Analisar o impacto do pré-processamento textual** na eficácia dos modelos de linguagem na detecção de toxicidade, considerando as particularidades da linguagem da comunidade gamer.
3. **Expor os possíveis riscos e vieses algorítmicos** dos modelos de linguagem na identificação de discurso tóxico/ofensivo mascarado como humor, discutindo as implicações éticas para a moderação de conteúdo, os limites do humor e a liberdade de expressão.
4. **Relacionar os padrões de toxicidade** detectados pelos modelos com o nível de engajamento e o perfil dos canais da subcomunidade gamer estudada.

4. Metodologia

4.1 Ordem Cronológica da Metodologia

- **Geetanjali e Kumar (2025)**, em sua revisão, detalha uma metodologia para detecção de discurso de ódio em etapas como: coleta de dados, rotulagem, pré-processamento e aplicação de diferentes técnicas.
- Similarmente, **Ramos et al. (2024)**, em sua revisão abrangente, descrevem a extração de dados, os métodos de classificação e as métricas de avaliação de desempenho.

4.1. Ordem Cronológica da Metodologia

Ambos os estudos demonstram que, uma abordagem estruturada, começando com a base de dados e progredindo para a modelagem e avaliação, é uma prática bem estabelecida na área.

1. Coleta de Dados
2. Rotulagem de Dados
3. Pré-processamento de Dados
4. Treinamento de Modelo
5. Classificação do Discurso Tóxico
6. Análise Quantitativa e Qualitativa Final

4.2. Antes da coleta de dados: Quais canais compõem essa sub comunidade?

- O YouTube não possui uma estrutura formal de **comunidade** baseada em tópicos/assuntos (como o Reddit).
- Dessa forma, por ora, chamaremos esses agrupamentos informais no YouTube de **"nichos"** ou **"sub comunidades"**.

4.3 Como foi feita essa caracterização

- Foi identificado um vídeo viral de 2024 que mapeia e caracteriza os principais streamers dessa sub comunidade gamer polêmica do YouTube. Esse vídeo teve cerca de 200.000 visualizações e foi reagido várias vezes alcançando um total aproximado de 5.000.000 de visualizações.
- Essa popularidade do vídeo é um indicativo da relevância e do reconhecimento da sub comunidade dentro do ecossistema do YouTube.
- A partir da transcrição desse vídeo, foi extraído o nome de todos os streamers explicitamente citados como membros dessa sub comunidade.
- Todo esse processo foi feito de forma sistematizada e reprodutível. Disponível em: Link para o Github do projeto

4.4 Resultado final da caracterização

Canal	Inscritos
luangameplay	1.440.000
renanplay	156.000
canaldoronaldinho	98.700
diegosheipado	39.100
fabiojunior	24.200
canaldocelinho	3.030
wallacegamer*	0

*Canal banido

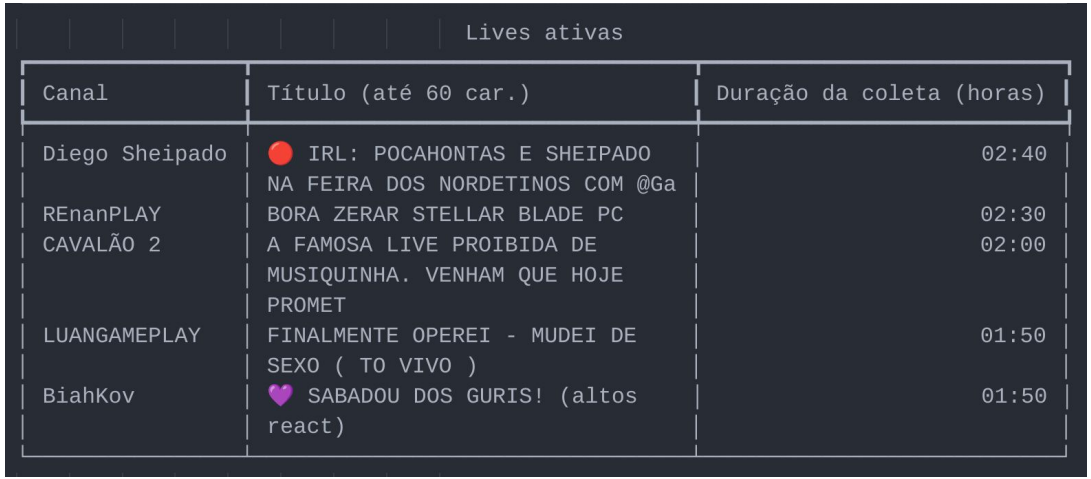
4.5 A Descoberta do "Dicionário da Comunidade"

- Em uma etapa subsequente de exploração, foram identificados e consultados **3 vídeos no formato de "dicionário da comunidade"**. Esses recursos, criados pelos próprios membros, forneceram **definições explícitas e contextuais** para diversas dessas expressões codificadas.
- Essa descoberta foi particularmente relevante, pois **o conhecimento subcultural é essencial para interpretar "corretamente" memes e o gênero do humor online**, que frequentemente requerem um entendimento específico para decifrar a mensagem real **(Schmid, 2025)**.

4.6 Metodologia de coleta dos dados

Para coletar o chat das transmissões ao vivo usando a API oficial do YouTube foi necessário criar um Monitor de Lives.

- Aplicação Python que detecta novas lives automaticamente;
- Inicia a coleta do chat ao vivo e metadados da transmissão.



Lives ativas

Canal	Título (até 60 car.)	Duração da coleta (horas)
Diego Sheipado	🔴 IRL: POCAHONTAS E SHEIPADO NA FEIRA DOS NORDETINOS COM @Ga	02:40
REnanPLAY	BORA ZERAR STELLAR BLADE PC	02:30
CAVALÃO 2	A FAMOSA LIVE PROIBIDA DE MUSIQUINHA. VENHAM QUE HOJE PROMET	02:00
LUANGAMEPLAY	FINALMENTE OPEREI - MUDEI DE SEXO (TO VIVO)	01:50
BiahKov	💜 SABADOU DOS GURIS! (altos react)	01:50

4.7 Resultado de coleta dos dados

- Período: **14/06/2025** 21:38 a **14/08/2025** 20:59
- Total de mensagens: **1.740.704**
- Total de lives: **231**
- Canais analisados: **7**

4.8 Quantidade de lives e mensagens por canal

Canal	Live Count	Total Mensagens
Canal do Celinho	15	61.294
Canal do Ronaldinho	33	7.550
Diego Sheipado	38	98.164
Fábio Streamer	55	39.784
LUANGAMEPLAY	35	532.186
REnanPLAY	48	984.899
Wallace Gamer	7	16.827
Total	**231**	**1.740.704**

5. Experimentos Iniciais

5.1 Rotulagem

- **Fonte:** Amostra de 3.000 comentários de chats ao vivo de uma comunidade gamer brasileira, bem ativa, conhecida pelo seu contexto único.
- **Rotulagem:** Realizada manualmente, seguindo um guia formal baseado nas diretrizes do YouTube para garantir consistência e qualidade.
- **Distribuição:**
  - Não Tóxico: 2.783 comentários (92.8%)
  - Tóxico: 217 comentários (7.2%)
- **Conclusão:** O dataset é altamente desbalanceado, refletindo um cenário real e tornando o F1-Score a métrica de avaliação mais adequada.

5.2 Experimento Inicial: Objetivos e Hipóteses

- **Objetivo Principal:** Comparar quantitativamente (via F1-Score[1]) o desempenho de uma API genérica (Perspective) com um modelo de linguagem especializado (BERT com fine-tuning) na tarefa.
- **Hipótese Principal:** O modelo BERT, após ser especializado (fine-tuned) nos dados da comunidade, apresentará um desempenho significativamente superior à API genérica.

[1](1) acertar os comentários tóxicos (2) sem acusar quem não é tóxico.

5.3 Experimento Inicial: Projeto Fatorial

Metodologia: Foi empregado um Projeto Fatorial de dois fatores, cada um com dois níveis.

Modelo Classificador	Tipo de Pré-processamento
Perspective API (genérico)	Texto Bruto (sem tratamento)
Modelo BERT (especializado)	Texto Padrão (minúsculas, sem quebras de linha)

Variável Resposta: F1-Score Médio (Binário, Foco na Classe "Tóxico"), obtido a partir de 30 replicações via Bootstrap para garantir a robustez estatística dos resultados.

5.4 Experimento Inicial: Resultados

Pré-processamento	Perspective API	Modelo BERT (Fine-tuned)
Bruto	**0.3273**	**0.6632**
Padrão	**0.3467**	**0.6857**

- **Efeito do Modelo:** Massivo. O BERT (F1 ≈ 0.67) superou a Perspective API (F1 ≈ 0.34) em 97.78%, validando a importância do fine-tuning.
- **Efeito do Pré-processamento:** Marginal, mas consistente. A melhora de 0.6632 para 0.6857 (+0.0225) no BERT é muito similar à observada na Perspective API.

5.5 Experimento Inicial: Conclusões

- **Hipótese Confirmada:** A especialização de um modelo de linguagem (fine-tuning) é uma abordagem drasticamente mais eficaz (**melhora de 97.78% no F1-Score**) para moderação em nichos com linguagem própria.
- **Impacto do Pré-processamento:** O efeito foi marginal, sugerindo que para modelos Transformer modernos, a qualidade e a especificidade dos dados de treino são muito mais importantes que limpezas simples de texto.
- **Implicação Prática:** A moderação de conteúdo eficaz em comunidades específicas demanda soluções de IA customizadas e adaptadas ao contexto cultural local.

6. Próximos Passos

6. Próximos Passos

- Melhorar o guia de rotulagem (Codebook), visando deixá-lo mais robusto.
- Pesquisar e explorar técnicas de rotulagem para mais de um anotador.
- Pesquisar e explorar técnicas para melhorar a performance do classificador de toxicidade (early stopping, enriquecimento da classe minoritária, dentre outros).
- Aplicar o classificador ao dataset completo de 1.740.704 comentários.
- Análise Quantitativa e Qualitativa Final.

7. Referências

7. Referências

FORTUNA, P.; NUNES, S. A Survey on Automatic Detection of Hate Speech in Text. **ACM Computing Surveys**, [S. l.], v. 51, n. 4, art. 85, p. 1-30, jul. 2018. Disponível em: https://doi.org/10.1145/3232676. Acesso em: 25 ago. 2025.

GEETANJALI; KUMAR, M. Exploring hate speech detection: challenges, resources, current research and future directions. **Multimedia Tools and Applications**, [S. l.], 2025. Disponível em: https://doi.org/10.1007/s11042-025-20716-2. Acesso em: 25 ago. 2025.

OLIVEIRA, A. S. et al. How Good Is ChatGPT For Detecting Hate Speech In Portuguese? *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 14., 2023, Porto Alegre. **Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS: SBC, 2023. p. 94–103.

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**, [S. l.], v. 14, n. 18, 2024. Disponível em: https://doi.org/10.1007/s13278-023-01183-9. Acesso em: 25 ago. 2025.

RAMOS, G. et al. A comprehensive review on automatic hate speech detection in the age of the transformer. **Social Network Analysis and Mining**, [S. l.], v. 14, n. 204, 2024. Disponível em: https://doi.org/10.1007/s13278-024-01361-3. Acesso em: 25 ago. 2025.

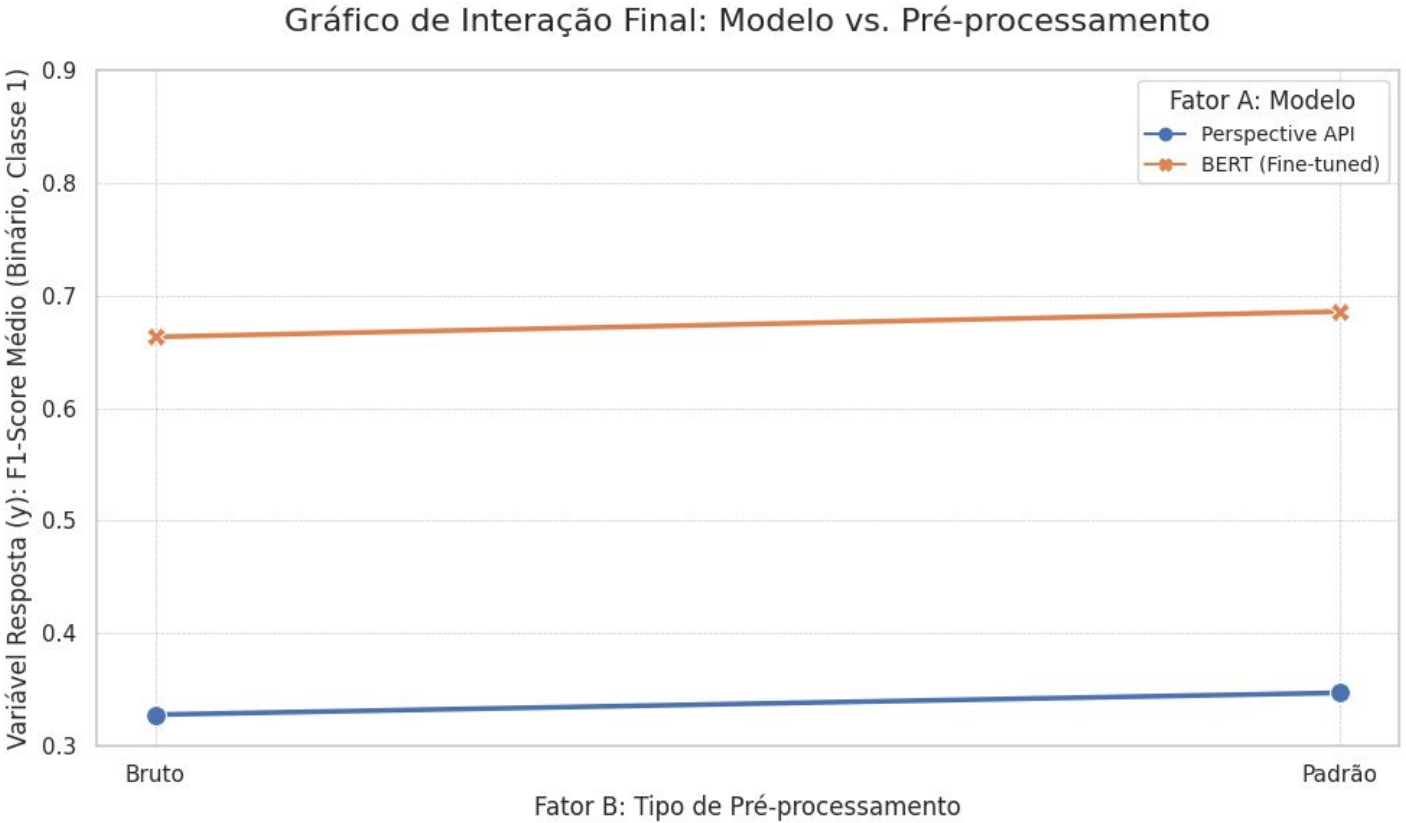
SCHMID, U. K. Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. **New Media & Society**, [S. l.], v. 27, n. 3, p. 1588–1606, 2025. Disponível em: https://doi.org/10.1177/14614448231198169. Acesso em: 25 ago. 2025.

Obrigado pela atenção!

“Análise de toxicidade na comunidade gamer do YouTube com o uso de modelos de linguagem”

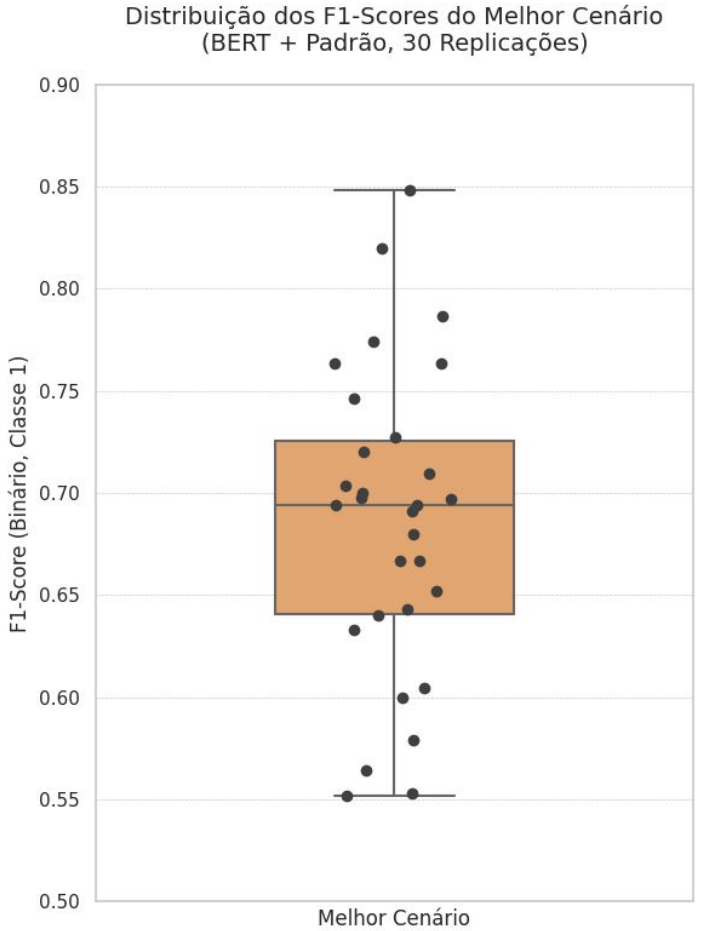
8. Apêndice (Gráficos)

8.1 Gráfico de Análise de Interação



8.2 Análise do melhor cenário

- **Melhor cenário:** Modelo BERT com fine-tuning e pré-processamento padrão, atingindo um **F1-Score médio de 0.6857** na tarefa de identificar comentários tóxicos.



fim.