

Apresentação 3: Relatório Final

Aluno: Israel Matias do Amaral

Tipo de projeto: TCC 1

Orientadora: Helen

TÍTULO PROVISÓRIO

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

Índice

1. Motivação (1–2 slides)
2. Problema, Objetivos e Hipóteses (1 slide)
3. Projeto Experimental (1 slide)
4. Dados (1 slide)
5. Resultados (2-3 slides)
6. Conclusões (1 slide)

1. Motivação

1. Motivação

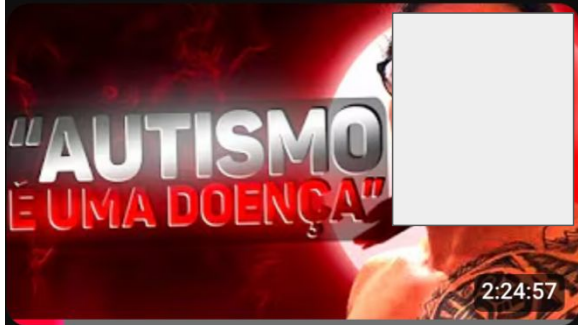
Condenação de Léo Lins reacende debate sobre limites do humor e da liberdade de expressão; veja o que dizem juristas

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. Motivação




PARA PARA MAIS UMA LIVIZINHA MASSA BIXO

1,3 mil visualizações • Transmitido há 3 semanas



LIVIZINHA RESENHA

1,8 mil visualizações • Transmitido há 3 semanas



LIVE COM UM MACACO

2,6 mil visualizações • Transmitido há 3 semanas

1. Motivação

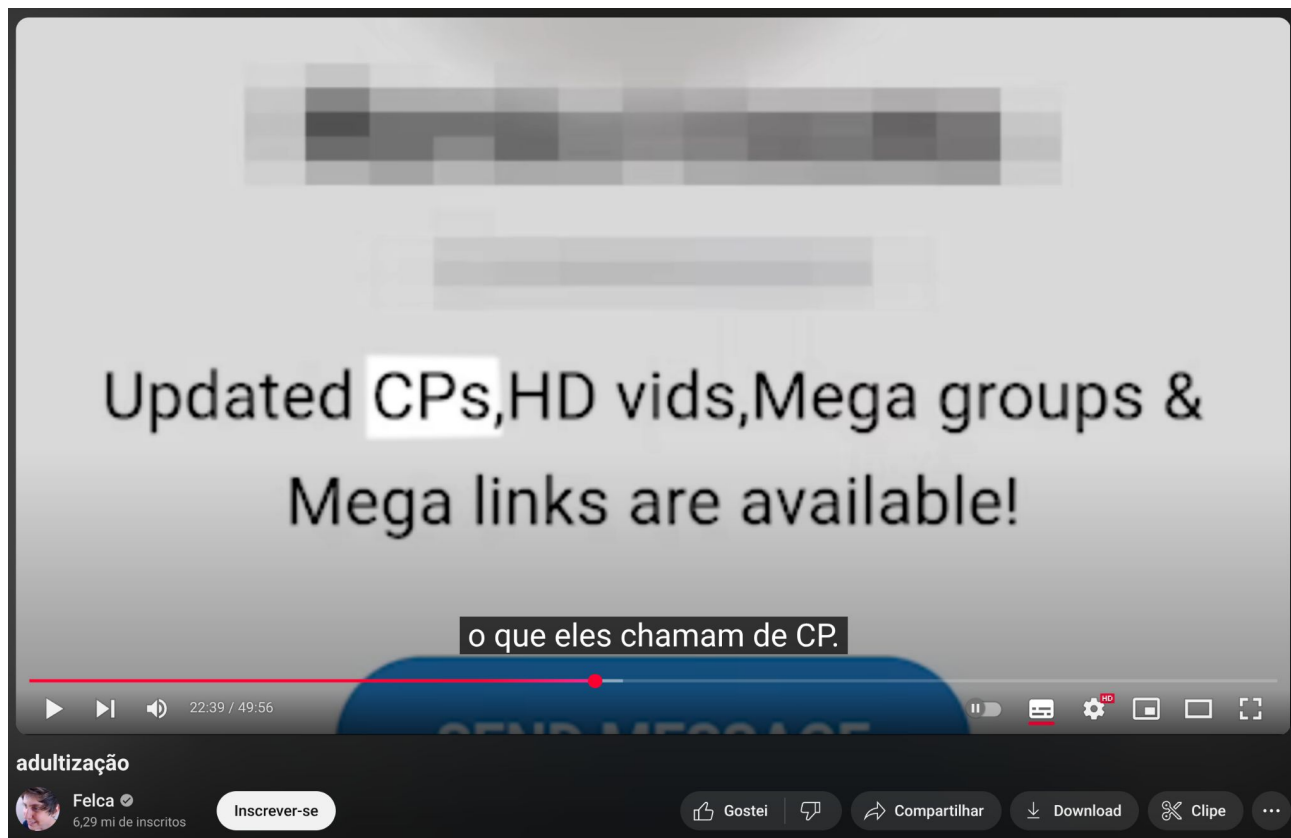
O DESAFIO DA MODERAÇÃO EM COMUNIDADES GAMER

- É evidente o problema do discurso de ódio e da toxicidade em ambientes online, especialmente em comunidades de jogos.
- A linguagem nessas comunidades é única: rápida, cheia de gírias, memes, ironia e "humor ácido".
- Ferramentas de moderação tradicionais falham em capturar o contexto, permitindo que toxicidade velada passe despercebida.

1. Motivação



1. Motivação



1. Motivação

PERGUNTA CENTRAL DA PESQUISA

- É possível adaptar uma Inteligência Artificial de ponta para que ela "aprenda" a linguagem específica de uma comunidade gamer brasileira e se torne mais eficaz que uma ferramenta genérica?
- Este trabalho investiga essa questão, comparando uma solução genérica com uma especializada.

2. Problema, Objetivos e Hipóteses

2. Problema, Objetivos e Hipóteses

- **Problema:** Ferramentas de moderação de conteúdo genéricas, como a Perspective API, podem ter baixa performance na detecção de toxicidade em nichos com linguagem altamente codificada e irônica.
- **Objetivo Principal:** Comparar quantitativamente (via F1-Score¹) 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.

¹Achar os comentários tóxicos: (1) sem errar muito e (2) sem acusar quem não é tóxico.

3. Projeto Experimental

3. Projeto Experimental

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.

4. Dados

4. Dados

- **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.

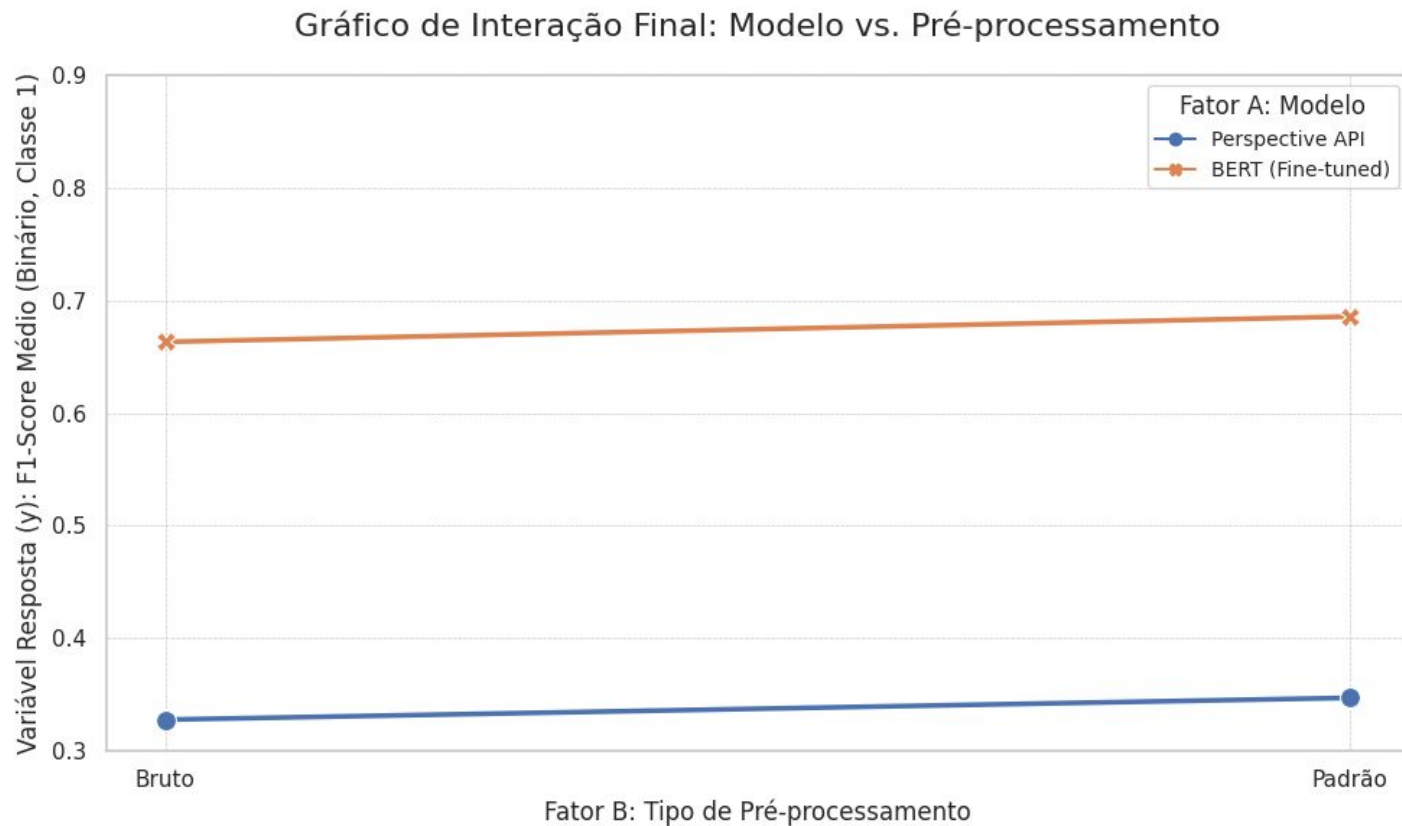
5. Resultados

5.1 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 ($F1 \approx 0.67$) superou a Perspective API ($F1 \approx 0.34$) 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.

5.2 Gráfico de Análise de Interação



5.3 Cálculo dos Efeitos (Principal e de Interação)

Efeito Principal do Modelo: 0.3374

- Insight: Em média, trocar da Perspective API para o modelo BERT causa um aumento de 0.3374 no F1-Score. Este é, de longe, o efeito mais forte.

Efeito Principal do Pré-processamento: 0.0212

- Insight: Em média, aplicar o pré-processamento "Padrão" causa um aumento de 0.0212 no F1-Score. É um efeito positivo, mas muito menor.

Efeito de Interação (Modelo: Pré-processamento): -0.0031

- Insight: Um efeito de interação tão próximo de zero indica que os fatores são independentes. O pequeno benefício do pré-processamento não muda significativamente de um modelo para o outro.

5.4 Testes de Significância e Influência

Abaixo estão os resultados da Análise de Variância - ANOVA, gerados a partir das 120 réplicas do experimento (30 para cada uma das 4 condições).

Usamos um nível de significância de 0.05. Se o p-value é menor que 0.05, o efeito é estatisticamente significativo.

Fator	Soma dos Quadrados	F-value	p-value (PR(>F))
Modelo	1.3664	553.16	< 0.000001
Pré-processamento	54	2.18	1.425
Interação	1	0.04	8.359

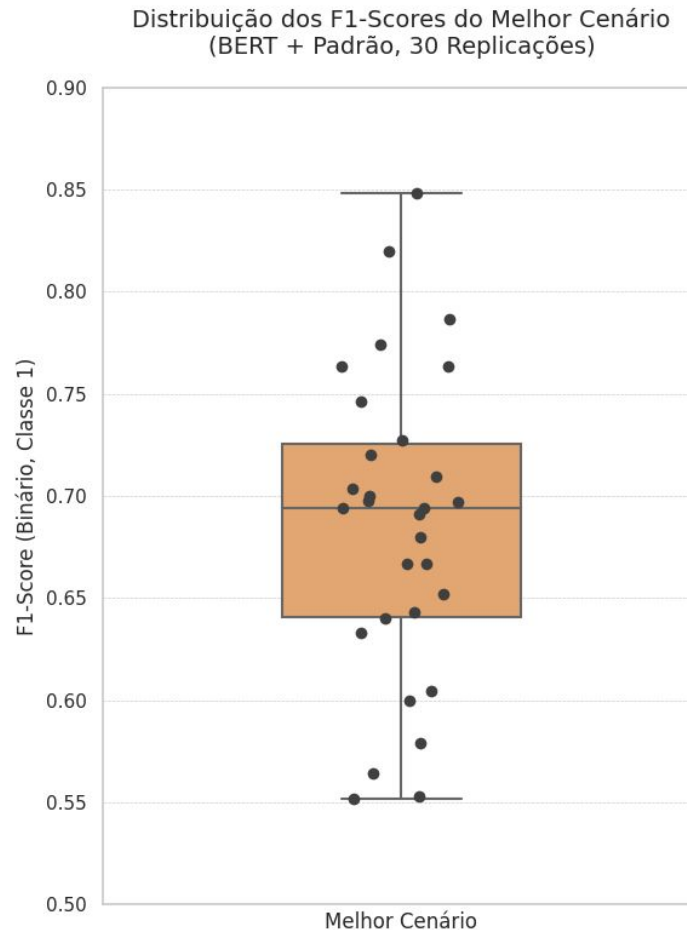
5.5 Influência (Alocação da Variação)

Calculando a proporção da Soma dos Quadrados (SQ), podemos ver a "influência" de cada fator na variação total dos resultados:

- O fator **Modelo** foi responsável por 97.2% da variação nos resultados.
- O fator **Pré-processamento** foi responsável por apenas 0.4% da variação.
- A **Interação** foi responsável por menos de 0.01% da variação.

5.6 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.



6. Conclusões

6. 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.

EXTRA: Limitações e Trabalhos Futuros

Limitações e Trabalhos Futuros

- Para trabalhos futuros, a metodologia pode ser significativamente robustecida com a implementação do **early stopping**. Isso envolveria a separação do dataset em três conjuntos distintos (treino, validação e teste). O conjunto de validação seria monitorado a cada época para interromper o treinamento assim que a performance parar de melhorar, garantindo que o modelo salvo seja o mais generalista possível antes de ser finalmente avaliado no conjunto de teste.
- Recomenda-se também, para futuras investigações, **aumentar o número de replicações do bootstrap (para 50, 100 ou mais)**. Embora isso represente um custo computacional significativamente maior, resultaria em uma estimativa da média e do desvio padrão do F1-Score ainda mais precisa, reduzindo a margem de erro e aumentando a certeza sobre a magnitude dos efeitos observados.

fim.