

# **Apresentação 2: Análise Exploratória dos Dados**

**Aluno:** Israel Matias do Amaral

**Tipo de projeto:** TCC 1

**Orientadora:** Helen

TÍTULO PROVISÓRIO

*“Análise de discurso de ódio na comunidade gamer do YouTube com o uso de modelos de linguagem”*

OU

*“Análise de chats de transmissões ao vivo no Youtube em uma sub comunidade gamer de humor negro”*

# Índice

1. Objetivos, Motivação e Atualizações do Projeto
2. Hipótese e Desenho do Experimento
3. Mudanças na Estratégia de Análise
4. Preparação e Caracterização dos Dados
5. Análise Exploratória dos Dados

# 1. Objetivos, Motivação e Atualizações do Projeto

## 1.1 Contextualização e 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.

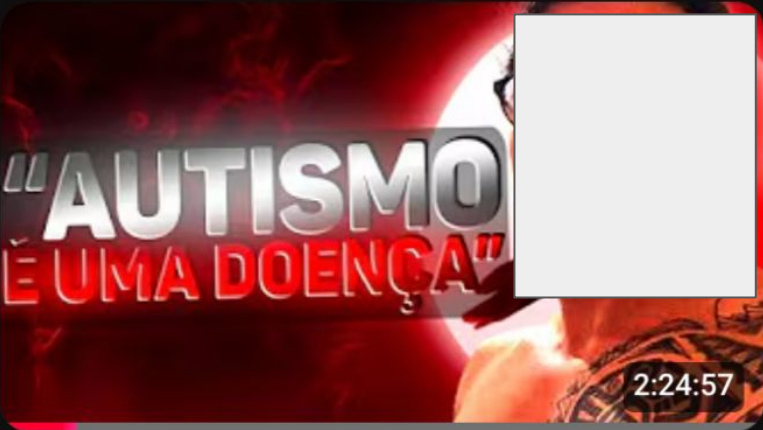


Por **Redação GloboNews e g1 SP**

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

# 1.1 Contextualização e Motivação

- A popularização das transmissões ao vivo no YouTube, especialmente na comunidade gamer, trouxe consigo um aumento de comentários ofensivos.
- O problema central é que muitos desses comentários são expressos com ironia e códigos, dificultando sua detecção automática.
- Este projeto foca em uma subcomunidade específica de streamers do Youtube conhecida pelo humor ácido, que combina alto volume de interações com uma alta incidência de linguagem ambígua, tornando-se um ambiente ideal para o estudo.

## 1.1 Contextualização e Motivação

 <p>2:24:57</p>	 <p>3:29:49</p>	 <p>3:32:46</p>
PARA PARA MAIS UMA LIVIZINHA MASSA BIXO	LIVIZINHA RESENHA	LIVE COM UM MACAC0
1,3 mil visualizações • Transmitido há 3 semanas	1,8 mil visualizações • Transmitido há 3 semanas	2,6 mil visualizações • Transmitido há 3 semanas

## 1.2 Objetivos do Estudo (Versão Atualizada)

1. Entender como um classificador de toxicidade do estado da arte classifica os comentários desta sub comunidade.
2. Entender como um classificador lida com linguagem ambígua e codificada.
3. Expor possíveis riscos do discurso tóxico/ofensivo mascarado como humor.
4. Contribuir para o entendimento dos limites do humor, moderação de conteúdo e liberdade de expressão.

## 2. Hipótese e Desenho do Experimento



## 2.1 Hipótese do Projeto

**Hipótese Central:** O resultado final depende não apenas do modelo de linguagem escolhido, mas também da estratégia de pré-processamento e do volume de dados utilizado.

Para testar esta ideia, as hipóteses foram formalizadas da seguinte forma:

- **Hipótese Nula ( $H_0$ ):** Os fatores do experimento (modelo, pré-processamento, tamanho da base) e suas interações não influenciam significativamente o desempenho da classificação.
- **Hipótese Alternativa ( $H_1$ ):** Pelo menos um dos fatores ou uma de suas interações influencia significativamente o desempenho da classificação.

## 2.2 Articulação dos Elementos do Projeto Experimental

- A hipótese será investigada por meio de um Projeto Fatorial  $2^3$  com 3 replicações.
- **Fatores e Níveis:** Foram escolhidos 3 fatores com 2 níveis cada.

Fator	Níveis
modelo_classificador	BERTimbau vs. LLM (Llama/Deepseek)
tipo_preprocessamento	Mínimo (Padrão) vs. Direcionado (Limpeza de Ruído)
tamanho_base_dados	Parcial (Amostra) vs. Completa

## 2.2 Articulação dos Elementos do Projeto Experimental

**Variável Resposta:** Será a medida de desempenho do classificador. As opções são:

- **Contínua:** Probabilidade de toxicidade (um valor entre 0 e 1).
- **Binária:** Rótulo da classificação (tóxico / não tóxico).

*A escolha final dependerá da natureza dos dados e dos modelos.*

## 2.2 Articulação dos Elementos do Projeto Experimental

**Replicações (3):** O experimento será repetido 3 vezes para cada uma das 8 combinações da tabela fatorial, totalizando 24 execuções.

**Justificativa:**

**Custo Computacional vs. Benefício Estatístico:**  $r=3$  é um ponto de equilíbrio: é um número baixo o suficiente para ser factível, mas já permite uma estimativa inicial da variabilidade (erro experimental), algo que com  $r=1$  ou  $r=2$  é muito impreciso.

## 2.2 Articulação dos Elementos do Projeto Experimental

### Conexão dos Elementos:

- A hipótese será testada executando o experimento fatorial  $2^3$ .
- Os fatores e níveis serão sistematicamente manipulados para observar o efeito na variável resposta.
- As replicações garantirão a validade dos resultados, permitindo aceitar ou rejeitar as hipóteses com base em evidências estatísticas (análise de variância) e, assim, cumprir os objetivos do estudo.

### 3. Mudanças na Estratégia de Análise

### 3. Mudanças na Estratégia de Análise

Desde a proposta inicial, algumas estratégias foram atualizadas:

- **Enriquecimento dos Dados:** O script de coleta foi aprimorado para capturar metadados adicionais, como likes, visualizações e comentários pós-live.
- **Expansão dos Modelos:** Além do BERTimbau e ChatGPT, agora consideramos o uso da Perspective API e modelos mais leves como Llama/Deepseek para execução local.
- **Aumento da Complexidade Experimental:** O projeto evoluiu de um fatorial  $2^2$  para um  $2^3$ , adicionando o fator tamanho\_base\_dados para avaliar o impacto da quantidade de dados no desempenho.

## 4. Preparação e Caracterização dos Dados



## 4.1 Fonte e Estrutura dos Dados

Os dados são comentários extraídos de chats ao vivo do YouTube, coletados via API oficial. Eles são divididos em:

*chat.csv* (mensagens) e *metadados.csv* (informações da transmissão).

- **Categóricas:** *autor, canal, id\_video, titulo*
- **Quantitativas Discretas:** *espectadores\_atuais, likes, visualizacoes, comentarios*
- **Quantitativas Contínuas:** *timestamp, data\_publicacao, data\_inicio\_live*
- **Texto Livre:** *mensagem, descricao*

## 4.2 Status de Prontidão dos Dados

O conjunto de dados está **parcialmente pronto** para a fase de experimentação, com os seguintes progressos e pendências:

- **Coleta e Volume:** A fase de coleta está em sua etapa final, com um volume total que se aproxima de 1 milhão de comentários de mais de 100 lives. Para a análise exploratória dos dados, fez-se um recorte de 10 dias.
- **Estruturação e Limpeza:** Os dados coletados possuem alta integridade, sem valores ausentes nas colunas essenciais. Os scripts para a limpeza de ruídos (remoção de links, menções de usuário e alongamentos de palavras), que correspondem a um dos níveis do fator "pré-processamento" no experimento, já foram desenvolvidos e estão prontos para serem aplicados.
- **Rotulagem (Pendente):** A etapa de rotulagem manual de um subconjunto dos dados, que servirá como base para o treinamento e teste dos modelos, é a principal pendência.

# 5. Análise Exploratória dos Dados

## 5. Análise Exploratória dos Dados

A análise a seguir foi feita no recorte de 10 dias (264.791 mensagens).

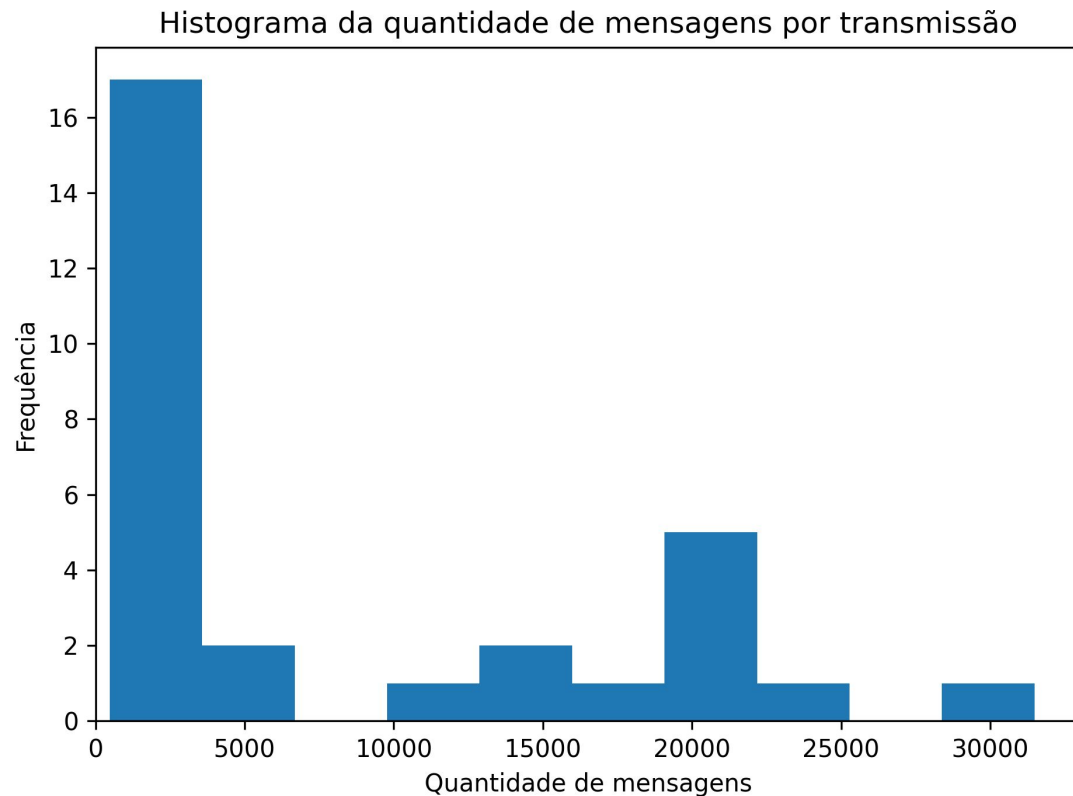
Canal	Live Count	Total Mensagens
REnanPLAY	6	133.031
LUANGAMEPLAY	4	59.992
Diego Sheipado	8	39.490
BiahKov	5	16.968
CAVALÃO 2	7	15.310

# Análise 1: Estatísticas Globais por Transmissão

Variável	Média	Mediana
quantidade_mensagens	8.826,30	3.332,50
tamanho_mensagem	31,81	29,89
tempo_entre_mensagens	27,35	15,70

- Variância significativa nas métricas, com médias e medianas indicando possível distribuição assimétrica.

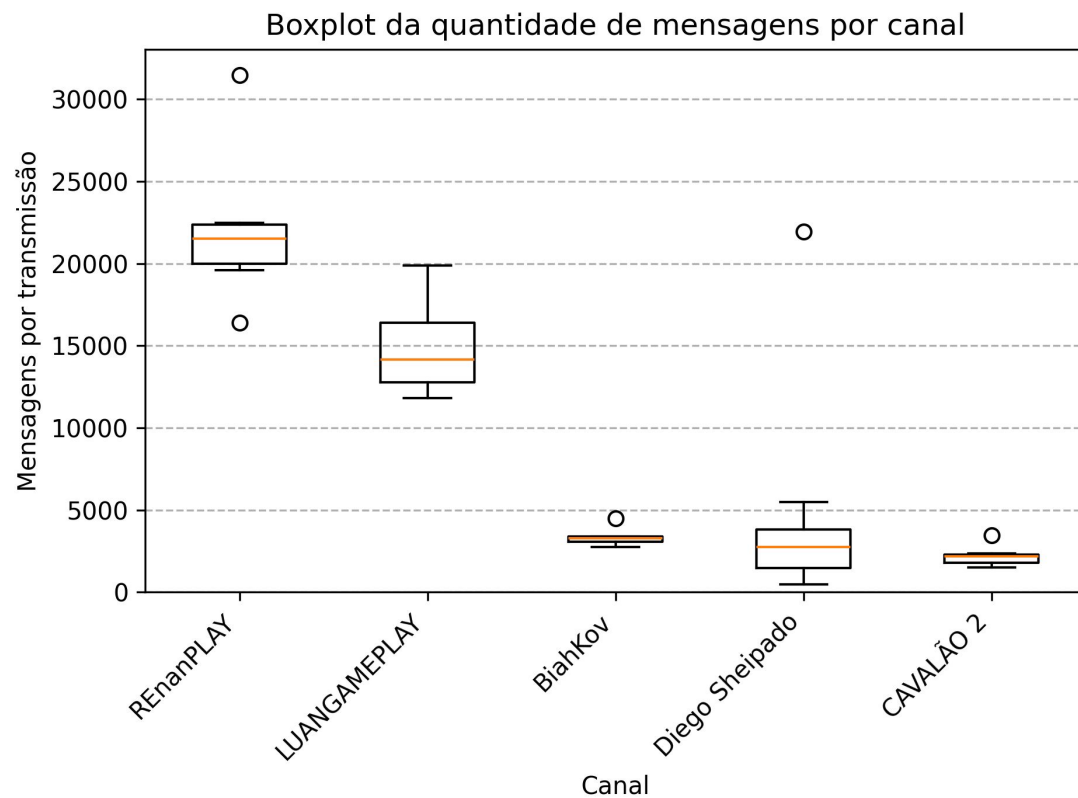
## Análise 2: Quantidade de mensagens por transmissão



Insight:

- A maioria das transmissões tem engajamento moderado, enquanto algumas apresentam volumes excepcionalmente altos, sugerindo a necessidade de considerar outliers em análises futuras.

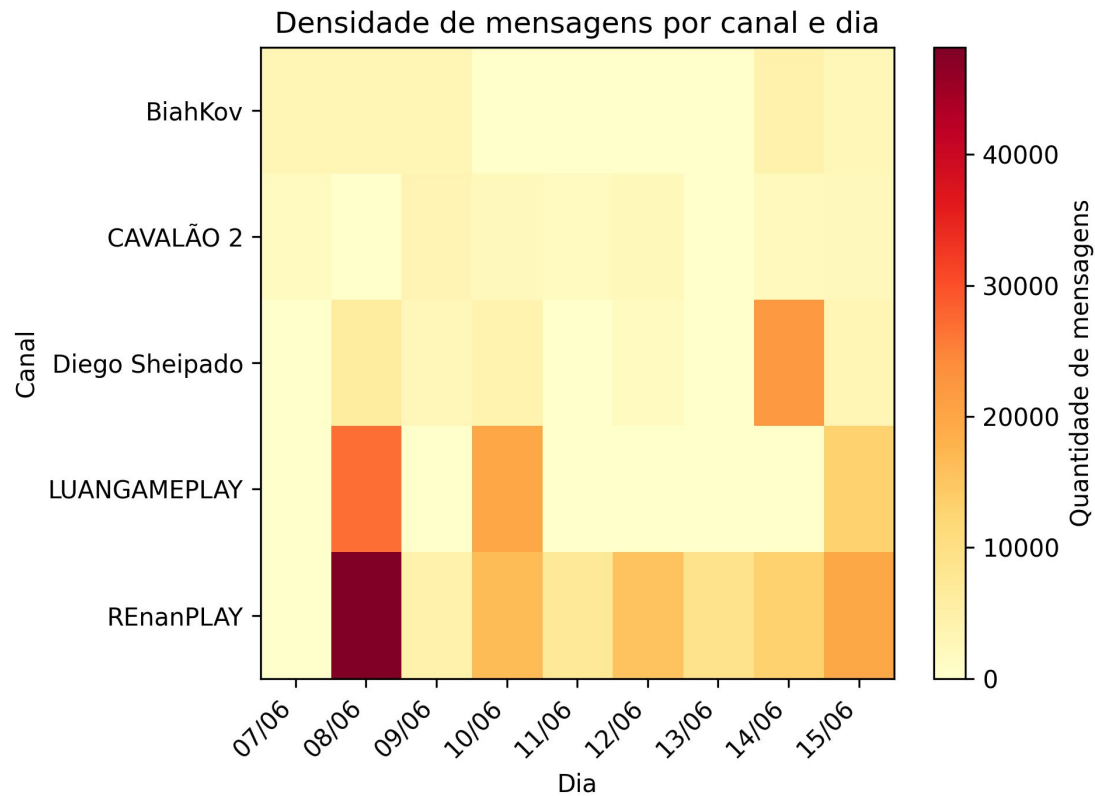
# Análise 3: Quantidade de mensagens por canal



Insight:

- Os canais mostram padrões variados de engajamento, com alguns exibindo maior volume e variabilidade, enquanto outros mantêm transmissões mais homogêneas.

## Análise 4: Volume de mensagens por canal e dia

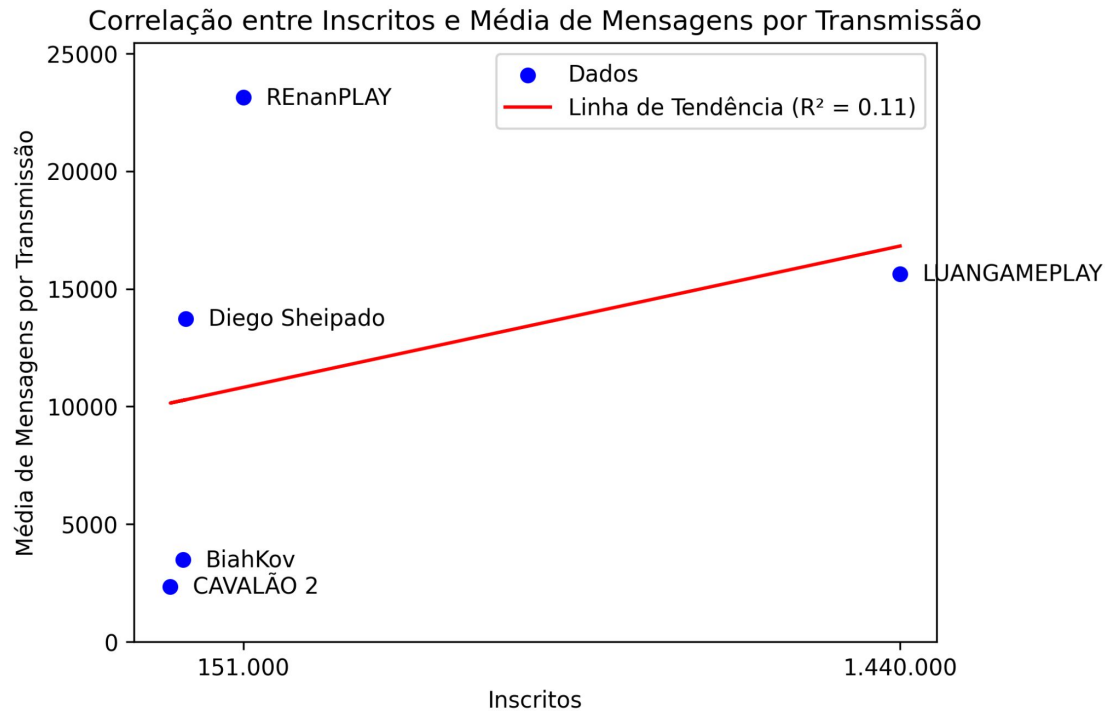


### Insight:

- A visualização revela picos de atividade concentrados em alguns canais e dias, com outros apresentando uma distribuição mais uniforme, sugerindo padrões recorrentes de engajamento.



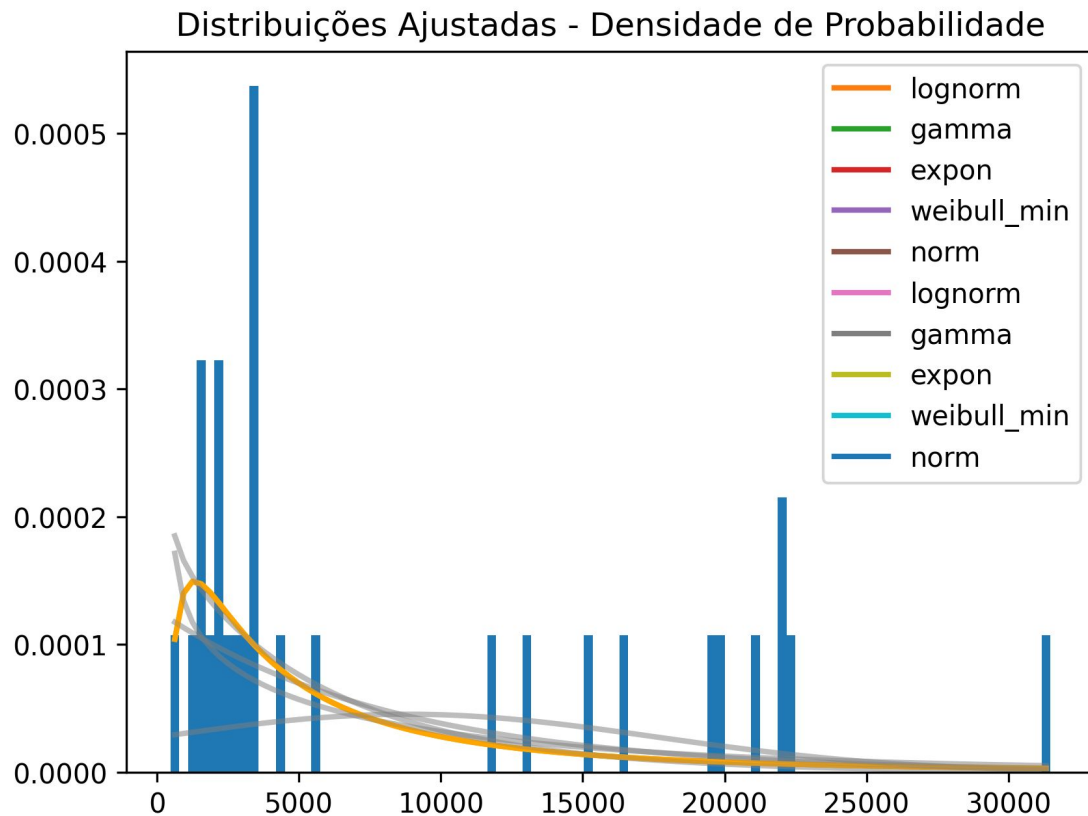
# Análise 5: Correlação entre Inscritos x Média de Mensagens por Transmissão



## Insight:

- Há uma correlação positiva fraca entre o número de inscritos e a média de mensagens por transmissão, com  $R^2$  de 0,11, indicando que apenas 11% da variabilidade nas mensagens pode ser explicada pelos inscritos.

# Análise 6: Verificação de Distribuição Teórica de *quantidade\_mensagens*



## Insight:

- Os dados têm uma cauda longa à direita, típica de fenômenos onde poucos eventos extremos dominam (ex.: lives com alto engajamento).

# Análise 7: Nuvem de palavras mais frequentes nos chats

- Após a criação de uma lista de stopwords customizada e agressiva (com 168 palavras), foi possível remover o ruído superficial (como "jogo", "live", nomes de streamers) e revelar os termos que caracterizam a cultura e a linguagem interna da comunidade analisada.

```
# Junta todas as listas de categorias em uma única lista final.
custom_stop_words = (
    stopwords_conectivos +
    stopwords_verbos +
    stopwords_adjetivos_adverbios +
    stopwords_interjeicoes_girias +
    stopwords_ofensas +
    stopwords_contexto_geral +
    stopwords_contexto_topicos +
    stopwords_contexto_streamers +
    stopwords_misc
)
```

A nuvem de palavras a seguir contém termos pejorativos, gírias e linguagem que podem ser considerados sensíveis. O objetivo é analisar de forma crítica a cultura de comunicação desses espaços, e não endossar o conteúdo.





# Conclusões da Análise Exploratória

- Diferenças claras de engajamento entre os canais.
- Alguns canais se destacam por alto volume de mensagens e devem ser considerados com cuidado na normalização das análises.
- A presença de transmissões com altíssima interação sugere que será importante:
  - Detectar e avaliar os outliers, tratando-os apenas quando forem inconsistentes com o comportamento esperado do conjunto de dados.
  - Levar em conta o canal nas análises futuras, já que ele pode influenciar os resultados e gerar diferenças no volume de mensagens entre as transmissões.
- Canais menores têm distribuições mais concentradas e previsíveis.
- A distribuição lognormal de *quantidade\_mensagens* reforça a necessidade de abordar a assimetria e os outliers, guiando a escolha de métodos estatísticos adequados na próxima fase.
- A análise do vocabulário valida a premissa de que a comunidade utiliza linguagem codificada, justificando a investigação sobre o pré-processamento como um fator crucial.