

UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

SocialDash: Análise Exploratória de Polarização, Toxicidade e Monetização em uma Rede Social de Vídeos

Luis Gustavo Almeida

Orientadores

Ana Cristina Bicharra Garcia

Jefferson Elbert Simões

RIO DE JANEIRO, RJ - BRASIL
MAIO de 2026

SocialDash: Análise Exploratória de Polarização, Toxicidade e Monetização em uma Rede Social de Vídeos

Luis Gustavo Almeida

TESE APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE DOUTOR PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Documento assinado digitalmente
gov.br
ANA CRISTINA BICHARRA GARCIA
Data: 29/05/2026 10:09:10-0300
Verifique em https://validar.iti.gov.br

Ana Cristina Bicharra Garcia - UNIRIO

Documento assinado digitalmente
gov.br
JEFFERSON ELBERT SIMOES
Data: 29/05/2026 10:15:41-0300
Verifique em https://validar.iti.gov.br

Jefferson Elbert Simões - UNIRIO

Documento assinado digitalmente
gov.br
SEAN WOLFGAND MATSUI SIQUEIRA
Data: 29/05/2026 10:48:48-0300
Verifique em https://validar.iti.gov.br

Sean Wolfgand Matsui Siqueira - UNIRIO

Documento assinado digitalmente
gov.br
CARLOS EDUARDO RIBEIRO DE MELLO
Data: 29/05/2026 10:42:27-0300
Verifique em https://validar.iti.gov.br

Carlos Eduardo Ribeiro de Mello - UNIRIO

Documento assinado digitalmente
gov.br
FABRICIO BENEVENUTO DE SOUZA
Data: 29/05/2026 10:28:02-0300
Verifique em https://validar.iti.gov.br

Fabrício Benevenuto de Souza - UFMG

Documento assinado digitalmente
gov.br
JULIANA BAPTISTA DOS SANTOS FRANCA
Data: 29/05/2026 07:54:21-0300
Verifique em https://validar.iti.gov.br

Juliana Baptista dos Santos França - UFRJ

RIO DE JANEIRO, RJ - BRASIL
MAIO de 2026

Catalogação informatizada pelo(a) autor(a)

A444 Almeida, Luis Gustavo

SocialDash: Análise Exploratória de Polarização, Toxicidade e Monetização em uma Rede Social de Vídeos
Luis Gustavo Almeida. – Rio de Janeiro :
UNIRIO, 2026.
165 f.

Orientador: Ana Cristina Bicharra Garcia.
Coorientador: Jefferson Elbert Simões.
Tese (Doutorado) – Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2026.

1. polarização. 2. toxicidade. 3. monetização. I. Garcia, Ana Cristina Bicharra, orient. II. Simões, Jefferson Elbert, coorient. III. Título.

Dedico este trabalho ao Gustavo do passado e ao Gustavo do futuro.

Ao Gustavo do passado, que após muitas fases da vida retornou de Fortaleza para São Paulo decidido a levar a informática a sério. Foi em uma escola simples, durante um curso de HTML, que percebi ter encontrado algo pelo qual era verdadeiramente apaixonado. Um curso levou a outro. Vieram o software livre, a programação, os servidores, as redes, os sistemas, os congressos e os eventos. Foi nesse período que um instrutor me disse uma frase simples:

”Se quiser saber mais, vai ter que fazer mestrado e doutorado.”

A frase ecoou durante anos. Ao mesmo tempo, outro sonho parecia distante: morar no Rio de Janeiro. Com o tempo, esses dois objetivos se fundiram em um só. Ao acompanhar o crescimento do Google, um projeto nascido dentro do meio acadêmico, pensei comigo mesmo:

”Se eles conseguiram criar algo dessa magnitude, é para lá que eu vou.”

Entre aquelas palavras e a entrega desta tese passaram-se 13 anos e 87 cursos. O que começou em uma sala de HTML tornou-se uma jornada que me levou ao mais alto grau da formação acadêmica.

Dedico também este trabalho ao Gustavo do futuro, que talvez leia estas palavras aos 99 anos, em 2075.

Espero que ele ainda mantenha a mesma curiosidade, a mesma paixão por aprender e a mesma vontade de construir coisas novas. Porque esta tese não é um ponto final. É apenas mais uma montanha. E, quando finalmente alcançamos o topo, descobrimos que ainda queremos alcançar a lua.

Agradecimentos

Agradeço, primeiramente, a Deus, pela proteção, pela saúde e pela força concedida ao longo desta caminhada.

À minha orientadora, professora Ana Cristina, agradeço pela orientação, pelas contribuições e pelo acompanhamento durante o desenvolvimento desta pesquisa.

Ao meu coorientador, professor Jefferson, agradeço pelas contribuições técnicas, pelas revisões e pelo apoio na consolidação deste trabalho.

Aos professores da banca, agradeço pela leitura atenta, pelas críticas, sugestões e contribuições que ajudaram a aprimorar esta tese.

Aos meus pais, agradeço pela vida, pela formação e por tudo aquilo que, direta ou indiretamente, contribuiu para que eu chegasse até aqui.

À minha irmã e ao meu irmão, agradeço pela presença familiar e pela história compartilhada ao longo da vida.

Ao meu sobrinho, por quem tenho grande admiração e carinho, agradeço pela inspiração, pela alegria e pela relação de respeito e afeto que construímos.

Aos amigos, colegas, professores e a todos que, em diferentes momentos, fizeram parte desta trajetória, deixo também meu sincero agradecimento.

Por fim, agradeço à própria caminhada. Aos desafios, às incertezas, aos recomeços e às conquistas que fizeram deste percurso não apenas uma etapa acadêmica, mas também uma profunda experiência de amadurecimento pessoal e profissional.

Almeida, Luis Gustavo. **SocialDash: Análise Exploratória de Polarização, Toxicidade e Monetização em uma Rede Social de Vídeos**. UNIRIO, 2026. 165 páginas. Tese de Doutorado. UNIRIO. Programa de Pós-Graduação em Informática.

RESUMO

Atualmente, pesquisadores enfrentam o desafio de analisar fenômenos complexos e interdependentes no YouTube, como polarização, toxicidade e monetização. Esse desafio decorre da fragmentação dos dados, da opacidade dos mecanismos algorítmicos e da ausência de ferramentas que integrem essas dimensões de forma sistemática, dificultando a correlação de indicadores econômicos, sociais e discursivos em larga escala. Nesse contexto, o problema de pesquisa consiste na dificuldade de estruturar e operacionalizar uma análise integrada dessas dimensões para apoiar a exploração de padrões e relações em ambientes digitais. Embora a literatura trate esses fenômenos de forma consolidada, eles são frequentemente analisados de maneira isolada, limitando a compreensão de suas interdependências. Esta tese adota a abordagem de Design Science Research para desenvolver um modelo analítico e um artefato computacional capazes de integrar essas dimensões. O artefato, denominado *SocialDash*, é uma aplicação Web que realiza coleta automatizada via API do YouTube e *web crawling*, incorporando métricas de toxicidade (Perspective API) e proxies de monetização. Os dados são organizados em widgets analíticos que permitem análises descritivas, comparativas e temporais, apoiando o processo de *sensemaking*. O *SocialDash* não busca estabelecer relações causais, mas oferecer uma infraestrutura para exploração empírica dessas dimensões. Como contribuição, propõe-se uma arquitetura integrada baseada em automação e *visual analytics*, demonstrando como a combinação de múltiplas fontes de dados reduz o custo cognitivo do pesquisador e viabiliza a análise de fenômenos sociotécnicos complexos.

Palavras-chave: Monetização, Polarização, Toxicidade, Youtube.

ABSTRACT

Currently, researchers face the challenge of analyzing complex and interdependent phenomena on YouTube, such as polarization, toxicity, and monetization. This challenge stems from data fragmentation, the opacity of algorithmic mechanisms, and the lack of tools capable of systematically integrating these dimensions, making it difficult to correlate economic, social, and discursive indicators at scale. In this context, the research problem lies in the difficulty of structuring and operationalizing an integrated analysis of these dimensions in order to support the exploration of patterns and relationships in digital environments. Although the literature addresses these phenomena in a consolidated manner, they are often analyzed in isolation, limiting the understanding of their interdependencies. This thesis adopts the Design Science Research approach to develop an analytical model and a computational artifact capable of integrating these dimensions. The artifact, named *SocialDash*, is a web application that performs automated data collection through the YouTube API and web crawling, incorporating toxicity metrics (Perspective API) and monetization proxies. The data are organized into analytical widgets that enable descriptive, comparative, and temporal analyses, supporting the *sensemaking* process. The *SocialDash* does not aim to establish causal relationships, but rather to provide an infrastructure for the empirical exploration of these dimensions. As a contribution, this research proposes an integrated architecture based on automation and *visual analytics*, demonstrating how the combination of multiple data sources reduces researchers' cognitive load and enables the analysis of complex sociotechnical phenomena.

Keywords: Polarization, Monetization, YouTube, Toxicity, Comments.

Sumário

1 Introdução **1**

1.1 Motivação 2

1.2 Questões de pesquisa 4

1.2.1 Objetivos da pesquisa 5

1.2.1.1 Objetivo geral 5

1.2.1.2 Objetivos específicos 5

1.3 Estrutura da tese 6

2 Fundamentação Teórica **7**

2.1 Relevância do YouTube 7

2.2 Polarização 9

2.3 Toxicidade 13

2.4 Economia de atenção 18

2.5 Monetização 20

2.5.1 Monetização em redes sociais 22

2.5.1.1 Monetização no YouTube 23

2.6 *Black-box auditing* 26

2.6.1 Estratégias metodológicas 28

2.7 *Sensemaking* 30
2.8 *Visual analytics* 35
2.9 Síntese da fundamentação teórica 37

3 Trabalhos Relacionados **38**
3.1 Caracterização da base bibliográfica 38

4 Metodologia de Pesquisa **49**
4.1 Design Science Research: problema, artefato e conjecturas 49
4.2 Coleta, análise e preparação dos dados 51
4.2.1 Ontologia, tipologia e taxonomia dos dados 51
4.3 Modelo 55
4.4 Justificativa para decisões de pesquisa 56
4.5 Implementação do artefato 65
4.6 Avaliação 66
4.6.1 Contexto experimental 67
4.6.2 Instrumento de avaliação 68

5 Benchmarks **70**
5.1 Dados de toxicidade 70
5.1.1 Validação experimental de métodos de detecção de toxicidade . . 72
5.1.2 Justificativa da escolha da Perspective API 79
5.2 Dados de polarização 81
5.3 Dados de monetização 94
5.3.1 URLs externas nas descrições de vídeos e canais 98
5.4 Prova de conceito e primeiros testes 100

6 Resultados da Avaliação **104**

6.1 Análise estatística 104

6.2 Feedbacks qualitativos 106

6.3 Interpretação por conjectura 108

6.4 Limitações observadas 111

6.5 Resultados gerais e síntese 112

7 Encerramento **114**

7.1 Contribuição científica 114

7.2 Limitações da pesquisa 115

7.3 Trabalhos futuros 116

7.4 Considerações finais 116

A Apêndice **125**

A.1 Diagramas do artefato 125

A.1.1 Diagrama de fluxo de dados 130

A.2 Tecnologias de implementação 132

A.3 Resumo das principais funções do *backend* do artefato 133

A.3.1 Conjectura — polarização e toxicidade 133

A.3.2 Conjectura — monetização e toxicidade 134

A.3.3 Conjectura — polarização e monetização 135

A.3.4 Conjectura — integração entre polarização, toxicidade e monetização 136

A.4 Prompts utilizados 137

A.5 Interfaces do artefato 140

Lista de Figuras

1.1 Interseções entre polarização, monetização e toxicidade 3

2.1 Esquema conceitual da dinâmica da toxicidade em ambientes digitais. . . 15

3.1 Distribuição temporal da base bibliográfica. 40

3.2 Indicadores de engajamento e homofilia no YouTube. 43

4.1 Estrutura epistemológica da pesquisa em DSR aplicada ao artefato *SocialDash*, adaptada de [72]. 50

4.2 Taxonomia conceitual dos dados utilizados pelo SocialDash, relacionando entidades, acessibilidade e níveis de inferência analítica. 53

4.3 Modelo conceitual para coleta, processamento e análise de sinais socio-técnicos em plataformas digitais. 56

4.4 Fluxo de normalização e enriquecimento no SocialDash. 58

4.5 Estratégia de segmentação temporal em *buckets*. 59

4.6 Métricas absolutas e relativas de engajamento. 61

4.7 Decomposição inferencial da monetização. 63

4.8 Funil inferencial de variáveis analíticas. 65

5.1 Escala ordinal utilizada para validação da mensuração de toxicidade . . . 72

5.2 Benchmark temporal do pipeline PMT. 103

5.3 Validação da persistência dos dados processados da POC no banco de dados, preservando o aspecto relacional dos dados. 103

A.1 Diagrama de caso de uso do SocialDash para análise PMT. 125

A.2 Diagrama de classes simplificado do SocialDash. 126

A.3 Diagrama de estados do fluxo de análise PMT no SocialDash. 127

A.4 Diagrama de sequência da análise P-M-T no SocialDash. 128

A.5 Esquema entidade-relacionamento simplificado da base de dados do SocialDash. 129

A.6 Diagrama de fluxo de dados do SocialDash 131

A.7 Fluxo simplificado de processamento da conjectura C1 no SocialDash. . . 134

A.8 Fluxo simplificado de processamento da conjectura C3 no SocialDash. . . 135

A.9 Fluxo simplificado de processamento da conjectura C2 no SocialDash. . . 136

A.10 Fluxo simplificado de processamento do Widget 4 no SocialDash. 137

A.11 Página inicial do SocialDash. 140

A.12 Interface conceitual do SocialDash. 141

A.13 Interface explicativa sobre toxicidade. 142

A.14 Interface explicativa sobre polarização. 143

A.15 Interface inicial da conjectura C1. 144

A.16 Interface de seleção de vídeos do SocialDash. 145

A.17 Resultado da conjectura C1. 146

A.18 Resultado da conjectura C2. 147

A.19 Resultado da conjectura C3. 148

A.20 Resultado integrado da conjectura C4. 149

Lista de Tabelas

2.1 Comparação entre redes sociais e sites mais populares globalmente em termos de usuários ativos e volume de visitas por mês. 9

2.2 Taxonomia de Polarização 11

2.3 Tipos de monetização no YouTube 25

3.1 Distribuição temática da base bibliográfica por eixo analítico 39

3.2 Distribuição geográfica da base bibliográfica 39

4.1 Taxonomia da base de dados do SocialDash. 54

4.2 Comparação entre coleta de dados via API oficial e técnicas de crawling utilizadas no contexto do artefato proposto. 57

4.3 Exemplos de métricas absolutas observáveis na plataforma e suas interpretações analíticas. 60

4.4 Métricas relativas derivadas da plataforma. 62

4.5 Síntese inferencial dos fenômenos PMT. 64

4.6 Instrumento de avaliação do SocialDash baseado nas conjecturas C1–C4 . 69

5.1 Síntese das estratégias de mensuração de toxicidade por tecnologia 71

5.2 Descrição das colunas do benchmark de toxicidade 73

5.3 Exemplo de cálculo das *proxies* derivadas do Google NLP 74

5.4 Detalhamento do processo de conversão das métricas do Google NLP em *proxies* derivadas de toxicidade. 74

5.5 Configuração do benchmark de sensibilidade de toxicidade 76

5.6 Amostra do benchmark de sensibilidade de toxicidade 77

5.7 Desempenho dos modelos na preservação da progressão de toxicidade . . 78

5.8 Atributos de toxicidade disponíveis na Perspective API 80

5.9 Formulações de polarização na literatura 81

5.10 Amostra de canais para benchmark de classificação ideológica via LLM . 82

5.11 Amostra de canais e estatísticas descritivas do dataset coletado 84

5.12 Configuração dos grupos de atributos utilizados nos prompts (*labs*) 87

5.13 Amostra do dataset utilizado no experimento inicial ($n = 10$) 89

5.14 Resultados com transcrição, metadados do vídeo e dados do canal 90

5.15 Resultados do experimento com $n = 100$ por combinação de dados . . . 91

5.16 Resultados dos experimentos com dataset completo (n = 1630) 93

5.17 Desempenho agregado por dimensão — dataset completo 93

5.18 Desempenho por dimensão e configuração (Labs × Dimensão) 93

5.19 Benchmark de abordagens de monetização na literatura 95

5.20 Benchmark empírico de monetização on-platform e estimativas externas . 96

5.21 URLs externas nas descrições dos vídeos: médias por canal 99

5.22 Prova de Conceito do SocialDash – Execução do Script PMT já no formato de uma página Web. 102

6.1 Síntese estatística das conjecturas avaliadas 105

6.2 Síntese estatística da avaliação de usabilidade do artefato 106

6.3 Temas observados nos feedbacks internos do SocialDash 107

6.4 Síntese qualitativa das questões abertas Q14 e Q15 108

6.5 Síntese interpretativa das conjecturas avaliadas 110

6.6 Relação entre resultados da avaliação e questões de pesquisa 110

6.7 Síntese técnico-operacional da avaliação do artefato 111

A.1 Estrutura relacional principal da base de dados do SocialDash 130

A.2 Requisitos de ambiente e tecnologias utilizadas no SocialDash 133

Lista de Nomenclaturas

BLM	*Black Lives Matter*
CPM	Custo Por Mil Impressões (*Cost Per Mille*)
CTR	*Click Through Rate*
DMCA	*Digital Millennium Copyright Act*
DSR	*Design Science Research*
IDW	*Intellectual Dark Web*
LLM	*Large Language Model*
MCN	*Multi-Channel Network*
OSN	*Online Social Networks*
RPM	Receita Por Mil (*Revenue Per Mille*)
RPA	*Robotic Process Automation*
SERP	*Search Engine Results Page*
TIC	Tecnologia da Informação e Comunicação
UGC	*User Generated Content*
WIPO	*World Intellectual Property Organization*
YPP	*YouTube Partner Program*

1. Introdução

Atualmente vivemos em um contexto amplamente mediado pela internet, no qual redes sociais e plataformas digitais desempenham papel central na circulação de informações, nas interações sociais e nos processos de produção cultural [1]. Essas plataformas não apenas conectam indivíduos, mas também influenciam comportamentos, dinâmicas discursivas, formação de opinião e mecanismos de engajamento coletivo. Estudos em psicologia social apontam ainda que fatores relacionados à agressividade, frustração e conflito podem influenciar interações coletivas e processos de hostilidade discursiva em ambientes digitais e politicamente polarizados [2].

Nesse ecossistema, o YouTube consolidou-se como uma das principais plataformas para circulação de conteúdo audiovisual e interação social online, sendo frequentemente investigado sob perspectivas relacionadas à recomendação algorítmica, monetização, engajamento e dinâmicas discursivas [3, 4, 5, 6]. Sua relevância decorre não apenas da escala global de utilização, mas também da combinação entre produção massiva de conteúdo, mecanismos algorítmicos de recomendação e estruturas econômicas baseadas em atenção. Estudos recentes mostram que a plataforma apresenta forte concentração de visualizações e engajamento, em que uma pequena parcela dos canais concentra grande parte da audiência total, configurando fenômenos associados ao efeito de rede e ao modelo *rich-get-richer*[1], além de vantagens estruturais relacionadas à antiguidade e consolidação dos canais [7]. Essas características tornam o YouTube um ambiente particularmente relevante para investigações sobre sistemas sociotécnicos mediados por algoritmos e economia da atenção.

Diante desse cenário, esta pesquisa busca ampliar o espaço de observação empírica sobre o funcionamento desse ecossistema. Ao integrar indicadores de monetização estimada, métricas de toxicidade e sinais de polarização em um sistema unificado, a tese

[1] O ‘Efeito Mateus”, proposto por Robert K. Merton (1968), descreve dinâmicas de vantagem acumulativa nas quais ‘os ricos ficam mais ricos”.

propõe um framework Web exploratório voltado à identificação de padrões temporais e associativos ainda pouco sistematizados na literatura.

1.1 Motivação

A literatura recente sobre plataformas digitais evidencia avanços significativos na análise de fenômenos como polarização, toxicidade e monetização. No entanto, esses fenômenos têm sido predominantemente investigados de forma isolada, com foco em dimensões específicas e com limitada integração analítica entre eles. Essa fragmentação dificulta a compreensão de dinâmicas mais amplas e interdependentes que caracterizam ambientes sociotécnicos complexos como o YouTube.

No caso da **polarização**, observa-se uma multiplicidade de abordagens e definições, incluindo dimensões ideológicas, afetivas, estruturais e algorítmicas [8, 9, 10]. Apesar desse avanço conceitual, persiste uma lacuna na operacionalização do fenômeno, que frequentemente depende de proxies indiretos, como formação de clusters ou padrões de consumo [10, 11]. Adicionalmente, a relação entre polarização e mecanismos econômicos das plataformas permanece pouco explorada.

No que se refere à **toxicidade**, a literatura apresenta avanços na sua formalização computacional, especialmente por meio de modelos baseados em aprendizado de máquina [12, 13]. Entretanto, essas abordagens enfrentam limitações relacionadas à dependência de proxies textuais, sensibilidade a contexto e dificuldades de generalização [12, 13]. Além disso, estudos indicam que conteúdos associados a emoções de alta ativação — como raiva, ansiedade e controvérsia — tendem a apresentar maiores níveis de engajamento e amplificação, sugerindo conexões com incentivos estruturais das plataformas [14, 15, 16].

Já a **monetização** tem sido analisada principalmente sob a ótica da economia da atenção e da performance de conteúdo, sendo frequentemente operacionalizada por métricas de engajamento, como visualizações e tempo de exibição [17, 18]. Embora existam evidências sobre estratégias off-platform e dinâmicas da economia dos criadores [19, 20], ainda há limitada integração entre monetização, padrões discursivos e efeitos sociotécnicos como polarização e toxicidade.

Para além dessas lacunas individuais, observa-se uma ausência mais profunda de integração entre essas dimensões.

Na interseção entre **polarização e toxicidade (P–T)**, estudos sugerem que ambientes polarizados tendem a intensificar interações hostis e discursos extremados [21, 22]. Con-

tudo, permanecem escassas as abordagens sistemáticas capazes de analisar empiricamente essa relação de forma contínua, integrada e comparativa em larga escala.

Na interseção entre **polarização e monetização (P–M)**, evidências indicam que conteúdos polarizados podem ser incentivados por dinâmicas de engajamento e visibilidade [3]. Ainda assim, permanece pouco compreendida a relação entre incentivos econômicos e a produção ou amplificação de conteúdo polarizado.

Na interseção entre **toxicidade e monetização (T–M)**, a literatura aponta para um possível *trade-off* entre engajamento e qualidade do discurso, no qual conteúdos mais intensos ou controversos tendem a gerar maior interação [14]. Consequentemente, essa relação ainda carece de instrumentos analíticos integrados que permitam sua exploração sistemática.

Por fim, na interseção tripla entre **polarização, toxicidade e monetização (P–T–M)**, identifica-se uma lacuna crítica relacionada à ausência de abordagens integradas capazes de combinar simultaneamente essas dimensões em um mesmo arcabouço analítico. Observa-se, de forma recorrente na literatura, a predominância de análises unidimensionais ou de relações pontuais entre pares de variáveis, o que limita a compreensão do fenômeno como um sistema sociotécnico interdependente.

A Figura 1.1 sintetiza graficamente esse problema de pesquisa, representando as três dimensões centrais — polarização (P), monetização (M) e toxicidade (T) — e suas interseções. As regiões de interseção parcial (P–T, P–M e T–M) correspondem a abordagens já exploradas de forma fragmentada e escassa na literatura, enquanto a interseção central (P–T–M) evidencia a lacuna analítica que esta pesquisa busca endereçar.

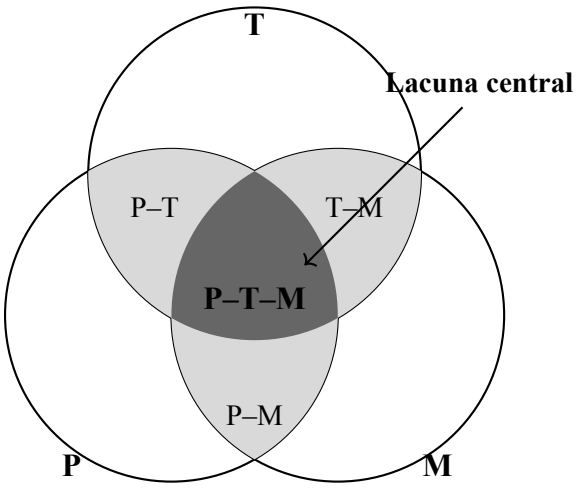


Figura 1.1: Representação conceitual das interseções entre polarização (P), monetização (M) e toxicidade (T), com destaque para a lacuna analítica na interseção tripla (P–T–M).

Essa lacuna é agravada pela opacidade dos sistemas algorítmicos que mediam essas dinâmicas. Abordagens de auditoria de caixa-preta demonstram a viabilidade de inferir padrões a partir de observação de outputs, mas também evidenciam limitações relacionadas à inferência indireta e à ausência de acesso aos mecanismos internos [8, 23, 24]. Nesse contexto, o desafio não é apenas técnico, mas também analítico e cognitivo, exigindo ferramentas que apoiem a integração e interpretação de múltiplas fontes de evidência.

Diante desse cenário, emerge a necessidade de abordagens que não apenas coletem e processem dados, mas que permitam integrar diferentes dimensões analíticas e apoiar o processo de interpretação. Nesse sentido, a literatura de *sensemaking* destaca a importância de representações que reduzam a complexidade e viabilizem a construção de significado a partir de dados heterogêneos [25]. De forma complementar, a *visual analytics* propõe a integração entre análise automatizada e exploração interativa como meio de apoiar processos de descoberta e inferência [26].

1.2 Questões de pesquisa

Com base nessas lacunas e fundamentos teóricos, esta pesquisa propõe investigar o seguinte problema central:

Como integrar e operacionalizar, em um ambiente analítico computacional, as dimensões de polarização, toxicidade e monetização em plataformas digitais, de forma a apoiar a exploração e interpretação de suas interdependências?

A partir desse problema, são formuladas as seguintes questões de pesquisa:

- **RQ1 (Fenômeno):** Como polarização, toxicidade e monetização se relacionam no ecossistema do YouTube a partir de sinais digitais observáveis?
- **RQ2 (Inferência):** Como proxies computacionais, APIs externas, fontes de dados públicos e modelos inferenciais podem operacionalizar essas dimensões sociotécnicas em plataformas digitais?
- **RQ3 (Artefato):** Como estruturar um ambiente analítico capaz de integrar coleta, inferência e visualização dessas dimensões de forma exploratória e escalável?

Em alinhamento com a abordagem de *Design Science Research* (DSR), esta tese apresenta três contribuições principais:

- **C1 – Modelo**: Proposição de um modelo conceitual e inferencial que articula polarização, toxicidade e monetização como dimensões interdependentes em ambientes sociotécnicos.
- **C2 – Metodologia**: Desenvolvimento de uma abordagem metodológica baseada em dados observáveis e inferência analítica, integrando técnicas de coleta, processamento e análise de dados em plataformas digitais.
- **C3 – Artefato**: Implementação do sistema *SocialDash*, como prova de conceito, que operacionaliza o modelo proposto por meio de uma arquitetura baseada em *Robot Process Automation* (RPA) e visual analytics, permitindo análises exploratórias e comparativas em larga escala.

1.2.1 Objetivos da pesquisa

Com base no problema de pesquisa e nas lacunas identificadas na literatura, esta tese tem como objetivo geral propor e avaliar uma abordagem integrada para análise de polarização, toxicidade e monetização em plataformas digitais.

1.2.1.1 Objetivo geral

Desenvolver um modelo conceitual e um artefato computacional capaz de integrar e operacionalizar as dimensões de polarização, toxicidade e monetização no YouTube, apoiando a análise exploratória e o processo de *sensemaking* em ambientes digitais complexos.

1.2.1.2 Objetivos específicos

- **OE1 – Modelagem conceitual**: Estruturar um modelo conceitual e inferencial capaz de representar as relações entre polarização, toxicidade e monetização a partir de sinais digitais observáveis e variáveis derivadas.
- **OE2 – Coleta e operacionalização**: Definir mecanismos de coleta, normalização e integração de dados provenientes de APIs oficiais, APIs de terceiros, Web crawling, transcrições, comentários e metadados do YouTube, convertendo diferentes dados em entidades canônicas e variáveis inferidas.
- **OE3 – Inferência computacional**: Implementar métodos de inferência para operacionalização das variáveis de toxicidade, polarização e monetização, incluindo uso de modelos de *Large Language Model* (LLM), APIs especializadas e métricas derivadas de engajamento e monetização.

- **OE4 – Desenvolvimento do artefato**: Projetar e implementar o sistema *SocialDash*, integrando módulos de aquisição, processamento, hidratação, inferência e visualização analítica de dados sociotécnicos.
- **OE5 – Visual analytics e exploração**: Desenvolver mecanismos interativos de exploração visual capazes de apoiar análises temporais, comparativas e exploratórias sobre polarização, toxicidade e monetização em plataformas digitais.
- **OE6 – Avaliação experimental**: Avaliar o artefato por meio de benchmarks computacionais, considerando desempenho operacional, escalabilidade, consistência inferencial e utilidade analítica da solução proposta.

1.3 Estrutura da tese

O Capítulo 1 apresenta a introdução da pesquisa, contextualizando o problema investigado, os objetivos gerais e específicos e as contribuições pretendidas.

O Capítulo 2 versa sobre a fundamentação teórica da tese. São discutidos conceitos relacionados à polarização, radicalização, toxicidade, monetização, economia da atenção e auditoria algorítmica.

O Capítulo 3 apresenta os trabalhos relacionados e a caracterização da base bibliográfica utilizada na pesquisa. A revisão sistemática da literatura permite identificar lacunas metodológicas e limitações nas abordagens existentes em plataformas digitais.

O Capítulo 4 descreve a metodologia de pesquisa adotada, fundamentada no paradigma de DSR. São apresentados o problema investigado, as conjecturas de pesquisa, os processos de coleta e preparação dos dados e a modelagem conceitual do artefato.

O Capítulo 5 aborda os experimentos, benchmarks computacionais e análises exploratórias realizadas durante o desenvolvimento do artefato. O capítulo contempla estudos relacionados à toxicidade, polarização e monetização.

O Capítulo 6 aborda os resultados quantitativos e qualitativos da avaliação do artefato.

O Capítulo 7 apresenta o encerramento da pesquisa, discutindo as contribuições científicas do trabalho, suas limitações e as oportunidades para pesquisas futuras.

O Apêndice A inclui diagramas do artefato, detalhes técnicos, *setup* de implementação, interfaces do sistema desenvolvido e materiais complementares utilizados para apoiar a compreensão da proposta metodológica.

2. Fundamentação Teórica

Este capítulo apresenta os principais fundamentos teóricos que sustentam a pesquisa, abordando conceitos relacionados ao YouTube, polarização, toxicidade, monetização, economia da atenção e métodos de análise de plataformas digitais. As seções seguintes discutem essas dimensões e suas relações no contexto de sistemas sociotécnicos mediados por algoritmos.

2.1 Relevância do YouTube

O YouTube constitui uma das plataformas centrais do ecossistema digital contemporâneo, desempenhando papel relevante na mediação da informação, da cultura e das interações sociais em larga escala [4]. Mais do que um repositório de vídeos, a plataforma opera como uma infraestrutura sociotécnica, na qual elementos técnicos e práticas sociais se articulam para moldar o comportamento dos usuários e as dinâmicas de circulação de conteúdo.

Lançado em 2005, o crescimento da plataforma foi rápido e sustentado, consolidando-se como um dos principais ambientes de consumo e produção de conteúdo online. Diferentemente de outras redes sociais, o YouTube se estrutura predominantemente em torno da linguagem audiovisual, operando como um sistema de distribuição de vídeos baseado em conteúdo gerado por usuários (*user-generated content*, UGC), sintetizado em seu lema original *Broadcast Yourself* [27].

Esse modelo representou uma inflexão significativa na lógica da produção midiática, ao reduzir barreiras de entrada e permitir que indivíduos sem acesso a infraestruturas tradicionais de mídia alcançassem grandes audiências. A consolidação da plataforma foi impulsionada por eventos-chave, como sua aquisição pelo Google em 2006, bem como pelo desenvolvimento de funcionalidades estruturais que ampliaram seu alcance e capa-

cidade de engajamento.

Entre essas funcionalidades, destacam-se os sistemas de recomendação, os mecanismos de compartilhamento, a seção de comentários, o botão de *like* e recursos de interação em tempo real, como transmissões ao vivo e chats. Tais elementos não apenas ampliam a difusão do conteúdo, mas também estruturam padrões de visibilidade, engajamento e interação, influenciando diretamente fenômenos como monetização e toxicidade no ambiente da plataforma [1].

Além de sua evolução histórica, a escala do YouTube reforça sua relevância como objeto de investigação acadêmica. Estimativas recentes indicam que a plataforma conta com aproximadamente 2.5 bilhões de usuários ativos mensais e cerca de 48 bilhões de visitas mensais[2], posicionando-se entre os maiores sistemas digitais do mundo, tanto em termos de audiência quanto de tráfego [7]. Sua presença global é evidenciada pela forte penetração em países como Índia, Estados Unidos, Indonésia e Brasil, o que a caracteriza como uma infraestrutura transnacional de circulação de conteúdo audiovisual.

A Tabela 2.1 evidencia essa centralidade ao situar o YouTube entre as principais plataformas digitais, destacando-se não apenas pelo número de usuários, mas também pelo volume de interações e pelo consumo de conteúdo. Em comparação com outras redes sociais, o YouTube apresenta uma combinação particular de escala, intensidade de uso e diversidade de formatos, o que o torna especialmente relevante para o estudo de dinâmicas complexas de produção, distribuição e recepção da informação no ecossistema digital contemporâneo [4, 5].

Essa centralidade não se limita à dimensão quantitativa. A literatura aponta que o YouTube constitui um sistema de mídia baseado em plataformas, no qual diferentes camadas — produção de conteúdo, mecanismos algorítmicos, modelos de monetização e interação social — se articulam de forma dinâmica [4]. Sob essa perspectiva, a plataforma não atua apenas como repositório de vídeos, mas como um ambiente sociotécnico que organiza e regula a visibilidade, o engajamento e a circulação de discursos.

Dessa forma, o YouTube se apresenta como um objeto privilegiado para a investigação de fenômenos contemporâneos relacionados à comunicação digital. Sua estrutura permite observar, de maneira integrada, processos como a distribuição algorítmica de conteúdo, a conversão de atenção em valor econômico, a formação de comunidades e a emergência de dinâmicas discursivas intensas [3, 5]. É nesse ambiente que se tornam particularmente visíveis fenômenos como polarização, monetização, toxicidade e disputas por atenção,

[2]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

que serão aprofundados nas seções seguintes.

Tabela 2.1: Comparação entre redes sociais e sites mais populares globalmente em termos de usuários ativos e volume de visitas por mês.

Plataforma	**Tipo**	**Usuários (mi)**	**Visitas (bi)**
Facebook	Rede social	3070	9.68
Instagram	Rede social	3000	5.55
WhatsApp	Rede social	3000	—
YouTube	Rede social / vídeo	**2580**	**48.1**
TikTok	Rede social	1990	—
WeChat	Rede social	1410	—
Telegram	Rede social	1000	—
Google	Website	—	98.2
ChatGPT	Website	—	5.38
Reddit	Website	—	4.88
Wikipedia	Website	—	4.67

Fonte: Statista (2026). Dados acessados em 06 mar. 2026.

2.2 Polarização

A polarização é definida na literatura como o processo pelo qual opiniões, crenças ou posicionamentos tornam-se progressivamente mais extremos e divergentes entre indivíduos ou grupos sociais. Em suas origens na psicologia social, esse fenômeno foi estudado a partir da ideia de polarização grupal, segundo a qual a interação em grupo, especialmente em contextos de discussão orientada ao consenso, tende a levar a deslocamentos das posições individuais em direção a extremos [9]. Esse resultado indica que a polarização não depende apenas das opiniões pré-existentes, mas pode emergir e se intensificar a partir da própria dinâmica de interação social. Para além da extremização de opiniões, a polarização pode também ser compreendida como um processo de concentração comportamental, no qual usuários passam a consumir, interagir e circular predominantemente dentro de um mesmo conjunto de conteúdos ou narrativas, independentemente do grau ideológico dessas posições [28].

Com o avanço dos estudos, a polarização passou a ser entendida não apenas como extremização de opiniões, mas como uma reorganização do espaço social em torno de

polos opostos. Em vez de um espectro contínuo, com posições intermediárias, observa-se a formação de estruturas mais rígidas e dicotômicas. Nesse contexto, a polarização envolve também identidade, pertencimento e percepção de ameaça entre grupos, o que contribui para tornar os debates mais hostis e menos deliberativos ao longo do tempo [9].

Uma distinção relevante na literatura é entre polarização ideológica e polarização afetiva [29]. A primeira refere-se a diferenças de posicionamento sobre temas específicos, enquanto a segunda está associada ao aumento de sentimentos negativos em relação ao grupo oposto. Em ambientes digitais, essa dimensão afetiva tende a se intensificar, uma vez que os indivíduos passam não apenas a discordar, mas a rejeitar simbolicamente o outro grupo, ampliando conflitos e antagonismos intergrupais [29].

Com a consolidação das redes sociais, a polarização passa a ser compreendida como um fenômeno sociotécnico, resultante da interação entre usuários, conteúdos e sistemas algorítmicos. Sob essa perspectiva, a compreensão desses sistemas é frequentemente limitada, uma vez que operam como caixas-pretas, o que tem levado ao desenvolvimento de abordagens de auditoria algorítmica (*black-box auditing*) voltadas à investigação de seus efeitos na amplificação de conteúdos e na formação de padrões de exposição [8].

Nesse contexto, é importante diferenciar os conceitos de filtro de bolha (*filter bubbles*) e câmaras de eco (*echo chambers*), ambos associados ao processo de homogeneização informacional em ambientes digitais. As bolhas de filtro estão mais relacionadas à exposição seletiva de conteúdo pelo próprio usuário, reforçando crenças prévias e o viés de confirmação [11, 30]. Já as câmaras de eco referem-se a dinâmicas coletivas de reforço social, nas quais informações e narrativas circulam de forma reiterada em grupos relativamente fechados [10, 28]. Em termos operacionais, esse processo pode ser compreendido como uma redução progressiva da diversidade de exposição a conteúdos ao longo do tempo, associada ao aumento da homogeneidade informacional e à intensificação das interações intragrupo [8]. Tal definição permite aproximar o conceito de polarização de métricas observáveis, como diversidade de recomendações, modularidade de redes e padrões de consumo de conteúdo em plataformas digitais.

No contexto das plataformas digitais, e especialmente no YouTube, a polarização ganha novas camadas associadas à lógica de funcionamento dessas plataformas. O YouTube combina mecanismos de busca, recomendação e consumo contínuo de conteúdo, criando trajetórias de exposição que podem reforçar determinadas narrativas [8, 23, 31, 33]. Evidências sugerem que usuários podem migrar progressivamente de conteúdos moderados para conteúdos mais extremos ao longo do tempo [31, 33].

Além disso, a polarização está associada à dinâmica de engajamento. Conteúdos que provocam reações mais intensas tendem a ser mais promovidos, criando um ciclo de amplificação. Esse processo pode ser interpretado como um ciclo de reforço entre o comportamento do usuário e os algoritmos de recomendação, no qual exposições anteriores condicionam recomendações futuras [21]. Esse processo não é apenas informacional, mas também econômico: em ecossistemas baseados na economia da atenção e na monetização, criadores podem ser incentivados a explorar formatos, narrativas e vínculos com a audiência que maximizem engajamento e sustentação financeira, inclusive por meio de estratégias alternativas à monetização nativa da plataforma [3, 19, 20].

Outro aspecto relevante é que a polarização não ocorre de forma isolada, mas integrada a um ecossistema informacional mais amplo. Conteúdos do YouTube são frequentemente compartilhados em outras plataformas, ampliando seu alcance e permitindo reinterpretações em diferentes contextos sociotécnicos [21, 22, 30]. Estudos indicam que o YouTube atua como uma importante fonte de conteúdo para a disseminação de narrativas controversas em diferentes ambientes digitais [34]. Além disso, mudanças nas regras de monetização e moderação de uma plataforma podem deslocar criadores, audiências e até padrões de toxicidade para outros ambientes, sugerindo que a polarização deve ser analisada também como uma dinâmica de circulação, adaptação e realocação interplataforma [62]. Nesse cenário, a polarização pode ser analisada sob diferentes dimensões complementares, conforme sintetizado na Tabela 2.2.

Tabela 2.2: Taxonomia de Polarização

Dimensão	Descrição	Referências
Ideológica	Diferenças de posicionamento sobre temas políticos, sociais ou morais	[9, 29]
Afetiva	Rejeição emocional entre grupos, incluindo hostilidade e deslegitimação	[21, 29]
Estrutural (rede)	Formação de clusters com baixa interação entre grupos	[10, 32]
Informacional	Exposição seletiva e reforço de crenças (*filter bubbles* e *echo chambers*)	[8, 10, 11]
Algorítmica	Influência dos sistemas de recomendação na visibilidade e na distribuição de conteúdo	[3, 8, 21, 23, 33, 50]
Discursiva	Intensificação do discurso e aumento de conteúdos polarizados e altamente engajadores	[34, 35]

Diante desse conjunto de estudos, a polarização pode ser compreendida como um fenômeno multidimensional, envolvendo a intensificação de posições, a rejeição entre grupos, a influência de sistemas algorítmicos e as dinâmicas de engajamento, o que demanda abordagens analíticas baseadas na integração de múltiplas dimensões observáveis.

Apesar dos avanços recentes, a literatura ainda apresenta limitações importantes. Uma das principais dificuldades refere-se à operacionalização da polarização. O fenômeno raramente é medido diretamente, sendo analisado por meio de *proxies* como a formação de *clusters* [10], padrões de consumo [11] ou métricas de engajamento. Essa diversidade de abordagens evidencia a ausência de um indicador consolidado.

Outro problema reside na ambiguidade conceitual entre polarização, radicalização, toxicidade e desinformação. A polarização refere-se ao aumento da distância entre grupos [9, 29]; a radicalização está associada ao deslocamento para posições mais extremas [33]; a toxicidade envolve o uso de linguagem agressiva ou hostil [22]; e a desinformação refere-se à disseminação de conteúdos falsos ou enganosos [21].

Além disso, muitos estudos se limitam a uma única plataforma, desconsiderando a circulação interplataforma. Evidências indicam que conteúdos circulam entre diferentes ambientes digitais, sugerindo que a polarização deve ser compreendida como um fenômeno ecossistêmico [21]. Também permanece insuficientemente explorada a dimensão econômica da polarização em plataformas, isto é, a forma como regras de monetização, desmonetização, financiamento alternativo e desigualdades entre criadores influenciam a produção, a circulação e a persistência de conteúdos controversos [3, 19, 20, 62].

Por fim, observa-se a ausência de ferramentas analíticas integradas capazes de combinar múltiplas dimensões do fenômeno. A maioria dos estudos adota abordagens isoladas e estáticas, o que limita a compreensão da polarização como um processo dinâmico e multifacetado.

Diante desse cenário, para os fins desta tese e do design do artefato *SocialDash*, a polarização é tratada como um construto multidimensional, operacionalizado a partir de múltiplos sinais observáveis do ambiente do YouTube. O artefato propõe uma abordagem baseada em LLMs, utilizando prompts estruturados e conjuntos de registros textuais para inferir estimativas de polarização. O processo parte de três etapas principais: (i) identificação do tema ideológico predominante associado ao conteúdo analisado; (ii) definição de polos discursivos coerentes com o tema identificado; e (iii) obtenção, a partir da própria LLM, de uma estimativa de classificação acompanhada de um índice de confiança da inferência realizada.

2.3 Toxicidade

Historicamente, a toxicidade pode ser compreendida como a manifestação de comportamentos hostis, agressivos ou abusivos em interações sociais, incluindo insultos, ameaças, assédio e discursos ofensivos. Em ambientes digitais, esse fenômeno adquire características próprias associadas à mediação tecnológica, à dinâmica das plataformas e às formas de interação online. Nesse contexto, a compreensão contemporânea da toxicidade não se origina inicialmente como uma categoria computacional ou métrica, mas como um fenômeno psicológico associado ao comportamento humano em contextos mediados por tecnologia. Um dos marcos fundamentais nesse sentido é o conceito de efeito de desinibição online, segundo o qual indivíduos tendem a expressar-se de forma mais intensa, aberta ou extrema no ambiente digital do que em interações presenciais [36].

A literatura distingue entre uma desinibição benigna — associada à autoexpressão e ao compartilhamento emocional — e uma forma negativa, denominada *toxic disinhibition*, caracterizada por linguagem rude, críticas agressivas, ódio, ameaças e outras formas de hostilidade interpessoal [36]. Dessa forma, a toxicidade emerge inicialmente como uma manifestação comportamental da desinibição digital, e não como uma propriedade linguística isolada.

A origem desse fenômeno está relacionada a características estruturais do ambiente online, como anonimato, invisibilidade, assincronicidade, dissociação identitária e redução da autoridade social, que reduzem barreiras normativas e facilitam a externalização de impulsos agressivos [36]. Dessa forma, a toxicidade pode ser compreendida, em sua base epistemológica, como um fenômeno sociopsicológico emergente da arquitetura sociotécnica dos ambientes digitais. Além disso, evidências empíricas indicam que ambientes de interação online, especialmente em redes sociais e sistemas de comentários, tendem a concentrar altos níveis de toxicidade, frequentemente associados a discussões políticas e conteúdos noticiosos, indicando que esse fenômeno não é apenas individual, mas também coletivo e estrutural [37].

Outrossim, evidências empíricas recentes indicam que a toxicidade pode emergir não apenas de interações espontâneas, mas também de ações coordenadas entre grupos de usuários, como no caso de ataques organizados (*raids*), nos quais comunidades externas mobilizam participantes para produzir grandes volumes de comentários hostis direcionados a conteúdos específicos. Esse tipo de comportamento evidencia que a toxicidade pode assumir uma dimensão estratégica e coletiva, indo além de manifestações individuais e reforçando sua natureza sociotécnica [38].

Com a expansão das plataformas digitais e a necessidade de moderação em larga escala, a toxicidade passou a ser progressivamente formalizada como um conceito operacional. Uma das definições mais difundidas descreve toxicidade como comentários rudes, desrespeitosos ou irrazoáveis que tendem a afastar participantes de uma discussão [13].

Essa definição representa uma mudança importante: a toxicidade deixa de ser entendida apenas como linguagem ofensiva e passa a ser definida também em termos de efeito social e impacto na interação, isto é, sua capacidade de deteriorar o ambiente discursivo. Não obstante, essa operacionalização implica simplificações, uma vez que o fenômeno permanece altamente dependente de contexto, intenção e interpretação subjetiva [13]. Paralelamente, estudos indicam que a toxicidade tende a se concentrar de forma desigual em determinados espaços e interações, sendo frequentemente produzida por uma fração limitada de usuários ou contextos, o que sugere a existência de padrões estruturais de distribuição da toxicidade [37]. Esse padrão também é corroborado por análises empíricas que identificam diferentes dimensões de toxicidade — como insultos, ameaças e ataques identitários — distribuídas de forma desigual entre usuários e interações, reforçando a natureza estrutural do fenômeno [39].

Igualmente, a literatura evidencia que a toxicidade é um fenômeno multidimensional, envolvendo elementos como agressividade verbal, desrespeito, estereótipos e ataques identitários. Em contextos aplicados, especialmente na moderação de conteúdo, a toxicidade é frequentemente operacionalizada por meio de indicadores práticos como insultos, discurso de ódio, assédio e uso de termos ofensivos [40]. A evolução do conceito de toxicidade pode ser compreendida em três dimensões principais.

A primeira dimensão, de natureza psicológica, corresponde à sua origem como expressão da desinibição e da agressividade em ambientes mediados, conforme descrito pelo efeito de desinibição online [36]. Nesse estágio, a toxicidade é interpretada como um comportamento emergente, dependente de fatores situacionais e das características do ambiente.

A segunda dimensão, de caráter sociotécnico, emerge com a consolidação das plataformas digitais, nas quais a toxicidade passa a ser reconhecida como um problema estrutural de interação em larga escala, especialmente em sistemas de comentários e redes sociais. Sob essa perspectiva, estudos indicam que criadores e plataformas enfrentam desafios constantes para lidar com comportamentos abusivos, equilibrando liberdade de expressão, engajamento e segurança [40]. Acrescenta-se que, a literatura recente sugere que os próprios modelos de monetização e engajamento das plataformas podem incentivar a produção de conteúdos emocionalmente intensos, controversos ou divisivos, criando con-

dições estruturais para o aumento da toxicidade [19]. Evidências indicam que níveis mais elevados de toxicidade em interações, especialmente em comentários, estão associados a maiores taxas de engajamento, como volume de comentários e interação do público, ainda que nem sempre resultem em maior monetização direta [14]. Ademais, a literatura aponta que a toxicidade está frequentemente associada a estados emocionais de alta ativação, como raiva, indignação e ansiedade, os quais tendem a aumentar significativamente a probabilidade de compartilhamento e difusão de conteúdo. Nesse sentido, conteúdos tóxicos não apenas emergem de dinâmicas sociopsicológicas, mas também apresentam maior potencial de propagação nas redes digitais, contribuindo para sua amplificação em larga escala [15, 16].

A terceira dimensão corresponde à operacionalização computacional da toxicidade por meio de modelos automatizados de detecção. Ferramentas como a Perspective API[3] atribuem escores probabilísticos a conteúdos textuais, viabilizando análises em larga escala [12, 13]. Nesse contexto, a toxicidade passa a ser tratada como uma variável mensurável e passível de integração em sistemas analíticos. A Figura 2.1 sintetiza essa dinâmica, articulando fatores psicológicos, estruturas sociotécnicas e mecanismos algorítmicos.



Figura 2.1: Esquema conceitual da dinâmica da toxicidade em ambientes digitais.

[3]https://perspectiveapi.com/

No contexto dessa dimensão computacional, a literatura recente apresenta diferentes abordagens para a mensuração da toxicidade, que podem ser organizadas em três categorias principais.

A primeira consiste em modelos baseados em aprendizado de máquina, que utilizam grandes bases de dados para atribuir probabilidades de toxicidade a textos, como no caso do Perspective API [13]. Esses modelos são amplamente empregados em pesquisas e sistemas de moderação automatizada.

A segunda abordagem envolve modelos *open-source* de classificação textual, como o Detoxify[8], que permitem maior controle metodológico, transparência e adaptação a diferentes contextos de análise [12].

A terceira abordagem corresponde a métodos baseados em regras, como filtros lexicais, *blocklists* e expressões regulares, amplamente utilizados na prática de moderação por criadores e plataformas [40]. Embora mais simples, esses métodos são úteis para reduzir a carga operacional, mas apresentam limitações significativas diante da complexidade e da natureza contextual do fenômeno.

Essas diferentes estratégias refletem tentativas de traduzir um fenômeno social complexo em métricas computacionais, evidenciando a tensão entre escala analítica e precisão interpretativa. Não obstante isso, tais métricas têm sido amplamente empregadas em estudos empíricos de larga escala para identificar padrões de toxicidade em eventos sociais e políticos, demonstrando como variações contextuais influenciam diretamente a intensidade e a distribuição da toxicidade nas interações online [37].

Nesse contexto, a utilização de modelos automatizados como o Perspective API deve ser compreendida não como solução definitiva, mas como um instrumento analítico sujeito a limitações que precisam ser explicitamente consideradas no desenho de sistemas de análise.

Apesar dos avanços metodológicos, a literatura identifica uma série de limitações na mensuração da toxicidade. Um dos principais problemas refere-se à dependência de *proxies* textuais simplificados, que podem não capturar nuances contextuais, ironia ou formas implícitas de hostilidade [12, 13]. Somado a isso, há evidências de que modelos automatizados podem incorporar atributos latentes, como informalidade ou estereótipos, confun-

[6]Introjeção solipsista refere-se à tendência de interpretar interações online como extensões do próprio pensamento interno, atribuindo voz e personalidade imaginada aos interlocutores digitais.

[7]Imaginação dissociativa refere-se à percepção do ambiente online como um espaço separado da realidade cotidiana, favorecendo comportamentos e expressões que dificilmente ocorreriam em interações presenciais.

[8]https://github.com/unitaryai/detoxify

dindo características linguísticas com toxicidade propriamente dita [13].

Outro desafio relevante é a baixa transferibilidade entre contextos, uma vez que modelos treinados em determinados domínios podem apresentar desempenho limitado quando aplicados a outros ambientes, culturas ou idiomas [13]. Esse problema torna-se particularmente crítico em plataformas globais, como o YouTube.

No plano aplicado, estudos indicam que ferramentas de moderação frequentemente são rudimentares, difíceis de configurar e insuficientes para lidar com formas sutis ou multimodais de toxicidade, exigindo constante intervenção humana [40]. Vale ressaltar ainda que, observa-se um *trade-off* importante entre a redução da toxicidade e a manutenção do engajamento, o que torna a moderação uma tarefa estratégica, e não apenas técnica [40]. Adicionalmente, evidências sugerem que políticas de moderação mais restritivas não eliminam necessariamente os problemas estruturais do ecossistema digital, uma vez que plataformas como o YouTube operam em um ambiente interconectado com outras redes e práticas sociais mais amplas, nas quais os comportamentos dos usuários se distribuem e se reconfiguram continuamente [41].

Esse *trade-off* revela uma tensão estrutural nos sistemas de plataforma, nos quais mecanismos voltados à maximização do engajamento podem, simultaneamente, incentivar a produção e a amplificação de conteúdos tóxicos. Em contextos baseados na economia da atenção, conteúdos que provocam respostas emocionais intensas tendem a gerar maior interação, sendo favorecidos pelos algoritmos de recomendação. Como resultado, a moderação de conteúdo não se configura apenas como um problema técnico, mas como uma decisão estratégica que envolve a gestão de externalidades negativas associadas ao próprio modelo de negócio das plataformas.

Por fim, a literatura aponta que a toxicidade não pode ser analisada isoladamente, pois está frequentemente associada a fenômenos mais amplos, como discurso de ódio, misoginia e exposição a conteúdos nocivos mediados por algoritmos de recomendação [31, 42]. Nesse sentido, estudos indicam que conteúdos conspiratórios no YouTube podem compor dinâmicas de circulação de conteúdo inter-relacionado, nas quais usuários são progressivamente expostos a narrativas relacionadas, podendo inclusive ser alvo de estratégias de monetização e publicidade direcionada [43].

Em síntese, a toxicidade deve ser compreendida como um fenômeno emergente da interação entre fatores psicológicos, estruturas sociotécnicas e mecanismos algorítmicos, cuja amplificação é mediada por dinâmicas de engajamento e atenção.

2.4 Economia de atenção

A economia da atenção parte do diagnóstico de que, em contextos de abundância informacional, o recurso escasso deixa de ser a informação e passa a ser a atenção humana [44]. Nessa perspectiva, a atenção é tratada como um recurso limitado e economicamente relevante, cuja captura, gestão e retenção se tornam centrais para organizações e plataformas. Essa formulação desloca o foco da economia tradicional para a disputa por visibilidade, engajamento e permanência.

Essa abordagem permite compreender a atenção como uma espécie de "moeda", isto é, um recurso que pode ser capturado, alocado e convertido em valor econômico [44]. Em ambientes digitais, essa lógica se intensifica, uma vez que múltiplos conteúdos competem simultaneamente pelo foco do usuário.

Estudos mais recentes expandem essa formulação ao compreender plataformas digitais como participantes de um mercado competitivo de atenção e atividade. Ribeiro e Faloutsos [45] propõem o conceito de *attention-activity marketplace*, no qual websites e redes sociais disputam continuamente o tempo, o engajamento e a atividade dos usuários. Nessa perspectiva, a atenção não apenas representa um recurso escasso, mas também um mecanismo dinâmico capaz de influenciar diretamente crescimento, popularidade e declínio de plataformas digitais. Os autores demonstram que alterações em mecanismos de interação e captura de atenção podem impactar significativamente a dinâmica competitiva entre plataformas, afetando padrões de permanência, engajamento e migração de usuários.

No contexto de plataformas como o YouTube, a economia da atenção se materializa na relação entre visibilidade, engajamento e receita. A literatura mostra que a monetização depende diretamente da capacidade de captar visualizações e gerar interação, sendo esses indicadores mais determinantes do que métricas estáticas como o número de inscritos [17].

Além dos mecanismos tradicionais de monetização via publicidade, especialmente por meio do *YouTube Partner Program* (YPP), criadores utilizam estratégias alternativas para converter atenção em receita, como links afiliados, doações, venda de produtos e redirecionamento para serviços externos [19, 20]. Essa distinção entre monetização institucional, mediada pela própria plataforma, e monetização alternativa, baseada em infraestruturas externas, tornou-se particularmente relevante após eventos como o chamado *Adpocalypse* (detalhado mais a frente), no qual mudanças nas políticas de moderação e publicidade do YouTube levaram à desmonetização de diversos conteúdos.

Como resposta, muitos criadores passaram a diversificar suas fontes de receita, re-

duzindo sua dependência direta da monetização nativa da plataforma. Esse movimento evidencia que a atenção capturada no YouTube pode ser convertida em valor em múltiplos circuitos econômicos, ampliando as estratégias de sobrevivência e crescimento dos criadores no ecossistema digital.

A centralidade da atenção reorganiza práticas de produção e trabalho digital. Criadores passam a atuar de forma estratégica, buscando maximizar exposição, engajamento e diversificação de receitas [20]. Sob essa perspectiva, a produção de conteúdo se insere em uma lógica de competição contínua por visibilidade, marcada por assimetrias estruturais na distribuição de audiência.

A literatura aponta que essa distribuição é altamente desigual, com forte concentração de visualizações e receitas em um número reduzido de canais [20]. Além disso, processos de monetização e desmonetização influenciam diretamente a visibilidade dos conteúdos, afetando sua promoção e recomendação [46]. A atenção, portanto, não é distribuída de forma neutra, mas mediada por mecanismos algorítmicos de ranqueamento e recomendação.

A transformação da atenção em valor econômico pode gerar incentivos problemáticos. Quando cliques, visualizações e tempo de permanência são recompensados, conteúdos que maximizam a atração — inclusive por meio de promessas enganosas ou estratégias manipulativas — tendem a se beneficiar dessa lógica.

Estudos sobre *spam* no YouTube indicam que conteúdos podem explorar a atenção dos usuários por meio de promessas de ganho ou redirecionamento para links externos, convertendo tempo e cliques em receita mesmo sem entregar valor informacional [47]. Isso evidencia que a economia da atenção não apenas organiza o mercado de conteúdo, mas também pode favorecer práticas oportunistas e formas de exploração da audiência.

A literatura também sugere que a atenção pode ser analisada a partir de diferentes dimensões: sua mensuração, sua base cognitiva, seu controle por tecnologias e sua relação com práticas tradicionais de captura de público, como a publicidade [44]. Essa classificação permite compreender a economia da atenção como um fenômeno que articula dimensões psicológicas, tecnológicas e econômicas de forma integrada.

Em contextos de abundância informacional, a atenção torna-se um recurso escasso, passando a ser disputada e convertida em valor econômico [44]. Em plataformas digitais, essa lógica estrutura a relação entre visibilidade, engajamento, monetização e governança algorítmica, influenciando tanto a circulação de conteúdos quanto as estratégias dos criadores [17, 19, 20, 46, 47].

Essa dinâmica está associada ao processo de plataformização da vida social, no qual plataformas digitais assumem papel central na organização das interações, da circulação de conteúdos e das atividades econômicas. Nesse contexto, algoritmos, sistemas de recomendação, mecanismos de coleta de dados e modelos de monetização passam a influenciar diretamente a visibilidade da informação e os comportamentos observados nesses ambientes.

No caso do YouTube, a dependência dos criadores em relação aos mecanismos de monetização, visibilidade e recomendação da plataforma influencia diretamente decisões sobre formatos, estilos discursivos e estratégias de engajamento. Dessa forma, a economia da atenção não opera apenas como um modelo abstrato, mas como um sistema de incentivos concretos que molda práticas produtivas e pode favorecer conteúdos mais intensos, controversos ou orientados à maximização de engajamento.

2.5 Monetização

A monetização, no contexto dos sistemas digitais, pode ser compreendida como o processo de conversão de recursos simbólicos — como atenção, informação e interação — em valor econômico. Diferentemente dos modelos tradicionais de mercado, nos quais o valor está diretamente associado à produção e troca de bens ou serviços, nos ambientes digitais a monetização emerge de dinâmicas indiretas, mediadas por plataformas e sustentadas pela captura e exploração da atenção dos usuários. Nesse sentido, a monetização está intrinsecamente relacionada à economia da atenção, na qual o tempo e o engajamento dos indivíduos se configuram como ativos centrais na geração de receita [11, 48].

Sob essa perspectiva, plataformas digitais operam como infraestruturas econômicas que organizam e regulam fluxos de interação entre múltiplos atores — usuários, criadores de conteúdo e anunciantes. A monetização, portanto, não ocorre de forma isolada, mas emerge como resultado de um ecossistema sociotécnico no qual algoritmos, métricas de engajamento e estratégias de visibilidade desempenham papel central na distribuição de valor. Esse arranjo implica que o valor econômico não está diretamente no conteúdo em si, mas na sua capacidade de gerar circulação, retenção e interação no interior da plataforma [20].

A literatura também aponta que a monetização em ambientes digitais apresenta caráter indireto e mediado, sendo frequentemente operacionalizada por meio de *proxies* como visualizações, tempo de permanência (*watch time*), cliques e outras métricas comportamentais. Essas métricas funcionam como indicadores da capacidade de um conteúdo de

capturar atenção, sendo posteriormente traduzidas em valor econômico por meio de sistemas publicitários, modelos de recomendação e mecanismos de ranqueamento [18, 49]. Dessa forma, a monetização depende menos do conteúdo em si e mais de sua performance no interior das dinâmicas algorítmicas da plataforma.

Além disso, observa-se que a monetização não se limita a mecanismos institucionais formais, mas inclui práticas informais e estratégias emergentes desenvolvidas pelos próprios atores do ecossistema digital. Nesse contexto, emergem formas alternativas de monetização, como redirecionamento para plataformas externas, marketing de afiliados, doações e venda de produtos ou serviços. Essas estratégias configuram um modelo distribuído de monetização, no qual o valor é capturado tanto dentro quanto fora das plataformas digitais [19, 20].

Um aspecto relevante destacado na literatura é que a monetização também pode estar associada a práticas oportunistas ou maliciosas, que exploram as próprias lógicas do sistema para gerar ganhos econômicos. Estudos indicam que conteúdos podem ser estruturados não necessariamente para informar ou entreter, mas para maximizar cliques, visualizações ou redirecionamentos externos, frequentemente por meio de promessas de ganhos financeiros ou incentivos enganosos [47]. Nesses casos, a monetização ocorre de forma parasitária, explorando as brechas do sistema de recomendação e a vulnerabilidade dos usuários, evidenciando que o processo de geração de valor em plataformas digitais também pode produzir externalidades negativas.

A partir desse conjunto de elementos, é possível compreender a monetização como um fenômeno multifacetado, que envolve: (i) a captura da atenção como recurso econômico; (ii) a mediação algorítmica na distribuição de visibilidade; (iii) a utilização de métricas comportamentais como *proxies* de valor; e (iv) a emergência de estratégias formais e informais de geração de receita. Essa abordagem permite tratar a monetização não apenas como um mecanismo econômico, mas como um processo sociotécnico complexo, resultante da interação entre infraestrutura, comportamento dos usuários e estratégias dos atores envolvidos.

A partir dessa base conceitual mais ampla, torna-se possível avançar para análises mais específicas, que considerem como essas dinâmicas se manifestam em redes sociais digitais e, de forma mais particular, em plataformas como o YouTube, onde a monetização assume formas específicas associadas à produção de conteúdo audiovisual, à recomendação algorítmica e à economia dos criadores.

2.5.1 Monetização em redes sociais

A monetização em redes sociais representa um desdobramento específico desse processo em ambientes digitais, caracterizado pela centralidade da interação social, da produção de conteúdo por usuários e da mediação algorítmica da exposição de conteúdo. Diferentemente de modelos tradicionais de monetização digital, nos quais a receita está majoritariamente associada ao tráfego ou à publicidade direta, nas redes sociais a geração de valor econômico está profundamente vinculada às dinâmicas de engajamento, circulação e amplificação de conteúdo.

Nesse contexto, as plataformas sociais operam como sistemas sociotécnicos que capturam, organizam e exploram a atenção coletiva em larga escala. A monetização emerge, portanto, como resultado da capacidade de determinados conteúdos ou perfis de gerar interação contínua — como visualizações, curtidas, comentários e compartilhamentos — que são posteriormente convertidas em valor econômico por meio de mecanismos como publicidade segmentada, ranqueamento e recomendação algorítmica de conteúdo [11, 18]. Ademais, a lógica econômica dessas plataformas está menos associada à produção de conteúdo em si e mais à sua performance no interior dos sistemas de distribuição e visibilidade.

Um dos elementos centrais desse modelo é a intermediação algorítmica, responsável por definir quais conteúdos serão expostos, para quais públicos e em que intensidade. Essa mediação influencia diretamente a monetização, uma vez que a visibilidade constitui um dos principais determinantes da geração de receita. Assim, conteúdos que conseguem se adaptar às dinâmicas algorítmicas — seja por meio de maior retenção, frequência de publicação ou capacidade de engajamento — tendem a obter maior retorno econômico. Esse processo reforça a ideia de que os modelos de monetização em redes sociais são, em grande medida, uma função da visibilidade algorítmica.

Além desse aspecto, a literatura evidencia que esse processo não se restringe aos mecanismos internos das plataformas. Criadores de conteúdo frequentemente utilizam estratégias externas e complementares, como redirecionamento para outras plataformas, marketing de afiliados, venda de produtos e sistemas de financiamento coletivo. Esse fenômeno caracteriza um modelo de monetização distribuída, no qual as redes sociais atuam como ponto de entrada e amplificação, enquanto a captura efetiva de valor pode ocorrer em múltiplos ambientes [19, 20]. Tal configuração indica que a monetização constitui um processo híbrido, que transcende os limites da própria plataforma. Nesse sentido, a Tabela 2.3 sintetiza as principais formas de monetização identificadas na literatura, organizadas de acordo com o nível de controle (plataforma ou criador) e os diferentes mecanismos de captura de valor.

Outro aspecto relevante diz respeito à profissionalização da produção de conteúdo nas redes sociais. A consolidação da chamada *creator economy* transformou usuários em agentes econômicos que operam de forma estratégica, buscando maximizar sua visibilidade e sua capacidade de conversão em receita. Esse processo é acompanhado por uma intensificação da competição por atenção, na qual poucos atores concentram grande parte da audiência e da monetização, enquanto a maioria enfrenta dificuldades para obter retorno econômico consistente [20]. Dessa forma, a lógica econômica em redes sociais está diretamente associada a dinâmicas estruturais de desigualdade e concentração.

Do ponto de vista metodológico, a mensuração da monetização nesse contexto apresenta desafios significativos. Em muitos casos, não há acesso direto aos dados de receita, o que leva os estudos a utilizarem *proxies* baseadas em métricas de atenção e engajamento. Indicadores como número de visualizações, tempo de exibição, taxa de interação e alcance são frequentemente empregados como aproximações do valor econômico gerado, embora essa relação não seja necessariamente linear nem direta [18, 50]. Isso reforça a natureza indireta e mediada da geração de receita em redes sociais.

Por fim, observa-se que as dinâmicas de monetização podem gerar incentivos que vão além da simples produção de conteúdo de qualidade. A busca pela maximização de visibilidade e engajamento pode levar à adoção de estratégias que priorizam apelo emocional, simplificação excessiva ou até mesmo práticas que incentivam o consumo contínuo, o redirecionamento externo e a amplificação de narrativas específicas, como conteúdos que prometem ganhos financeiros ou direcionam usuários para sistemas externos com fins monetários. Esse tipo de prática evidencia que a monetização não apenas estrutura a economia das redes sociais, mas também influencia diretamente os padrões de produção e circulação de conteúdo.

Dessa forma, a monetização em redes sociais pode ser compreendida como um processo sociotécnico complexo, estruturado pela interação entre infraestrutura tecnológica, mediação algorítmica, comportamento dos usuários e estratégias econômicas dos criadores. Mais do que um mecanismo econômico, trata-se de um sistema baseado em relações de visibilidade e incentivos comportamentais, cujas dinâmicas se materializam de forma particularmente evidente em plataformas como o YouTube.

2.5.1.1 Monetização no YouTube

A monetização no YouTube representa uma das formas mais consolidadas e complexas de monetização em plataformas digitais, caracterizando-se pela integração entre produção de conteúdo audiovisual, mediação algorítmica e múltiplos mecanismos de captura de

valor. Diferentemente de outras redes sociais, o YouTube estruturou, desde cedo, um modelo econômico baseado no compartilhamento de receita com criadores, consolidando um ecossistema no qual a produção de conteúdo está diretamente associada a incentivos financeiros [20, 51].

Sob essa perspectiva, a monetização institucional (*on-platform*) assume papel central, sendo operacionalizada principalmente por meio do YPP, que permite aos criadores receberem parte da receita gerada por anúncios exibidos em seus vídeos. Esse modelo baseia-se em métricas como visualizações, tempo de exibição e perfil da audiência, sendo mediado por sistemas de recomendação cuja lógica interna não é transparente, caracterizando um cenário de opacidade algorítmica [52].

Assim, a capacidade de um vídeo gerar retenção e engajamento torna-se determinante para sua performance econômica, uma vez que influencia diretamente sua distribuição algorítmica e, consequentemente, sua exposição a anúncios [11, 18], ainda que os critérios específicos dessa distribuição permaneçam em grande parte opacos para criadores e pesquisadores [52].

Além da publicidade tradicional, o YouTube incorporou mecanismos adicionais de monetização institucional, como assinaturas de canais, conteúdos exclusivos e formas de interação monetizada, como *Super Chat* e "Valeu Demais". Esses mecanismos reforçam a monetização direta da audiência (*fan funding*), estabelecendo uma relação mais próxima entre criadores e público e reduzindo, ainda que parcialmente, a dependência exclusiva da receita publicitária [20]. Nesse sentido, o YouTube amplia seu modelo econômico ao integrar diferentes fluxos de monetização dentro da própria plataforma, conforme sistematizado na Tabela 2.3.

No entanto, a monetização no YouTube não se limita aos mecanismos internos da plataforma. Criadores frequentemente adotam estratégias *off-platform*, como marketing de afiliados, venda de produtos, patrocínios e redirecionamento para plataformas externas de financiamento coletivo [19, 20, 41]. Esses mecanismos podem complementar a receita obtida via publicidade e, em determinados contextos — especialmente em cenários de instabilidade, mudanças de políticas ou desmonetização —, reduzir a dependência em relação à monetização institucional da plataforma [19, 20]. Isso evidencia que o YouTube funciona simultaneamente como ambiente de monetização e como canal de aquisição e redirecionamento de audiência para outros fluxos econômicos, operando como um intermediário central na circulação de valor no ecossistema digital [19, 20].

Tabela 2.3: Tipos de monetização no YouTube

Tipo	Controle	Mecanismo	Descrição	Exemplos	Ref
Institucional (*on-platform*)	Plataforma	Publicidade (CPM, CPC)	Receita gerada por anúncios exibidos junto ao conteúdo	YouTube Ads, pre-roll	[20]
		Revenue sharing	Compartilhamento de receita entre plataforma e criador	YouTube Partner Program	[20, 51]
		Assinatura / Premium	Acesso pago a conteúdos ou funcionalidades exclusivas	YouTube Premium, memberships	[20, 41]
		Interação monetizada (*fan funding*)	Pagamentos diretos da audiência dentro da plataforma	Super Chat, "Valeu Demais", stickers	[20]
Estratégica (*off-platform*)	Criador	Afiliados / Redirecionamento	Ganhos por direcionamento de usuários a produtos ou serviços externos	Links, apps, e-commerce	[19]
		Doações / Crowdfunding	Contribuições diretas da audiência aos criadores (fora da plataforma)	Patreon, PayPal	[19, 20]
		Venda de produtos/serviços	Monetização por meio de bens físicos ou digitais próprios	Cursos, merch	[19]
		Patrocínio / *branded content*	Parcerias comerciais diretas entre marcas e criadores	Influenciadores	[19]

Outro aspecto relevante da monetização no YouTube diz respeito ao papel da moderação e das políticas da plataforma. A adoção de sistemas automatizados de detecção de conteúdo, baseados em técnicas de *machine learning*, introduz mecanismos de classificação cujo funcionamento apresenta baixa transparência e limitações interpretativas, especialmente em função da dependência de contexto para a avaliação da toxicidade [41, 53]. Além disso, preocupações com *brand safety*[9] e episódios de retirada de anunciantes levaram a mudanças nas práticas de governança da plataforma, evidenciando a vulnerabilidade dos criadores em relação às decisões algorítmicas e institucionais [41]. Nesse cenário, a monetização no YouTube não depende apenas da performance do conteúdo, mas também da capacidade dos criadores de se adaptar às regras, restrições e incertezas do ambiente regulado pela plataforma.

Não obstante isso, a literatura aponta que a monetização no YouTube pode assumir formas desviantes, nas quais atores exploram as dinâmicas da plataforma para obter ganhos econômicos de maneira oportunista. Isso inclui conteúdos que prometem ganhos financeiros fáceis, redirecionamento de usuários para aplicativos ou serviços externos e incentivam interações artificiais — como comentários ou compartilhamento de dados pessoais — com o objetivo de aumentar visualizações e cliques [47]. Tais práticas evidenciam que

[9]Conjunto de práticas e mecanismos utilizados por plataformas e anunciantes para evitar que anúncios sejam associados a conteúdos considerados inadequados, ofensivos, extremistas ou potencialmente prejudiciais à reputação das marcas.

a monetização baseada em métricas de engajamento pode gerar distorções no ecossistema, favorecendo conteúdos enganosos ou de baixa qualidade e comprometendo a qualidade da experiência dos usuários [47].

Por fim, observa-se que a monetização no YouTube está inserida em um ambiente altamente competitivo e estruturalmente desigual, no qual uma pequena parcela de criadores concentra a maior parte da audiência e da receita, enquanto a maioria enfrenta dificuldades para alcançar sustentabilidade econômica [20]. Essa dinâmica evidencia a existência de uma distribuição altamente assimétrica de visibilidade e renda, marcada por incerteza e precariedade, na qual apenas uma fração dos participantes consegue obter sucesso econômico consistente [20]. Nesse contexto, a monetização em plataformas digitais deve ser compreendida não apenas como um mecanismo técnico, mas como um fenômeno condicionado por dinâmicas de visibilidade, competição e relações assimétricas entre plataforma e criadores [20].

Igualmente, a monetização no YouTube pode ser compreendida como um sistema híbrido, que combina mecanismos institucionais, estratégias externas e práticas emergentes, todos mediados por algoritmos e métricas de atenção, cuja opacidade dificulta a compreensão dos mecanismos que estruturam a visibilidade, os incentivos econômicos e a circulação de conteúdo na plataforma [3, 52]. Essa complexidade torna o YouTube um objeto particularmente relevante para o estudo das relações entre economia, tecnologia e produção de conteúdo digital, configurando-se como um caso emblemático da monetização em plataformas digitais contemporâneas.

2.6 *Black-box auditing*

A crescente centralidade de mecanismos algorítmicos na mediação da informação digital tem intensificado o problema da opacidade, especialmente em plataformas baseadas em recomendação, como o YouTube. Nesses ambientes, decisões sobre visibilidade, associação e distribuição de conteúdo são realizadas por algoritmos proprietários, cuja lógica interna não é acessível a pesquisadores externos, configurando o que a literatura denomina sistemas de "caixa-preta" [6, 70]. Esse problema de opacidade não é apenas técnico, mas estrutural, na medida em que tais sistemas tornam-se progressivamente menos acessíveis ao escrutínio público, mesmo quando exercem funções centrais na mediação informacional [70].

No contexto desta tese, o *SocialDash* é posicionado não apenas como uma ferramenta de coleta de dados, mas como uma infraestrutura de suporte à auditoria de sistemas algo-

rítmicos. Em particular, o artefato permite que o pesquisador observe e analise padrões de recomendação, visibilidade e engajamento a partir de dados públicos e métricas derivadas. Dessa forma, o sistema viabiliza a inferência de políticas implícitas de moderação e monetização permitindo que o pesquisador observe e analise padrões de recomendação, visibilidade e engajamento a partir de dados públicos e métricas derivadas. Dessa forma, o sistema viabiliza a inferência de possíveis políticas implícitas de moderação e monetização a partir de sinais observáveis como metadados, interações e indicadores de toxicidade, contribuindo para reduzir a opacidade estrutural das plataformas.

A necessidade de investigar sistemas algorítmicos opacos contribuiu para o desenvolvimento do campo da responsabilização algorítmica (*algorithmic accountability*) [6], dedicado à análise dos impactos sociais e das estruturas de poder associadas aos algoritmos. Nesse contexto, auditorias de caixa-preta (*black-box auditing*) e técnicas de engenharia reversa tornaram-se estratégias relevantes para investigar relações entre entradas e saídas dos sistemas, mesmo sem acesso ao funcionamento interno dos modelos [6, 70].

Nesse cenário, a personalização desempenha papel central: evidências indicam que o histórico de consumo influencia diretamente as recomendações recebidas, ampliando a exposição a conteúdos similares e reforçando padrões preexistentes de comportamento informacional [24]. Como consequência, fenômenos como *filter bubbles* devem ser compreendidos não como estados binários, mas como processos graduais e cumulativos, cuja intensidade pode ser observada ao longo do tempo por meio da composição das recomendações [23, 24]. Adicionalmente, estudos sugerem que sistemas de recomendação podem atuar como vetores de radicalização, conduzindo usuários a conteúdos progressivamente mais extremos ao longo de trajetórias de navegação [42], enquanto conteúdos carregados de emoção moral tendem a apresentar maior capacidade de difusão, contribuindo para a amplificação de narrativas polarizadas [15].

Historicamente, o estudo de algoritmos em plataformas digitais evoluiu de análises descritivas para abordagens empíricas mais robustas, voltadas à identificação de vieses e padrões de comportamento. Trabalhos iniciais concentraram-se na análise de recomendações e rankings, inferindo estruturas a partir da observação de dados disponíveis [4, 52]. Com o avanço da pesquisa, passaram a ser empregadas metodologias mais sofisticadas, incluindo auditorias sistemáticas baseadas em coleta controlada de dados e simulação de comportamento de usuários. Estudos sobre recomendação no YouTube demonstram a viabilidade de identificar padrões de viés, polarização e formação de bolhas informacionais mesmo na ausência de acesso direto ao algoritmo [8, 23].

O princípio central das auditorias de caixa-preta consiste em inferir propriedades do

sistema a partir da análise de suas saídas. Tal estratégia permite investigar fenômenos como viés ideológico em recomendações, formação de *filter bubbles*, propagação de desinformação e dinâmicas de radicalização. Por exemplo, auditorias baseadas em agentes simulados (*sock puppets*) possibilitam isolar variáveis específicas, como histórico de consumo, e observar como essas variáveis afetam as recomendações recebidas [23, 24]. Estudos mais recentes indicam que perfis simulados treinados ocupam uma posição metodológica intermediária entre agentes sem histórico e usuários reais, pois acionam efeitos de personalização ao mesmo tempo em que preservam controle experimental sobre a exposição observada [8].

2.6.1 Estratégias metodológicas

A análise dos artigos selecionados permite identificar diferentes estratégias empíricas no âmbito das auditorias de caixa-preta, sem recorrer a taxonomias externas à literatura.

1. Auditoria por agentes simulados: Essa abordagem consiste na criação de perfis artificiais com comportamentos controlados, com o objetivo de observar como o sistema responde a diferentes entradas (*inputs*). Um aspecto metodológico relevante refere-se à granularidade da análise: estudos recentes sugerem que, no caso do YouTube, avaliações realizadas no nível do vídeo — e não apenas do canal — reduzem distorções interpretativas, uma vez que canais podem conter conteúdos heterogêneos [8]. Essa abordagem permite isolar com maior precisão os efeitos do algoritmo sobre as recomendações.

Exemplos:

- Auditoria de desinformação e *filter bubbles* [23, 24]
- Análise de viés ideológico [8]

2. Auditoria por trajetórias de consumo: Esse tipo de auditoria baseia-se na análise de sequências de navegação entre conteúdos, permitindo identificar padrões de progressão e associação entre diferentes tipos de informação. Essa abordagem possibilita observar como usuários transitam de conteúdos mais moderados para conteúdos progressivamente mais extremos ao longo do tempo, caracterizando trajetórias de radicalização [33]. Além disso, permite investigar a reversibilidade dessas trajetórias, indicando que mudanças no comportamento de consumo — como a exposição a conteúdos de refutação (*debunking*) — podem reduzir a intensidade de bolhas informacionais [23].

Exemplo:

- Caminhos de radicalização no YouTube [33]

3. Auditoria estrutural baseada em redes: Ese método utiliza análise de grafos para mapear relações entre conteúdos, permitindo identificar padrões de conexão, formação de clusters e comunidades. Essa abordagem possibilita compreender a estrutura do ecossistema informacional, evidenciando níveis de segregação, modularidade e concentração de conteúdos, frequentemente associados à formação de bolhas informacionais e dinâmicas de polarização.

Exemplos:

- Redes de vídeos e clusters de conteúdo [4, 52]

4. Auditoria baseada em outputs (recomendação e ranking): Ea técnica foca na análise direta das recomendações e resultados apresentados pelo sistema, sem necessidade de simulação explícita de comportamento. Essa abordagem permite avaliar padrões de exposição, diversidade de conteúdo e possíveis vieses nas recomendações, sendo particularmente útil para identificar a presença de *filter bubbles* e efeitos de personalização.

Exemplos:

- Avaliação de recomendações e bolhas informacionais [23, 24]

5. Auditoria orientada a fenômenos específicos

Nesse caso, a auditoria é direcionada à investigação de fenômenos particulares, como desinformação, radicalização, monetização e moderação [3, 21, 23, 33, 34, 50]. Essa perspectiva evidencia que a auditoria de caixa-preta não se limita a uma técnica específica, mas constitui um instrumento analítico flexível, capaz de ser aplicado a diferentes problemas relacionados aos impactos sociais dos sistemas algorítmicos.

Apesar de sua relevância, a auditoria de caixa-preta apresenta limitações importantes. Primeiramente, a inferência é indireta: não é possível observar diretamente os mecanismos internos do sistema, o que restringe a capacidade de estabelecer relações causais robustas [30]. Além disso, sistemas de recomendação são dinâmicos e adaptativos, implicando que as recomendações evoluem em função do histórico acumulado de interação. Essa característica gera efeitos cumulativos que dificultam a replicabilidade experimental e a comparabilidade entre estudos [24].

Para além dessas limitações, outro desafio central consiste na interpretação dos padrões observados em ambientes digitais, uma vez que o conteúdo acessado pelos usuários

resulta da interação entre compartilhamentos em redes sociais e mecanismos das plataformas, dificultando a identificação precisa das causas subjacentes [21].

Esse problema é agravado pelo fato de que o ambiente online favorece comportamentos de desinibição, nos quais indivíduos tendem a expressar opiniões e emoções de forma mais intensa do que em interações presenciais [36]. Adicionalmente, restrições técnicas e legais — incluindo limitações de APIs e restrições associadas aos termos de uso das plataformas — impõem barreiras à coleta de dados, afetando a escala, a replicabilidade e até mesmo a transparência dos estudos [49].

A análise dos artigos revela lacunas importantes no campo, incluindo a ausência de padronização metodológica, limitações na inferência causal, fragmentação temática e baixa articulação com teorias sociais mais amplas.

A auditoria de caixa-preta consolida-se como uma abordagem fundamental para o estudo de sistemas algorítmicos opacos, permitindo investigar empiricamente seus efeitos mesmo na ausência de transparência. Consequentemente, seus resultados devem ser interpretados como inferências condicionadas e contextuais, e não como descrições determinísticas do funcionamento interno dos algoritmos.

Embora seu foco seja metodológico, essa abordagem fornece evidências empíricas sistemáticas sobre os efeitos observáveis dos sistemas algorítmicos para a compreensão de fenômenos associados à economia da atenção. Em particular, parte da literatura aponta para um conflito estrutural entre iniciativas de mitigação de vieses algorítmicos e o modelo de negócios das plataformas, uma vez que mecanismos orientados ao engajamento tendem a amplificar conteúdos mais aderentes às preferências e predisposições dos usuários [8]. Esse aspecto torna-se ainda mais relevante ao se considerar que criadores podem recorrer a estratégias de monetização externa para contornar restrições impostas pela própria plataforma, mantendo incentivos econômicos mesmo em contextos de moderação ou desmonetização [19].

2.7 *Sensemaking*

Na literatura, o *sensemaking* é compreendido como um processo de construção de significado em contextos marcados por ambiguidade, complexidade e incerteza [54]. Na tradição associada a Karl Weick, o conceito está vinculado à necessidade de tornar inteligíveis eventos confusos e experiências organizacionais equívocas, sendo recorrentes temas como equivocalidade, promulgação (*enactment*) e construção retrospectiva de sentido.

Nessa perspectiva, a realidade não é tratada como algo inteiramente dado e transparente, mas como algo que precisa ser interpretado, organizado e tornado compreensível pelos atores.

Essa abordagem desloca o foco de uma concepção passiva de recepção de informação para uma concepção ativa de interpretação. O *sensemaking* passa, assim, a designar o processo pelo qual indivíduos e coletivos reduzem ambiguidade, estabilizam entendimentos provisórios e constroem referências que orientam a ação [54].

Além dessa tradição organizacional, a literatura também formula o *sensemaking* em termos mais operacionais. Russell, Stefik, Pirolli e Card [25] definem *sensemaking* como o processo de buscar uma representação e codificar dados nessa representação para responder a questões específicas da tarefa. Essa definição evidencia que a produção de sentido depende da escolha, da construção e da transformação de representações analíticas.

Nesse enquadramento, o sentido não surge de modo imediato. Ele é produzido por movimentos sucessivos de busca de representação, instanciação, ajuste e uso analítico dessas representações. Os autores descrevem esse padrão como um ciclo de aprendizado (*learning loop*) complexo, no qual o *sensemaking* alterna movimentos orientados por esquemas prévios e objetivos analíticos com movimentos de revisão induzidos por lacunas, resíduos e inadequações encontradas nos próprios dados [25]. O processo, portanto, é iterativo, progressivo e dependente de reformulações contínuas.

No contexto desta pesquisa, o problema de *sensemaking* assume uma forma específica: o pesquisador do YouTube não enfrenta apenas escassez de informações, mas excesso de dados heterogêneos, distribuídos entre diferentes níveis (comentários, vídeos, canais, métricas de engajamento e recomendações algorítmicas). Nessa situação, o desafio central deixa de ser o acesso aos dados e passa a ser a sua organização em representações que permitam análise integrada. O *SocialDash* é concebido precisamente como um suporte a esse processo, estruturando os dados em diferentes formas de representação (tabelas, séries temporais, gráficos e widgets analíticos) que permitem reduzir a ambiguidade e apoiar a construção progressiva de interpretações.

Uma contribuição importante dessa literatura é mostrar que o *sensemaking* possui uma estrutura de custos. Diferentes operações analíticas exigem diferentes recursos cognitivos e externos, e as representações são escolhidas, modificadas ou substituídas justamente para reduzir o custo das operações necessárias à tarefa [25]. Com isso, o conceito adquire não apenas densidade epistemológica, mas também uma dimensão instrumental.

O ponto central é que compreender algo não depende apenas da existência de da-

dos, mas também da disponibilidade de formas adequadas de organizá-los. Quanto mais apropriada a representação, menor tende a ser o custo de busca, comparação, inferência, reorganização e uso analítico. O sentido, nesse caso, não é apenas um produto interno da cognição, mas também o resultado de suportes materiais e representacionais que tornam a análise viável [25].

Essa discussão é aprofundada por Baker, Jones e Burkman ao situarem o *sensemaking* no contexto das tarefas exploratórias de dados. Para os autores, *sensemaking* é a capacidade de compreender informação complexa, assimilá-la, criar ordem e desenvolver um modelo mental da situação como precursor da resposta [55]. Essa formulação articula compreensão, organização e modelagem mental, indicando que o processo excede a simples identificação de fatos isolados.

No contexto exploratório, os autores destacam subtarefas como observar pontos de dados, procurar padrões ou *outliers*, fazer inferências, comparar o observado com conhecimento prévio, gerar hipóteses e produzir analogias [55]. Essas atividades têm em comum o fato de que o analista ainda não dispõe de uma explicação pronta do fenômeno, sendo necessário construí-la progressivamente. Por essa razão, o *sensemaking* é tratado como objetivo central da exploração de dados [55].

Uma das contribuições mais sistemáticas dessa literatura consiste em explicitar as condições sob as quais representações visuais favorecem o sensemaking. Segundo Baker, Jones e Burkman [55], tais representações tendem a facilitar a produção de sentido ao permitir que usuários estabeleçam associações, diferenciem padrões, organizem informações e percebam relações quantitativas. Além disso, sua efetividade aumenta quando apresentam propriedades gestálticas adequadas, mantêm consistência com conhecimentos prévios do usuário e favorecem raciocínio analógico.

Essa formulação é relevante porque mostra que o *sensemaking* não depende apenas do conteúdo dos dados, mas também da forma como eles são apresentados e organizados visualmente. As quatro abordagens perceptuais definem a base pela qual o observador percebe similaridades, diferenças, ordens e proporções; as propriedades gestálticas favorecem o agrupamento e a percepção de padrões; a consistência com o conhecimento prévio reduz o custo de tradução e interpretação; e o raciocínio analógico permite compreender domínios menos familiares a partir de domínios mais familiares [55]. O sentido, assim, emerge da articulação entre percepção, memória, comparação e inferência.

A relação entre *sensemaking* e *visual analytics* é estreita. Keim e colaboradores definem *visual analytics* como a combinação de técnicas automatizadas de análise com vi-

sualizações interativas para promover entendimento, raciocínio e tomada de decisão em conjuntos de dados amplos e complexos [56]. Nessa perspectiva, a visualização não constitui um estágio final de apresentação, mas parte constitutiva do processo analítico.

O próprio campo de *visual analytics* é organizado em torno da ideia de um ciclo de produção de conhecimento no qual análise, visualização, percepção, exploração e refinamento se articulam iterativamente [56]. Em vez de tratar o usuário como mero receptor de resultados, essa abordagem o posiciona como parte ativa de um processo de interpretação assistido por recursos computacionais. O *sensemaking*, nesse contexto, constitui o fundamento cognitivo que permite transformar dados em entendimento. A análise automatizada amplia a capacidade de processamento, enquanto a visualização interativa amplia a capacidade de inspeção, comparação, formulação de hipóteses e revisão [56].

Embora o *sensemaking* não substitua a fundamentação substantiva de fenômenos como polarização, monetização e toxicidade, ele oferece uma camada analítica transversal para a sua investigação. Em contextos de opacidade, como aqueles estudados por abordagens de auditorias de caixa-preta, o pesquisador precisa interpretar saídas parciais, *proxies*, padrões e indícios sem acesso completo aos mecanismos internos. O problema deixa de ser apenas técnico e passa a ser também cognitivo e representacional: como organizar sinais dispersos para construir uma interpretação plausível.

Nessa direção, o *sensemaking* contribui para compreender o trabalho analítico envolvido na leitura integrada de métricas, padrões, indícios e evidências produzidos em contextos digitais complexos. Em situações marcadas por múltiplos sinais, relações não evidentes e mecanismos parcialmente opacos, a produção de entendimento depende da organização de informações dispersas em estruturas interpretáveis, capazes de sustentar comparação, exploração, formulação de hipóteses e revisão analítica. O *sensemaking* pode ser compreendido como o processo que organiza e integra diferentes evidências para análise conjunta de fenômenos, sem substituir as teorias específicas que explicam cada um deles.

A literatura examinada sugere algumas limitações e tensões relevantes. Em primeiro lugar, há uma dispersão entre contribuições de matriz mais organizacional, voltadas à interpretação, à ambiguidade e à construção de sentido em contextos sociais, e contribuições mais operacionais, voltadas à representação, ao custo cognitivo, à exploração de dados e à interação analítica [25, 54, 55, 56]. Embora essas vertentes sejam complementares, sua integração ainda não se encontra plenamente consolidada.

Em segundo lugar, o campo apresenta definições fortes e compatíveis, mas ainda

não uma formulação única e suficientemente unificada para contextos digitais complexos. Parte da literatura enfatiza o caráter interpretativo e retrospectivo da construção de sentido [54]; outra enfatiza a busca e transformação de representações [25]; outra detalha os requisitos perceptuais e cognitivos das representações visuais [55]; e a literatura de *visual analytics* insere esse processo em ambientes híbridos de interação entre usuário e análise automatizada [56]. O resultado é um campo conceitualmente rico, mas distribuído em tradições parcialmente distintas.

Em terceiro lugar, a própria agenda de *visual analytics* indica a necessidade de avançar na caracterização de tipos de tarefa, tipos de dados, formas de interação e critérios de avaliação [56]. Isso sugere que a articulação entre *sensemaking* e sistemas analíticos interativos ainda constitui um espaço aberto para sistematizações adicionais, sobretudo quando se trata de fenômenos digitais heterogêneos, dinâmicos e de alta complexidade.

Nessa perspectiva, o *SocialDash* pode ser compreendido como uma instanciação parcial do modelo de *sensemaking* proposto por Pirolli e Card [25]. O artefato integra processos de exploração informacional (*foraging*), por meio da coleta automatizada de dados via API e *crawling*, e processos de organização analítica (*schematizing*), estruturando os dados em representações analíticas como *dashboards* e *widgets* interativos. Essa integração reduz o custo de estruturação das informações (*cost of structure*), diminuindo o esforço necessário para organizar, comparar e interpretar dados provenientes de múltiplas fontes. Como resultado, o sistema não apenas apresenta dados, mas atua como suporte à formulação e revisão de hipóteses relacionadas a fenômenos como polarização, monetização e toxicidade.

Com base nesses trabalhos, o *sensemaking* pode ser compreendido como um processo iterativo de construção de significado diante de dados complexos, ambíguos e incompletos, mediado por representações que organizam a informação, reduzem o custo cognitivo e sustentam operações de comparação, inferência, formulação de hipóteses e revisão interpretativa [55, 54, 25, 56]. Nessa formulação, articulam-se a perspectiva interpretativa e retrospectiva associada à tradição de Weick [54], a perspectiva representacional e iterativa desenvolvida por Russell e colaboradores [25], a abordagem visual-cognitiva sistematizada por Baker, Jones e Burkman [55], e a dimensão interativa e híbrida entre usuário e análise automatizada enfatizada pela literatura de *visual analytics* [56].

2.8 *Visual analytics*

A *visual analytics* emerge como resposta ao crescimento acelerado do volume, da variedade e da dinamicidade dos dados, em um contexto no qual a capacidade de coleta e armazenamento supera a capacidade humana de análise e interpretação. Nesse cenário, o problema central deixa de ser o acesso à informação e passa a ser sua transformação em conhecimento confiável, interpretável e acionável [26, 56].

A consolidação do campo ocorre em meados dos anos 2000, especialmente a partir da agenda de pesquisa sistematizada em *Illuminating the Path* [26]. Esse documento, produzido no âmbito do *National Visualization and Analytics Center* (NVAC), pode ser compreendido não apenas como um livro, mas como uma agenda fundacional do campo, funcionando como uma espécie de manifesto científico que organiza problemas, conceitos e direções de pesquisa em *visual analytics* [26]. Sob essa perspectiva, "P&D" (pesquisa e desenvolvimento) refere-se à articulação entre investigação teórica e criação de sistemas analíticos aplicados.

Esse movimento marca uma inflexão importante: a visualização deixa de ser tratada apenas como ferramenta de apresentação e passa a ser integrada a processos analíticos mais amplos, voltados à exploração de dados complexos, dinâmicos e incertos [26, 56].

Nesse contexto, *visual analytics* é definida como a integração entre técnicas automatizadas de análise e visualizações interativas, com o objetivo de promover entendimento, raciocínio e tomada de decisão [56]. Diferentemente da *information visualization*, cujo foco está na representação visual de dados, a *visual analytics* incorpora explicitamente o processo analítico.

Essa distinção pode ser ilustrada com um exemplo aplicado. Em um cenário como o do *SocialDash*, a *information visualization* corresponderia, por exemplo, à exibição de um gráfico com o volume de comentários por vídeo ou a distribuição de *likes* e *dislikes*. Já a *visual analytics* envolveria um sistema que, além de exibir esses dados, aplica algoritmos para identificar padrões — como grupos de comentários tóxicos ou polarizados — e permite ao usuário explorar esses padrões de forma interativa, ajustando filtros, comparando grupos e investigando relações entre engajamento, monetização e conteúdo.

Do ponto de vista estrutural, a literatura organiza o campo em dimensões como raciocínio analítico, representações visuais e interação, transformações de dados e comunicação dos resultados [26]. Essas dimensões compõem um processo integrado, no qual dados são transformados, analisados, visualizados e reinterpretados de forma iterativa. Esse pro-

cesso é frequentemente descrito como um ciclo analítico, no qual a análise automatizada orienta a exploração visual, que, por sua vez, gera novas hipóteses e direciona novas análises [56]. É nesse ponto que a aproximação com o *sensemaking* se torna mais evidente: a produção de conhecimento não ocorre de forma linear, mas por meio de ciclos de interpretação, comparação e revisão.

Uma diferença conceitual importante em relação à *information visualization* aparece na reformulação de um princípio clássico do campo. Enquanto a visualização tradicional se organiza em torno da lógica "visão geral primeiro, ampliação e filtragem em seguida, detalhes sob demanda" (*overview first, zoom and filter, details on demand*), a *visual analytics* propõe uma abordagem orientada pela análise, resumida pela lógica "analisar primeiro, destacar o que é relevante, ampliar, filtrar e aprofundar a análise, detalhes sob demanda" (*analyze first, show the important, zoom, filter and analyze further, details on demand*) [56].

Essa diferença reflete uma mudança de foco. Na visualização tradicional, o usuário explora os dados de forma predominantemente manual, buscando identificar padrões visualmente. Já em *visual analytics*, técnicas analíticas e computacionais auxiliam previamente na identificação de padrões, anomalias ou regiões potencialmente relevantes do espaço de dados, direcionando a atenção do analista para aspectos considerados mais significativos.

Apesar de seus avanços, a *visual analytics* apresenta limitações teóricas e metodológicas. Entre os principais desafios estão a ausência de taxonomias consolidadas de tarefas e tipos de dados, a dificuldade de avaliação de sistemas analíticos complexos e limitações na compreensão dos processos cognitivos envolvidos na interpretação visual [26, 56]. Além disso, reconhece-se que representações e transformações de dados não são neutras, podendo introduzir vieses que afetam as conclusões analíticas [26].

Em síntese, a *visual analytics* pode ser compreendida como um campo interdisciplinar voltado à integração entre análise automatizada, representação visual e cognição humana, organizado em ciclos iterativos de análise e interpretação [26, 56]. Sua relevância reside na capacidade de apoiar a investigação de fenômenos complexos, especialmente em contextos caracterizados por grande volume de dados, ambiguidade interpretativa e necessidade de inferência — como aqueles associados à polarização, monetização e toxicidade em plataformas digitais.

2.9 Síntese da fundamentação teórica

A convergência dos temas explorados nesta fundamentação revela um desafio central para a análise de plataformas digitais contemporâneas: a opacidade sociotécnica dos sistemas que mediam a produção, a circulação e a monetização de conteúdo. No caso do YouTube, essa opacidade se manifesta na atuação de algoritmos proprietários que estruturam visibilidade, recomendação e engajamento, dificultando a observação direta dos mecanismos que influenciam fenômenos como polarização, toxicidade e desempenho econômico dos criadores.

A literatura analisada indica que tais fenômenos não ocorrem de forma isolada. Estudos sobre polarização e toxicidade mostram que esses processos são frequentemente intensificados por dinâmicas de engajamento e incentivos econômicos associados à economia da atenção. Ao mesmo tempo, pesquisas sobre monetização evidenciam que a visibilidade e o alcance dos conteúdos estão diretamente relacionados à capacidade de capturar e reter atenção, criando um ambiente no qual fatores discursivos, comportamentais e econômicos se inter-relacionam de forma complexa.

Apesar desse reconhecimento, observa-se uma lacuna importante: as abordagens existentes tendem a tratar essas dimensões de forma fragmentada, e as ferramentas disponíveis não permitem analisar, de maneira integrada e em tempo quase real, as relações entre discurso (toxicidade), comportamento (engajamento), exposição (recomendação) e desempenho econômico (monetização). Além disso, a natureza de "caixa-preta" dos sistemas algorítmicos impõe limitações adicionais, restringindo o acesso a variáveis internas e exigindo estratégias indiretas de observação.

3. Trabalhos Relacionados

O presente capítulo apresenta uma síntese dos principais trabalhos relacionados utilizados como fundamentação desta pesquisa. O objetivo consiste em compreender como a literatura tem abordado — frequentemente de forma fragmentada — fenômenos associados ao YouTube, incluindo polarização, toxicidade, monetização, auditoria algorítmica.

3.1 Caracterização da base bibliográfica

A análise da base bibliográfica permitiu identificar que o campo se encontra fragmentado em diferentes eixos analíticos frequentemente investigados de forma isolada. Embora existam avanços relevantes nas áreas de polarização, toxicidade, monetização, auditoria algorítmica e *sensemaking*/*visual analytics*, observa-se baixa integração metodológica entre essas dimensões. A partir da análise de títulos, resumos e palavras-chave dos 70 trabalhos selecionados, foi possível organizar a literatura em cinco categorias principais. A Tabela 3.1 apresenta a distribuição temática identificada, e a planilha com os metadados dos 70 trabalhos analisados foi disponibilizada[10] para reutilização em pesquisas futuras.

Essa distribuição evidencia não apenas concentração significativa de estudos em determinados temas, mas também a relativa ausência de abordagens sociotécnicas integradoras capazes de articular simultaneamente polarização, monetização, toxicidade e observabilidade algorítmica. Tal lacuna constitui uma das principais motivações desta pesquisa e fundamenta a proposta metodológica apresentada nos capítulos seguintes.

[10] `https://bit.ly/4mCgHQT`

Tabela 3.1: Distribuição temática da base bibliográfica por eixo analítico

Eixo/Tema	**Referências**	**Total**
Polarização / *filter bubbles* / radicalização	[8, 9, 10, 11, 12, 14, 15, 21, 22, 23, 24, 28, 29, 30, 31, 33, 34, 35, 39, 42, 50, 57, 58, 59, 60, 61]	26
Toxicidade / moderação / discurso hostil	[12, 13, 14, 15, 16, 21, 27, 28, 31, 36, 38, 39, 40, 50, 57, 62, 63, 64]	18
Monetização / *creator economy*	[1, 3, 5, 14, 17, 19, 20, 21, 32, 35, 41, 43, 44, 46, 47, 51, 62, 65, 66, 67, 68, 69, 71]	22
Auditoria algorítmica / *black-box* / recomendações	[3, 6, 8, 12, 18, 23, 24, 31, 33, 41, 42, 46, 49, 52, 58, 59, 60, 61, 70]	20
Sensemaking / Visual analytics	[25, 26, 35, 48, 54, 55, 56]	7
Outros	[45, 72]	2
Total geral	**72 trabalhos analisados**	**72**

A distribuição geográfica da base bibliográfica mostra predominância de trabalhos vinculados aos Estados Unidos, seguidos por contribuições do Brasil. Também aparecem, com menor frequência, estudos oriundos de diferentes países europeus e asiáticos. Essa concentração sugere que o debate sobre YouTube, recomendação, monetização e moderação é fortemente impulsionado por centros acadêmicos norte-americanos, embora haja contribuição relevante de outras regiões, especialmente em estudos aplicados e análises contextuais. Ressalta-se, contudo, que parte dos registros não apresenta país informado, de modo que a distribuição geográfica deve ser interpretada como indicativa.

Tabela 3.2: Distribuição geográfica da base bibliográfica

País	**Quantidade**
Estados Unidos	35
Brasil	7
Reino Unido	2
França	2
Outros países	9
Não informado	14
Total	**70**

A distribuição temporal da base bibliográfica evidencia a presença de alguns trabalhos clássicos, sobretudo de natureza conceitual, publicados entre as décadas de 1960 e 2000,

ao lado de uma concentração expressiva de estudos recentes. Observa-se crescimento acentuado a partir de 2016, com pico em 2022, o que sugere a consolidação recente do campo e o aumento do interesse por temas como recomendação algorítmica, monetização, toxicidade e polarização em plataformas digitais.

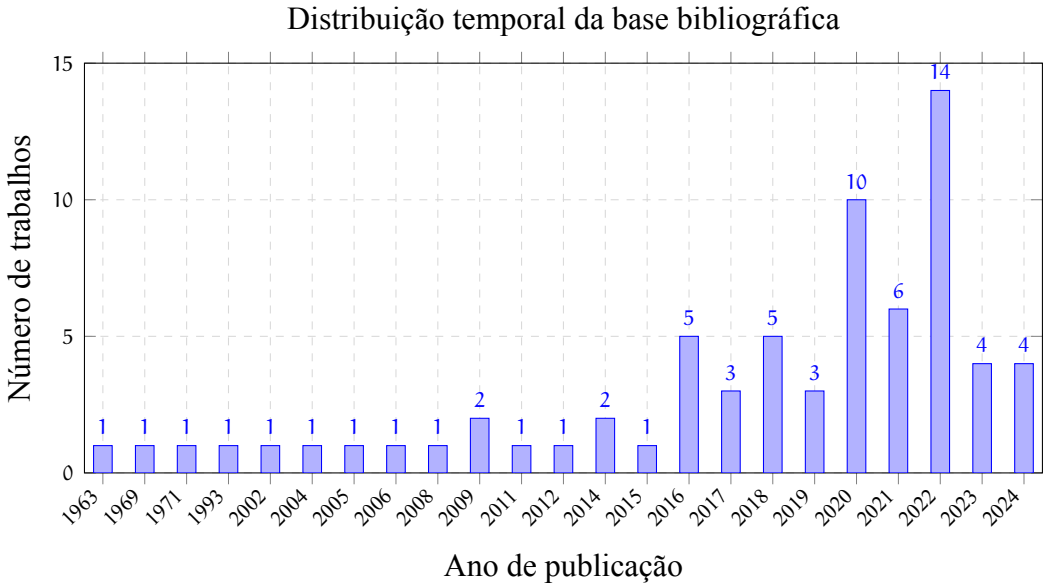


Figura 3.1: Distribuição temporal da base bibliográfica.

O YouTube consolidou-se como uma das plataformas digitais mais influentes da atualidade, tanto em número de usuários quanto em impacto sociopolítico. A combinação entre produção massiva de UGC, mecanismos de recomendação algorítmica e estruturas de monetização confere à plataforma uma dimensão de alcance e influência que ultrapassa o entretenimento, afetando debates públicos e dinâmicas informacionais e políticas [4, 19].

Apesar dessa centralidade, a literatura acadêmica ainda apresenta fragmentação na análise das diferentes dimensões que estruturam o ecossistema da plataforma. Estudos têm investigado aspectos específicos, como recomendação algorítmica, polarização discursiva, toxicidade em comentários ou estratégias de monetização, mas raramente integram essas variáveis de forma conjunta e sistemática.

De maneira geral os trabalhos na academia abrangem questões ligadas ao sistema de recomendação do YouTube, por se tratar de código fechado, muitos trabalhos procuraram analisar o viés de recomendação do algoritmo do YouTube, através de auditorias de conteúdo. Esse viés costuma ser questionado pelo fato de alguns autores classificarem o YouTube como “maquina de dependência de longo prazo” [4, 19], assim, usuários podem ser induzidos a fenômenos conhecidos como “tocas de coelho”, filtros de bolha e câmaras de eco. Esses fenômenos estão geralmente ancorados na ideia de que as pessoas tendem a

confirmar suas crenças pre-existentes reforçando o viés de confirmação. De acordo com nossas pesquisas fenômenos como fake news, desinformação, polarização e conteúdo *borderline* costumam chamar a atenção de pesquisadores e o YouTube acaba sendo apontado como um grande vetor de disseminação de conteúdos digitais [27].

Muitos estudos abordam isoladamente a monetização, a polarização ou a toxicidade em plataformas de vídeo online, mas poucos oferecem um panorama abrangente que conecta essas grandezas. Além disso, muitos estudos referentes a monetização em sua maioria não procuram entender como o engajamento se relaciona com a monetização [67].

Em nossa pesquisa priorizamos trabalhos de cunho quantitativo, recentes, escritos nos idiomas inglês ou português e preferencialmente publicados nos últimos 8 anos. Como dito anteriormente, há uma certa carência na literatura sobre nosso problema de pesquisa e por isso incluímos trabalhos relacionados a outras redes sociais.

Em 2009, o primeiro trabalho reconhecido sobre o YouTube, Joshua Green e Jean Burgess publicaram "YouTube: Online Video and Participatory Culture" [1]. A obra aborda as diversas maneiras como a plataforma é empregada pelas indústrias de mídia, produtores de conteúdo e suas audiências, além de comunidades.

No ano de 2011 o livro "The Filter Bubble: What the Internet Is Hiding from You" [11], Eli Pariser aborda a crescente tendência de personalização dos algoritmos de busca, enfatizando a importância da conscientização sobre a transparência algorítmica em nossas interações online que muitas vezes podem se aproveitar dos dados de histórico e demografia para oferecer mais do mesmo e colocar usuários de maneira não consentida em filtros de bolha reforçando crenças pre-existentes e agindo sob o viés de confirmação.

Leonardo de Marchi em seu livro "Como os algoritmos do YouTube calculam valor? Uma análise da produção de valor para vídeos digitais de música através da lógica social de derivativo" [68], propõe uma abordagem de monetização para vídeos de música no YouTube à luz da teoria da lógica social de derivativos. Ele argumenta em sua hipótese, que o método utilizado para criar valor em vídeos digitais incorpora técnicas originárias do mercado de derivativos.

Uma investigação conduzida por Chu et al. [51] analisou práticas de monetização ilícita no YouTube por meio da construção de dois conjuntos de dados: um composto por discussões extraídas de 11 fóruns online sobre monetização na plataforma e outro formado por informações coletadas em 5 sites dedicados à compra e venda de contas do YouTube. A partir dessa análise, os autores identificaram seis métodos de exploração monetária ilícita, sendo quatro praticados por usuários individuais e dois associados a redes de multica-

nais (MCN). O estudo oferece uma visão detalhada dessas práticas, contribuindo para a compreensão do fenômeno e incentivando o desenvolvimento de pesquisas voltadas ao combate dessas atividades prejudiciais.

O estudo conduzido por Ballard et al. [43] analisa a monetização de criadores de conteúdo de teorias da conspiração no YouTube, destacando a presença de anúncios predatórios ou enganosos nesses canais em comparação com conteúdo mainstream. Foram coletados anúncios de mais de 6000 anunciantes únicos em canais de conspiração e mainstream, revelando que vídeos de conspiração apresentam quase 11 vezes mais anúncios predatórios. Além disso, canais de conspiração são mais propensos a utilizar formas alternativas de monetização, com mais da metade dos canais desmonetizados recorrendo a sites de terceiros para gerar receita. O estudo conclui que teóricos da conspiração possuem diversas oportunidades de monetização, apresentando uma frequência significativamente maior de anúncios predatórios em seus vídeos.

O trabalho realizado por Zappin et al. [46] retrata o aumento no consumo de conteúdo digital e sua consequente maior censura por plataformas como o YouTube, por meio da desmonetização de vídeos, processo conhecido como *Adpocalypse*. Nesse contexto, vídeos menos promovidos ou recomendados poderiam resultar em uma espécie de censura indireta. O artigo aborda quatro técnicas de aprendizado de máquina — C4.5, random forest, regressão linear e máquina de vetores de suporte — para prever a possibilidade de censura (via desmonetização) com base em metadados dos vídeos, alcançando uma precisão de até 87%, o que pode auxiliar criadores de conteúdo a otimizar sua produção para maximizar a receita.

Dunna et al. [3] investigam fatores associados às transições de desmonetização e remonetização de vídeos no YouTube, analisando sua correlação com atributos como taxa de crescimento do canal, palavras-chave e categoria dos conteúdos. As evidências indicam que eventos de desmonetização são relativamente raros, ocorrendo em aproximadamente 0,5% dos casos, enquanto a remonetização, embora mais lenta e burocrática, foi observada em cerca de 25% dos casos, sugerindo uma possível melhoria nos processos de moderação da plataforma. Os resultados também indicam que canais maiores parecem ser favorecidos por taxas diferenciadas de desmonetização e remonetização, levantando a hipótese de que o YouTube possa privilegiar criadores com maior audiência. Além disso, os autores apontam uma forte influência da desmonetização na taxa de crescimento dos canais.

Rieder et al. [4] realizaram uma ampla análise quantitativa do YouTube em escala global, investigando padrões estruturais e comportamentais da plataforma a partir de aproximadamente 56 milhões de canais. Entre os resultados, destacam-se métricas relacionadas

à intensidade de engajamento, taxa proporcional de curtidas e homofilia[11] entre países. O estudo identificou que o Brasil apresentou simultaneamente elevados níveis de intensidade de reação e forte homofilia, indicando uma comunidade nacional altamente ativa e predominantemente orientada ao consumo de conteúdo doméstico. Esses achados reforçam a relevância sociotécnica do YouTube no contexto comunicacional brasileiro, particularmente em discussões relacionadas à economia da atenção, recomendação algorítmica e propagação discursiva.

A Figura 3.2 sintetiza visualmente os principais achados relacionados ao Brasil, destacando sua posição de proeminência em indicadores proporcionais de engajamento e interação doméstica. Em conjunto, esses resultados sugerem um ambiente de elevada participação da audiência, potencialmente favorável à amplificação algorítmica de conteúdos intensos, polarizados ou altamente reativos, aspecto posteriormente explorado por estudos sobre radicalização, bolhas informacionais e auditoria algorítmica [8, 22, 33].

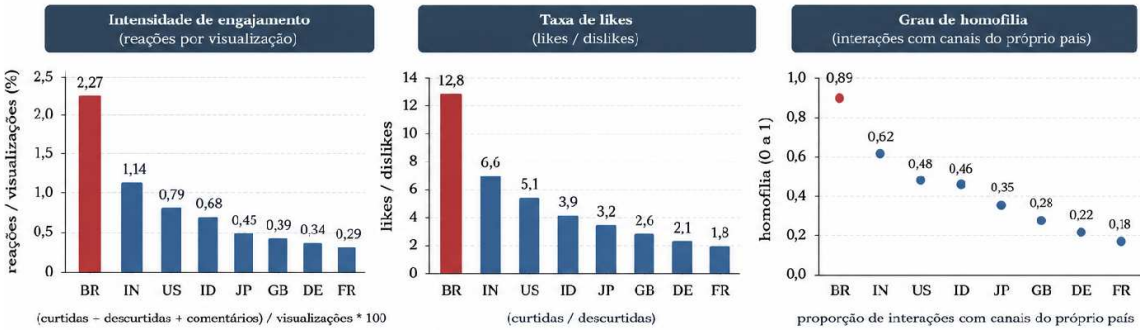


Figura 3.2: Indicadores de engajamento, taxa de curtidas e homofilia associados ao cenário brasileiro no YouTube. Adaptado de Rieder, Coromina e Matamoros-Fernández (2020) [4].

Hua et al. [19] analisam estratégias de monetização alternativa e divulgação de links em descrições de vídeos utilizadas por comunidades e canais considerados problemáticos, como *Alt-right*[12], *Alt-lite*[13] e *manosphere*[14] [33]. O estudo utiliza, entre outras fontes, o conjunto de dados *Youniverse*, composto por aproximadamente 71 milhões de canais com mais de 10 mil inscritos. Os resultados indicam que a monetização externa é utilizada por cerca de 61% dos canais analisados, especialmente em situações nas quais o conteúdo é

[11] Homofilia refere-se à tendência de indivíduos se conectarem e interagirem preferencialmente com outros que possuem características, crenças ou posicionamentos semelhantes.

[12] *Alt-right* refere-se a movimentos digitais associados a posições nacionalistas, extremistas ou ultraconservadoras, frequentemente vinculadas a discursos radicais e conspiratórios.

[13] *Alt-lite* refere-se a grupos e influenciadores digitais com discurso conservador e provocativo, porém geralmente menos explicitamente extremista que a *alt-right*.

[14] *Manosphere* refere-se a comunidades digitais centradas em masculinidade, relações de gênero e antifeminismo, frequentemente associadas a discursos misóginos.

desmonetizado pela plataforma, levando os criadores a recorrerem à publicação de links externos como forma alternativa de geração de receita.

A partir dessa análise, os autores propõem uma taxonomia de monetização alternativa estruturada em quatro categorias principais: marketing de afiliados (*affiliate marketing*), pedidos de doações, uso de criptomoedas — especialmente Bitcoin — e comercialização de produtos vinculados ao próprio canal, como cursos, mercadorias e serviços personalizados. Essas estratégias evidenciam que criadores de conteúdo não dependem exclusivamente do programa oficial de monetização da plataforma, recorrendo frequentemente a fontes externas de receita para sustentar suas atividades e preservar maior autonomia econômica.

Além disso, o estudo apresenta um ranqueamento dos principais canais e atores envolvidos nessas práticas. Os autores também destacam o papel do YouTube como um possível mecanismo de moderação, uma vez que a desmonetização pode atuar como forma de restringir conteúdos associados a determinados léxicos ou discursos considerados problemáticos. Nesse contexto, argumenta-se que criadores pertencentes a esses segmentos recorrem de forma intensiva a estratégias de monetização alternativa.

Pavlopoulos et al. [53] utilizam comentários da Wikipédia para avaliar a influência do contexto na identificação de toxicidade em postagens. Os resultados indicam que a inclusão do contexto da trilha de comentários não melhora significativamente o desempenho de algoritmos classificadores de toxicidade.

Andres, Rossi e Tremblay [62] analisam o fenômeno conhecido como *Adpocalypse*. Sob essa perspectiva, em 2017, diversos anunciantes passaram a recusar a associação de suas marcas a determinados vídeos da plataforma, especialmente aqueles relacionados a conteúdos considerados nocivos ou controversos. Como consequência, grandes empresas retiraram investimentos publicitários do YouTube, provocando uma redução significativa na receita obtida por criadores de conteúdo. Esse cenário levou muitos YouTubers a buscar fontes alternativas de financiamento, como o Patreon e outras formas de monetização externa.

Airoldi, Beraldo e Gandini [52] exploram o funcionamento do algoritmo de recomendação do YouTube no contexto do conteúdo musical por meio da construção de uma rede de associações entre aproximadamente 22 mil vídeos musicais da plataforma. A partir dessa rede, os autores identificaram 50 grupos homogêneos que podem ser interpretados como categorias musicais emergentes formadas a partir do comportamento da audiência. O estudo também analisa características como estilos musicais e elementos textuais pre-

sentes nos títulos dos vídeos, utilizando esses dados para formar *clusters* que representam categorias musicais derivadas dos próprios dados.

Guimarães et al. [27] analisam a toxicidade em comentários publicados em páginas de notícias no Facebook durante um evento político relevante no Brasil: a libertação de Luiz Inácio Lula da Silva em novembro de 2019. Para isso, os autores coletaram posts e comentários de páginas consideradas relevantes e aplicaram a Perspective API para estimar os níveis de toxicidade nas interações. Os dados foram organizados em três grupos — figuras públicas, mídias tradicionais e mídias alternativas. Os resultados indicam que comentários em resposta a outros comentários tendem a apresentar níveis mais elevados de toxicidade, além de revelar que páginas com maior incidência de mensagens tóxicas também concentram discussões igualmente mais tóxicas. O estudo também aponta que eventos políticos de grande repercussão podem provocar aumentos momentâneos nos níveis de toxicidade nas interações online.

Velho, Mendes e Azevedo [37] examinam a popularidade de vídeos de ciência no YouTube ao investigar a influência de nove atributos sobre o desempenho desses conteúdos na plataforma. Entre os fatores relacionados ao conteúdo estão o tema do vídeo, o formato de apresentação e os recursos de edição utilizados. Já entre os fatores não diretamente relacionados ao conteúdo encontram-se a duração do vídeo, a idade do vídeo, a produtividade do canal, o número de curtidas, o número de comentários e o canal ao qual o vídeo pertence. A análise, baseada em canais brasileiros dedicados à divulgação científica, indica que formatos como *vlogs* e animações tendem a apresentar maior popularidade. Os resultados também apontam que o número de curtidas, a produtividade do canal, a antiguidade do vídeo e o formato do conteúdo estão entre os fatores mais determinantes para o desempenho desses vídeos.

Van Kessel, Toor e Smith [65] apresentam um relatório que mapeia canais do YouTube com pelo menos 250 mil inscritos ao final de 2018 e analisa os vídeos publicados na primeira semana de 2019. O estudo identificou aproximadamente 43 mil canais que, nesse período, publicaram quase 250 mil vídeos, totalizando cerca de 48 mil horas de conteúdo. Em média, os vídeos desses canais tinham duração aproximada de 12 minutos e alcançavam cerca de 58 mil visualizações em sua primeira semana na plataforma. Embora o relatório não pertença diretamente à literatura acadêmica, ele foi incluído nesta revisão por apresentar informações estatísticas relevantes que contribuem para a compreensão do funcionamento da plataforma.

No trabalho de Rieder et al. [20], os autores investigam estratégias de monetização no YouTube por meio de um estudo em larga escala sobre práticas de postagem de links nas

descrições de vídeos. A pesquisa analisa como criadores de conteúdo buscam expandir sua presença em outras plataformas digitais, ampliando suas possibilidades de monetização alternativa. Para isso, foram examinados aproximadamente 153 mil canais com mais de 100 mil inscritos, observando-se a presença de URLs nas descrições dos vídeos, cuja frequência aumentou ao longo do tempo. O estudo também considera variações desse fenômeno entre diferentes tamanhos de canais, categorias de conteúdo e localizações geográficas.

Já no trabalho de Arthurs, Drakopoulou e Gandini [41], apresenta-se uma análise descritiva sobre a evolução do YouTube, destacando o papel do Google após a aquisição da plataforma por US$1,65 bilhão em 2006. Segundo os autores, essa aquisição impulsionou significativamente o desenvolvimento do serviço ao fornecer recursos financeiros, infraestrutura tecnológica e estratégias de monetização que favoreceram sua expansão global. O estudo também discute o crescimento exponencial do conteúdo publicado na plataforma, estimando que o número total de vídeos tenha alcançado aproximadamente 4 bilhões até 2016. Além disso, os autores destacam os desafios enfrentados pelos criadores de conteúdo em um ambiente altamente competitivo, no qual uma pequena parcela dos vídeos concentra a maior parte das visualizações, bem como a evolução das categorias de conteúdo que apresentam melhor desempenho.

O estudo de Shevtsov et al. [57] investiga relações entre polarização política e comportamento discursivo em diferentes plataformas digitais durante as eleições norte-americanas de 2020. Os autores combinam análise de sentimento, grafos de interação e detecção de comunidades para comparar padrões observados no Twitter e em comentários do YouTube relacionados a conteúdos políticos. Utilizando algoritmos de detecção de comunidades, como *Louvain*, o trabalho identifica forte homofilia entre usuários com posicionamentos semelhantes, evidenciando a formação de grupos relativamente coesos e segmentados. Os resultados sugerem que padrões de polarização não se restringem a plataformas isoladas, mas tendem a persistir entre diferentes ecossistemas digitais interconectados. Além disso, o artigo reforça a relevância analítica dos comentários do YouTube como espaço de manifestação discursiva e interação sociopolítica, aproximando-se das discussões desta tese sobre polarização afetiva, comportamento coletivo e análise de comunidades online.

No artigo de Postigo [32], o autor analisa comentaristas de conteúdo de videogames no YouTube para discutir as chamadas arquiteturas de trabalho digital associadas à produção de conteúdo online. A principal lente analítica utilizada é a das *affordances*, empregada para compreender como características tecnológicas incorporadas ao design da plataforma criam um conjunto específico de práticas para os usuários, ao mesmo tempo em que atendem aos interesses comerciais do próprio YouTube. O estudo também sugere que esse tipo

de análise pode ser aplicado a outras plataformas digitais baseadas em conteúdo gerado por usuários.

O trabalho de Cakmak et al. [58] analisa vieses no algoritmo de recomendação do YouTube a partir de diferentes narrativas sociopolíticas, utilizando métricas de *emotion drift* (mudança gradual do perfil emocional dos conteúdos ao longo das recomendações), *topic drift* (afastamento progressivo do tema original durante cadeias de recomendação), análise moral e análise de redes ao longo das profundidades de recomendação da plataforma, os autores identificam evidências de *filter bubbles*, *echo chambers* e desvios progressivos de conteúdo ao longo das recomendações, observando que conteúdos mais polarizadores apresentam maior diversidade emocional e maior afastamento temático em relação aos vídeos iniciais. O estudo também mostra que vídeos altamente influentes atuam como elementos de atração nas cadeias de recomendação, contribuindo para deslocamentos discursivos ao longo da navegação.

Han [17] desenvolve um modelo analítico para investigar como diferentes métricas de desempenho — como número de visualizações, inscrições no canal, taxa de comentários após a visualização, taxa de expressão de atitude em relação ao vídeo e idade do canal — podem influenciar as receitas anuais de um YouTuber. O estudo analisa uma amostra de 57 canais e demonstra que indicadores como número de visualizações, taxa de comentários após a visualização e taxa de expressão de atitude apresentam efeitos positivos significativos sobre as receitas anuais. Por outro lado, o número de inscrições no canal não apresentou impacto estatisticamente significativo nesse indicador. Além disso, os resultados sugerem que as receitas anuais tendem a diminuir à medida que o canal envelhece.

O estudo de Buntain et al. [59] investiga os efeitos das mudanças implementadas pelo YouTube em 2019 para reduzir recomendações de conteúdos considerados potencialmente nocivos, conspiratórios ou extremistas, analisando impactos não apenas dentro da plataforma, mas também no compartilhamento desses vídeos em redes externas como Twitter e Reddit. Os autores utilizam modelos de *Interrupted Time Series* (ITS)[15] para comparar padrões de compartilhamento antes e depois da política de *de-recommendation*, identificando reduções significativas na circulação de vídeos conspiratórios e de canais associados à *Alternative Influence Network* (AIN)[16]. Os resultados sugerem que intervenções algorítmicas no sistema de recomendação do YouTube podem produzir efeitos mensuráveis em ecossistemas informacionais externos, reduzindo parcialmente a propagação de con-

[15] *Interrupted Time Series* é uma técnica estatística utilizada para analisar efeitos de intervenções ao longo do tempo.

[16] *Alternative Influence Network* refere-se a uma rede de influenciadores digitais associados a conteúdos políticos polarizados e narrativas controversas.

teúdos considerados problemáticos. O trabalho também discute limitações relacionadas à opacidade algorítmica e ausência de transparência da plataforma, reforçando o aspecto da opacidade algorítmica.

Kim et al. [66] analisam o impacto das interações na seção de comentários sobre o engajamento dos usuários no YouTube. Para isso, os autores coletaram aproximadamente 87 mil comentários provenientes de 647 vídeos do segmento de beleza coreano. Utilizando técnicas de análise de redes sociais e regressão hierárquica, o estudo demonstra que interações equilibradas entre usuários, bem como respostas dos próprios criadores de conteúdo, aumentam significativamente o engajamento da audiência, especialmente em contextos com menor número de comentaristas. Os resultados também sugerem implicações para o design da seção de comentários da plataforma e para estratégias de interação adotadas por criadores de conteúdo.

Wang et al. [67] apresentam um modelo empírico para analisar os fatores que influenciam o sucesso da monetização de vídeos produzidos por YouTubers, buscando preencher uma lacuna na literatura sobre o impacto de métricas como visualizações e inscrições nos ganhos dos canais. A partir da análise de dados de 116 dos canais mais visualizados da plataforma, os autores identificaram que o número de novas visualizações diárias, o crescimento diário das visualizações e o total acumulado de visualizações apresentam correlação positiva com a receita diária dos criadores de conteúdo. Em contrapartida, o número de inscrições no canal não demonstrou impacto direto sobre a receita. O estudo contribui para a compreensão dos fatores associados ao desempenho econômico de criadores no YouTube, além de oferecer insights relevantes para estratégias de monetização em plataformas de compartilhamento de vídeo.

Por fim, o estudo de Menezes et al. [61] investiga como o sistema de recomendações do YouTube pode produzir dinâmicas de confinamento informacional semelhantes a *filter bubbles*, mesmo em cenários não personalizados. Modelando o YouTube como um grafo de recomendações, os autores demonstram que os conjuntos de vídeos sugeridos apresentam elevada estabilidade temporal e forte tendência de reforço temático, especialmente em conteúdos de música e política. Além disso, observam que vídeos mais populares tendem a gerar grafos menores, mais densos e semanticamente mais confinados, sugerindo que a plataforma amplia a navegação dentro de espaços temáticos relativamente homogêneos em vez de promover diversidade ampla de conteúdo. Os resultados indicam ainda que os grafos mais confinados tendem a se organizar justamente em torno de vídeos com maior audiência e potencial de retenção, reforçando mecanismos de concentração algorítmica da atenção.

4. Metodologia de Pesquisa

Este capítulo apresenta a metodologia de pesquisa adotada nesta tese, estruturada segundo o paradigma DSR, que caracteriza-se pela construção e avaliação de artefatos voltados à resolução de problemas relevantes, buscando simultaneamente utilidade prática e contribuição científica. No contexto desta pesquisa, a abordagem é utilizada para fundamentar o desenvolvimento do *SocialDash* e orientar a investigação empírica das relações entre polarização, monetização e toxicidade em plataformas algorítmicas de vídeo.

4.1 Design Science Research: problema, artefato e conjecturas

Esta pesquisa adota o paradigma DSR (*Design Science Research*), voltado ao desenvolvimento e avaliação de artefatos capazes de contribuir simultaneamente para problemas práticos e produção de conhecimento científico. Neste trabalho, a DSR é utilizada para investigar, de forma integrada, relações entre polarização, monetização e toxicidade em plataformas algorítmicas de vídeo.

A literatura sobre plataformas digitais frequentemente aborda essas dimensões de maneira isolada, apesar da forte relação entre recomendação algorítmica, engajamento, concentração de audiência e incentivos econômicos [4, 7, 8]. Ainda assim, permanecem limitações na compreensão integrada desses fenômenos em ambientes mediados por algoritmos.

Nesse contexto, esta pesquisa propõe o desenvolvimento do *SocialDash*, um artefato computacional voltado à coleta, integração e análise de dados relacionados à polarização, monetização e toxicidade. Mais do que um *dashboard*, o artefato busca oferecer um ambiente metodológico capaz de apoiar investigações empíricas integradas sobre essas dimensões no YouTube.

A Figura 4.1 sintetiza a estrutura epistemológica desta pesquisa em DSR [72]. O mo-

delo articula problema em contexto, artefato, conjecturas e avaliação empírica, evidenciando o papel do *SocialDash* como mecanismo de integração entre teoria e investigação aplicada.

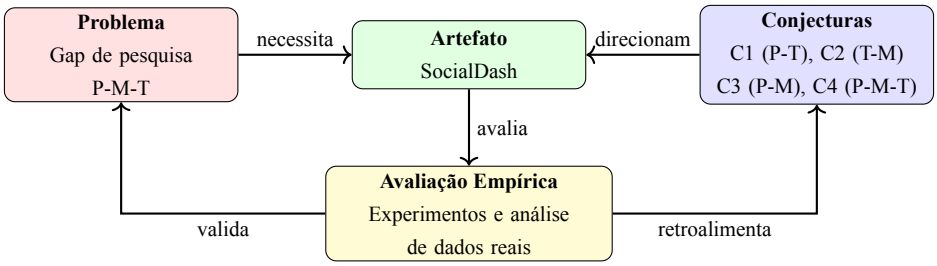


Figura 4.1: Estrutura epistemológica da pesquisa em DSR aplicada ao artefato *SocialDash*, adaptada de [72].

A partir dessa estrutura, as conjecturas desta pesquisa emergem da interseção entre três dimensões centrais discutidas na literatura sobre plataformas digitais: polarização (P), monetização (M) e toxicidade (T). O *SocialDash* foi concebido justamente para operacionalizar empiricamente essas relações de maneira integrada, permitindo investigar simultaneamente padrões discursivos, engajamento da audiência e incentivos econômicos no ecossistema do YouTube.

Nesse contexto, a pesquisa organiza-se em quatro conjecturas centrais:

- **C1 — Polarização e Toxicidade**: conteúdos associados a maior intensidade polarizadora podem apresentar níveis mais elevados de hostilidade discursiva e toxicidade nos comentários.
- **C2 — Polarização e Monetização**: conteúdos polarizados podem atrair maior engajamento, alcance e potencial de monetização dentro da economia da atenção.
- **C3 — Monetização e Toxicidade**: ambientes altamente engajados e potencialmente mais monetizáveis podem estar associados a padrões discursivos mais intensos ou hostis.
- **C4 — Polarização, Monetização e Toxicidade**: a interação simultânea entre essas três dimensões pode revelar dinâmicas sociotécnicas amplificadas por mecanismos algorítmicos de recomendação e engajamento.

As conjecturas são exploradas de forma progressiva, partindo de relações bivariadas até a interseção simultânea entre polarização, monetização e toxicidade. Maiores detalhes

operacionais, métricas utilizadas e resultados preliminares encontram-se descritos nas seções subsequentes.

4.2 Coleta, análise e preparação dos dados

A definição dos dados a serem coletados no contexto desta pesquisa não é arbitrária, mas deriva diretamente das exigências analíticas impostas pela literatura sobre plataformas digitais, recomendação algorítmica e comportamento online. Em particular, a literatura indica que fenômenos como polarização e desinformação não são diretamente observáveis em plataformas como o YouTube, sendo frequentemente inferidos a partir de sinais indiretos, como padrões de engajamento, comentários, metadados, interações discursivas e trajetórias de recomendação [22].

4.2.1 Ontologia, tipologia e taxonomia dos dados

A modelagem dos dados do artefato constitui uma etapa central desta pesquisa, pois estabelece a estrutura conceitual utilizada para transformar fenômenos complexos do ecossistema do YouTube em objetos analíticos observáveis. No contexto da DSR, essa organização é particularmente relevante, uma vez que o artefato depende da integração entre diferentes tipos de entidades, fontes de dados, níveis de observabilidade e métodos de inferência.

Embora *query*, vídeo, canal e comentário tenham sido definidos como entidades analíticas centrais do *SocialDash*, essas estruturas não operam de forma isolada. Existe um encadeamento natural entre elas, refletindo a própria organização sociotécnica do YouTube. Em geral, uma *query* (busca) conduz à recuperação de vídeos e canais relacionados ao tema investigado. A partir dessa coleta inicial, novas entidades associadas tornam-se acessíveis de forma encadeada.

Nesse contexto, vídeos encontram-se vinculados a canais, canais agregam múltiplos vídeos e vídeos podem conter zero ou muitos comentários. Esse processo produz uma estrutura parcialmente hierárquica e parcialmente relacional, permitindo expandir progressivamente a coleta de dados e estabelecer conexões entre conteúdo, produtores, audiência e interação discursiva. Tal organização foi incorporada desde as etapas iniciais de modelagem do artefato, orientando tanto a coleta quanto a estruturação analítica dos dados.

Tal estrutura é consistente com estudos que demonstram que vídeos constituem a principal unidade operacional dos sistemas de recomendação algorítmica [23], canais atuam

como agentes centrais de produção e monetização de conteúdo na plataforma [20], e comentários configuram importantes espaços de manifestação discursiva e interação social [39].

Além da organização ontológica das entidades, esta pesquisa propõe uma taxonomia epistemológica dos dados, classificando as informações de acordo com sua forma de obtenção, nível de acessibilidade e dependência de inferência analítica. Tal distinção torna-se particularmente relevante em estudos sobre plataformas digitais, frequentemente caracterizadas por opacidade algorítmica, acesso parcial aos dados e mecanismos internos de funcionamento não diretamente observáveis pelo pesquisador [6].

Sob essa perspectiva, os dados foram classificados em quatro categorias principais:

- **Dados diretos**: informações explicitamente disponibilizadas pela plataforma e acessíveis diretamente via API, HTML ou interface pública. São dados observáveis sem necessidade de inferência adicional, como visualizações, títulos, comentários, descrições e datas de publicação.
- **Dados indiretos**: informações potencialmente acessíveis, porém dependentes de disponibilidade parcial, processamento adicional ou limitações impostas pela plataforma. Nessa categoria enquadram-se, por exemplo, transcrições automáticas, links externos, HTML contextual e elementos removidos ou parcialmente ocultos pela API.
- **Dados inferidos**: variáveis que não existem explicitamente na plataforma e precisam ser estimadas a partir de múltiplos sinais observáveis, modelos analíticos, *proxies* estatísticos ou serviços externos. Variáveis como toxicidade, monetização e polarização pertencem a essa categoria, uma vez que dependem de modelos de inferência, processamento textual, APIs de terceiros ou relações matemáticas derivadas.
- **Dados inacessíveis**: informações estruturalmente indisponíveis ao pesquisador e restritas à infraestrutura interna da plataforma ou ao proprietário do canal. Exemplos incluem receita real do canal, custo por mil (CPM), receita por mil (RPM), métricas internas de distribuição algorítmica e compartilhamentos não expostos publicamente.

A Figura 4.2 sintetiza essa organização conceitual, relacionando entidades, níveis de acesso aos dados e processos de inferência utilizados pelo artefato.

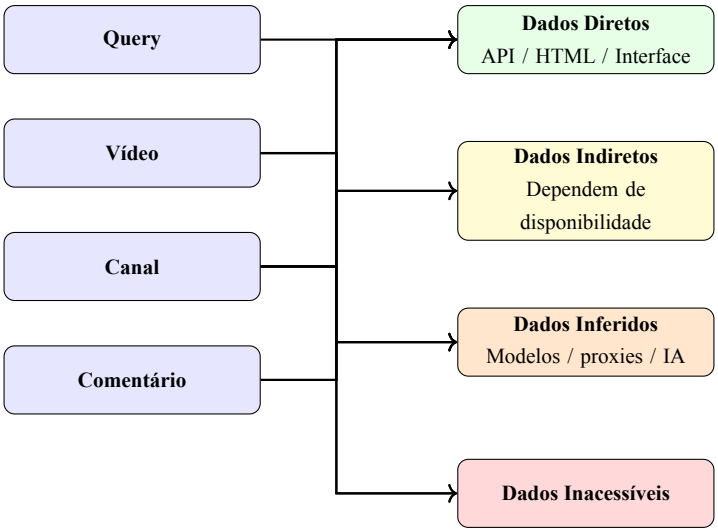


Figura 4.2: Taxonomia conceitual dos dados utilizados pelo SocialDash, relacionando entidades, acessibilidade e níveis de inferência analítica.

A partir dessa estrutura, foi construída a taxonomia relacional apresentada na Tabela 4.1, que organiza os atributos do sistema segundo entidade, origem, natureza epistemológica, método de obtenção e observações analíticas. Essa organização permite distinguir claramente:

- o que é diretamente observável;
- o que depende de processamento adicional;
- o que precisa ser inferido analiticamente;
- e o que permanece inacessível devido à opacidade estrutural da plataforma.

Tal distinção é particularmente importante em pesquisas de auditoria algorítmica e análise de plataformas digitais, nas quais grande parte dos fenômenos investigados emerge da interpretação indireta de sinais observáveis e *proxies* inferenciais.

A Tabela 4.1 apresenta a tipologia completa dos dados utilizados no sistema, organizando-os segundo entidade, origem, natureza e método de obtenção. Essa tabela sintetiza a ontologia e a taxonomia propostas, evidenciando a heterogeneidade dos dados e a presença de múltiplas camadas analíticas no artefato.

Tabela 4.1: Taxonomia da base de dados do SocialDash.

Tipo	**Ent.**	**Orig.**	**Nat.**	**Mét.**	**Observações**
Query	In	User	Direto	Interface	Termo inicial que define o tema e direciona a coleta.
Video ID	V	User/YT	Direto	API	Identificador único do vídeo.
Título	V	YT	Direto	API	Base textual para léxico e classificação.
Descrição	V	YT	Direto	API	Base textual e indicador de monetização *off-platform*.
Tags	V	YT	Direto	API	Palavras-chave que podem classificar um vídeo.
Categoria	V	YT	Direto	API	Categoria informada pelo produtor.
Data publicação	V	YT	Direto	API	Data de upload para análise temporal.
Duração	V	YT	Direto	API	Duração do vídeo.
Transcrição	V	YT	Indireto	API/URL	Obtida via API externa; nem sempre disponível.
Views	V	YT	Direto	API	Proxy de alcance e atenção.
Likes	V	YT	Direto	API	Engajamento positivo do público.
Comentários	V	YT	Direto	API	Proxy de interação social.
Recomendações	V	YT	Indireto	URL	Vídeos relacionados em uma página de vídeo.
Channel ID	Can	User/YT	Direto	API	Identificador único do canal.
Descrição	Can	YT	Direto	API	Pode conter links externos e posicionamento do canal.
Keywords canal	Can	YT	Direto	API	Palavras-chave associadas ao canal.
Categ. canal	Can	YT	Direto	API	Categoria do canal.
Data criação	Can	YT	Direto	API	Base temporal para análise longitudinal.
País de origem	Can	YT	Direto	API	Proxy geopolítico.
Inscritos	Can	YT	Direto	API	Indicador de base de audiência.
Total views canal	Can	YT	Direto	API	Alcance acumulado do canal.
Total vídeos	Can	YT	Direto	API	Indicador de produtividade.
Texto comentário	Com	YT	Direto	API	Base textual para análise de toxicidade.
Autor comentário	Com	YT	Direto	API	Permite análise de comportamento do usuário.
Timestamp	All	YT	Direto	API	Data do evento.
HTML	All	YT	Direto	URL	Registro auditável para rastreabilidade.
Toxicidade	Com	Sys	Inferível	API	Escore contínuo [0,1] derivado de modelo externo.
Sentimento	Com	Sys	Inferível	NLP	Classificação emocional derivada do texto.
Polarização	V/Can	Sys	Inferível	LLM/NLP	Indicador inferido a partir de múltiplos sinais.
Monetização	Can	Sys	Inferível	API/URL	Estimativa pontual e indícios externos.
Receita On	–	3	Inacessível	–	Receita de anúncios em US$/mês.
Receita Off	–	3	Indireto	NLP/URL	Inferida via links externos, afiliados e plataformas paralelas.

In: Entrada; V: Vídeo; Can: Canal; Com: Comentário; All: Todos; User: Usuário; YT: YouTube; Sys: Sistema; 3: Terceiros.

A base de dados do SocialDash foi projetada para preservar a estrutura hierárquica do fenômeno observado no YouTube, representando relacionalmente a conexão entre buscas, canais, vídeos e comentários.

De forma geral, a modelagem da base de dados busca equilibrar três exigências do artefato:

- **fidelidade ao fenômeno observado**, preservando a estrutura hierárquica do You-

Tube;

- **robustez de coleta**, evitando duplicidades e inconsistências;
- **flexibilidade analítica**, permitindo agregar dados brutos, variáveis inferidas e evidências contextuais.

4.3 Modelo

A Figura 4.3 apresenta o modelo conceitual utilizado nesta pesquisa para coleta, processamento e análise de sinais sociotécnicos em plataformas digitais. O objetivo do modelo não é descrever detalhes específicos de implementação, mas representar, de forma abstrata, o fluxo geral de transformação de dados brutos em fenômenos analíticos observáveis.

O modelo foi concebido de forma modular e parcialmente agnóstica em relação à plataforma analisada e às tecnologias utilizadas. Essa abordagem busca reduzir dependências de APIs, linguagens, bancos de dados ou mecanismos específicos de coleta, permitindo maior flexibilidade metodológica diante da constante mudança estrutural das plataformas digitais.

A arquitetura inicia-se a partir de entradas iniciais (*seeds*), como consultas textuais, URLs, IDs, canais ou outros identificadores. Essas entradas alimentam mecanismos de coleta responsáveis pela recuperação de entidades e metadados por meio de APIs, requisições HTTP ou captura de páginas.

Após a coleta inicial, ocorre uma etapa de enriquecimento (*hydrate*), na qual entidades parcialmente observadas podem receber dados adicionais provenientes de serviços externos, processamento textual ou integração entre diferentes fontes. Os dados enriquecidos passam então por processos de limpeza, normalização e organização estrutural.

Na sequência, o modelo realiza a extração de sinais observáveis associados a fenômenos de interesse da pesquisa, como polarização, monetização e toxicidade. Esses sinais alimentam módulos inferenciais responsáveis pela geração de métricas derivadas, classificações e indicadores analíticos.

Por fim, os resultados são convertidos em variáveis inferíveis voltadas à exploração visual e ao *sensemaking*, incluindo dashboards, tabelas, gráficos, filtros e mecanismos interativos de análise. Essa separação entre coleta, processamento e visualização busca favorecer modularidade, rastreabilidade e extensibilidade metodológica.

A principal contribuição do modelo reside na separação conceitual entre plataforma, coleta, processamento e fenômeno analisado. Isso permite que a mesma estrutura metodológica possa ser adaptada para diferentes contextos, reduzindo acoplamentos tecnológicos e favorecendo reutilização em pesquisas futuras.

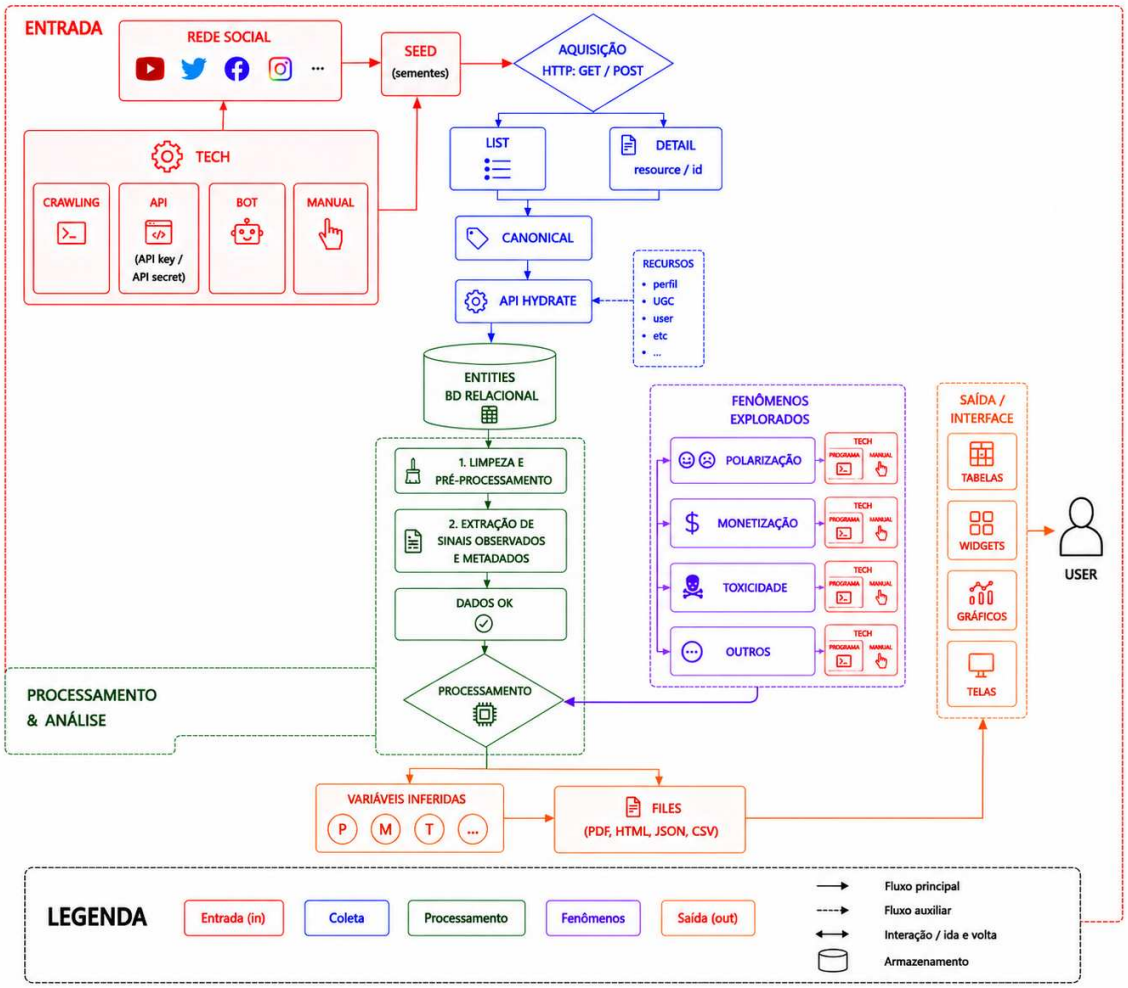


Figura 4.3: Modelo conceitual para coleta, processamento e análise de sinais sociotécnicos em plataformas digitais.

4.4 Justificativa para decisões de pesquisa

Em plataformas digitais baseadas em algoritmos, marcadas por opacidade estrutural e dinâmicas da economia da atenção, variáveis como polarização, toxicidade e monetização não são disponibilizadas diretamente pela plataforma, exigindo o uso de *proxies* computacionais, auditorias algorítmicas e inferências analíticas [5, 6, 8, 23].

A utilização prioritária da YouTube Data API v3 ocorreu devido à elevada instabilidade estrutural do HTML da plataforma. Durante o desenvolvimento do *SocialDash*, observou-

se que o YouTube altera frequentemente elementos de interface, estruturas DOM e mecanismos de renderização, além de empregar carregamento dinâmico e mecanismos *anti-bot*, dificultando abordagens exclusivamente baseadas em *Web crawling*. A Tabela 4.2 sintetiza as principais diferenças entre coleta via API oficial e *Web crawling*.

Dessa forma, optou-se pela API oficial como principal fonte de coleta, em razão de sua maior estabilidade, padronização e previsibilidade semântica dos dados retornados. Ainda assim, reconhece-se que a API disponibiliza apenas parte das informações efetivamente utilizadas pelos mecanismos internos da plataforma.

Mesmo utilizando a YouTube Data API v3, diversos sinais potencialmente relevantes permanecem estruturalmente inacessíveis aos pesquisadores. Informações como transcrições completas, CPM, RPM, métricas históricas detalhadas, descurtidas, sinais internos de recomendação e outros indicadores econômicos ou algorítmicos não são plenamente disponibilizados pela plataforma. Essa limitação justificou a adoção complementar de fontes externas de dados, obtidas tanto via APIs adicionais quanto por mecanismos de *Web crawling*, funcionando como estratégias de contorno para aproximação das variáveis investigadas.

Entretanto, mesmo essas variáveis derivadas não possuem natureza exata ou determinística. A pesquisa opera sobre sinais digitais observáveis, utilizando *proxies* computacionais para inferir fenômenos sociotécnicos mais amplos. Assim, não se observa diretamente ideologia real, intenção psicológica, receita efetiva ou regras internas de recomendação algorítmica, mas apenas indícios computacionais associados a esses fenômenos.

Tabela 4.2: Comparação entre coleta de dados via API oficial e técnicas de crawling utilizadas no contexto do artefato proposto.

Aspecto	**API**	**Crawling**
Estabilidade estrutural	Alta	Baixa
Padronização	Alta	Variável
Resistência anti-bot	Menor	Alta
Confiabilidade	Alta	Média
Acesso a dados internos	Limitado	Parcial
Custo de manutenção	Baixo	Alto

A partir do modelo proposto, a coleta é iniciada por entradas semelhantes às formas usuais de acesso às plataformas digitais, como *query*, URL, `@canal`, `videoId` ou `channelId`. Essas entradas não são tratadas diretamente como dados finais, mas como

sementes de busca que precisam ser convertidas em identificadores canônicos.

No caso do YouTube, diferentes formas de entrada — como URLs de vídeos, identificadores de canais (`@canal`), URLs encurtadas ou customizadas — precisam ser normalizadas em IDs canônicos da plataforma, como `videoId` e `channelId`. Essa etapa reduz ambiguidades, evita duplicidades e permite que os identificadores sejam posteriormente hidratados via API oficial e fontes complementares, enriquecendo entidades como vídeos, canais e comentários.

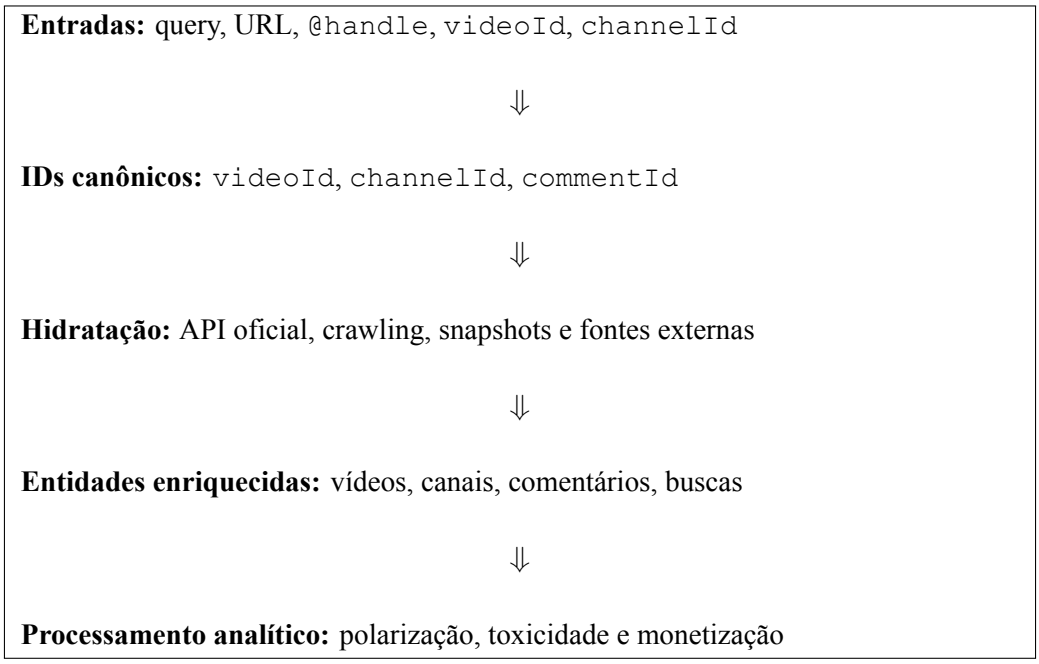


Figura 4.4: Fluxo simplificado do processo de normalização, hidratação e enriquecimento das entidades no *SocialDash*, ilustrando o processo de inferência de polarização, toxicidade e monetização.

Essa estratégia permite desacoplar a entrada inicial do processo analítico. Independentemente de o usuário informar uma busca textual, um canal ou um vídeo específico, o sistema converte a entrada em entidades identificáveis, hidrata seus metadados e integra sinais provenientes de múltiplas fontes para posterior processamento.

Uma vez estabelecida essa estrutura de entidades enriquecidas, surge também a necessidade de lidar com limitações operacionais relacionadas ao volume de dados disponibilizados pelas APIs e ao custo computacional envolvido na coleta contínua da plataforma.

As APIs da plataforma retornam volumes limitados por chamada, geralmente 50 ou 100 registros (somente para comentários). Para reduzir consumo de quota, custo computacional e tempo de processamento, esta pesquisa abstrai o número de requisições, adotando

inicialmente uma coleta mínima por entidade, sem impedir ampliações posteriores.

Como a API do YouTube prioriza o retorno de registros segundo critérios internos de relevância — e não necessariamente por distribuição temporal uniforme — adotou-se uma estratégia complementar de segmentação em *buckets* temporais. Para cada entidade analisada, considera-se o intervalo entre sua data inicial (por exemplo, upload do vídeo, criação do canal ou primeiro registro disponível) e a data atual. Esse intervalo é convertido em faixas temporais equivalentes, permitindo distribuir os registros coletados ao longo do tempo a partir de seus respectivos *timestamps*. Essa abordagem busca maximizar a dimensão temporal da análise, reduzindo parcialmente os efeitos da priorização algorítmica imposta pela própria plataforma.

Distribuição temporal por buckets

$$T_{upload} \quad T_1 \quad T_2 \quad T_3 \quad T_4 \quad T_{hoje}$$

$$|------||------||------||------||------|$$

Exemplo: $video_X$ com upload em 12/12/2021; data atual: 18/05/2026

$$T_0 = ts(12/12/2021) = 1639267200 \qquad T_h = ts(18/05/2026) = 1779062400$$

$$\Delta = \frac{T_h - T_0}{5} = \frac{1779062400 - 1639267200}{5} = 27959040$$

Bucket	Início	Fim
B1	T_0	$T_0 + \Delta$
B2	$T_0 + \Delta$	$T_0 + 2\Delta$
B3	$T_0 + 2\Delta$	$T_0 + 3\Delta$
B4	$T_0 + 3\Delta$	$T_0 + 4\Delta$
B5	$T_0 + 4\Delta$	T_h

$$comentario_Y \rightarrow data \rightarrow timestamp \rightarrow bucket$$

$$ts(comentario_Y) \in [inicio(B_i), fim(B_i)] \Rightarrow comentario_Y \in B_i$$

A distribuição depende dos registros retornados pela API; por isso, alguns buckets podem concentrar mais dados.

Figura 4.5: Estratégia simplificada de segmentação temporal em *buckets*, utilizada para distribuir registros ao longo do intervalo temporal da entidade analisada, permitindo comparações proporcionais entre diferentes períodos de atividade.

A estratégia de distribuição temporal em *buckets* foi concebida principalmente como um mecanismo de apoio ao *sensemaking* temporal e à amostragem analítica dos dados

coletados. Em um cenário ideal, sem restrições de quota, custo computacional ou limitações operacionais das APIs, todos os registros históricos poderiam ser armazenados integralmente no banco de dados juntamente com seus respectivos timestamps originais. Entretanto, diante das limitações práticas observadas, a divisão temporal em buckets permite preservar parcialmente a noção de evolução longitudinal dos fenômenos analisados mesmo com amostras reduzidas.

A mesma lógica é aplicada, por exemplo, à distribuição temporal de vídeos de um canal. Nesse caso, considera-se como T_0 a data de criação do canal e como T_h a data atual. O intervalo temporal é dividido em subconjuntos equivalentes, e cada vídeo coletado é alocado visualmente no bucket correspondente ao seu timestamp de upload. Posteriormente, cálculos agregados realizados dentro desses intervalos — como volume de vídeos, intensidade de comentários, toxicidade média ou *proxies* de monetização — permitem produzir uma visão aproximada de *visual analytics* sobre a evolução temporal do fenômeno observado.

Além da distribuição temporal, os próprios metadados observáveis da plataforma também permitem derivar relações proporcionais potencialmente relevantes para interpretação sociotécnica dos conteúdos. Embora esta pesquisa não tenha como objetivo central construir um modelo completo de análise de engajamento, tais relações representam uma dimensão metodológica importante que não poderia deixar de ser observada.

Inicialmente, o sistema opera sobre métricas absolutas diretamente disponibilizadas pela plataforma, como visualizações, inscritos, comentários, likes e quantidade de vídeos publicados. Esses valores representam atributos estruturais básicos dos canais e conteúdos analisados. Segue a Tabela 4.3 com exemplos de métricas absolutas observáveis.

Tabela 4.3: Exemplos de métricas absolutas observáveis na plataforma e suas interpretações analíticas.

Métrica	**`Metadado`**	**Possível interpretação**
Total de views	`viewCount`	Alcance bruto do conteúdo
Total de likes	`likeCount`	Aprovação agregada da audiência
Total de comentários	`commentCount`	Intensidade geral de interação
Total de inscritos	`subscriberCount`	Tamanho estrutural da audiência
Total de vídeos	`videoCount`	Volume histórico de produção

Entretanto, valores absolutos nem sempre refletem intensidade proporcional de participação. Um canal estruturalmente grande pode apresentar milhões de visualizações, mas

baixo engajamento relativo. Em contrapartida, canais menores podem produzir conteúdos altamente mobilizadores quando observadas relações proporcionais entre views, likes, comentários e inscritos. A Figura 4.6 ilustra conceitualmente a diferença entre desempenho absoluto e desempenho relativo.

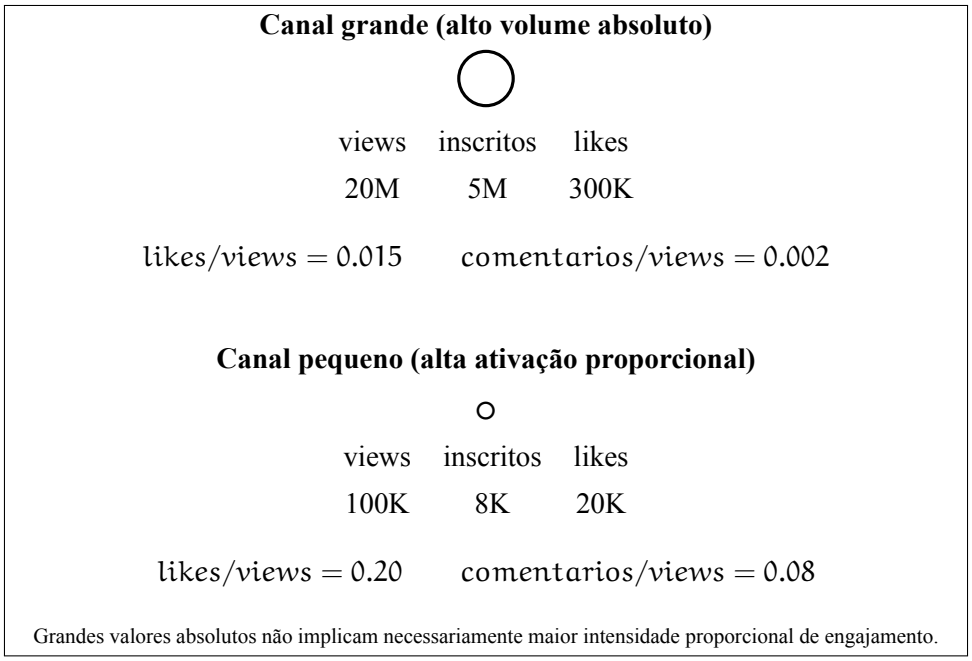


Figura 4.6: Exemplo conceitual da diferença entre métricas absolutas e relativas de engajamento, evidenciando que grandes volumes de audiência não implicam necessariamente maior intensidade proporcional de interação.

A partir desses sinais observáveis, torna-se possível derivar métricas relativas envolvendo ativação de audiência, intensidade discursiva e estrutura conversacional. Embora tais relações estejam parcialmente fora do escopo principal desta tese, elas constituem uma camada analítica complementar relevante para investigações em plataformas digitais.

Segue a Tabela 4.4 com exemplos de métricas derivadas de engajamento proporcional e participação discursiva.

Essas métricas permitem ampliar a granularidade interpretativa dos sinais digitais coletados, oferecendo mecanismos adicionais de *sensemaking* e *visual analytics* sobre mobilização, participação e dinâmica discursiva dentro das plataformas digitais.

Tabela 4.4: Exemplos de métricas relativas derivadas de sinais observáveis da plataforma, utilizadas para inferir padrões proporcionais de engajamento, mobilização e interação discursiva.

Métrica	**Relação**	**Interpretação**
Likes / views	likeCount / viewCount	Aprovação proporcional
Comentários / views	commentCount / viewCount	Ativação da audiência
Likes / comentários	likeCount / commentCount	Aprovação versus debate
Views / inscritos	viewCount / subscriberCount	Capacidade de mobilização
Comentários / inscritos	commentCount / subscriberCount	Participação proporcional da base
Likes / inscritos	likeCount / subscriberCount	Engajamento da comunidade
Comentários / duração	commentCount / durationSeconds	Intensidade discursiva temporal
Views / duração	viewCount / durationSeconds	Eficiência temporal de consumo
Replies / comentários raiz	replies / rootComments	Debate interno
Replies / comentários totais	replies / totalComments	Horizontalidade conversacional
Likes em comentários / replies	commentLikes / replies	Validação social das interações
Coment. autor / coment. totais	ownerReplies / totalComments	Participação ativa do criador

De maneira semelhante, a monetização também pode ser analisada sob uma perspectiva de granularidade inferencial. Embora esta pesquisa não tenha como objetivo construir um modelo econométrico preciso, arquiteturas orientadas a séries temporais e sinais digitais permitem decompor monetização agregada em unidades analíticas menores associadas ao conteúdo produzido.

Partindo da monetização instantânea estimada de um canal ao longo do tempo, seria possível integrar essa curva temporal e relacionar o valor monetizado acumulado com unidades observáveis derivadas do UGC. Nesse cenário simplificado, assume-se um ambiente relativamente estável, sem exclusão significativa de vídeos, sem alterações estruturais relevantes de monetização e mantendo mecanismos convencionais de publicidade da plataforma.

A Equação $M(t)$ representa a monetização acumulada ao longo do tempo. O termo $r(t)$ corresponde à taxa instantânea estimada de monetização em determinado instante temporal t, enquanto dt representa uma variação infinitesimal no eixo temporal. Assim, a integral acumulada de $r(t)$ ao longo do período analisado fornece uma aproximação da monetização agregada do canal.

A ideia central consiste em considerar que o próprio conteúdo produzido pelo usuário representa o ativo que gera valor econômico. Dessa forma, a monetização acumulada poderia ser relacionada a diferentes granularidades observáveis do UGC, como vídeos, duração temporal, palavras transcritas ou categorias temáticas. A Figura 4.7 ilustra conceitualmente esse cenário.

Granularidade inferencial da monetização

$$\frac{\int_0^t r(t)\,dt}{\sum \text{unidade}_{UGC}}$$

Integral da monetização estimada ao longo do tempo dividida pelo somatório de unidades observáveis do conteúdo.

Unidade	Exemplo	Interpretação
Vídeos	US$ / vídeo	valor médio por publicação
Tempo	US$ / segundo	valor associado à duração
Texto	US$ / palavra	valor associado ao conteúdo textual
Tema	US$ / categoria	valor associado ao enquadramento temático

Cenário simplificado baseado em monetização contínua, estabilidade temporal e ausência de exclusão relevante de conteúdo.

Figura 4.7: Exemplo conceitual de decomposição inferencial da monetização em diferentes unidades observáveis do conteúdo, permitindo análises proporcionais associadas a vídeos, duração temporal, texto e categorias temáticas.

Sob essa perspectiva, a monetização agregada poderia ser parcialmente decomposta em unidades menores associadas ao próprio conteúdo produzido. Entretanto, tais relações possuem natureza fortemente inferencial, uma vez que diversos fatores internos da plataforma permanecem inacessíveis, incluindo retenção, perfil demográfico, leilões publicitários e mecanismos algorítmicos não observáveis. Assim, o objetivo desta discussão não é propor um modelo econométrico preciso, mas apenas ilustrar possibilidades analíticas futuras envolvendo monetização, séries temporais e economia da atenção.

De maneira geral, essas limitações reforçam uma característica central desta pesquisa: os fenômenos investigados não são observados diretamente, mas inferidos a partir de sinais digitais acessíveis computacionalmente. Sob essa perspectiva, as decisões metodológicas adotadas foram fundamentadas tanto na literatura quanto em estudos exploratórios preliminares realizados durante o desenvolvimento do artefato.

A toxicidade, por exemplo, foi inferida a partir da análise textual computacional de comentários utilizando técnicas automatizadas de classificação, abordagem recorrente em estudos anteriores [39, 63]. Já a polarização foi tratada como um construto inferencial operacionalizado a partir de sinais discursivos presentes em títulos, descrições, metadados e transcrições dos vídeos, considerando que a plataforma não fornece indicadores ideológicos explicitamente observáveis [22, 23, 33].

Da mesma forma, a monetização foi analisada por meio de *proxies* externas, uma vez que dados financeiros reais dos canais não são publicamente acessíveis. Para isso, foram

utilizados indicadores derivados de plataformas como VidIQ[17] e métricas públicas de engajamento e propagação de conteúdo, seguindo abordagens exploratórias semelhantes às encontradas na literatura [17, 35, 67].

Os estudos preliminares realizados durante a pesquisa também contribuíram para orientar essas decisões metodológicas. Experimentos iniciais indicaram, por exemplo, que transcrições completas dos vídeos tendem a oferecer sinais discursivos mais consistentes para inferência de polarização do que apenas estatísticas de engajamento ou metadados isolados. De maneira semelhante, análises exploratórias de toxicidade auxiliaram na definição das estratégias de processamento textual utilizadas no artefato.

Assim, as decisões adotadas nesta pesquisa não foram definidas exclusivamente por critérios tecnológicos, mas por um processo iterativo envolvendo fundamentação teórica, experimentação preliminar e viabilidade operacional. O objetivo consistiu em equilibrar limitações computacionais, disponibilidade de dados e capacidade analítica do modelo proposto.

Sob essa perspectiva, os três fenômenos centrais investigados — toxicidade, polarização e monetização — não são tratadas como verdades diretas da plataforma, mas como construções analíticas inferidas a partir de dados disponíveis, chamadas a APIs, crawling e mecanismos automatizados de processamento textual e inferência computacional.

Tabela 4.5: Síntese da operacionalização inferencial dos fenômenos de toxicidade, polarização e monetização a partir de dados observáveis e *proxies* computacionais.

Fenômeno	**Entrada observável**	**Processamento**	**Saída**
Toxicidade	Comentários coletados via YouTube API v3, incluindo texto e emojis	Cada comentário é enviado à Perspective API; os escores individuais são agregados por vídeo	Valor entre 0 e 1; média de toxicidade; pontos em gráfico
Polarização	Título, descrição, tags e transcrição do vídeo	Prompt estruturado enviado à LLM para classificação temática, polo, escore, confiança e justificativa	Categoria, polo, escore, confiança e explicação curta
Monetização *on-platform*	`channelId`	Consulta/crawling no VidIQ para estimativa pública de receita mensal	Valor estimado em US$/mês
Monetização *off-platform*	URLs presentes na descrição	Extração e contagem de links externos como proxy de estratégias externas de monetização	Quantidade de URLs e indício de monetização externa

A Figura 4.8 sintetiza a lógica inferencial adotada. Os dados observáveis entram no processo analítico e são transformados em variáveis aproximadas. Com exceção do escore retornado pela API de toxicidade, os demais resultados possuem natureza fortemente

[17] `https://vidiq.com/pt/`

inferencial, pois a ideologia real, a receita efetiva e os mecanismos internos da plataforma permanecem inacessíveis ao pesquisador.

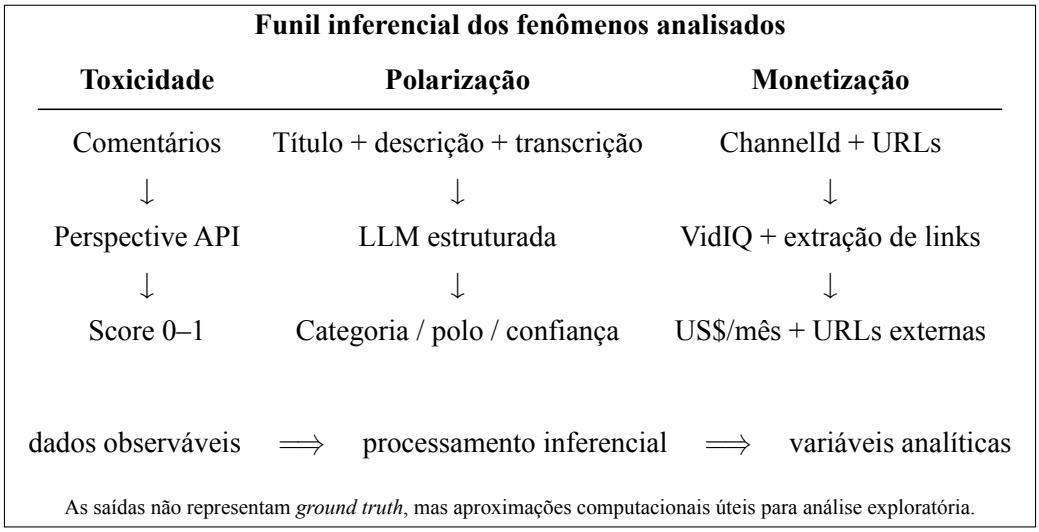


Figura 4.8: Funil inferencial simplificado utilizado para transformar sinais observáveis da plataforma em variáveis analíticas aproximadas relacionadas à toxicidade, polarização e monetização.

4.5 Implementação do artefato

A implementação do *SocialDash* constituiu uma das principais contribuições desta pesquisa, não apenas como instrumento técnico de coleta e visualização de dados, mas como uma infraestrutura analítica voltada à investigação integrada de fenômenos sociotécnicos associados ao YouTube. Diferentemente de abordagens tradicionais, que normalmente analisam polarização, monetização ou toxicidade de forma isolada, o sistema foi concebido para permitir a observação simultânea dessas dimensões dentro de um mesmo ambiente computacional.

O desenvolvimento do artefato ocorreu de forma incremental e experimental, envolvendo etapas de coleta, hidratação, processamento, inferência e visualização de dados provenientes de múltiplas fontes. Durante a implementação, foram adotadas práticas relacionadas à segurança e organização da infraestrutura computacional, incluindo armazenamento de chaves de APIs e segredos de autenticação em arquivos de configuração protegidos e excluídos do controle de versionamento por meio de mecanismos apropriados de isolamento de credenciais. As escolhas relacionadas ao ambiente computacional, linguagens utilizadas, bibliotecas, sistema operacional e infraestrutura de desenvolvimento encontram-se descritas de maneira mais detalhada no capítulo 5.

O sistema foi estruturado como uma aplicação Web responsiva acessível por navegador, permitindo utilização tanto em ambientes acadêmicos quanto exploratórios. Além dos módulos analíticos, foram desenvolvidas páginas explicativas destinadas à contextualização dos fenômenos investigados, buscando tornar conceitos relacionados à polarização, toxicidade, monetização e plataformas digitais mais acessíveis também para usuários não especialistas.

Cada conjectura investigada nesta pesquisa passou a possuir uma interface própria dentro do sistema. Apesar das diferenças analíticas entre as conjecturas, a lógica operacional foi mantida de maneira relativamente uniforme. Em geral, o pesquisador realiza consultas por palavras-chave, URLs ou identificadores canônicos, seleciona os registros desejados e posteriormente executa o processamento analítico correspondente por meio da própria interface.

A implementação também incorporou mecanismos de inferência computacional capazes de transformar dados não estruturados em variáveis quantitativas observáveis. Comentários, títulos, descrições e transcrições são processados por serviços especializados de análise textual, permitindo atribuição de escores contínuos de toxicidade, classificação discursiva e identificação de padrões relacionados à monetização e polarização.

Paralelamente, o sistema integra indicadores associados ao desempenho econômico e à monetização dos canais, utilizando tanto dados públicos obtidos via APIs quanto mecanismos complementares de coleta e crawling. Embora tais informações possuam natureza inferencial e não representem diretamente os mecanismos internos da plataforma, elas permitem construir sinais analíticos úteis para investigação exploratória da economia da atenção no contexto do YouTube.

4.6 Avaliação

Embora o ambiente experimental tenha sido disponibilizado de forma pública via Web, o processo de avaliação foi conduzido em contexto controlado de pesquisa, com participantes previamente orientados sobre os objetivos das tarefas e sobre o caráter exploratório do artefato. Essa abordagem buscou equilibrar condições mais próximas do uso real do sistema com um nível mínimo de padronização metodológica necessário à comparação das respostas e interpretações obtidas durante o experimento.

4.6.1 Contexto experimental

A avaliação do *SocialDash* foi conduzida em um ambiente Web independente, disponibilizado em domínio próprio[18], com o objetivo de simular uma situação de uso próxima ao contexto real de investigação empírica em plataformas digitais. A opção por um ambiente externo, desvinculado de laboratórios, instituições ou grupos de pesquisa, buscou reduzir possíveis interferências contextuais na percepção dos participantes durante a interação com o artefato.

O procedimento avaliativo foi estruturado a partir da execução de quatro tarefas analíticas, cada uma associada a uma conjectura de design proposta nesta pesquisa. Essas tarefas orientaram os participantes a explorar relações entre polarização, toxicidade e monetização no YouTube, utilizando os recursos de coleta, visualização, comparação e interpretação disponibilizados pelo sistema.

A avaliação contou com dez participantes, com diferentes níveis de formação acadêmica, incluindo graduação, mestrado e doutorado. Esse perfil foi considerado adequado ao propósito da avaliação, uma vez que o uso do artefato exige capacidade mínima de interpretação de métricas, sinais discursivos, indicadores de engajamento e dinâmicas próprias de plataformas digitais.

Durante a execução das tarefas, o próprio *SocialDash* disponibilizou campos opcionais de feedback textual em cada fluxo analítico. Esses registros permitiram capturar impressões dos participantes no momento da interação com o sistema, isto é, durante o uso efetivo do artefato e antes do preenchimento do formulário externo. Essa camada complementar de avaliação foi utilizada como evidência qualitativa auxiliar, permitindo observar como os usuários formulavam hipóteses, justificavam escolhas e interpretavam os sinais apresentados pelo sistema.

Após a realização das tarefas, os participantes preencheram um formulário online de avaliação[19]. O formulário reuniu questões fechadas em escala Likert de cinco pontos, variando de 1 (Discordo totalmente) a 5 (Concordo totalmente), além de questões abertas voltadas à identificação de limitações, sugestões de melhoria e possibilidades de expansão do artefato.

Dessa forma, a avaliação combinou três fontes principais de evidência: a interação dos participantes com as tarefas analíticas no ambiente experimental, os feedbacks qualitativos

[18] `https://www.tesedoutorado.com.br`

[19] Disponível em: `https://docs.google.com/forms/d/e/1FAIpQLScObWCf8dgPE2LfM-xIOfFBNgvQE6tUTh7m_bAAUrNqFZlc2g/viewform`.

registrados durante o uso do sistema e as respostas estruturadas do formulário externo. Essa combinação permitiu avaliar não apenas o funcionamento técnico do artefato, mas também sua capacidade de apoiar processos de investigação exploratória, interpretação contextual e formulação de hipóteses sobre fenômenos sociotécnicos no YouTube.

4.6.2 Instrumento de avaliação

O instrumento de avaliação foi estruturado para analisar diferentes dimensões relacionadas ao uso investigativo do *SocialDash*, incluindo aprendizagem conceitual, investigabilidade das conjecturas analíticas propostas, usabilidade do sistema e percepções qualitativas dos participantes durante a interação com o artefato.

As questões foram organizadas em blocos temáticos (SQ0–SQ5), cada um associado a objetivos específicos da avaliação. O bloco SQ0 concentrou questões relacionadas à compreensão conceitual das dimensões de polarização, monetização e toxicidade, buscando identificar se o sistema auxiliava a familiarização inicial dos participantes com os constructos investigados na pesquisa.

Os blocos SQ1–SQ4 foram diretamente associados às conjecturas de design propostas nesta tese. Cada bloco avaliou a capacidade do artefato de tornar investigáveis diferentes relações entre polarização, monetização e toxicidade, correspondendo também aos fluxos analíticos executados pelos participantes durante as tarefas experimentais.

O bloco SQ5 concentrou questões relacionadas à usabilidade, facilidade de aprendizado e percepção operacional do sistema, permitindo avaliar aspectos associados à experiência de uso do artefato em contexto exploratório.

As questões Q1–Q13 foram avaliadas por meio de escala Likert de cinco pontos, esse formato permitiu consolidar medidas estatísticas relacionadas à percepção dos participantes, incluindo média, moda, mediana e desvio padrão das respostas obtidas.

Além das questões fechadas, o formulário incluiu duas questões abertas (Q14 e Q15), destinadas ao registro de limitações percebidas, sugestões de melhoria e possíveis extensões analíticas do artefato. Essas respostas foram posteriormente organizadas em categorias qualitativas, utilizadas como evidência complementar na interpretação dos resultados.

A Tabela 4.6 apresenta a estrutura consolidada do instrumento de avaliação utilizado durante o experimento.

Tabela 4.6: Instrumento de avaliação do SocialDash baseado nas conjecturas C1–C4

#	Questão	Tipo
SQ0 – Aprendizado conceitual sobre as dimensões analisadas		
Q1	O SocialDash auxiliou você a compreender o conceito de polarização em conteúdos do YouTube	Likert
Q2	O SocialDash auxiliou você a compreender o conceito de toxicidade em interações e comentários	Likert
Q3	O SocialDash auxiliou você a compreender os mecanismos de monetização e engajamento na plataforma	Likert
SQ1 – Investigabilidade da relação Polarização ↔ Toxicidade (C1)		
Q4	O artefato permitiu observar relações entre polarização e toxicidade nos conteúdos analisados	Likert
Q5	O artefato tornou mais clara a relação entre engajamento polarizado e reações hostis da audiência	Likert
SQ2 – Investigabilidade da relação Polarização ↔ Monetização (C2)		
Q6	O sistema permitiu analisar conjuntamente indicadores de polarização e *proxies* de monetização	Likert
Q7	As métricas apresentadas facilitaram a observação de padrões entre polarização e desempenho econômico	Likert
SQ3 – Investigabilidade da relação Monetização ↔ Toxicidade (C3)		
Q8	O sistema possibilitou observar relações entre monetização e toxicidade nos conteúdos analisados	Likert
Q9	O sistema permitiu observar padrões temporais de toxicidade e monetização	Likert
SQ4 – Investigabilidade da interseção Polarização ↔ Monetização ↔ Toxicidade (C4)		
Q10	O sistema permitiu observar simultaneamente polarização, monetização e toxicidade	Likert
SQ5 – Usabilidade e utilidade científica		
Q11	A navegação entre as funcionalidades ocorreu de forma clara e intuitiva	Likert
Q12	O sistema apresentou facilidade de aprendizado durante o uso inicial	Likert
Q13	O sistema minimizou erros ou dificuldade operacional durante a utilização	Likert
Q14	Quais limitações ou melhorias você identificou no sistema?	Aberta
Q15	Quais outras sugestões ou dimensões analíticas deveriam ser incorporadas?	Aberta

5. Benchmarks

Este capítulo apresenta os experimentos preliminares, benchmarks computacionais e análises exploratórias realizados durante o desenvolvimento do artefato *SocialDash*. Inicialmente, são apresentados os experimentos relacionados à toxicidade, incluindo comparação entre diferentes modelos e estratégias de normalização dos escores. Em seguida, são discutidos os testes envolvendo inferência de polarização a partir de modelos de linguagem e sinais discursivos extraídos dos vídeos. Posteriormente, são exploradas as abordagens preliminares relacionadas à monetização. Por fim, o capítulo apresenta provas de conceito e primeiros testes integrando simultaneamente polarização, toxicidade e monetização em um mesmo ambiente, evidenciando o potencial do artefato para apoiar análises sociotécnicas exploratórias em plataformas digitais.

5.1 Dados de toxicidade

A coleta de comentários constitui o núcleo empírico da análise de toxicidade e polarização afetiva, por representar o principal espaço de interação discursiva entre usuários no YouTube. A literatura associa a toxicidade ao uso de linguagem hostil ou desrespeitosa, frequentemente amplificada em ambientes online [39]. Além de refletirem reações individuais, os comentários funcionam como espaço de construção coletiva de sentido, no qual se manifestam conflitos, alinhamentos ideológicos e processos de reforço identitário [2]. Sua utilização como unidade de análise justifica-se pela expressividade discursiva, granularidade analítica e disponibilidade em larga escala.

A mensuração da toxicidade representa um desafio metodológico, por tratar-se de um construto subjetivo e dependente de contexto linguístico e cultural. Por essa razão, a literatura recorre a abordagens automatizadas como *proxies*, especialmente em cenários de larga escala [39].

Entre essas abordagens, destaca-se o uso da Perspective API, amplamente adotada para atribuição de scores probabilísticos contínuos de toxicidade a comentários online [13, 27]. Essa API produz valores entre 0 e 1 associados à probabilidade de um comentário ser considerado tóxico, permitindo sua incorporação em análises comparativas, temporais e exploratórias. No entanto, estudos apontam limitações relacionadas à interpretação contextual, informalidade textual, adaptação entre diferentes ambientes sociotécnicos e restrições metodológicas dos modelos [13, 27, 53], indicando que tais scores devem ser interpretados como indicadores probabilísticos operacionais, e não como medidas absolutas.

Além da Perspective API, a literatura apresenta um conjunto diversificado de abordagens, incluindo modelos supervisionados de machine learning [14], ferramentas open-source como Detoxify [12], abordagens lexicais [31, 63], filtros heurísticos [40] e *proxies* baseados em sinais linguísticos e emocionais [63, 64]. Essa diversidade reforça o caráter multidimensional da toxicidade, que envolve diferentes formas de expressão discursiva.

A Tabela 5.1 sintetiza essas estratégias, destacando suas principais características e limitações, permitindo situar a abordagem adotada nesta pesquisa no contexto do estado da arte.

Tabela 5.1: Síntese das estratégias de mensuração de toxicidade por tecnologia

Tecnologia	**Referências**	**Observação metodológica**
Perspective API (Google Jigsaw)	[13, 27, 39, 53, 62]	Tecnologia mais recorrente na literatura. Baseada em score contínuo [0,1], amplamente validada empiricamente e com bom alinhamento com avaliações humanas, embora apresente limitações relacionadas a contexto e possíveis vieses linguísticos.
Modelos baseados em Transformers (BERT, BiLSTM, etc.)	[14, 53]	Abordagens mais sofisticadas que incorporam contexto e aprendizado supervisionado, com melhor capacidade de generalização, porém maior custo computacional e complexidade de implementação.
Modelos open-source (Detoxify)	[12]	Alternativa baseada em Transformers com controle local e transparência metodológica. Permite replicabilidade, mas com menor validação empírica em larga escala comparada à Perspective API.
Abordagens lexicais e linguísticas	[31, 63]	Identificação baseada em vocabulário, padrões linguísticos e comunidades específicas. Alta interpretabilidade, porém limitada na captura de nuances e contexto.
Filtros e regras heurísticas	[40]	Estratégias simples baseadas em listas de palavras e regras definidas manualmente. Baixo custo e alta transparência, mas baixa sensibilidade semântica.
Modelos conceituais e indicadores indiretos	[28, 36, 64]	Toxicidade tratada como construto teórico ou inferida por emoções e padrões de interação. Útil para fundamentação conceitual, mas sem operacionalização direta.
Moderação e governança algorítmica	[42, 71]	Abordagens que tratam toxicidade como fenômeno sistêmico ligado à moderação, desmonetização e recomendação. Expande a análise para além do nível textual.

5.1.1 Validação experimental de métodos de detecção de toxicidade

Para avaliar a adequação das tecnologias de mensuração de toxicidade, foi definido um teste simples de consistência ordinal. A ideia é submeter a cada tecnologia um conjunto de enunciados organizados em níveis progressivos de intensidade, partindo de uma discordância leve até uma ofensa explícita. Como todas as ferramentas analisadas foram normalizadas para a mesma escala de interpretação, com valores entre 0 e 1, espera-se que os escores atribuídos respeitem a ordem crescente dos exemplos.

A Figura 5.1 ilustra essa escala ordinal, construída com base em níveis progressivos de intensidade discursiva, utilizada como referência para os testes de consistência dos modelos.

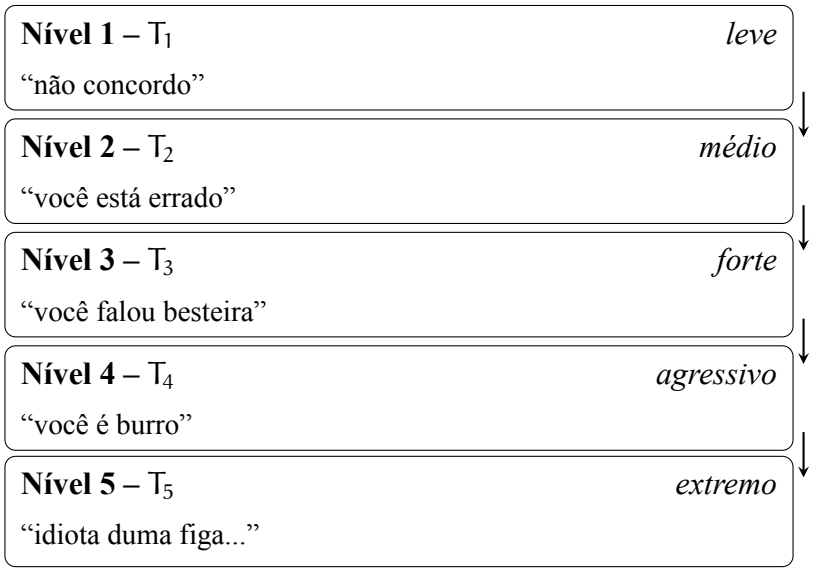


Figura 5.1: Escala ordinal utilizada para validação da mensuração de toxicidade

Assim, a tecnologia é considerada mais consistente quando seus escores preservam a monotonicidade esperada entre os níveis. Em outras palavras, o valor atribuído ao enunciado do nível 2 deve ser maior ou igual ao do nível 1, e assim sucessivamente até o nível 5. Esse procedimento não busca medir a toxicidade absoluta de cada frase, mas verificar se a ferramenta consegue capturar a intensificação semântica da toxicidade em uma escala padronizada de 0 a 1.

A partir dessa estrutura, torna-se possível definir um procedimento de normalização dos escores, conforme sintetizado na Tabela 5.2. Para viabilizar a comparação entre di-

ferentes tecnologias de mensuração de toxicidade, foi necessário estabelecer uma escala comum de análise. As ferramentas avaliadas apresentam saídas heterogêneas, com diferentes intervalos e interpretações. Enquanto modelos como Perspective API, LLM e Detoxify já retornam valores diretamente no intervalo $[0, 1]$, o Google NLP fornece métricas sentimentais em escalas distintas, exigindo transformações adicionais. Adicionalmente, foram definidas *proxies* derivadas capazes de converter todas as saídas para uma mesma escala comparável de toxicidade.

Tabela 5.2: Descrição das colunas do benchmark de toxicidade

Coluna	**Origem**	**Range**	**Interpretação / cálculo**
Persp	Perspective API	$[0, 1]$	Valores próximos de 0 indicam baixa toxicidade; valores próximos de 1 indicam alta probabilidade de toxicidade.
LLM	Modelo de linguagem	$[0, 1]$	Score estimado por prompt controlado, solicitando ao modelo a probabilidade do texto ser um comentário tóxico.
GSS	Google NLP	$[-1, 1]$	*Google Sentiment Score*. Mede polaridade sentimental: -1 indica sentimento muito negativo, 0 indica neutralidade e $+1$ indica sentimento muito positivo. Não mede toxicidade diretamente.
GSM	Google NLP	$[0, +\infty)$	*Google Sentiment Magnitude*. Mede intensidade emocional do texto. Não possui teto fixo; em comentários curtos tende a permanecer em valores baixos ou moderados.
GTP	Derivado de GSS	$[0, 1]$	*Google Toxicity Proxy*. Conversão da negatividade em proxy comparável de toxicidade: $\text{GTP} = \max(0, -\text{GSS})$. Sentimentos positivos resultam em 0.
GTPP	Derivado de GSS e GSM	$[0, 1]$	*Google Toxicity Proxy Ponderada*. Proxy de toxicidade ponderada pela intensidade emocional: $\text{GTPP} = \max(0, -\text{GSS}) \times \min(1, \text{GSM})$.
Detoxify	Modelo open-source	$[0, 1]$	Score de toxicidade produzido por modelo open-source. Valores próximos de 1 indicam maior toxicidade textual.

Dentre as tecnologias analisadas, o Google NLP requer tratamento específico, uma vez que não fornece diretamente uma medida de toxicidade. No caso do Google NLP, a métrica de sentimento (GSS) varia entre -1 e +1, sendo que valores negativos indicam polaridade negativa. Considerando que a toxicidade está mais associada à negatividade do que à positividade, foi adotada a transformação $\text{GTP} = \max(0, -\text{GSS})$, convertendo o sentimento negativo em uma proxy de toxicidade.

Adicionalmente, foi considerada uma versão ponderada, denominada GTPP, que incorpora a magnitude (GSM) como fator de intensidade emocional, limitada ao intervalo $[0, 1]$. Dessa forma, todas as métricas — incluindo Persp, LLM, Detox, GTP e GTPP — passam a operar na mesma escala, permitindo comparação direta entre os modelos.

A Tabela 5.3 apresenta exemplos simplificados desse procedimento de conversão, ilustrando como valores de sentimento são transformados em *proxies* comparáveis de toxicidade.

Tabela 5.3: Exemplo de cálculo das *proxies* derivadas do Google NLP

Texto	**GSS**	**GSM**	**GTP**	**GTPP**
"Acho que sua explicação científica precisa de mais dados e menos opinião."	-0.2	0.2	0.2	0.04
"Você interpretou mal o estudo e chegou a uma conclusão fraca."	-0.5	0.5	0.5	0.25
"Você está espalhando desinformação científica com muita confiança."	-0.9	0.9	0.9	0.81

A Tabela 5.4 apresenta exemplos concretos desse procedimento de conversão, ilustrando como os valores de sentimento são transformados em *proxies* comparáveis de toxicidade.

Tabela 5.4: Detalhamento do processo de conversão das métricas do Google NLP em *proxies* derivadas de toxicidade.

Texto	**GSS**	**GSM**	**GTP**	**GTPP**
Acho que sua explicação científica precisa de mais dados e menos opinião.	-0.2	0.2	$\max(0, -(-0.2)) = 0.2$	$0.2 \times \min(1, 0.2) = 0.04$
Você interpretou mal o estudo e chegou a uma conclusão fraca.	-0.5	0.5	$\max(0, -(-0.5)) = 0.5$	$0.5 \times \min(1, 0.5) = 0.25$
Você está espalhando desinformação científica com muita confiança.	0.9	0.9	$\max(0, -0.9) = 0$	$0 \times \min(1, 0.9) = 0$
Que explicação ridícula, você claramente não entende nada de ciência.	-0.7	0.7	$\max(0, -(-0.7)) = 0.7$	$0.7 \times \min(1, 0.7) = 0.49$

Com base na estrutura de normalização definida, foi desenvolvido um benchmark de sensibilidade de toxicidade com o objetivo de avaliar a capacidade discriminatória das diferentes abordagens computacionais na detecção de variações graduais de toxicidade em comentários online. Diferentemente de benchmarks tradicionais baseados em classificação binária, o experimento foi estruturado para testar a sensibilidade semântica dos modelos frente a pequenas variações linguísticas.

Para isso, foi construído um conjunto de dados sintético composto por 100 comentários, sendo 50 em português e 50 em inglês (traduções equivalentes). Os registros foram gerados com auxílio de LLM, a partir de instruções controladas que definiam tema, idioma, estilo textual e nível esperado de toxicidade (1 a 5). O procedimento utilizou uma escala ordinal composta por cinco níveis progressivos de intensidade discursiva, variando de comentários neutros ou levemente críticos até manifestações explicitamente ofensivas.

A geração buscou reproduzir características típicas de interações em plataformas digitais, incluindo linguagem informal, internetês, abreviações, emojis, sarcasmo e variações coloquiais, permitindo posteriormente contrastar os diferentes modelos de detecção quanto à sua sensibilidade semântica e capacidade de preservar a progressão esperada de toxicidade.

Cabe destacar que, por se tratar de um dataset sintético gerado por um modelo de linguagem, existe a possibilidade de viés favorecendo o próprio LLM na tarefa de classificação. Ainda assim, optou-se por essa abordagem devido à sua capacidade de controlar progressões semânticas de forma precisa, sendo essa limitação considerada na interpretação dos resultados.

O tamanho dos comentários foi calibrado para refletir a média observada em conteúdos gerados por usuários em plataformas digitais, tipicamente variando entre 10 e 25 palavras. Essa escolha visa garantir maior aderência ecológica ao comportamento real dos usuários em ambientes como o YouTube, onde interações tendem a ser curtas, diretas e frequentemente informais.

Além da progressão de intensidade, o dataset incorporou diferentes dimensões linguísticas relevantes para a literatura de toxicidade e discurso online, incluindo sarcasmo, linguagem coloquial, abreviações típicas de internet (internetês), palavrões e emojis. A inclusão de emojis é particularmente relevante, pois estudos indicam que esses elementos estão associados à expressão emocional e ao engajamento em interações digitais, podendo representar sinais importantes para interpretação de estados emocionais e dinâmicas discursivas em comentários online [63, 64].

Outro aspecto metodológico importante foi a inclusão deliberada de comentários positivos e neutros, com o objetivo de avaliar a taxa de falsos positivos dos modelos — isto é, situações em que conteúdos não tóxicos são erroneamente classificados como ofensivos.

No que se refere às tecnologias avaliadas, o benchmark comparou diferentes abordagens amplamente utilizadas na literatura e na indústria:

- **Perspective API (Google/Jigsaw)** — modelo supervisionado com saída probabilística contínua;
- **LLM (prompt-based)** — classificação via modelo de linguagem com instrução explícita;
- **Detoxify** — modelo open-source baseado em redes neurais profundas;

- **Google NLP** — análise de sentimento (score e magnitude), posteriormente transformada em *proxies* de toxicidade.

O modelo de LLM é o *gpt-4o-mini* foi utilizado com *prompt* explícito para classificação de toxicidade em escala contínua entre 0 e 1, tendo como instrução principal: "Classifique a probabilidade deste comentário ser tóxico em um valor entre 0 e 1." A execução utilizou baixa temperatura (aproximadamente 0.3) para reduzir variabilidade estocástica e aumentar consistência entre execuções. Testes com diferentes temperaturas não produziram alterações relevantes nos padrões observados. Ainda assim, os valores inferidos devem ser interpretados como estimativas heurísticas, e não probabilidades estatísticas calibradas. Além disso, os 50 registros foram traduzidos totalizando ao final 100 registros.

A configuração completa do experimento é apresentada na Tabela 5.5, detalhando os principais parâmetros do benchmark.

Tabela 5.5: Configuração do benchmark de sensibilidade de toxicidade

Dimensão	**Descrição**
Volume de dados	50 comentários em português + 50 comentários em inglês (traduções equivalentes), totalizando 100 registros.
Geração dos textos	Comentários sintéticos gerados via LLM, controlando nível de toxicidade (1–5), tema, polaridade e estilo linguístico.
Tamanho dos textos	Aproximadamente 10–25 palavras por comentário, alinhado à média observada em comentários de plataformas sociais (YouTube, Twitter), conforme literatura sobre UGC e análise textual.
Idioma	Português e inglês, permitindo avaliar robustez cross-lingual das APIs.
Polaridade	Inclusão de comentários positivos, neutros e negativos para avaliação de falso positivo.
Progressão semântica	Cada grupo contém 5 níveis progressivos de toxicidade (escala ordinal controlada).
Variação linguística	Inclusão de: sarcasmo, linguagem informal, abreviações (internetês), palavrões e emojis
Modelos avaliados	Perspective API, LLM (prompt-based), Detoxify, Google NLP (sentiment + *proxies*).
Normalização	Todos os outputs convertidos para escala comum [0,1] para comparabilidade.
Métricas de avaliação	Ordem preservada e consistência intra-grupo.
Disponibilidade	Resultados completos e código disponíveis em: `https://github.com/lga37/dr2`

A implementação utilizou integração entre Laravel (PHP) para orquestração dos testes e Python (versão 3.10+) para execução do modelo Detoxify, garantindo compatibilidade

com bibliotecas modernas de processamento de linguagem natural. Uma amostragem reduzida dos resultados esta representandos na Tabela 5.6.

Tabela 5.6: Amostra do benchmark de sensibilidade de toxicidade

N	Texto	Persp	LLM	Detox	GSS	GSM	GTP	GTPP
1	Acho que sua explicação científica precisa de mais dados e menos opinião.	0.033	0.100	0.00160	-0.2	0.2	0.2	0.04
2	Você interpretou mal o estudo e chegou a uma conclusão fraca.	0.071	0.100	0.00383	-0.5	0.5	0.5	0.25
3	Você está espalhando desinformação científica com muita confiança.	0.146	0.100	0.00388	0.9	0.9	0	0
4	Que explicação ridícula, você claramente não entende nada de ciência.	0.396	0.250	0.91443	-0.7	0.7	0.7	0.49
5	Você é um completo imbecil científico e esse vídeo é uma vergonha.	0.899	0.750	0.98730	-0.7	0.7	0.7	0.49
1	A explicação científica ficou clara e ajudou a entender melhor o assunto.	0	0	0.00055	0.9	0.9	0	0
2	Gostei da forma como você usou dados para sustentar o argumento.	0.013	0	0.00046	0.9	0.9	0	0
3	O vídeo trouxe uma análise científica equilibrada e bem organizada.	0	0	0.00068	0.9	0.9	0	0
4	Excelente explicação, simples sem perder o rigor científico.	0.009	0	0.00089	0.9	0.9	0	0
5	Conteúdo brilhante, didático e muito útil para combater desinformação.	0.003	0	0.00059	0.9	0.9	0	0
1	Não concordo com você, mas ok □□	0.026	0	0.00051	0.2	0.2	0	0
2	Você viajou um pouco nessa opinião kkk	0.162	0.100	0.00170	0.1	0.1	0	0
3	Você falou besteira demais □□	0.641	0.100	0.85897	-0.3	0.3	0.3	0.09
4	Você é burro demais □□	0.751	0.650	0.98057	-0.3	0.3	0.3	0.09
5	Tu é um imbecil do carxxxlho □□□□	0.925	0.850	0.99361	0.3	0.3	0	0
1	Não gostei da atuação do time, mas ainda dá para melhorar.	0.017	0	0.00056	-0.4	0.4	0.4	0.16
2	O time jogou mal e parecia sem organização em campo.	0.021	0.050	0.02532	-0.6	0.6	0.6	0.36
3	Esse técnico é fraco e não sabe montar o time direito.	0.241	0.250	0.46885	-0.6	0.6	0.6	0.36
4	Time ridículo, técnico burro e torcida sofrendo com essa porcaria.	0.674	0.650	0.99472	-0.6	0.6	0.6	0.36
5	Esse time é um lixo completo, bando de inúteis sem vergonha.	0.659	0.850	0.99382	-0.7	0.7	0.7	0.49

Legenda: Persp = Perspective API; LLM = modelo de linguagem via prompt; Detox = Detoxify; GSS = Google Sentiment Score; GSM = Google Sentiment Magnitude; GTP = Google Toxicity Proxy; GTPP = Google Toxicity Proxy Ponderada. As colunas em verde correspondem às saídas originais e *proxies* derivadas do Google NLP.

Observa-se que o LLM preservou integralmente a ordem esperada nesse exemplo,

enquanto a Perspective API também apresentou crescimento monotônico e gradual dos scores. O Detoxify demonstrou alta sensibilidade a insultos explícitos, mas com tendência à saturação rápida, saltando de valores muito baixos para valores próximos de 1. Já o Google NLP, por medir sentimento e não toxicidade diretamente, precisou ser convertido para uma proxy comparável: na versão sem amplificação, considera-se apenas a negatividade do score; na versão com amplificação, essa negatividade é ponderada pela magnitude emocional do texto.

Embora o LLM tenha apresentado desempenho perfeito na preservação da ordem esperada, esse resultado deve ser interpretado com cautela, pois o dataset sintético foi gerado com auxílio de modelo de linguagem. Essa condição pode favorecer avaliações semânticas realizadas por LLMs. Ainda assim, os testes com diferentes configurações de temperatura indicaram estabilidade das respostas, reduzindo a probabilidade de que o resultado decorra apenas de variabilidade estocástica, conforme Tabela 5.7.

Tabela 5.7: Desempenho dos modelos na preservação da progressão de toxicidade

Modelo	**Português**	**Inglês**	**Avaliação**
LLM	10/10	10/10	Excelente
Perspective	9/10	10/10	Muito bom
Detoxify	6/10	7/10	Médio
GTP	4/10	3/10	Fraco
GTPP	5/10	4/10	Fraco

A avaliação da capacidade dos modelos em preservar progressões controladas de toxicidade evidenciou diferenças significativas entre as abordagens. O modelo baseado em LLM apresentou desempenho perfeito, mantendo monotonicidade em todos os grupos analisados, tanto em português quanto em inglês, o que indica elevada sensibilidade semântica ao aumento gradual de hostilidade textual. A Perspective API também demonstrou desempenho consistente, com pequenas limitações pontuais em português, confirmando sua robustez como proxy contínuo de toxicidade amplamente validado na literatura.

Por outro lado, o modelo Detoxify apresentou comportamento menos estável, caracterizado por saturação precoce e saltos abruptos nos scores, o que reduz sua capacidade de discriminação em níveis intermediários de toxicidade. Já as *proxies* derivadas do Google NLP (GTP e GTPP) apresentaram desempenho significativamente inferior, com falhas recorrentes na preservação da ordem esperada. Essas falhas decorrem principalmente da natureza da análise de sentimento, que não foi projetada para capturar toxicidade dire-

tamente, levando a zeramento indevido em textos negativos com polaridade positiva ou neutra.

Esses resultados reforçam que, embora múltiplas abordagens possam ser utilizadas como aproximações, modelos explicitamente treinados para detecção de toxicidade (como Perspective) ou avaliações semânticas diretas via LLM apresentam maior aderência ao fenômeno investigado. Consequentemente, a escolha metodológica do SocialDash privilegia a Perspective API como proxy principal, utilizando as demais abordagens como mecanismos complementares de análise exploratória.

Com base nos resultados empíricos e na análise das características metodológicas das abordagens avaliadas, torna-se possível justificar a escolha da tecnologia principal adotada nesta pesquisa.

5.1.2 Justificativa da escolha da Perspective API

A escolha da Perspective API como tecnologia principal para operacionalização da toxicidade foi orientada pela combinação entre aderência à literatura, desempenho empírico nos benchmarks realizados e adequação aos objetivos analíticos do *SocialDash*. Embora o LLM tenha apresentado melhor desempenho ordinal no experimento controlado, sua avaliação é parcialmente favorecida pelo fato de o dataset sintético ter sido gerado com auxílio de modelo de linguagem, além de seus scores não constituírem probabilidades estatisticamente calibradas. Por outro lado, a Perspective API consiste em uma solução especificamente projetada para detecção de toxicidade, amplamente utilizada em estudos sobre moderação, discurso online e plataformas digitais.

A literatura revisada converge ao tratar a toxicidade como um fenômeno multifacetado e contextual, expresso por diferentes formas de discurso hostil, como insultos, linguagem ofensiva, ameaças e ataques identitários [39, 63, 40, 71, 42, 31]. Nesse contexto, a Perspective API se destaca por modelar a toxicidade como uma probabilidade contínua e decompor o construto em múltiplos atributos específicos, permitindo análises mais granulares e multidimensionais [27, 39, 13].

No contexto desta pesquisa, sua adoção foi orientada principalmente por quatro fatores:

- **Desempenho consistente no benchmark**: a Perspective API apresentou elevada capacidade de preservar progressões graduais de toxicidade, mantendo comportamento monotônico estável em diferentes grupos linguísticos e níveis de intensidade

discursiva;

- **Aderência metodológica à literatura**: a ferramenta é amplamente utilizada em pesquisas sobre toxicidade, moderação e discurso online, favorecendo comparabilidade com trabalhos prévios e maior alinhamento metodológico com o estado da arte;
- **Adequação operacional ao artefato**: a API oferece integração simples, processamento escalável e saída padronizada, permitindo sua incorporação direta ao pipeline analítico do *SocialDash*;
- **Decomposição multidimensional da toxicidade**: além do atributo principal de toxicidade (*TOXICITY*), a API fornece subatributos específicos relacionados a agressividade severa, insultos, palavrões, ameaças e ataques identitários, possibilitando análises mais detalhadas sobre diferentes manifestações discursivas.

A Tabela 5.8 apresenta os principais atributos disponibilizados pela Perspective API e seus respectivos usos metodológicos no contexto desta pesquisa. No contexto do *SocialDash*, a Perspective API foi adotada como proxy principal de toxicidade.

Tabela 5.8: Atributos de toxicidade disponíveis na Perspective API

Atributo	**Descrição**	**Uso metodológico**
TOXICITY	Probabilidade geral de o comentário ser percebido como tóxico ou hostil.	Proxy principal de toxicidade contínua (0–1).
SEVERE_TOXICITY	Versão mais restritiva da toxicidade, focada em casos extremos.	Identificação de conteúdos altamente agressivos.
INSULT	Linguagem ofensiva direcionada a indivíduos ou grupos.	Análise de ataques diretos em interações.
PROFANITY	Uso de palavrões e linguagem vulgar.	Proxy lexical de agressividade explícita.
THREAT	Indicação de ameaça ou intenção de violência.	Identificação de risco elevado em discurso.
IDENTITY_ATTACK	Ataques baseados em identidade (raça, gênero, religião, etc.).	Análise de discurso de ódio e discriminação.

5.2 Dados de polarização

Diante das limitações metodológicas e de acesso aos mecanismos internos da plataforma, esta pesquisa operacionaliza a polarização a partir de sinais discursivos e comportamentais observáveis no conteúdo produzido pelos usuários. Nesse contexto, títulos, descrições, transcrições e métricas de engajamento são utilizados como indicadores inferenciais para identificação de posicionamentos ideológicos e padrões aproximados de polarização. Essa abordagem está alinhada à literatura que trata polarização como um construto multifacetado, indireto e não diretamente observável em sistemas como o YouTube [8, 22].

Além disso, diferentes estudos propõem métricas distintas para mensuração da polarização, envolvendo proporções comportamentais, divergência entre usuários, análise de conteúdo e classificação ideológica. Essas formulações reforçam a ausência de consenso metodológico e evidenciam a natureza inferencial do fenômeno [8, 10, 22, 60, 64].

A Tabela 5.9 sintetiza diferentes estratégias de mensuração da polarização identificadas na literatura. Observa-se que as abordagens variam entre métricas comportamentais, classificações ideológicas, medidas de divergência entre usuários e *proxies* baseados em conteúdo textual ou afetivo. Em comum, essas propostas tratam a polarização como um fenômeno inferido, dependente do tipo de dado disponível e do recorte analítico adotado, não existindo uma métrica única amplamente aceita [10, 22]. O *SocialDash* adota LLM como mecanismo inferencial de polarização.

Tabela 5.9: Formulações de polarização na literatura

Ref	Tipo de medida	Fórmula / Regra	Variáveis	Interpretação
[64]	Proporção comportamental	$p_u = \frac{\text{interações em um grupo}}{\text{total de interações}}, \quad p_u \in [0,1]$	p_u: polarização do usuário	Polarização como concentração de comportamento
[10]	Classificação por limiar	Usuário polarizado se $\geq 95\%$ das interações em um grupo	interações, likes	Polarização como comportamento extremo
[10]	Distribuição ideológica	$\rho \in [-1,1]$ (com bimodalidade)	posição ideológica	Polarização como separação entre grupos
[8]	Score ideológico contínuo	$\text{Slant} = \frac{C-L}{C+L}$	C: conservador, L: liberal	Inclinação ideológica do conteúdo
[60]	Divergência entre usuários	$R_{pol} = \frac{1}{n^2 d} \sum_{k,l} \|\hat{x}_k - \hat{x}_l\|^2$	vetores de preferência	Polarização como discordância entre usuários
[60]	Variância agregada	$R_{pol} = \frac{1}{d} \sum \sigma_j^2$	variância dos itens	Polarização como dispersão estatística
[10, 39]	Proxy afetivo (NLP)	Sentimento: negativo, neutro, positivo	comentários / NLP	Polarização como conflito emocional
[22]	Meta-abordagem	Sem fórmula única (uso de múltiplos proxies)	múltiplas variáveis	Polarização como construto multidimensional

Em resposta às limitações apontadas na literatura [22], esta pesquisa utiliza LLMs como mecanismo inferencial capaz de integrar múltiplos sinais observáveis em um processo unificado de classificação. Por meio de prompts estruturados (Apêndice A.4), combinando metadados, transcrições e contexto discursivo, o modelo infere presença, direção e intensidade da polarização, atuando como integrador semântico de diferentes *proxies* analíticas.

Para operacionalizar essa abordagem, são considerados diferentes domínios de polarização — política, religião e ciência — em português e inglês, refletindo a natureza heterogênea do YouTube [4]. Parte-se do pressuposto de que canais com posicionamentos explícitos tendem a produzir conteúdos alinhados a esses vieses, permitindo utilizar tais conteúdos como *benchmark* empírico para validação inferencial dos modelos. Ainda que essas classificações possuam natureza probabilística, a consistência agregada dos resultados fornece indícios relevantes sobre padrões latentes de polarização discursiva.

Conforme apresentado na Tabela 5.10, os canais selecionados para este estudo foram escolhidos de forma a representar posições ideológicas previamente conhecidas, com inclinações declaradas nas dimensões de política, religião e ciência. Essa seleção foi realizada a partir de pesquisas na internet e também com o auxílio de modelos de linguagem, buscando garantir diversidade e representatividade nas categorias analisadas.

Tabela 5.10: Amostra de canais para benchmark de classificação ideológica via LLM

Canal	***Handle***	**Lang**	**Dim.**	**Polo**	**Subs.**	**Views**	**Videos**
Meteoro Brasil	@MeteoroBrasil	BR	Pol.	Esq.	1.1M	120M	1.2K
Brasil Paralelo	@brasilparalelo	BR	Pol.	Dir.	1.2M	200M	700
TV 247	@tv247	BR	Pol.	Esq.	1.3M	300M	20K
Jovem Pan News	@JovemPanNews	BR	Pol.	Dir.	3.5M	2B	60K
Democracy Now!	@democracynow	US	Pol.	Esq.	1.5M	500M	20K
Fox News	@FoxNews	US	Pol.	Dir.	10M	10B+	200K
Padre Paulo Ric	@padrepauloricardo	BR	Rel.	Cristão	1.5M	150M	2K
Umbanda EAD	@UmbandaEAD	BR	Rel.	Afro	250K	20M	500
Islam Brasil	@islambrasil	BR	Rel.	Islâmico	150K	10M	300
Pirula	@Pirula25	BR	Rel./Sci.	Ateu	1M	100M	1K
Nerdologia	@Nerdologia	BR	Sci.	Científico	3M	400M	1.2K
Ciência Todo Dia	@CienciaTodoDia	BR	Sci.	Científico	1.5M	150M	600
Globebusters	@Globebusters	US	Sci./Cons.	Consp.	100K	10M	500
The Highersid	@TheHighersideChats	US	Sci./Cons.	Consp.	300K	50M	800
Dr. John Campbell	@Campbellteaching	US	Saúde	Anti-vac	3M	500M	1K

Essa estratégia permite a construção de um benchmark de validação, no qual a orientação ideológica dos canais funciona como uma *proxy* de referência para avaliação das inferências realizadas pela LLM no nível dos vídeos. Embora nem todo vídeo reflita integralmente o posicionamento ideológico do canal, espera-se que conteúdos provenientes de canais com orientação explícita apresentem padrões consistentes de polarização, possibilitando verificar se o modelo é capaz de identificar corretamente tanto a presença quanto a direção dessa polarização. Dessa forma, a abordagem não assume determinismo absoluto, mas permite avaliar tendências agregadas e a robustez do modelo na captura de sinais ideológicos no conteúdo.

A coleta dos dados foi realizada por meio da API oficial do YouTube, seguindo uma estratégia temporal. Para cada canal, o intervalo entre a data de criação e o momento atual foi dividido em cinco *buckets* (intervalos de tempo) de tamanho equivalente. Em cada intervalo, foram coletados até 50 vídeos, respeitando o limite da API, resultando em um máximo teórico de 250 vídeos por canal. Entretanto, em alguns casos, esse limite não foi atingido. Isso ocorre devido a fatores como a escassez de vídeos em determinados períodos, limitações inerentes à API e características de caixa-preta da plataforma. Por exemplo, alguns canais podem não retornar vídeos durante a coleta, seja por inconsistências da API ou por configurações do próprio produtor de conteúdo, que pode restringir a disponibilidade de dados.

Após a coleta, foram realizadas análises estatísticas preliminares com o objetivo de caracterizar o dataset. Para cada canal, foram calculadas médias de métricas relevantes, incluindo visualizações, curtidas, comentários e duração dos vídeos (em segundos). Além disso, foram consideradas características textuais, como o comprimento médio dos títulos, das descrições e a quantidade média de *tags* por vídeo.

Esses indicadores presentes na Tabela 5.11 permitem uma visão agregada do comportamento dos canais e contribuem para a normalização das comparações entre eles.

A partir dessa base estruturada, torna-se possível avançar para análises mais profundas envolvendo modelos de linguagem. Considerando que cada canal possui uma inclinação previamente definida, espera-se que as respostas das LLMs, ao serem aplicadas sobre esse conteúdo, reflitam tendências alinhadas às dimensões e rótulos esperados. Ainda que nem todos os vídeos apresentem posicionamento explícito, a hipótese é de que, em nível agregado, haja convergência significativa.

Tabela 5.11: Amostra de canais e estatísticas descritivas do dataset coletado

@	Dim.	Polo	Proc.	Tot.	%	V	L	C	Dur.	TL	DL	Tg
@islmbra	relig	islamico	6	6	100	15790	48	0	165.5	23.8	22.7	8.5
@Globebu	cienc	conspir.	59	59	100	4476	15	1	170.8	56.3	584.7	27.1
@Ciencia	cienc	cientif.	232	1239	18.7	757615	69849	1380	469.8	36.5	852.6	24.3
@Nerdolo	cienc	cientif.	201	1109	18.1	295536	27867	751	598.6	46.8	1674.7	14.3
@TheHigh	cienc	conspir.	250	1387	18.02	15083	360	59	3967.2	66.0	1187.1	1.8
@Campbel	cienc	anti-ci.	250	3087	8.10	184928	7116	1742	1094.1	29.5	976.9	8.0
@padrepa	relig	cristao	184	7357	2.50	72173	6121	112	935.9	59.5	526.9	5.0
@brasilp	polit	direita	164	8199	2.00	364272	19157	848	1110.4	56.9	1007.7	19.2
@Meteoro	polit	esquerda	200	11900	1.68	112438	15909	1112	1115.8	47.3	712.8	5.1
@democra	polit	esquerda	28	18947	0.15	118569	3927	470	543.0	61.5	714.0	5.4
@FoxNews	polit	direita	100	137389	0.07	106576	2280	689	304.9	56.5	1211.2	19.1
@JovemPa	polit	direita	50	199244	0.03	28788	922	89	1978.0	75.9	830.0	20.4

Legenda: Proc. = vídeos processados; Tot. = total de vídeos do canal; % = percentual coletado; V = visualizações médias; L = curtidas médias; C = comentários médios; Dur. = duração média em segundos; TL = comprimento médio do título; DL = comprimento médio da descrição; Tg = quantidade média de tags.

Para viabilizar esse aprofundamento, foi incorporada uma nova dimensão analítica baseada no conteúdo textual das falas dos vídeos. Utilizou-se uma API de terceiros para extração automática de transcrições do YouTube[20], que retorna dados estruturados contendo segmentos textuais associados a informações temporais, como *start* e *duration*. Esse enriquecimento permite alimentar os *prompts* com maior densidade semântica, ampliando a capacidade inferencial das LLMs.

```
{text: ”muito bem Voltamos para tentar então”,
 start: 16.44,
 duration: 2.72}
```

Inicialmente, esses procedimentos são conduzidos em amostras reduzidas, possibilitando ajustes iterativos nos *prompts* e validação preliminar dos resultados, antes da escalabilidade para o dataset completo.

A validação dos dados foi realizada manualmente em amostras, comparando-se trechos iniciais e finais das transcrições com os respectivos vídeos, confirmando a consistência e aderência do conteúdo extraído.

Entretanto, para os objetivos desta pesquisa, optou-se por realizar um pré-processamento adicional, no qual todos os segmentos foram concatenados em um único campo textual contínuo. Adicionalmente, as informações temporais (*start* e *duration*) foram descartadas, uma vez que o foco da análise recai sobre o conteúdo semântico agregado do discurso, e não sobre sua estrutura temporal.

[20] `https://www.searchapi.io/api/v1/search`

Cada transcrição foi armazenada individualmente em formato JSON, utilizando como identificador o próprio *videoId*. Esse procedimento foi adotado com o objetivo de garantir rastreabilidade, reprodutibilidade e auditoria futura dos dados, sendo todo o conjunto de transcrições e scripts de processamento disponibilizado juntamente com o artefato da pesquisa.

Em termos de volume textual, as transcrições apresentaram uma dimensão significativa. Em média, cada vídeo resultou em um corpus equivalente a aproximadamente quatro a cinco páginas de texto em formato *Word*, o que corresponde, de forma aproximada, a um intervalo entre 2.000 e 3.000 palavras por vídeo, dependendo da duração e da densidade discursiva. Esse volume é naturalmente proporcional ao tempo de duração dos vídeos, o que abre possibilidades analíticas adicionais, como a investigação da relação entre duração e quantidade de conteúdo textual.

A incorporação das transcrições representa uma ampliação relevante do modelo analítico, permitindo a integração entre sinais estruturais (metadados) e conteúdo discursivo. Adicionalmente, além das variáveis previamente coletadas, o modelo passa a considerar também o texto integral das falas, possibilitando análises mais profundas por meio de modelos de linguagem.

A partir dessa base enriquecida, a próxima etapa consiste na construção e calibração de *prompts* para modelos de linguagem, com o objetivo de avaliar a capacidade desses modelos em identificar padrões ideológicos alinhados às dimensões previamente definidas (política, religião e ciência). A hipótese subjacente é que, embora nem todos os vídeos apresentem posicionamento explícito, o conjunto agregado de conteúdos de um canal tende a refletir sua inclinação ideológica predominante.

Para garantir maior controle experimental, os testes são conduzidos de forma incremental, partindo de amostras reduzidas (e.g., 10 vídeos com suas respectivas transcrições), avançando para conjuntos intermediários (e.g., 100 vídeos) e, posteriormente, para escalas maiores (e.g., 1.000 ou o dataset completo com 1.724 vídeos). Essa estratégia permite avaliar progressivamente o impacto das variáveis e identificar quais atributos contribuem para a inferência ideológica e quais introduzem ruído no processo.

Sob essa perspectiva, os dados são organizados em agrupamentos estruturados de atributos, denominados *labs*, que permitem a combinação controlada de diferentes conjuntos de informações na entrada dos modelos de linguagem. A construção dos *prompts* segue, portanto, uma lógica modular, na qual cada execução pode ativar diferentes *labs*, viabilizando a análise comparativa entre configurações distintas. Cada uma desta configurações

de atributos irá constituir um *cenário* experimental distinto, que será enumerado pra fins de consolidação dos resultados.

O conjunto mínimo obrigatório, denominado *base*, é sempre incluído em todas as execuções e contém apenas o título do vídeo e o idioma esperado, funcionando como uma linha de base (*baseline*) para avaliar o desempenho do modelo sob condições de informação reduzida. Utilizar apenas este conjunto de atributos constitui o *cenário 1*.

A partir desse núcleo, outros grupos de atributos podem ser adicionados de forma incremental. O grupo *video* inclui informações descritivas do vídeo, como descrição textual, categoria (convertida de identificador numérico para representação textual), palavras-chave (*tags*) e *hashtags* extraídas automaticamente da descrição. O uso destas informações constitui o *cenário 2*.

O grupo *stats* incorpora métricas quantitativas associadas ao vídeo, incluindo número de visualizações, curtidas, comentários e duração em segundos. Esses atributos funcionam como *proxies* de engajamento e relevância dentro da plataforma. Ao incorporar estas informações, obtemos nosso *cenário 3*.

O grupo *channel* adiciona informações contextuais do canal ao qual o vídeo pertence, como nome, descrição, palavras-chave, número de inscritos, visualizações totais e quantidade de vídeos publicados. Esses dados permitem capturar características mais amplas do produtor de conteúdo. O *cenário 4* corresponde ao uso deste grupo em adição aos grupo *video* e *base*.

Por fim, o grupo *transcript* inclui a transcrição completa do vídeo, previamente processada e concatenada em um único campo textual contínuo. Esse componente representa a dimensão discursiva mais rica do modelo, permitindo análises baseadas no conteúdo semântico das falas. Incorporando este grupo, constituímos três cenários distintos: no *cenário 5*, utilizamos *base* e *transcript*; no *cenário 6*, incluímos também o grupo *vídeo*, e no *cenário 7*, incluímos também o grupo *channel* (sendo excluído apenas o grupo *stats* neste caso.

A combinação desses grupos permite a realização de experimentos controlados, nos quais diferentes configurações de entrada são testadas para avaliar como o modelo de linguagem responde à presença ou ausência de determinados tipos de informação. Essa abordagem possibilita identificar quais atributos contribuem efetivamente para a inferência de padrões ideológicos e quais introduzem ruído ou redundância no processo, conforme Tabela 5.12.

Essa estratégia experimental é particularmente relevante no contexto desta pesquisa, pois permite analisar não apenas o desempenho do modelo, mas também o impacto relativo de diferentes dimensões informacionais (estrutural, contextual e discursiva) na tarefa de classificação ideológica.

Tabela 5.12: Configuração dos grupos de atributos utilizados nos prompts (*labs*)

Grupo (Lab)	**Atributos incluídos**
Base	Título do vídeo; Idioma
Video	Descrição; Categoria (string); Tags; Hashtags
Stats	Visualizações; Curtidas; Comentários; Duração
Channel	Nome do canal; Descrição; Keywords; Inscritos; Views totais; Total de vídeos
Transcript	Transcrição completa concatenada

A etapa seguinte da metodologia consiste na definição dos *prompts* utilizados para interação com os modelos de linguagem. Devido à complexidade e extensão desses prompts, optou-se por apresentar sua estrutura de forma modular, destacando seus principais componentes e diretrizes analíticas. Os prompts completos encontram-se no capitulo A.

O modelo é inicialmente instruído a assumir o papel de um classificador acadêmico, com foco na análise de conteúdo. Essa definição é fundamental para orientar o comportamento do modelo em direção a uma postura analítica e não opinativa:

```
"Você é um classificador acadêmico especializado em
análise de conteúdo."
```

Em seguida, é especificada a dimensão analítica de interesse (política, religião ou ciência), bem como o conjunto de dados fornecido ao modelo, derivado da configuração dos *labs*:

```
"A dimensão analítica esperada é: {dimensao}"
"Os dados disponíveis são: {labs}"
```

Um aspecto central do prompt é a instrução explícita de que a classificação deve ser realizada sobre o vídeo, e não sobre o canal, evitando viés de generalização:

```
"Classifique o vídeo, não o canal."
```

A classificação é restrita a um conjunto fechado de categorias (polos), previamente definidos conforme a dimensão analisada. Esses polos funcionam como classes discretas de saída:

- **Política**: esquerda, centro, direita, indefinido
- **Religião**: cristão, afro-brasileiro, islâmico, ateu-secular, indefinido
- **Ciência**: científico, conspiratório, anti-ciência, indefinido

Além das instruções básicas, o prompt incorpora um conjunto de diretrizes metodológicas que orientam o processo decisório do modelo. Essas diretrizes incluem:

- Hierarquia de evidência: priorização de sinais discursivos sobre metadados superficiais;
- Regras de decisão: definição de critérios mínimos para classificação;
- Tratamento de ambiguidade: possibilidade de classificar como "indefinido" em casos inconclusivos;
- Interpretação dos campos: orientação sobre como utilizar cada tipo de dado disponível.

Outro elemento relevante é a exigência de evidências explícitas, obrigando o modelo a justificar suas classificações com base em trechos ou sinais identificáveis nos dados fornecidos.

Por fim, o modelo é instruído a retornar sua resposta em formato JSON estruturado, permitindo posterior processamento automatizado. A estrutura de saída inclui:

```
{
  "dimension": "...",
  "predicted_label": "...",
  "confidence": "...",
  "ambiguous": "...",
  "evidence": "...",
  "sentiment": "..."
}
```

O campo *predicted_label* representa o resultado principal da classificação, enquanto os demais campos auxiliam na interpretação do resultado, incluindo o grau de confiança, a presença de ambiguidade, evidências utilizadas e análise de sentimento.

Essa abordagem permite não apenas avaliar a acurácia da classificação, mas também compreender o processo inferencial do modelo, possibilitando a análise de padrões de decisão, inconsistências e potenciais vieses. A utilização de diferentes configurações de entrada (via *labs*) torna possível investigar, de forma sistemática, como distintos tipos de informação influenciam o desempenho na tarefa de inferência ideológica.

Nesse contexto, a avaliação experimental foi iniciada com um subconjunto reduzido de dados ($n = 10$ vídeos), selecionado a partir do conjunto previamente coletado e enriquecido com transcrições completas. Essa etapa possui caráter exploratório e de calibração do pipeline, permitindo validar a estrutura dos *prompts* e observar o comportamento inicial do modelo sob diferentes configurações de entrada, tendo como referência a linha de base estabelecida.

Conforme descrito anteriormente, todos os registros utilizados nesta etapa já possuem transcrição disponível, garantindo que, em níveis mais avançados do experimento, o conteúdo discursivo completo possa ser incorporado. Consequentemente, neste primeiro teste, utiliza-se apenas a configuração *baseline* (base), composta exclusivamente pelo título do vídeo e pelo idioma.

Essa configuração representa o nível mínimo de informação fornecida ao modelo, funcionando como um controle experimental. A expectativa é que o desempenho seja limitado, uma vez que sinais discursivos mais profundos não estão disponíveis.

A Tabela 5.13 apresenta uma amostra dos registros utilizados no experimento.

Tabela 5.13: Amostra do dataset utilizado no experimento inicial ($n = 10$)

VideoId	Título	L	Handle	Dim.	Polo	Pred.	Hit	C	Amb	Sent.	S_I
1gD42J9PEhY	Ryan Patrick Blu	US	@TheHigher	cien	conspirat.	indefinido	0	0.7	1	indefinido	0.0
7PYgBLazEU	John Perkins	US	@TheHigher	cien	conspirat.	indefinido	0	0.8	0	neutro	0.1
8x360wjVFXU	Papa Francisco	BR	@padrepaulo	relig	cristao	cristao	1	0.8	0	neutro	0.1
EibdNFLJkkc	Daniel Immerwahr	US	@TheHigher	cien	conspirat.	indefinido	0	0.7	1	indefinido	0.0
OSwBRZJZyC	Clip 5: The Ride	US	@Globebusters	cien	conspirat.	indefinido	0	0.8	0	neutro	0.1
Rrhipnouil	MAGA/Trump	BR	@JovemPan	pol	direita	indefinido	0	0.8	0	neutro	0.1
a6i7nOgphzU	Matteo Messina	BR	@Nerdologia	cien	cientifico	indefinido	0	0.8	0	neutro	0.1
o9BymdD7Z4	Great interview	US	@Campbell	cien	anti-cienc.	indefinido	0	0.6	1	indefinido	0.0
r08DZSXG1nQ	Ansiedade	BR	@Nerdologia	cien	cientifico	indefinido	0	0.7	1	neutro	0.1
vslgKMFqITQ	Memória de San	BR	@padrepaulo	relig	cristao	indefinido	0	0.6	1	indefinido	0.0

Seguindo uma estratégia incremental de análise, foram realizados múltiplos experimentos combinando progressivamente diferentes conjuntos de dados (*labs*), incluindo metadados do vídeo, estatísticas de engajamento, dados do canal e transcrições. O objetivo foi avaliar a contribuição relativa de cada grupo de informações para a inferência de orientação discursiva.

Os experimentos foram executados sequencialmente e serão explicados nos próximos parágrafos. A Tabela 5.14 apresenta o resultado consolidado dos 7 cenários delimitado para a amostra inicial, com $n = 10$ vídeos.

Os resultados do cenário 1 confirmam a hipótese esperada para o baseline. A taxa de acerto observada (Hit = 0.10) é significativamente baixa, indicando que o modelo apresenta dificuldade em inferir a orientação discursiva com base apenas no título do vídeo. Apesar de a confiança média (Conf = 0.73) e a confiança nos acertos (C_H = 0.80) serem relativamente elevadas, observa-se também um alto nível de desvio (Dev = 0.90), sugerindo que o modelo tende a produzir respostas confiantes, porém frequentemente desalinhadas com o rótulo esperado. Esse comportamento reforça a limitação do conjunto *base*, evidenciando a necessidade de incorporação de informações adicionais para melhorar a capacidade inferencial do modelo.

Tabela 5.14: Resultados com transcrição, metadados do vídeo e dados do canal

i	B	V	S	C	T	Hit	Conf	C_H	C_D	Amb	Dev	S+	S-	S_I
1	x					0.10	0.73	0.80	0.72	0.50	0.90	0.00	0.00	0.06
2	x	x				0.50	0.83	0.84	0.82	0.10	0.50	0.20	0.10	0.215
3	x	x	x			0.50	0.815	0.83	0.80	0.10	0.50	0.30	0.10	0.275
4	x	x		x		0.50	0.815	0.86	0.77	0.10	0.50	0.30	0.10	0.23
5	x				x	0.60	0.88	0.90	0.85	0.00	0.40	0.40	0.20	0.40
6	x	x			x	0.60	0.88	0.90	0.85	0.00	0.40	0.40	0.20	0.405
7	x	x		x	x	0.60	0.88	0.90	0.85	0.00	0.40	0.40	0.20	0.395

Legenda da Tabela 5.14: i = cenário; B = base (título e idioma); V = dados do vídeo; S = estatísticas do vídeo; C = dados do canal; T = transcrição; Hit = taxa de alinhamento entre rótulo previsto e rótulo esperado; Conf = confiança média; C_H = confiança média nos acertos; C_D = confiança média nos desvios; Amb = taxa de ambiguidade; Dev = taxa de desvio intra-canal; S+ = proporção de sentimento positivo; S- = proporção de sentimento negativo; S_I = intensidade média do sentimento.

Esses achados reforçam a necessidade de incorporação de informações adicionais — como metadados descritivos e conteúdo textual mais rico — para aprimorar a capacidade de inferência.

Os resultados no cenário 2 mostram que a inclusão dos metadados do vídeo produziu melhora expressiva em relação ao baseline. A taxa de alinhamento subiu de 0.10 para 0.50, enquanto a ambiguidade caiu de 0.50 para 0.10. Esse resultado sugere que descrição, categoria e tags funcionam como *proxies* semânticos relevantes, oferecendo ao modelo mais elementos para inferir o enquadramento discursivo do vídeo.

Para os experimentos com $n = 100$ vídeos, delimitamos novamente um conjunto de cenários, utilizando diferentes combinações possíveis dos conjuntos de dados (*labs*), incluindo metadados do vídeo, estatísticas, dados do canal e transcrições. O objetivo foi verificar a consistência dos padrões observados anteriormente em uma amostra ampliada. A Tabela 5.15 apresenta os resultados consolidados para todas as combinações testadas.

Tabela 5.15: Resultados do experimento com $n = 100$ por combinação de dados

i	B	V	S	C	T	Hit	Conf	C_H	C_D	Amb	Dev	S+	S-	S_I
1	x					0.19	0.7215	0.8395	0.6938	0.52	0.81	0.08	0.05	0.136
2	x	x				0.45	0.817	0.8556	0.7855	0.15	0.55	0.20	0.16	0.259
3	x	x	x			0.46	0.818	0.8467	0.7935	0.11	0.54	0.17	0.14	0.2503
4	x	x		x		0.46	0.8315	0.8696	0.7991	0.14	0.54	0.19	0.14	0.2505
5	x	x	x	x		0.47	0.835	0.8649	0.8085	0.12	0.53	0.25	0.14	0.26
6	x				x	0.56	0.8715	0.8732	0.8693	0.03	0.44	0.24	0.25	0.3497
7	x	x			x	0.56	0.8755	0.8804	0.8693	0.03	0.44	0.23	0.25	0.3506
8	x	x	x		x	0.53	0.8745	0.8821	0.866	0.03	0.47	0.24	0.27	0.3551
9	x	x		x	x	0.56	0.881	0.8866	0.8739	0.03	0.44	0.25	0.25	0.352
10	x	x	x	x	x	0.55	0.873	0.8791	0.8656	0.03	0.45	0.25	0.25	0.3472

Os resultados evidenciam uma hierarquia relativamente estável de contribuição informacional entre os diferentes conjuntos de dados analisados.

O baseline apresentou desempenho limitado, indicando que título e idioma, isoladamente, possuem baixo poder discriminativo para inferência discursiva.

A inclusão dos metadados do vídeo produziu o primeiro ganho substancial, sugerindo que descrição, categoria, *tags* e *hashtags* funcionam como importantes *proxies* semânticos do conteúdo.

Por outro lado, métricas quantitativas de engajamento apresentaram impacto reduzido e, em alguns cenários, potencialmente ruidoso, indicando baixa capacidade explicativa direta para classificação ideológica.

O principal ganho ocorreu com a introdução da transcrição completa dos vídeos, que elevou significativamente a taxa de alinhamento, reduziu a ambiguidade e aumentou a estabilidade das inferências.

O modelo foi então aplicado ao dataset completo, originalmente composto por 1726 vídeos. Durante a etapa de extração de transcrições, ocorreram 94 falhas, resultando em um conjunto final de aproximadamente 1630 vídeos válidos para análise. A maior parte dos erros esteve associada à indisponibilidade de transcrições (*"There is no available translation for this video"*), havendo poucos casos de falhas técnicas, o que indica robustez geral do pipeline de coleta e processamento. Todos os erros foram registrados para fins de rastreabilidade.

Considerando os resultados das etapas exploratórias anteriores, observou-se que os maiores ganhos inferenciais estavam concentrados nos cenários 1, 2 e 3. Dessa forma, o processamento do dataset ampliado foi direcionado prioritariamente a essas três configurações, permitindo avaliar a estabilidade dos padrões observados em maior escala.

Após a inclusão das transcrições, os ganhos adicionais decorrentes da incorporação de outras variáveis tornaram-se marginais, indicando um possível fenômeno de saturação informacional. Esse comportamento sugere que a principal capacidade inferencial do modelo está concentrada no conteúdo discursivo explícito presente nas falas dos vídeos, enquanto variáveis complementares tendem a atuar como sinais auxiliares de menor impacto incremental.

Além disso, o comportamento relativo entre os conjuntos de variáveis permaneceu consistente tanto nas simulações exploratórias ($n = 10$ e $n = 100$) quanto no dataset ampliado ($n = 1630$). Observa-se que os cenários responsáveis pelos maiores saltos na taxa de acerto mantiveram a mesma hierarquia de contribuição ao longo dos diferentes experimentos, sugerindo estabilidade e robustez dos padrões identificados.

De forma geral, os resultados obtidos nos três experimentos podem ser considerados satisfatórios, evidenciando padrões consistentes entre as diferentes configurações testadas. Observa-se que a inclusão da transcrição representa o principal ganho informacional, com impacto direto na taxa de acerto, na confiança e na redução da ambiguidade. A combinação de transcrição com metadados do vídeo produz ganhos adicionais marginais, indicando que grande parte da capacidade inferencial do modelo está concentrada

no conteúdo textual completo. Os resultados dessa ultima analise para o conjunto total de registros, estão expressos nas Tabelas 5.16, 5.17 e 5.18.

Os resultados mostram que o desempenho do modelo varia significativamente entre as dimensões analisadas. A classificação em Religião apresentou os maiores índices de acerto, seguida por Ciência, enquanto Política obteve os resultados mais modestos. Além disso, observa-se que as configurações contendo transcrição alcançaram desempenho superior na maioria dos cenários, reforçando seu papel como principal fonte de informação para a inferência discursiva.

Tabela 5.16: Resultados dos experimentos com dataset completo (n = 1630)

i	B	V	S	C	T	Hit	Conf	C_H	C_D	Amb	Dev	S+	S-	S_I
1	x	x				0.4227	0.8152	0.8494	0.7902	0.1669	0.5773	0.2025	0.1074	0.2261
2	x				x	0.5221	0.8656	0.8788	0.8510	0.0479	0.4779	0.2613	0.1871	0.3258
3	x	x			x	0.5331	0.8718	0.8831	0.8588	0.0393	0.4669	0.2528	0.1859	0.3217

Tabela 5.17: Desempenho agregado por dimensão — dataset completo

Dimensão	Hit	Confiança
Ciência	0.5059	0.8511
Política	0.2861	0.8282
Religião	0.9982	0.9125

Tabela 5.18: Desempenho por dimensão e configuração (Labs × Dimensão)

Grupo	Hit	Confiança
base + transcript | ciência	0.5382	0.8652
base + transcript | política	0.3204	0.8415
base + transcript | religião	1.0000	0.9339
base + transcript + video | ciência	0.5318	0.8730
base + transcript + video | política	0.3670	0.8477
base + transcript + video | religião	1.0000	0.9325
base + video | ciência	0.4478	0.8150
base + video | política	0.1709	0.7955
base + video | religião	0.9946	0.8712

Ressalta-se, entretanto, que não se trata de um modelo supervisionado no sentido estrito, uma vez que não há rótulos objetivos e verificáveis para o conteúdo interno dos vídeos e canais. A própria noção de polarização e ideologia envolve elementos subjetivos, frequentemente associados a construções discursivas e interpretações contextuais. Nesse sentido, os resultados devem ser compreendidos como aproximações inferenciais, e não como classificações definitivas.

Ainda assim, os experimentos fornecem subsídios relevantes para a operacionalização dos construtos derivados dos dados do YouTube, permitindo avançar na construção de modelos mais específicos para inferência de polarização. Esses resultados orientam as próximas etapas da pesquisa, nas quais serão exploradas abordagens mais refinadas para modelagem dessas dimensões.

5.3 Dados de monetização

A análise da monetização no YouTube apresenta desafios metodológicos significativos, sobretudo em função da ausência de acesso direto aos dados financeiros da plataforma. Diferentemente de métricas como visualizações, comentários e inscrições, os valores de receita, CPM e taxas efetivas de monetização são internos ao sistema e não são disponibilizados ao pesquisador externo. Nesse sentido, a monetização configura-se como uma variável latente, cuja observação direta é inviável, exigindo abordagens inferenciais baseadas em sinais observáveis.

A literatura tem abordado essa limitação por meio da utilização de *proxies* e indicadores indiretos. Como apresentado na Tabela 5.19, diferentes estudos empregam métricas como visualizações, retenção, engajamento e presença de anúncios para inferir a capacidade de geração de receita dos canais. Evidências empíricas indicam que a monetização está fortemente associada à dinâmica de atenção e engajamento dos usuários.

O trabalho de [67] demonstra que a relação entre número de inscritos e receita é limitada, destacando que o fator crítico reside na conversão de inscritos em visualizações efetivas. De forma complementar, [51] identifica que o número de visualizações possui efeito positivo e estatisticamente significativo sobre a receita, enquanto [66] evidencia que a interação por meio de comentários constitui um componente relevante do engajamento.

Variáveis como frequência de publicação e duração média dos vídeos são consideradas, uma vez que influenciam diretamente a distribuição algorítmica e o tempo de exibição (*watch time*), elementos centrais na lógica de monetização da plataforma.

Tabela 5.19: Benchmark de abordagens de monetização na literatura

Artigo	Tipo de Monetização	Proxy Utilizado	Valor em Dinheiro	Método
[51]	Monetização nativa + ecossistema	CPM, views, retenção, YPP	Parcial	Revisão e modelagem conceitual
[43]	Ads + monetização off-platform	Presença de anúncios, links externos	Não	Coleta e classificação de anúncios
[46]	Demonetização	Status monetizado/demonetizado	Não	Machine Learning
[3]	Monetização algorítmica	Transições de monetização e crescimento	Não	Análise longitudinal
[19]	Monetização alternativa	Links (Patreon, cripto, afiliados)	Sim	Scraping + bases externas
[62]	Mudança de modelo econômico	Migração ads → patronage	Não	Análise temporal
[20]	Estratégias econômicas	URLs em descrições	Não	Large-scale crawling
[17]	Receita direta	Views, comentários, engajamento	Sim	Modelo estatístico
[67]	Monetização por atenção	Views vs inscritos	Não	Análise quantitativa
[14]	Monetização vs toxicidade	Monetization cues, comentários	Não	NLP + proxies
[5]	Economia da atenção	Views, audiência	Não	Base teórica
[16]	Viralidade → valor	Emoção, compartilhamento	Não	Experimento comportamental
[69]	Entrada na monetização	Watch hours, inscritos	Não	Estudo de caso
[47]	Monetização exploratória	Spam, links, tráfego	Não	Análise de conteúdo
[71]	Demonetização econômica	Advertiser-friendly	Parcial	Análise qualitativa

Diante desse cenário, a monetização pode ser entendida como uma função indireta de variáveis observáveis, tais como visualizações, tempo de exibição e engajamento (comentários, curtidas e interações), associadas à economia da atenção e à dinâmica de distribuição algorítmica de conteúdo. Dessa forma, a monetização é operacionalizada nesta pesquisa como um índice composto baseado na convergência desses sinais.

Para complementar essa abordagem, são utilizadas ferramentas externas como VidIQ e SocialBlade[21], que estimam a monetização a partir de métricas públicas da plataforma e de modelos proprietários de inferência. Assim como o próprio YouTube, essas ferramentas também operam parcialmente como sistemas *black-box*, uma vez que seus critérios exatos de cálculo não são integralmente divulgados. Ainda assim, elas fornecem estimativas amplamente utilizadas por criadores de conteúdo e estudos exploratórios como indicadores aproximados de monetização *on-platform*.

No caso do VidIQ, parte dessa inferência é reforçada pelo fato de a plataforma atuar

[21] `www.socialblade.com`

como parceira certificada do YouTube, tendo acesso a dados agregados que complementam as métricas disponíveis na API pública. Entretanto, os valores retornados devem ser interpretados como estimativas analíticas e não como medidas exatas de receita.

Com o objetivo de avaliar a confiabilidade dessas *proxies*, foi realizada uma triangulação empírica utilizando dados reais obtidos em plataformas de negociação de canais, como o Flippa[22], nas quais proprietários disponibilizam capturas do YouTube Analytics contendo evidências de monetização. Esses dados permitem comparar a receita observada com as estimativas fornecidas pelas ferramentas externas, possibilitando uma análise da consistência entre monetização real e monetização inferida.

Além disso, evidências mostram que criadores frequentemente utilizam estratégias externas de monetização, como links afiliados, redirecionamento para outras plataformas, Patreon e venda de produtos [20]. Nesse contexto, a presença de URLs externas nas descrições dos vídeos constitui um indicador relevante de monetização *off-platform* [19], sendo incorporada como variável complementar na análise. A Tabela 5.20 apresenta o resultado do benchmark empírico, incluindo canais em diferentes estados de atividade, permitindo avaliar o comportamento das estimativas em cenários distintos.

Tabela 5.20: Benchmark empírico de monetização on-platform e estimativas externas

i	Canal	País	Id.	Subs.	Views	Hist.	Recente	VidIQ	SB	Tend.	Status
1	EuSou	BR	3a	667K	28.7M	$290/m	2026: 0	$0	$0	▬	Inat.
2	SanYa	RU	5a	347K	182M	$3K/m	jan/26: $2.2K	$5K	$2.5–40K	▟	Ativo
3	Travel	US	5a	26K	8.4M	$3.1K/m	mar/26: $3.7K	$1K	$100–1.5K	▟	Semi
4	Swish	US	6a	109K	31M	$10K/m	mar/26: $663	$1K	$94–1.5K	▙	Ativo
5	Biblio	PE	1a	190K	11M	$7.2K/m	dez/25: queda	$0	$0	▙	Inat.
6	Docum.	BR	4a	2.6M	1.22B	$565/m	nov/25: $478	$0	$0	▙	Inat.
7	Hist.	BR	2a	38K	15M	$2.2K/m	out/25: $240	$41	$3–48	▙	Inat.
8	**Jovem**	**BR**	**15a**	**163K**	**26.8M**	**híbr.**	**mar/26: R$2–6K**	**$411**	**$133–2.1K**	▟	**Ativo**

[22] www.flippa.com

Descrição das colunas do benchmark:

- **Id.:** idade do canal em anos.
- **Hist.:** receita média histórica reportada em plataformas externas, dashboards ou declarações públicas, normalmente associada a períodos anteriores de maior atividade do canal.
- **Recente:** última receita mensal observada nos registros disponíveis, representando o estado mais próximo do momento atual do canal.
- **VidIQ:** estimativa atual de monetização mensal baseada em métricas de desempenho do canal, utilizada como proxy de monetização *on-platform*.
- **SB:** intervalo de monetização mensal do SocialBlade.
- **Tend.:** representação visual simplificada da trajetória recente de monetização, indicando crescimento, estabilidade ou queda.
- **Status:** classificação qualitativa do estado atual do canal em termos de atividade: ativo, semi-ativo ou inativo.

A análise dos dados apresentados na Tabela 5.20 evidencia padrões consistentes no comportamento das ferramentas de estimativa. Em canais inativos ou em declínio, as estimativas tendem rapidamente a zero, independentemente da existência de monetização histórica significativa, indicando que tais ferramentas capturam predominantemente o fluxo recente de receita.

Por outro lado, em canais ativos ou semi-ativos, observa-se maior proximidade entre valores estimados e observados, ainda que com variações. Esse resultado reforça a hipótese de que a monetização no YouTube é fortemente dependente da atividade recente do canal, especialmente da frequência de publicação, volume de visualizações e engajamento.

Um caso ilustrativo é o canal “Jovem Tranquilão”, no qual o próprio criador divulga publicamente dados detalhados de monetização, incluindo receitas provenientes tanto da plataforma (AdSense) quanto de fontes externas. A comparação entre esses valores e as estimativas fornecidas pelo VidIQ revela uma correspondência consistente no componente *on-platform*, sugerindo que a ferramenta captura adequadamente a parcela da monetização diretamente associada à dinâmica interna do YouTube.

Sob essa perspectiva, ferramentas como VidIQ devem ser interpretadas não como estimadores absolutos de receita, mas como indicadores relativos da monetização *on-platform*, refletindo a dinâmica de atenção e distribuição algorítmica do conteúdo. Observa-se ainda que, embora baseadas em modelos independentes, as estimativas de VidIQ e SocialBlade tendem a convergir para a mesma ordem de grandeza, sugerindo a captura de padrões consistentes da monetização na plataforma.

Ao mesmo tempo, essas ferramentas apresentam limitações estruturais, especialmente por não capturarem receitas provenientes de estratégias *off-platform*. Adicionalmente, há diferenças operacionais relevantes, como a impossibilidade de coleta automatizada no SocialBlade, enquanto o VidIQ permite extração sistemática de dados, justificando sua adoção nesta pesquisa.

Em síntese, a monetização no YouTube deve ser compreendida como uma variável inferida a partir de sinais observáveis distribuídos em diferentes níveis. Os resultados indicam que a unidade do vídeo constitui a principal fonte de informação para capturar padrões de monetização, especialmente no contexto de estratégias *off-platform*. Adicionalmente, esta pesquisa integra métricas *on-platform* e *off-platform* por meio de *proxies*, reconhecendo o caráter opaco da plataforma e a impossibilidade de acesso direto aos dados internos.

5.3.1 URLs externas nas descrições de vídeos e canais

Foi realizada uma análise exploratória das URLs presentes nas descrições dos vídeos e dos canais do dataset (1726 registros), com o objetivo de identificar sinais de monetização externa alternativa, redes sociais, encurtadores e outras formas de direcionamento externo, conforme Tabela 5.21.

A inspeção das descrições dos canais revelou baixa incidência de URLs, sendo utilizadas predominantemente para fins institucionais, como apresentação, identidade editorial e informações de contato. Apenas alguns casos específicos apresentaram links diretos, como *@democracynow*, *@globebusters* e *@campbellteaching*. Em contraste, as descrições dos vídeos exibem maior densidade de URLs, indicando que o direcionamento externo ocorre principalmente no nível do vídeo.

Esse padrão sugere uma distinção funcional relevante: enquanto a descrição do canal atua como espaço institucional, a descrição dos vídeos funciona como mecanismo dinâmico de conversão, divulgação e monetização alternativa.

Tabela 5.21: URLs externas nas descrições dos vídeos: médias por canal

Handle	Lang	Vídeos	URLs/Víd.	Social	Money	Short	Other	Money
@MeteoroBrasil	BR	200	4.53	2.19	0.52	0.87	0.96	0.94
@CienciaTodoDia	BR	232	6.45	2.85	0.24	0.40	2.97	0.55
@TheHighersideChats	US	250	2.88	0.53	0.09	0.05	2.21	0.34
@Campbellteaching	US	250	1.78	0.22	0.05	0.02	1.49	0.02
@Nerdologia	BR	201	22.50	4.21	0.00	14.77	3.51	0.16
@brasilparalelo	BR	164	3.98	1.90	0.00	1.08	1.01	0.57
@JovemPanNews	BR	50	7.18	4.96	0.00	0.00	2.22	0.02
@democracynow	US	28	1.36	0.00	0.00	0.00	1.36	0.68
@FoxNews	US	100	7.46	1.98	0.00	0.50	4.98	0.00
@padrepauloricardo	BR	184	0.39	0.02	0.00	0.36	0.01	0.02
@islambrasil	BR	6	0.00	0.00	0.00	0.00	0.00	0.00
@Globebusters	US	59	3.02	1.92	0.00	0.00	1.10	0.00

Legenda: URLs/Víd. = média de URLs encontradas por vídeo; Social = média de URLs classificadas como redes sociais ou divulgação; Money = média de URLs associadas a monetização externa direta, como PayPal, Padrim, Patreon, Apoia.se, Buy Me a Coffee, Ko-fi, PicPay, Mercado Pago e similares; Short = média de URLs encurtadas, como Bitly, TinyURL, t.co, goo.gl, Linktree e similares; Other = média de URLs não classificadas nas categorias anteriores; Money = média de ocorrências textuais relacionadas a monetização, como Pix, doação, donate, support, contribuição, financiamento coletivo, bitcoin, crypto, wallet e termos similares.

Os dados indicam que alguns canais utilizam intensamente a descrição dos vídeos para inserir links externos. O caso de @Nerdologia se destaca pela alta média de URLs por vídeo, especialmente associadas a encurtadores. Já o canal @MeteoroBrasil apresenta presença expressiva de links sociais e sinais textuais de monetização, coerente com a presença recorrente de campanhas de financiamento coletivo nas descrições. Canais como @JovemPanNews e @FoxNews apresentam maior volume de links, mas com predominância de URLs sociais ou genéricas, não necessariamente associadas à monetização direta.

Assim, a análise sugere que a monetização externa alternativa não aparece prioritariamente na descrição institucional dos canais, mas nas descrições dos vídeos. Esse resultado é metodologicamente relevante, pois indica que a unidade de análise adequada para capturar estratégias de monetização externa é o vídeo, e não apenas o canal.

Cabe destacar que os resultados desta etapa dependem diretamente dos conjuntos de

domínios e padrões textuais definidos manualmente para classificação das URLs. Para redes sociais, foram considerados domínios como Facebook, Instagram, Twitter/X, YouTube, TikTok, LinkedIn, Telegram, WhatsApp e Discord. Para monetização externa, foram incluídos domínios como PayPal, Patreon, Padrim, Apoia.se, Buy Me a Coffee, Ko-fi, PicPay, Catarse, Benfeitoria, Substack, Stripe, Mercado Pago e PagSeguro, totalizando aproximadamente 15 serviços de referência. Também foram identificados encurtadores de URL, como Bitly, TinyURL, t.co, goo.gl, Cutt.ly, Rebrandly, Linktree e outros.

Esses encurtadores foram mantidos como categoria própria, pois podem redirecionar posteriormente para redes sociais, serviços de monetização ou páginas genéricas. Assim, os valores apresentados devem ser interpretados como estimativas conservadoras, uma vez que a resolução das URLs encurtadas poderá redistribuir parte desses links entre as demais categorias.

Além dos domínios, foram utilizados padrões textuais associados a monetização, contemplando termos em português e inglês, como ”Pix, chave Pix, doação, doar, contribua, financiamento coletivo, donate, donation, support, membership, crowdfunding, bitcoin, crypto, wallet” e termos similares. Portanto, os números apresentados não representam uma medida absoluta, mas uma aproximação operacional baseada nos critérios definidos nesta etapa da pesquisa.

Em síntese, a análise realizada indica que a monetização no YouTube deve ser compreendida a partir de sinais observáveis distribuídos em diferentes níveis, sendo a unidade do vídeo a principal fonte de informação para capturar estratégias de monetização, especialmente no contexto de monetização alternativa (*off-platform*). Os resultados mostram que, enquanto a descrição dos canais possui caráter predominantemente institucional, a descrição dos vídeos concentra a maior parte dos mecanismos de conversão, engajamento e direcionamento externo. Dessa forma, as análises desta tese adotam o vídeo como unidade central para inferência tanto de monetização quanto de outras dimensões analíticas. Adicionalmente, o trabalho contempla simultaneamente métricas de monetização *on-platform* e *off-platform*, reconhecendo, contudo, que ambas são tratadas como aproximações baseadas em *proxies*, dado o caráter opaco (“caixa preta”) da plataforma e a ausência de acesso direto aos dados internos do YouTube.

5.4 Prova de conceito e primeiros testes

Nesta fase, o objetivo é demonstrar a viabilidade prática do modelo proposto por meio da implementação de uma prova de conceito funcional. Para isso, foi desenvolvido um

script executável em ambiente Web, responsável por integrar e acionar os diversos módulos do sistema, incluindo a coleta de dados via API do YouTube, a extração de transcrições, a análise de toxicidade por meio da Perspective API, a estimativa de monetização via VidIQ e a classificação de polarização utilizando modelos de linguagem.

A escolha por um ambiente Web não é arbitrária: ela permite observar diretamente o tempo de resposta do sistema sob condições reais de uso, um aspecto crítico na avaliação de usabilidade. Em sistemas interativos, o feedback rápido e contínuo ao usuário é considerado um fator central na percepção de qualidade da interação. Nesse contexto, o tempo de processamento observado — aproximadamente dois minutos — evidencia uma limitação importante do protótipo inicial, especialmente em cenários de uso exploratório e interação contínua.

A Tabela 5.22 apresenta os resultados obtidos a partir da execução do script sobre um vídeo de natureza política. Observa-se que o sistema foi capaz de coletar, estruturar e persistir dados referentes ao vídeo, canal e comentários, além de calcular métricas derivadas como toxicidade, polarização e monetização. Esse resultado demonstra a coerência entre o modelo conceitual proposto e sua implementação prática, validando a integração das diferentes fontes de dados e serviços externos.

Em particular, o alto tempo de processamento está associado à natureza sequencial das chamadas a APIs externas — especialmente no cálculo de toxicidade — o que reforça a necessidade de estratégias de otimização, como paralelização, cache e processamento assíncrono.

Tabela 5.22: Prova de Conceito do SocialDash – Execução do Script PMT já no formato de uma página Web.

Vídeo					
ID YouTube	**Título**	**Publicado em**	**Views**	**Likes**	**Comentários**
LDG8GFqVnKQ	Conheça o maior esquema de corrupção do planeta	2022-10-19	80243	4040	689

Canal					
ID YouTube	**Nome**	**Criado em**	**Inscritos**	**Views**	**Vídeos**
UCvO2BExvkAbGMsTGnEnI_Ng	Lula	2018-06-05	1600000	517219069	4849

Comentários (Amostra)					
ID	**Usuário**	**Data**	**Texto**	**Likes**	**Toxicidade**
UgyVWSZh...	@IMPERADORIVS	2023-05-05	Quem diz é a “alma mais honesta do Brasil”...	33	0.0444
Ugwg-8Op...	@LiphNeo	2022-10-29	E o PETROLÃO??? 30 bilhões...	23	0.3604
UgwUMzT...	@marcos7084	2022-11-28	Acuse-os do que vc faz...	161	0.0705
UgxPFC...	@LucasSantos	2024-04-16	De acordo com o departamento...	27	0.0153
Ugz5Dt...	@lucianaviana	2022-10-24	QUEREMOS O VÍDEO COMPLETO...	266	0.0911

Análise PMT								
Palavras	**Tox. média**	**Tox. máx**	**Taxa alta**	**Categoria**	**Score**	**Polo**	**Confiança**	**VidIQ (US$)**
309	0.0897	0.602	0	política	1	esquerda	0.95	1

Assim, esta prova de conceito cumpre um papel central na pesquisa: além de validar tecnicamente o modelo, fornece subsídios empíricos para o refinamento do artefato, orientando as próximas etapas de desenvolvimento e avaliação.

A Figura 5.2 apresenta o benchmark de processamento das principais etapas do protótipo, evidenciando a distribuição temporal entre chamadas às APIs externas e operações internas do sistema. Observa-se que a análise de toxicidade via Perspective API concentra aproximadamente 87,88% do tempo total de execução. Esse resultado decorre da necessidade de múltiplas requisições — cerca de 100 chamadas, uma para cada comentário analisado — enquanto integrações como YouTube Data API, SearchAPI e modelos de linguagem operam, em geral, com apenas uma requisição por vídeo. A etapa de persistência foi realizada com sucesso conforme figura A.4.

Palavras transcript	Toxicidade média	Toxicidade máxima	Taxa alta tox.	Categoria polarização	Score polarização
309	0.0897	0.602	0	politica	1

URLs externas detectadas

-

Tempo de processamento por etapa

Etapa	Tempo (s)	% do total
YouTube API - vídeo	0.275	0.2%
YouTube API - canal	0.301	0.21%
YouTube API - comentários	0.482	0.34%
Banco de dados - persistência	0.095	0.07%
YouTube API - vídeos do canal	7.466	5.3%
Perspective API - toxicidade	123.714	87.88%
SearchAPI - transcrição	4.138	2.94%
VidIQ - monetização estimada	0.753	0.53%
Extração de URLs externas	0	0%
OpenAI LLM - polarização	3.545	2.52%
Total	**140.77**	**100%**

Figura 5.2: Benchmark temporal das etapas de processamento da prova de conceito (PoC) do modelo PMT, evidenciando a distribuição do tempo de execução entre chamadas externas, integrações via APIs e operações internas do protótipo.

```
3   SELECT
4       'VIDEO' COLLATE utf8mb4_general_ci AS bloco,
5       'ID DB' COLLATE utf8mb4_general_ci AS campo,
6       CAST(v.id AS CHAR) COLLATE utf8mb4_general_ci AS valor
7   FROM videos v WHERE v.cod = 'LDG8GFqVnKQ'
8   UNION ALL
9   SELECT
10      'VIDEO' COLLATE utf8mb4_general_ci,
11      'Título' COLLATE utf8mb4_general_ci,
```

lt Grid | Filter Rows: | Export: | Wrap Cell Content:

bloco	campo	valor
VIDEO	ID DB	2831
VIDEO	Título	Conheça o maior esquema de corrupção do planeta
CANAL	Nome	Lula
COMENTARIOS	Comentário #739	Usuário: @IMPERADORIVS | Likes: 33 | Data: 2023-05-05 15:35:41 | Texto: Quem diz é a "alm…
COMENTARIOS	Comentário #740	Usuário: @LiphNeo | Likes: 23 | Data: 2022-10-29 18:02:35 | Texto: E o PETROLAO!?? 30 bilhõ…
COMENTARIOS	Comentário #741	Usuário: @marcos7084 | Likes: 161 | Data: 2022-11-28 06:10:23 | Texto: Acuse-os do que vc faz... Chame-os do que vc é... Conhecereis a verdade e a verdade vos libertará.... 🙏
COMENTARIOS	Comentário #742	Usuário: @LucasSantos-xc7mv | Likes: 27 | Data: 2024-04-16 01:03:17 | Texto: De acordo com …
COMENTARIOS	Comentário #743	Usuário: @lucianaviana8397 | Likes: 266 | Data: 2022-10-24 06:33:55 | Texto: QUEREMOS O V…
COMENTARIOS	Comentário #744	Usuário: @sergiovalentini9510 | Likes: 14 | Data: 2023-10-09 18:28:18 | Texto: Acuse a direita a…
COMENTARIOS	Comentário #745	Usuário: @GabrielAlves-tl1vc | Likes: 14 | Data: 2023-05-04 11:50:57 | Texto: E o mensalão foi e…
COMENTARIOS	Comentário #746	Usuário: @alexandredelacalle4876 | Likes: 17 | Data: 2022-10-19 15:28:54 | Texto: Essa coisa p…
COMENTARIOS	Comentário #747	Usuário: @santosII | Likes: 15 | Data: 2022-10-19 16:13:02 | Texto: E tem investigar as PROPIN…

Figura 5.3: Validação da persistência dos dados processados da POC no banco de dados, preservando o aspecto relacional dos dados.

6. Resultados da Avaliação

Este capítulo apresenta os resultados da avaliação experimental do *SocialDash*, incluindo análises quantitativas e qualitativas relacionadas à utilidade analítica, usabilidade e capacidade investigativa do artefato. São discutidos os resultados gerais, análises estatísticas, feedbacks qualitativos, interpretações associadas às conjecturas C1–C4, limitações observadas e uma síntese consolidada das evidências obtidas durante o experimento.

6.1 Análise estatística

A análise estatística foi conduzida a partir das respostas obtidas nas questões fechadas do formulário (Q1–Q13), organizadas em blocos correspondentes às conjecturas analíticas avaliadas e aos aspectos de usabilidade do artefato. Para cada conjunto de questões foram calculadas medidas descritivas de tendência central e dispersão, incluindo média, moda, mediana e desvio padrão.

Os blocos C0–C4 foram utilizados para consolidar estatisticamente as dimensões relacionadas à aprendizagem conceitual e à investigabilidade das conjecturas propostas nesta pesquisa. O bloco C0 correspondeu à compreensão inicial dos constructos de polarização, monetização e toxicidade, enquanto os blocos C1–C4 avaliaram diretamente as relações investigativas entre essas dimensões no contexto do *SocialDash*.

Valores reduzidos de desvio padrão foram interpretados como indicativos de maior convergência entre os participantes, enquanto desvios mais elevados sugerem maior variabilidade de percepção ou influência de fatores operacionais durante a interação com o sistema. Da mesma forma, a predominância de moda igual a 5 foi utilizada como indicador de elevada frequência de avaliações positivas entre os respondentes.

A Tabela 6.1 apresenta a consolidação estatística das conjecturas, incluindo medidas por bloco temático e interpretações exploratórias derivadas das respostas obtidas.

Tabela 6.1: Síntese estatística das conjecturas avaliadas

Conj.	Questão	Média	Moda	Mediana	DP	Consol.	Síntese
C0	Q1	4,0	5	5	1,33	4,1	Avaliação positiva da utilidade inicial do artefato.
	Q2	4,2	5	5	1,03		
	Q3	4,1	5	4,5	1,29		
	Consolidado	4,1	5	5	1,22		
C1	Q4	3,9	5	4,5	1,60	3,9	Reconhecimento de utilidade analítica na integração proposta.
	Q5	3,9	5	4,5	1,60		
	Consolidado	3,9	5	4,5	1,60		
C2	Q6	3,8	5	4,5	1,62	3,8	Percepção favorável sobre relações entre monetização e polarização.
	Q7	3,8	5	4,5	1,55		
	Consolidado	3,8	5	4,5	1,58		
C3	Q8	4,1	5	5	1,45	4,1	Avaliação positiva das análises entre monetização e toxicidade.
	Q9	4,0	5	4,5	1,25		
	Consolidado	4,05	5	4,75	1,35		
C4	Q10	4,2	5	5	1,23	4,2	Integração tripla percebida como contribuição relevante.
	Consolidado	4,2	5	5	1,23		

Do ponto de vista estatístico, os resultados indicam tendência predominantemente positiva em todas as dimensões analisadas, com médias variando entre 3,8 e 4,2 em uma escala Likert de cinco pontos. Observa-se predominância de moda igual a 5 em todas as conjecturas, sugerindo elevada frequência de avaliações máximas entre os participantes. As medianas permaneceram entre 4,5 e 5, indicando concentração das respostas em níveis altos de concordância e percepção favorável do artefato.

Os desvios padrão variaram aproximadamente entre 1,03 e 1,62, revelando dispersão moderada, porém sem comprometer a tendência geral de avaliação positiva. Em conjunto, os resultados sugerem consistência estatística na percepção de utilidade analítica do *SocialDash*, especialmente na integração simultânea entre polarização, monetização e toxicidade, além de evidências favoráveis quanto à capacidade do sistema de apoiar compreensão conceitual e investigação exploratória dos fenômenos analisados.

Além das conjecturas analíticas, realizou-se também uma consolidação estatística específica relacionada à usabilidade do artefato, abrangendo clareza de navegação, facilidade de aprendizado e estabilidade operacional percebida pelos participantes.

A Tabela 6.2 apresenta os resultados estatísticos consolidados das questões de usabilidade (Q11–Q13).

Tabela 6.2: Síntese estatística da avaliação de usabilidade do artefato

Bloco	Questão	Média	Moda	Mediana	DP	Interpretação estatística
UX	Q11	3,8	5	5	1,93	Os resultados indicam percepção positiva da usabilidade geral do sistema, embora acompanhada de maior dispersão estatística associada a limitações operacionais e desempenho assíncrono durante algumas execuções.
	Q12	3,8	5	5	1,93	
	Q13	3,7	5	5	1,89	
	Consolidado	3,77	5	5	1,91	

Os resultados relacionados à usabilidade indicam percepção geral positiva do sistema, embora acompanhada de maior dispersão estatística quando comparada às questões diretamente associadas às conjecturas analíticas. Esse comportamento sugere que fatores operacionais, como tempo de processamento, dependência de APIs externas e execução assíncrona de tarefas, impactaram parcialmente a experiência de uso durante algumas interações.

Ainda assim, a predominância de mediana e moda iguais a 5 sugere que, mesmo diante das limitações operacionais observadas, os participantes perceberam o ambiente como funcionalmente adequado para fins de exploração analítica e investigação empírica.

6.2 Feedbacks qualitativos

Além das questões estruturadas em escala Likert, a avaliação incorporou mecanismos complementares de coleta qualitativa, compostos tanto pelos feedbacks internos (um campo textarea em cada página do artefato) registrados diretamente no *SocialDash* quanto pelas respostas abertas das questões Q14 e Q15 do formulário externo.

Os feedbacks internos foram coletados durante a execução das tarefas analíticas associadas às conjecturas C1–C4, permitindo registrar impressões contextuais produzidas no próprio momento de interação com o sistema. Essa estratégia possibilitou observar não apenas avaliações posteriores dos participantes, mas também aspectos relacionados ao processo interpretativo desenvolvido durante a utilização efetiva do artefato.

Os feedbacks registrados ao longo das tarefas experimentais, contendo identificador único, tarefa associada, timestamp de início e duração em segundos, enquanto a Tabela 6.3 consolida os principais temas identificados nos comentários internos do sistema.

Tabela 6.3: Temas observados nos feedbacks internos do SocialDash

Tema identificado	**Freq.**	**Exemplo observado**
Inferência baseada em título/tema	6	Participantes relataram selecionar vídeos a partir de títulos e temática percebida.
Inferência de posicionamento político	5	Usuários relacionaram canais e conteúdos a espectros ideológicos.
Interpretação contextual dos resultados	5	Os participantes utilizaram os resultados como apoio inferencial, e não como resposta definitiva.
Expectativa de toxicidade	4	Participantes relataram expectativa prévia de comentários tóxicos ou hostis.
Uso de *proxies* de engajamento	3	Métricas como inscritos, engajamento e descrições foram utilizadas como sinais analíticos.

Os registros qualitativos indicam que os participantes utilizaram o *SocialDash* predominantemente de forma interpretativa e exploratória, formulando hipóteses a partir de títulos, temática dos vídeos, posicionamento político percebido, métricas de engajamento, comportamento discursivo e características contextuais dos canais analisados.

Observou-se também que diversos participantes demonstraram compreender parcialmente o caráter inferencial dos *proxies* computacionais utilizados pelo sistema, utilizando os resultados como mecanismos de apoio analítico e não como classificações automáticas definitivas. Esse comportamento reforça a proposta metodológica da pesquisa, baseada em investigabilidade, interpretação contextual e exploração de sinais sociotécnicos observáveis.

De forma complementar, as questões abertas Q14 e Q15 permitiram identificar percepções relacionadas a limitações operacionais, melhorias de usabilidade e possibilidades de expansão metodológica do artefato. As respostas foram agrupadas em categorias qualitativas recorrentes, posteriormente consolidadas na Tabela 6.4.

As respostas relacionadas à Q14 concentraram-se principalmente em aspectos associados ao tempo de processamento, dependência de APIs externas, organização visual da interface e necessidade de otimizações arquiteturais, como paralelização de tarefas e redução de gargalos operacionais. Esses resultados são coerentes com a natureza experimental da prova de conceito desenvolvida nesta pesquisa.

Já os resultados da Q15 evidenciaram interesse dos participantes na ampliação das capacidades analíticas do sistema, incluindo inclusão de novas métricas, maior detalhamento temporal, mecanismos de explicabilidade inferencial, exportação de resultados e expansão futura para outras plataformas digitais.

Tabela 6.4: Síntese qualitativa das questões abertas Q14 e Q15

Q	Tema identificado	Freq.	Interpretação
Q15	Integração das dimensões PMT	5	Os participantes destacaram positivamente a integração simultânea entre polarização, monetização e toxicidade como principal diferencial analítico do artefato.
Q14	Complexidade interpretativa	4	Parte dos respondentes indicou que conceitos relacionados à polarização, monetização e toxicidade exigem contextualização analítica durante o uso do sistema.
Q14	Interface e organização visual	4	Foram identificadas sugestões relacionadas à organização da navegação, posicionamento das abas e melhoria da experiência visual do artefato.
Q15	Inclusão de novas métricas	4	Os respondentes sugeriram expansão do conjunto de métricas, rankings e indicadores comparativos entre canais e vídeos.
Q14	Timeout e lentidão	3	Os participantes relataram lentidão operacional, dependência de APIs externas e necessidade de otimização no processamento assíncrono das tarefas.
Q15	Maior detalhamento temporal	3	Usuários demonstraram interesse em análises temporais mais refinadas e maior granularidade nas visualizações.
Q15	Explicabilidade dos modelos	3	Os participantes sugeriram maior transparência sobre inferências, *proxies* computacionais e mecanismos analíticos utilizados pelo sistema.
Q14	Necessidade de paralelização	2	Os feedbacks sugerem otimizações relacionadas à execução simultânea de tarefas e redução de gargalos operacionais.
Q15	Expansão para outras plataformas	2	Foi sugerida adaptação futura do modelo analítico para outras plataformas digitais e diferentes contextos sociotécnicos.
Q15	Exportação e reprodutibilidade	2	Foram sugeridas funcionalidades relacionadas à exportação de dados, comparação entre múltiplos canais e apoio à reprodutibilidade analítica.

Em conjunto, os feedbacks qualitativos reforçam a percepção de que o *SocialDash* foi compreendido como ambiente de apoio à formulação de hipóteses e interpretação exploratória de fenômenos sociotécnicos complexos, e não como mecanismo determinístico de classificação automática.

6.3 Interpretação por conjectura

A interpretação das conjecturas deve ser compreendida a partir da função central do *SocialDash*: não produzir classificações definitivas sobre polarização, monetização e toxicidade, mas estruturar um ambiente analítico capaz de organizar sinais observáveis e apoiar processos de investigação exploratória conduzidos por pesquisadores humanos.

Nesse contexto, o artefato foi concebido para favorecer processos graduais de interpretação e contextualização dos dados, aproximando-se de abordagens de *sensemaking*, nas quais a compreensão emerge da exploração iterativa de múltiplos sinais analíticos. Para isso, o sistema incorpora mecanismos visuais e interativos voltados à interpretação contextual dos fenômenos observados, incluindo organização temporal dos dados, uso de cores associadas a diferentes entidades analíticas (vídeos, canais e comentários) e *affordances* visuais que facilitam a navegação entre elementos relacionados do ecossistema analisado.

Na conjectura C1 (Polarização $\leftrightarrow$ Toxicidade), os participantes indicaram que o sistema tornou perceptíveis relações entre alinhamento ideológico, intensidade discursiva e toxicidade nos comentários. A combinação entre classificação temática, métricas de toxicidade e visualizações temporais permitiu observar padrões de concentração discursiva e hostilidade em conteúdos específicos. Os resultados sugerem suporte exploratório à conjectura, especialmente no que se refere à capacidade do artefato de integrar sinais analíticos distintos em uma mesma interface interpretativa.

Na conjectura C2 (Polarização $\leftrightarrow$ Monetização), os participantes reconheceram a utilidade analítica dos *proxies* de monetização utilizados pelo sistema, incluindo métricas de engajamento, URLs externas e indicadores financeiros derivados de plataformas complementares. Ao mesmo tempo, os resultados evidenciaram compreensão de que tais métricas possuem natureza inferencial e não representam diretamente receita financeira real. Ainda assim, os participantes perceberam que a integração entre indicadores de alcance, engajamento e alinhamento discursivo possibilitou investigar possíveis relações entre conteúdos polarizados e dinâmicas associadas à economia da atenção.

Na conjectura C3 (Monetização $\leftrightarrow$ Toxicidade), os participantes perceberam utilidade nos mecanismos de análise temporal e nos agrupamentos analíticos empregados pelo artefato para investigar padrões discursivos associados à monetização. Entretanto, observou-se maior variabilidade nas respostas e relatos de limitações operacionais relacionadas ao tempo de processamento e dependência de tarefas síncronas. Esses fatores impactaram parcialmente a fluidez da exploração analítica, sugerindo que a conjectura apresentou suporte exploratório moderado, porém sensível às limitações técnicas do ambiente experimental.

Na conjectura C4 (Polarização $\leftrightarrow$ Monetização $\leftrightarrow$ Toxicidade), os resultados indicaram a percepção mais forte de contribuição analítica do artefato. Os participantes relataram que a visualização integrada das três dimensões ampliou a capacidade interpretativa do sistema, favorecendo a formulação de hipóteses exploratórias sobre dinâmicas socio-

técnicas no YouTube. Elementos como organização temporal, cores por entidade analítica e navegação entre vídeos, canais e comentários foram percebidos como relevantes para processos de interpretação contextualizada e identificação de padrões emergentes.

A Tabela 6.5 apresenta uma síntese interpretativa consolidada do suporte observado para cada conjectura avaliada.

Tabela 6.5: Síntese interpretativa das conjecturas avaliadas

Conj.	**Suporte**	**Síntese interpretativa**
C1	Forte	Os participantes perceberam relações consistentes entre alinhamento ideológico, intensidade discursiva e toxicidade nos comentários, especialmente por meio da integração entre métricas analíticas e visualizações temporais.
C2	Moderado/Forte	Os *proxies* de monetização foram considerados úteis para investigação exploratória, embora os participantes tenham reconhecido sua natureza inferencial e dependente de contextualização analítica.
C3	Moderado	A análise temporal entre monetização e toxicidade apresentou utilidade investigativa, porém limitações operacionais e dependência de processamento síncrono impactaram parcialmente a experiência analítica.
C4	Muito forte	A integração simultânea entre polarização, monetização e toxicidade foi percebida como uma das principais contribuições do artefato, ampliando a capacidade interpretativa e a formulação de hipóteses exploratórias.

Os resultados obtidos também permitem relacionar empiricamente a avaliação do artefato às questões de pesquisa propostas nesta tese. A Tabela 6.6 sintetiza as principais evidências observadas durante o experimento em relação às RQs investigadas.

Tabela 6.6: Relação entre resultados da avaliação e questões de pesquisa

RQ	**Evidências observadas na avaliação**
RQ1	Os participantes conseguiram identificar relações entre polarização, toxicidade e monetização a partir da integração de dados observáveis e inferíveis no ambiente experimental.
RQ2	Os resultados indicaram que *proxies* computacionais, APIs externas e métricas inferenciais foram compreendidos como mecanismos viáveis para operacionalizar dimensões sociotécnicas complexas em plataformas digitais.
RQ3	A avaliação evidenciou que o artefato foi capaz de integrar coleta, inferência e visualização em uma interface exploratória considerada útil para investigação analítica de fenômenos em rede social.

De maneira complementar, a avaliação também permitiu analisar aspectos técnico-operacionais relacionados ao funcionamento do sistema, capacidade investigativa, suporte às conjecturas e potencial de escalabilidade analítica do artefato. A Tabela 6.7 consolida essa visão integrada dos resultados observados durante o experimento.

Tabela 6.7: Síntese técnico-operacional da avaliação do artefato

Aspecto	Síntese
Desempenho técnico	O sistema apresentou funcionamento considerado adequado para exploração analítica, embora limitações relacionadas ao tempo de processamento e tarefas síncronas tenham sido relatadas em cenários específicos.
Capacidade investigativa	Os participantes conseguiram utilizar o ambiente para investigar relações entre polarização, monetização e toxicidade de forma integrada e contextualizada.
Suporte às conjecturas	As quatro conjecturas receberam suporte exploratório em diferentes intensidades, com destaque para a integração simultânea das dimensões P-M-T.
Escalabilidade analítica	O uso de métricas automatizadas, APIs externas e visualizações integradas foi percebido como adequado para exploração de grandes volumes de dados sociotécnicos.

Em conjunto, os resultados sugerem que o *SocialDash* apresentou capacidade consistente de integrar coleta, inferência e visualização analítica em um ambiente único, apoiando processos de interpretação contextualizada, formulação de hipóteses exploratórias e investigabilidade de fenômenos sociotécnicos complexos no contexto do YouTube.

6.4 Limitações observadas

Durante a avaliação, não foram identificados erros críticos de implementação capazes de inviabilizar o uso do sistema. As principais limitações observadas estiveram relacionadas ao tempo de processamento, ocorrência eventual de *timeouts* e dependência de múltiplas integrações externas utilizadas para coleta, inferência e enriquecimento analítico dos dados.

Parte dessas limitações decorre diretamente da natureza distribuída das operações realizadas pelo *SocialDash*, que depende da integração entre APIs externas, mecanismos de coleta pública de dados, processamento síncrono e serviços de inferência automatizada. Em determinados momentos, variações de latência, indisponibilidade temporária de serviços externos ou limitação de requisições impactaram parcialmente a fluidez operacional de algumas tarefas analíticas executadas pelos participantes.

As limitações relacionadas ao desempenho também devem ser interpretadas à luz da natureza experimental do artefato desenvolvido nesta pesquisa. O *SocialDash* foi concebido como uma prova de conceito funcional voltada à validação metodológica e investiga-

tiva das conjecturas propostas, e não como uma infraestrutura distribuída otimizada para operação contínua em larga escala.

Nesse contexto, algumas decisões arquiteturais priorizaram flexibilidade analítica, integração de múltiplas fontes de dados e capacidade exploratória em detrimento de otimizações avançadas de desempenho. Essa escolha permitiu acelerar o processo iterativo de desenvolvimento e validação do artefato, favorecendo experimentação analítica, investigabilidade e adaptação dos fluxos exploratórios ao longo da pesquisa.

Assim, mecanismos como paralelização extensiva, filas independentes, balanceamento de carga, cache distribuído, persistência intermediária de resultados e processamento assíncrono avançado permanecem como possibilidades de evolução futura do sistema.

Apesar dessas limitações, os resultados indicam que o ambiente experimental apresentou estabilidade suficiente para execução das tarefas propostas e coleta das evidências analíticas utilizadas na avaliação do artefato, mantendo coerência com os objetivos exploratórios e metodológicos definidos nesta pesquisa.

6.5 Resultados gerais e síntese

De maneira geral, os resultados da avaliação[23] indicam percepção predominantemente positiva do artefato, tanto em relação à sua utilidade analítica quanto à capacidade de apoiar investigações exploratórias envolvendo polarização, monetização e toxicidade no contexto do YouTube. As médias observadas permaneceram próximas ou superiores a 4 na maior parte dos blocos temáticos, enquanto a predominância de moda igual a 5 e medianas concentradas entre 4,5 e 5 reforçam a tendência de elevada concordância quanto à relevância das funcionalidades propostas.

Os maiores níveis de concordância foram observados nas questões relacionadas à integração simultânea entre polarização, monetização e toxicidade, sugerindo que os participantes perceberam valor analítico especialmente na articulação conjunta entre múltiplos sinais observáveis da plataforma. Também foram registradas avaliações positivas quanto à capacidade do sistema de apoiar compreensão conceitual inicial dos fenômenos investigados e estruturar análises comparativas entre diferentes entidades analíticas.

Por outro lado, os resultados também evidenciaram dispersão moderada em dimensões relacionadas à usabilidade, estabilidade operacional e interpretação de *proxies* infe-

[23] https://docs.google.com/spreadsheets/d/162BZ_sWr8DCCq2m6GtCoVte_L4MEy5fv6Yy5f-HP3KQ/edit?usp=sharing

renciais. Os desvios padrão mais elevados sugerem que parte dos participantes encontrou dificuldades associadas ao tempo de processamento, dependência de APIs externas e necessidade de maior contextualização analítica durante o uso do sistema.

As respostas abertas e os feedbacks qualitativos reforçam que o *SocialDash* foi compreendido predominantemente como ferramenta de apoio interpretativo e inferencial, e não como mecanismo automático de classificação definitiva. Esse aspecto é particularmente relevante do ponto de vista epistemológico, pois evidencia que o artefato foi percebido como instrumento de mediação analítica voltado à exploração, contextualização e formulação de hipóteses sobre fenômenos sociotécnicos complexos.

Os resultados também sugerem níveis distintos de maturidade operacional entre as dimensões investigadas. A toxicidade mostrou-se mais diretamente operacionalizável devido à disponibilidade de métricas contínuas mais consolidadas, enquanto polarização e monetização dependeram mais intensamente de inferências contextuais, sinais computacionais indiretos e interpretação semântica conduzida pelos participantes.

Em síntese, a avaliação indica que o *SocialDash* apresentou capacidade consistente de integrar coleta, inferência, visualização e interação analítica em um único ambiente, oferecendo suporte à análise contextualizada de sinais digitais associados à polarização, monetização e toxicidade. Assim, os resultados obtidos fornecem evidências favoráveis à viabilidade do método integrador proposto e ao potencial investigativo do artefato desenvolvido nesta pesquisa.

7. Encerramento

7.1 Contribuição científica

A literatura sobre plataformas digitais tem avançado significativamente na análise de fenômenos como polarização, toxicidade e monetização. Consequentemente, tais dimensões são predominantemente investigadas de forma isolada, com abordagens metodológicas específicas para cada fenômeno, o que pode limitar a compreensão de possíveis interdependências entre essas dimensões no contexto de sistemas sociotécnicos complexos como o YouTube. Estudos sobre polarização concentram-se em dinâmicas de recomendação e formação de comunidades [8, 33], enquanto trabalhos sobre toxicidade exploram principalmente análise textual e comportamental em comentários [39]. Já a monetização é frequentemente tratada sob a ótica econômica e algorítmica, focando em métricas de engajamento e receita [17]. Essa fragmentação evidencia um **gap teórico-metodológico**: a ausência de um modelo integrado capaz de articular essas dimensões em uma estrutura analítica unificada.

Diante desse cenário, esta tese propõe como contribuição científica central o desenvolvimento e a validação de um método integrador capaz de operacionalizar e correlacionar, de forma sistemática, as dimensões de polarização, toxicidade e monetização em plataformas de vídeo inseridas na economia da atenção [5]. O foco da contribuição não reside apenas no artefato computacional desenvolvido, mas principalmente na formalização de um processo analítico que utiliza sinais observáveis — como visualizações, comentários, interações e metadados — como *proxies* para a inferência de variáveis latentes interpretáveis, permitindo a análise conjunta de fenômenos tradicionalmente investigados de forma separada.

Essa contribuição se desdobra em três níveis, alinhados às diretrizes da abordagem de DSR:

- **(1) Modelo conceitual**: Propõe uma estrutura teórica que articula a relação entre engajamento, algoritmos de recomendação e dinâmicas discursivas, explicando como a economia da atenção e os mecanismos de visibilidade influenciam simultaneamente a monetização e a amplificação de discursos potencialmente polarizados ou tóxicos. Esse modelo responde à lacuna apontada pela literatura sobre a falta de integração entre essas dimensões.

- **(2) Método analítico**: Define um pipeline metodológico para coleta, processamento e transformação de dados provenientes do YouTube (via API e crawling), incorporando métricas de toxicidade (como a Perspective API) e *proxies* de monetização. Este método explicita *como* operacionalizar empiricamente o modelo conceitual, respondendo à necessidade de reprodutibilidade e sistematização destacada na literatura.

- **(3) Artefato computacional (SocialDash)**: Implementa o método proposto em uma aplicação Web interativa, que permite explorar visualmente correlações entre polarização, toxicidade e monetização. O artefato materializa a contribuição metodológica e possibilita sua aplicação por outros pesquisadores.

7.2 Limitações da pesquisa

Apesar dos avanços apresentados, esta pesquisa possui limitações que devem ser consideradas na interpretação dos resultados.

Em primeiro lugar, destaca-se a dependência de *proxies* computacionais para a inferência de variáveis não diretamente observáveis, como monetização e polarização. Embora essa estratégia seja amplamente utilizada na literatura de auditoria algorítmica e plataformas digitais, ela impõe limitações relacionadas à precisão e à validade inferencial das análises realizadas [12].

Em segundo lugar, existem restrições associadas ao acesso a dados internos da plataforma, especialmente no que se refere aos mecanismos de recomendação e a informações financeiras efetivamente utilizadas pelo ecossistema de monetização. Essas limitações reforçam o caráter indireto das análises e evidenciam os desafios estruturais envolvidos na investigação de sistemas algorítmicos proprietários [6, 8, 12].

Do ponto de vista técnico, o sistema foi implementado em um ambiente com recursos limitados, o que exigiu a adoção de estratégias de amostragem e redução do volume de dados processados. Como resultado, as análises realizadas devem ser interpretadas como

aproximações exploratórias, e não como representações completas do comportamento da plataforma.

Por fim, a natureza exploratória do artefato implica que os resultados obtidos não devem ser interpretados como evidências causais, mas como indícios empíricos que podem orientar investigações futuras.

7.3 Trabalhos futuros

Os resultados desta pesquisa abrem diversas possibilidades para investigações futuras, tanto no plano metodológico quanto no desenvolvimento do artefato.

No que se refere ao modelo analítico, futuras pesquisas podem explorar a incorporação de novas dimensões, como desinformação, credibilidade e dinâmica de recomendação, ampliando o escopo do modelo proposto.

Em termos de operacionalização, há potencial para o uso de técnicas mais avançadas de processamento de linguagem natural, incluindo modelos de linguagem de grande escala, para aprimorar a inferência de polarização e a análise de conteúdo.

Do ponto de vista técnico, melhorias na infraestrutura podem permitir o processamento de volumes maiores de dados, reduzindo as limitações atuais relacionadas a amostragem e desempenho. Adicionalmente, a implementação de mecanismos de armazenamento e reuso de dados pode aumentar a eficiência do sistema.

Outra linha promissora é a ampliação do conjunto de widgets analíticos, incluindo novas formas de visualização e análise, bem como a integração com outras fontes de dados, como redes sociais e plataformas externas.

Por fim, destaca-se o potencial do artefato como ferramenta de apoio à pesquisa empírica. A disponibilização pública do sistema e dos dados utilizados pode contribuir para a replicabilidade dos estudos e para o avanço do campo, em consonância com princípios de ciência aberta.

7.4 Considerações finais

Esta tese apresentou uma abordagem integrada para análise de polarização, toxicidade e monetização em plataformas digitais, combinando modelagem conceitual, inferência

analítica e desenvolvimento de artefato.

A principal contribuição reside na proposição de um modelo que articula essas dimensões de forma interdependente, bem como na sua operacionalização por meio do sistema *SocialDash*. Ao integrar coleta de dados, processamento analítico e visualização interativa, o artefato permite reduzir o custo cognitivo da análise e apoiar a construção de conhecimento em ambientes complexos.

Assim, a pesquisa contribui tanto para o avanço teórico quanto para o desenvolvimento de ferramentas práticas, evidenciando o potencial de DSR como abordagem para investigação de fenômenos sociotécnicos contemporâneos.

Referências Bibliográficas

[1] BURGESS, Jean; GREEN, Joshua. The entrepreneurial vlogger: participatory culture beyond the professional-amateur divide. In: SNICKARS, Pelle; VONDERAU, Patrick (eds.). *The YouTube Reader*. Stockholm: National Library of Sweden, 2009. p. 89–107.

[2] MCNEIL, E. B. *Psychology and War: A Review*. Michigan: Department of Psychology, The University of Michigan, 1963.

[3] DUNNA, Arun et al. Paying attention to the algorithm behind the curtain: bringing transparency to YouTube's demonetization algorithms. *Proceedings of the ACM on Human-Computer Interaction*, v. 6, n. CSCW2, p. 1–31, 2022.

[4] RIEDER, Bernhard; COROMINA, Òscar; MATAMOROS-FERNÁNDEZ, Ariadna. Mapping YouTube: a quantitative exploration of a platformed media system. *First Monday*, v. 25, n. 8, 2020.

[5] FRANCK, Georg. The economy of attention. *Journal of Sociology*, 2018.

[6] DIAKOPOULOS, Nicholas. Algorithmic accountability. *Digital Journalism*, v. 3, n. 3, p. 398–415, 2015.

[7] BÄRTL, Mathias. YouTube channels, uploads and views: a statistical analysis of the past 10 years. *Convergence*, v. 24, n. 1, p. 16–32, 2018.

[8] HAROON, Muhammad et al. YouTube, the great radicalizer? Auditing and mitigating ideological biases in YouTube recommendations. *arXiv preprint*, arXiv:2203.10666, 2022.

[9] MOSCOVICI, Serge; ZAVALLONI, Marisa. The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, v. 12, n. 2, p. 125–135, 1969.

[10] QUATTROCIOCCHI, Walter; SCALA, Antonio; SUNSTEIN, Cass R. Echo chambers on Facebook. *SSRN Electronic Journal*, 2016.

[11] PARISER, Eli. *The Filter Bubble: What the Internet Is Hiding from You*. London: Penguin Books, 2011.

[12] PAPADAMOU, Kostantinos; ZANNETTOU, Savvas; BLACKBURN, Jeremy; DE CRISTOFARO, Emiliano; STRINGHINI, Gianluca; SIRIVIANOS, Michael. "How over is it?" Understanding the incel community on YouTube. *Proceedings of the ACM on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2021.

[13] MURALIKUMAR, Meena Devii; YANG, Yun Shan; MCDONALD, David W. A human-centered evaluation of a toxicity detection API: testing transferability and unpacking latent attributes. *ACM Transactions on Social Computing*, 2023.

[14] BERTAGLIA, Thales; GOANTA, Catalina; IAMNITCHI, Adriana. The monetisation of toxicity: analysing YouTube content creators and controversy-driven engagement. In: *Proceedings of the 4th International Workshop on Open Challenges in Online Social Networks*. 2024. p. 1–9.

[15] BRADY, William J.; WILLS, Julian A.; JOST, John T.; TUCKER, Joshua A.; VAN BAVEL, Jay J. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, v. 114, n. 28, p. 7313–7318, 2017.

[16] BERGER, Jonah; MILKMAN, Katherine L. What makes online content viral? *Journal of Marketing Research*, v. 49, n. 2, p. 192–205, 2012.

[17] HAN, Bo. How do YouTubers make money? A lesson learned from the most subscribed YouTuber channels. *International Journal of Business Information Systems*, v. 33, n. 1, p. 132–143, 2020.

[18] COVINGTON, Paul; ADAMS, Jay; SARGIN, Emre. Deep neural networks for YouTube recommendations. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. 2016. p. 191–198.

[19] HUA, Yiqing et al. Characterizing alternative monetization strategies on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, v. 6, n. CSCW2, p. 1–30, 2022.

[20] RIEDER, Bernhard et al. Making a living in the creator economy: a large-scale study of linking on YouTube. *Social Media + Society*, v. 9, n. 2, 2023.

[21] PEREIRA, Érica Verônica et al. Analyzing YouTube videos shared on WhatsApp and Telegram political public groups. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. 2022. p. 28–37.

[22] ALMEIDA, Luis Gustavo; GARCIA, Ana Cristina Bicharra; SIMÕES, Jefferson Elbert. Polarisation, filter bubbles and radicalisation on YouTube: a systematic literature review. *International Journal of Web Based Communities*, v. 21, n. 4, p. 324–347, 2025.

[23] SRBA, Ivan et al. Auditing YouTube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems*, v. 1, n. 1, p. 1–33, 2023.

[24] HUSSEIN, Eslam; JUNEJA, Prerna; MITRA, Tanushree. Measuring misinformation in video search platforms: an audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 2020.

[25] RUSSELL, Daniel M.; STEFIK, Mark J.; PIROLLI, Peter; CARD, Stuart K. The cost structure of sensemaking. In: *Proceedings of the INTERCHI Conference on Human Factors in Computing Systems*. 1993.

[26] THOMAS, James J.; COOK, Kristin A. (eds.). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos: IEEE Computer Society, 2005.

[27] GUIMARÃES, Samuel S. et al. Characterizing toxicity on Facebook comments in Brazil. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. 2020. p. 253–260.

[28] BESSI, Alessandro; ZOLLO, Fabiana; DEL VICARIO, Michela; PULIGA, Michelangelo; SCALA, Antonio; CALDARELLI, Guido; UZZI, Brian; QUATTROCIOCCHI, Walter. Users polarization on Facebook and YouTube. *PLOS ONE*, v. 11, n. 8, e0159641, 2016.

[29] IYENGAR, Shanto; LELKES, Yphtach; LEVENDUSKY, Matthew; MALHOTRA, Neil; WESTWOOD, Sean J. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, v. 22, p. 129–146, 2019.

[30] HOSSEINMARDI, Homa; GHASEMIAN, Amir; CLAUSET, Aaron; MOBIUS, Markus; ROTHSCHILD, David M.; WATTS, Duncan J. Examining the consump-

tion of radical content on YouTube. *Proceedings of the National Academy of Sciences*, 2021.

[31] PAPADAMOU, Kostantinos; ZANNETTOU, Savvas; BLACKBURN, Jeremy; DE CRISTOFARO, Emiliano; STRINGHINI, Gianluca; SIRIVIANOS, Michael. "How over is it?" Understanding the incel community on YouTube. *arXiv preprint*, arXiv:2001.08293, 2021.

[32] POSTIGO, Hector. The socio-technical architecture of digital labor: converting play into YouTube money. *New Media & Society*, v. 18, n. 2, p. 332–349, 2016.

[33] RIBEIRO, Manoel Horta et al. Auditing radicalization pathways on YouTube. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020. p. 131–141.

[34] NOGARA, Gianluca et al. The disinformation dozen: an exploratory analysis of COVID-19 disinformation proliferation on Twitter. In: *Proceedings of the 14th ACM Web Science Conference 2022*. 2022. p. 348–358.

[35] ALMEIDA, Luis Gustavo; GARCIA, Ana Cristina Bicharra; SIMÕES, Jefferson Elbert. Polarização no YouTube: interação colaborativa e a atração dos discursos fortes e monetização – proposta de uma ferramenta online. In: *Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*. Porto Alegre: SBC, 2025. p. 127–138.

[36] SULER, John. The online disinhibition effect. *CyberPsychology & Behavior*, v. 7, n. 3, p. 321–326, 2004.

[37] VELHO, Raphaela Martins; MENDES, Amanda Merian Freitas; AZEVEDO, Caio Lucidius Naberezny. Communicating science with YouTube videos: how nine factors relate to and affect video views. *Frontiers in Communication*, v. 5, p. 567606, 2020.

[38] MARICONTI, Enrico et al. "You know what to do": proactive detection of YouTube videos targeted by coordinated hate attacks. *Proceedings of the ACM on Human-Computer Interaction*, v. 3, n. CSCW, p. 1–21, 2019.

[39] OBADIMU, Adewale et al. Identifying toxicity within YouTube video comments. In: *Social, Cultural, and Behavioral Modeling: 12th International Conference SBP-BRiMS 2019*. Cham: Springer, 2019. p. 214–223.

[40] TRIPATHI, Ashutosh; BHARTI, Kusum Kumari; GHOSH, Mohona. A study on characterizing the ecosystem of monetizing video spams on YouTube platform. In:

Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS). 2019.

[41] ARTHURS, Jane; DRAKOPOULOU, Sophia; GANDINI, Alessandro. Researching YouTube. *Convergence*, v. 24, n. 1, p. 3–15, 2018.

[42] ALBADI, Nuha; KURDI, Maram; MISHRA, Shivakant. Deradicalizing YouTube: characterization, detection, and personalization of religiously intolerant Arabic videos. *ACM Transactions on Social Computing*, 2022.

[43] BALLARD, Cameron et al. Conspiracy brokers: understanding the monetization of YouTube conspiracy theories. In: *Proceedings of the ACM Web Conference 2022*. 2022. p. 2707–2718.

[44] DAVENPORT, Thomas H.; BECK, John C. *The Attention Economy: Understanding the New Currency of Business*. Boston: Harvard Business School Press, 2001.

[45] RIBEIRO, Bruno; FALOUTSOS, Christos. Modeling website popularity competition in the attention-activity marketplace. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. Shanghai: ACM, 2015. p. 389–398.

[46] ZAPPIN, Anthony et al. YouTube monetization and censorship by proxy: a machine learning perspective. *Procedia Computer Science*, v. 198, p. 23–32, 2022.

[47] TRIPATHI, Ashutosh; BHARTI, Kusum Kumari; GHOSH, Mohona. A study on characterizing the ecosystem of monetizing video spams on YouTube platform. Germany, 2019.

[48] SIMON, Herbert A. Designing organizations for an information-rich world. In: *Computers, Communications, and the Public Interest*. Baltimore: Johns Hopkins Press, 1971.

[49] ROTH, Camille; MAZIÈRES, Antoine; MENEZES, Telmo. Tubes and bubbles: topological confinement of YouTube recommendations. *PLOS ONE*, 2020.

[50] NIU, Shuo et al. Understanding hate group videos on YouTube. In: *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. 2022. p. 32–36.

[51] CHU, Andrew et al. Behind the tube: exploitative monetization of content on YouTube. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022. p. 2171–2188.

[52] AIROLDI, Massimo; BERALDO, Davide; GANDINI, Alessandro. Follow the algorithm: an exploratory investigation of music on YouTube. *Poetics*, v. 57, p. 1–13, 2016.

[53] PAVLOPOULOS, John et al. Toxicity detection: does context really matter? *arXiv preprint*, arXiv:2006.00998, 2020.

[54] SUTCLIFFE, Kathleen M.; BROWN, Andrew D.; PUTNAM, Linda L. Introduction to the special issue "Making Sense of Organizing: in Honor of Karl Weick". *Organization Studies*, v. 27, n. 11, p. 1573–1578, 2006.

[55] BAKER, Jeff; BURKMAN, Jim; JONES, Donald R. Using visual representations of data to enhance sensemaking in data exploration tasks. *Journal of the Association for Information Systems*, v. 10, 2009.

[56] KEIM, Daniel A.; ANDRIENKO, Gennady; FEKETE, Jean-Daniel; GÖRG, Carsten; KOHLHAMMER, Jörn; MELANÇON, Guy. Visual analytics: definition, process, and challenges. Berlin: Springer, 2008.

[57] SHEVTSOV, Alexander; OIKONOMIDOU, Maria; ANTONAKAKI, Despoina; PRATIKAKIS, Polyvios; IOANNIDIS, Sotiris. What tweets and YouTube comments have in common? Sentiment and graph analysis on data related to US elections 2020. *PLOS ONE*, v. 18, n. 1, e0270542, 2023.

[58] CAKMAK, Mert Can; OKEKE, Obianuju; ONYEPUNUKA, Ugochukwu; SPANN, Billy; AGARWAL, Nitin. Analyzing bias in recommender systems: a comprehensive evaluation of YouTube's recommendation algorithm. In: *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2023.

[59] BUNTAIN, Cody; BONNEAU, Richard; NAGLER, Jonathan; TUCKER, Joshua A. YouTube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

[60] RASTEGARPANAH, Bashir; GUMMADI, Krishna P.; CROVELLA, Mark. Fighting fire with fire: using antidote data to improve polarization and fairness of recommender systems. In: *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*. 2019.

[61] THOMAS, James J.; COOK, Kristin A. (eds.). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Los Alamitos: IEEE Computer Society, 2005.

[62] ANDRES, Raphaela; ROSSI, Michelangelo; TREMBLAY, Mark J. YouTube "Adpocalypse": the YouTubers' journey from ad-based to patron-based revenues. *ZEW-Centre for European Economic Research Discussion Paper*, n. 59, 2023.

[63] JIANG, Shan; WILSON, Christo. Linguistic signals under misinformation and fact-checking: evidence from user comments on social media. 2018.

[64] ALDOUS, Kholoud Khalil; AN, Jisun; JANSEN, Bernard J. Measuring 9 emotions of news posts from 8 news organizations across 4 social media platforms for 8 months. *ACM Transactions on Social Computing*, v. 4, n. 4, Article 15, 2022.

[65] VAN KESSEL, Patrick; TOOR, Skye; SMITH, Aaron. A week in the life of popular YouTube channels. Washington, DC: Pew Research Center, 2019.

[66] BYUN, Unji; JANG, Moonkyoung; BAEK, Hyunmi. The effect of YouTube comment interaction on video engagement: focusing on interactivity centralization and creators' interactivity. *Online Information Review*, v. 47, n. 6, p. 1083–1097, 2023.

[67] HAN, Bo. Views versus subscriptions: which one matters to a YouTuber's monetisation success? *International Journal of Web Based Communities*, v. 14, n. 4, p. 325–334, 2018.

[68] DE MARCHI, Leonardo. Como os algoritmos do YouTube calculam valor? Uma análise da produção de valor para vídeos digitais de música através da lógica social de derivativo. *MATRIZes*, v. 12, n. 2, p. 193–215, 2018.

[69] WALUYOHADI, Waluyohadi; LOPATKA, Agnieszka. 90 days to achieve monetization: implementing design thinking in a YouTube channel. *Procedia Computer Science*, v. 207, p. 4207–4216, 2022.

[70] MARICONTI, Enrico; SUAREZ-TANGIL, Guillermo; BLACKBURN, Jeremy; DE CRISTOFARO, Emiliano et al. "You know what to do": proactive detection of YouTube videos targeted by coordinated hate attacks. *Proceedings of the ACM on Human-Computer Interaction*, v. 3, n. CSCW, Article 207, 2019.

[71] ALDOUS, Kholoud Khalil; AN, Jisun; JANSEN, Bernard J. Measuring 9 emotions of news posts from 8 news organizations across 4 social media platforms for 8 months. *ACM Transactions on Social Computing*, 2022.

[72] PIMENTEL, Mariano; FILIPPO, Denise; SANTOS, Thiago Marcondes dos. Design Science Research: pesquisa científica atrelada ao design de artefatos. *RE@D – Revista de Educação a Distância e Elearning*, v. 3, n. 1, 2020.

A. Apêndice

A.1 Diagramas do artefato

Esta subseção apresenta diagramas estruturais e comportamentais do *SocialDash*, com o objetivo de complementar a descrição conceitual e metodológica do artefato. Os modelos ilustram os principais componentes do sistema, os fluxos operacionais e a integração entre coleta de dados, inferência analítica e visualização.

Os diagramas concentram-se na instanciação computacional do artefato, evidenciando a articulação entre serviços externos, módulos internos e mecanismos utilizados nas tarefas investigativas relacionadas à polarização, monetização e toxicidade.

A Figura A.1 apresenta o diagrama de caso de uso do sistema. A partir de URLs, vídeos, canais ou consultas textuais, o *SocialDash* realiza coleta de dados, obtenção de transcrições, estimativas de monetização, análise de toxicidade e inferência discursiva, retornando visualizações analíticas integradas.

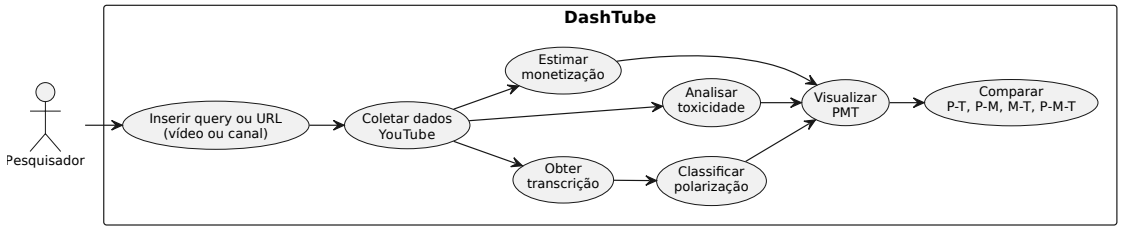


Figura A.1: Diagrama de caso de uso do SocialDash para análise PMT.

A Figura A.2 apresenta o diagrama de classes simplificado do *SocialDash*. O modelo organiza entidades persistentes — como usuários, buscas, canais, vídeos e comentários — juntamente com componentes responsáveis por coleta, processamento e inferência analítica.

A separação entre persistência de dados e lógica de análise contribui para a modularidade da implementação, permitindo integrar diferentes serviços externos, rotinas de processamento e mecanismos analíticos ao fluxo operacional do artefato.

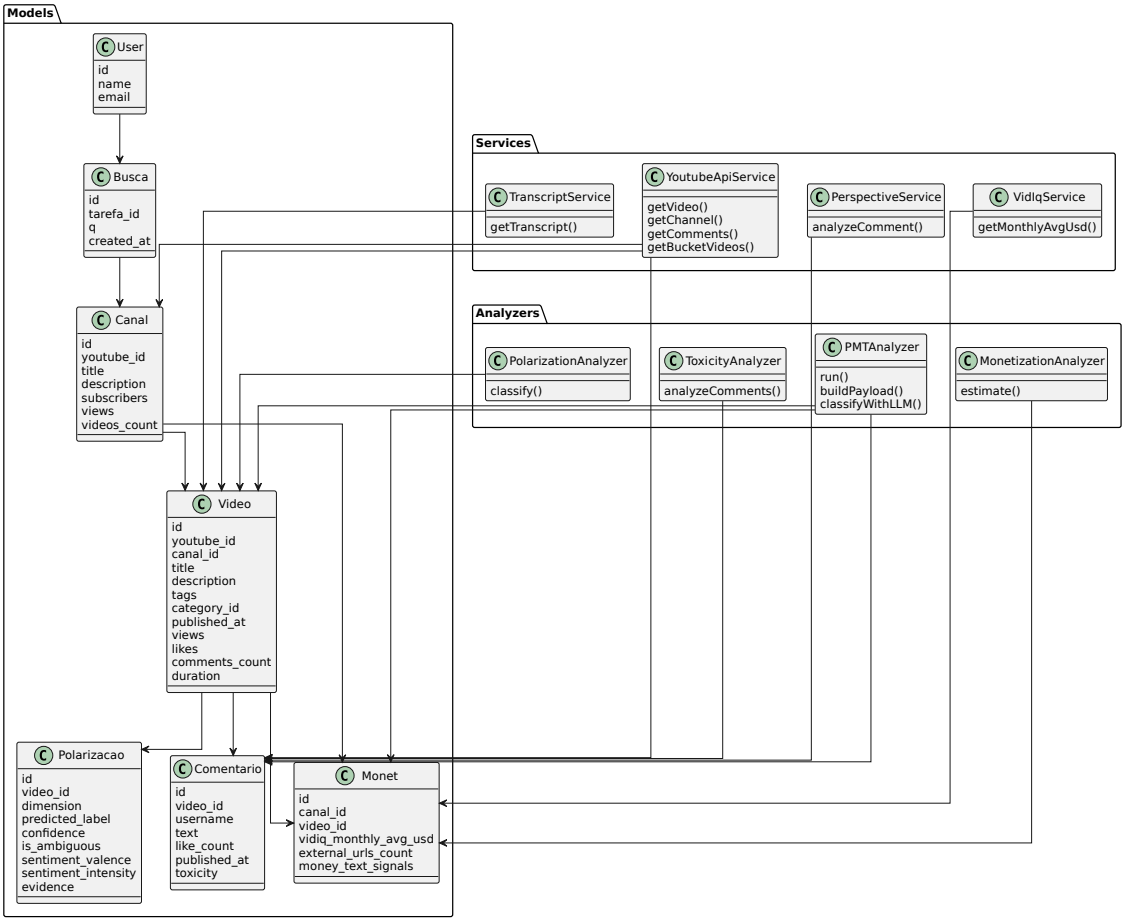


Figura A.2: Diagrama de classes simplificado do SocialDash.

A Figura A.3 apresenta o ciclo de estados do *SocialDash* durante a execução das análises. O fluxo inicia a partir de entradas fornecidas pelo pesquisador, como URLs, vídeos, canais ou consultas textuais, evoluindo para estados relacionados à coleta de dados, transcrição, análise de toxicidade, estimação de monetização e inferência de polarização.

O modelo também prevê estados parciais e situações de falha associadas a dependências externas, permitindo a geração de dashboards incompletos, porém ainda informativos para fins exploratórios.

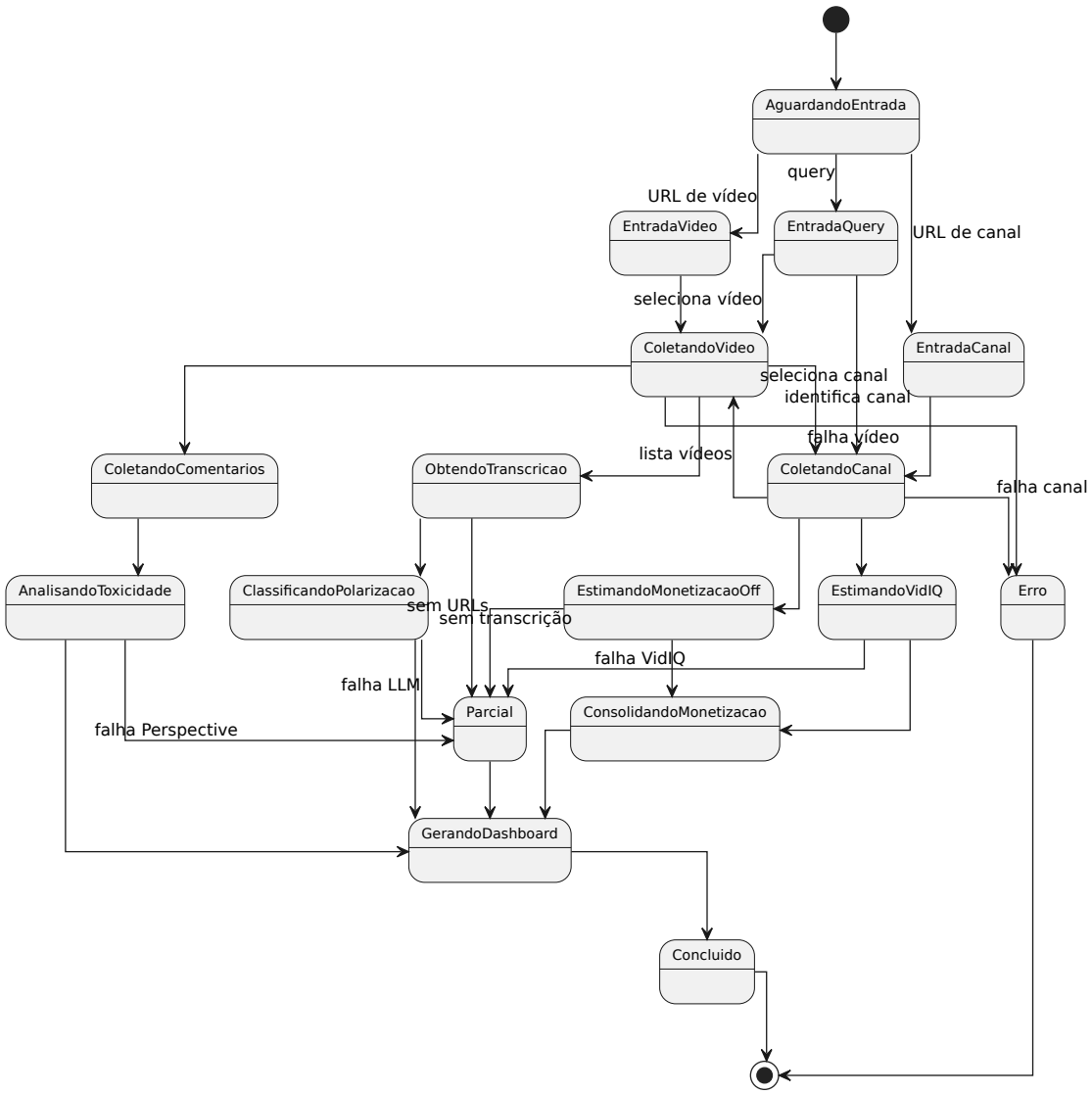


Figura A.3: Diagrama de estados do fluxo de análise PMT no SocialDash.

A Figura A.4 apresenta a sequência operacional do *SocialDash*. O fluxo descreve a interação entre interface, módulos de coleta, mecanismos de processamento e visualização analítica.

O diagrama evidencia a ordem das operações de aquisição de dados, obtenção de transcrições e comentários, execução das análises e retorno dos resultados integrados ao usuário.

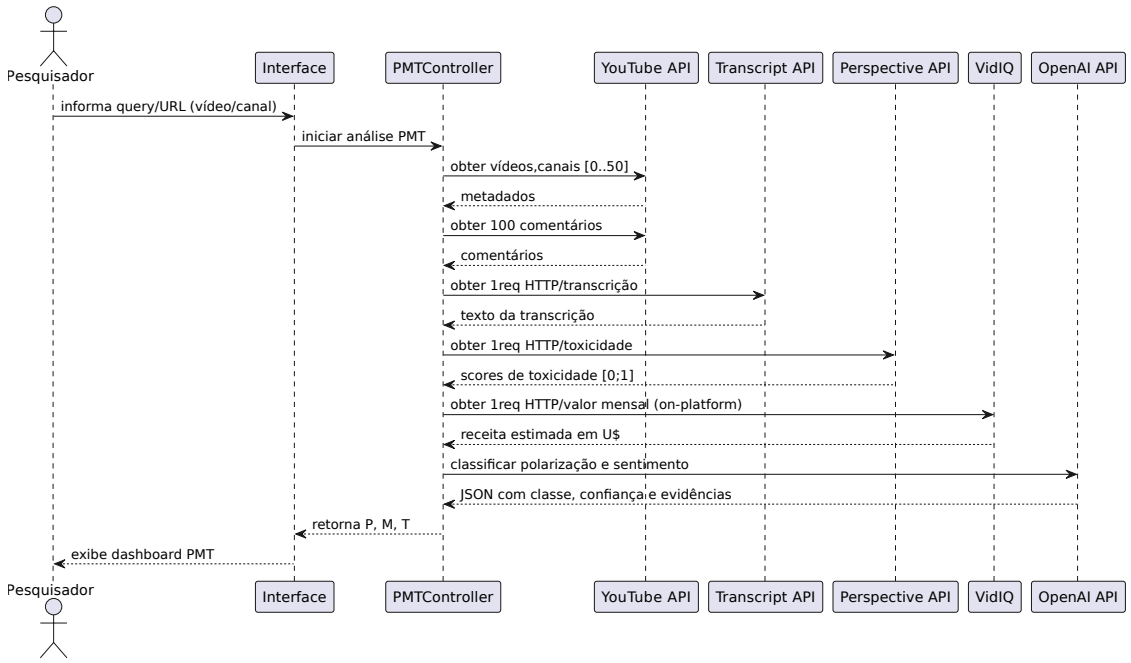


Figura A.4: Diagrama de sequência da análise P-M-T no SocialDash.

A modelagem do *SocialDash* utiliza chaves primárias, chaves estrangeiras e restrições de unicidade para preservação da integridade relacional e prevenção de duplicidades. Os registros também incorporam marcação temporal (*timestamp*) e campos normalizados, como *slug*, visando padronização textual e rastreabilidade das coletas.

Os tipos de dados foram definidos conforme a natureza das informações persistidas, incluindo estruturas textuais extensas, dados semiestruturados e métricas numéricas contínuas.

A Figura A.5 apresenta o modelo entidade-relacionamento simplificado da base de dados, organizado em torno das entidades de buscas, canais, vídeos, comentários, registros históricos e sinais de monetização.

Essa estrutura é complementada por tabelas auxiliares relacionadas a monetização e histórico temporal de canais. Embora parte das análises possa ser executada em memória, a persistência relacional permite auditoria, reprocessamento e rastreabilidade dos dados coletados.

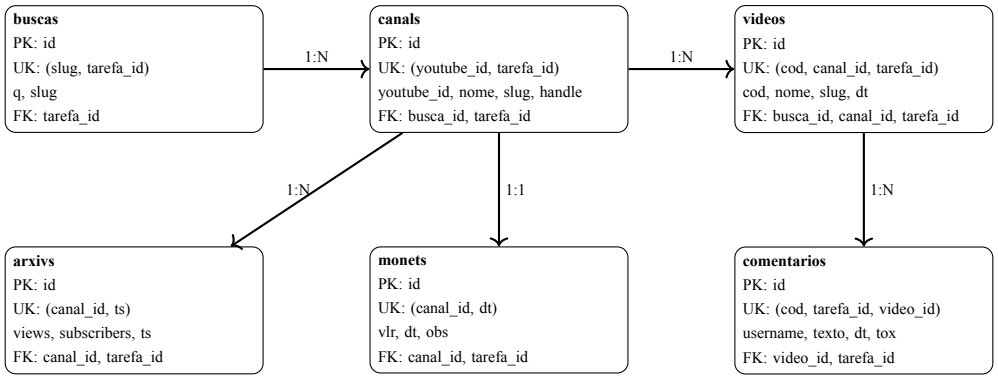


Figura A.5: Esquema entidade-relacionamento simplificado da base de dados do SocialDash.

A escolha dos tipos de dados seguiu critérios de adequação semântica e operacional. Campos textuais, métricas numéricas, estruturas flexíveis e registros temporais foram modelados conforme sua função analítica e requisitos de processamento.

A Tabela A.1 apresenta a estrutura relacional simplificada da base de dados, incluindo tipos, tamanhos, restrições e justificativas associadas aos principais campos utilizados pelo *SocialDash*.

Em síntese, a modelagem relacional do *SocialDash* busca representar os principais elementos analíticos associados ao ecossistema do YouTube, incorporando mecanismos de integridade, rastreabilidade e extensibilidade.

O uso de chaves relacionais, restrições de unicidade e campos normalizados contribui para evitar duplicidades e facilitar o reprocessamento dos dados coletados. Além disso, estruturas auxiliares relacionadas a histórico temporal e monetização permitiram explorar limitações operacionais e desafios metodológicos associados à coleta em plataformas externas.

Tabela A.1: Estrutura relacional principal da base de dados do SocialDash

Tabela	Campo	Tipo	Chave	Tamanho	Justificativa
buscas	id	bigint	PK	8 bytes	Identificador interno da busca
buscas	q	string	–	255	Armazena a query original digitada pelo pesquisador
buscas	slug	string	UK comp.	255	Normalização textual para rotas, filtros e unicidade contextual
buscas	tarefa_id	bigint	FK	8 bytes	Vincula a busca à coleta/tarefa executada
canals	id	bigint	PK	8 bytes	Identificador interno do canal
canals	youtube_id	string	UK comp.	255	Identificador externo do canal no YouTube
canals	nome	string	–	255	Nome visível do canal
canals	slug	string	–	255	Recurso normalizado para rota e indexação
canals	handle	string	–	255	Identificador público do canal (@handle)
canals	desc	text	–	longo	Descrição textual do canal
canals	keywords	json	–	variável	Estrutura flexível para palavras-chave e metadados
canals	inscritos	bigInteger	–	8 bytes	Escala elevada justifica tipo numérico amplo
canals	views	bigInteger	–	8 bytes	Total de visualizações do canal
canals	dt	date	–	3 bytes	Data relevante para análise temporal
canals	local	string	–	255	País/localização do canal
canals	categ	string	–	255	Categoria do canal
canals	videos	bigInteger	–	8 bytes	Número total de vídeos
canals	busca_id	bigint	FK	8 bytes	Relaciona canal à busca que o recuperou
canals	tarefa_id	bigint	FK	8 bytes	Contextualiza a coleta
videos	id	bigint	PK	8 bytes	Identificador interno do vídeo
videos	cod	string	UK comp.	255	Identificador externo do vídeo; recomendável renomear para *youtube_id*
videos	nome	string	–	255	Título curto do vídeo
videos	slug	string	–	255	Recurso normalizado para rota e indexação
videos	desc	text	–	longo	Descrição textual do vídeo
videos	caption	text	–	longo	Campo de apoio para transcrição/legenda
videos	hashtags	json	–	variável	Armazena hashtags com flexibilidade estrutural
videos	duration	unsignedInteger	–	4 bytes	Duração em segundos para cálculo e comparação
videos	categ_id	unsignedInteger	–	4 bytes	Categoria numérica do YouTube
videos	lang	string	–	255	Idioma do conteúdo, relevante para NLP
videos	dt	datetime	–	8 bytes	Data/hora de publicação ou coleta
videos	nlp1 / nlp2	float	–	4 bytes	Scores analíticos; recomendável renomear semanticamente
videos	gpt	text	–	longo	Saída textual/classificatória de LLM
videos	busca_id/canal_id	bigint	FK	8 bytes	Mantêm o contexto hierárquico do vídeo
comentarios	id	bigint	PK	8 bytes	Identificador interno do comentário
comentarios	cod	string	UK comp.	255	Numero total de comentários
comentarios	username	string	–	255	Autor do comentário; pode ser anonimizável
comentarios	texto	text	–	longo	Unidade discursiva principal de análise
comentarios	likes / dislikes	unsignedInteger	–	4 bytes	Métricas de reação da audiência
comentarios	dt	timestamp	–	4 bytes	Temporalidade explícita da interação
comentarios	tox	float	–	4 bytes	Score contínuo de toxicidade
comentarios	video_id / tarefa_id	bigint	FK	8 bytes	Vínculo com vídeo e coleta
monets	id	bigint	PK	8 bytes	Identificador do registro de monetização
monets	vlr	unsignedInteger	–	4 bytes	Proxy de valor monetário agregado ou estimado
monets	dt	datetime	UK comp.	8 bytes	Marca temporal do proxy por canal
monets	obs	text	–	longo	Observações contextuais sobre monetização
monets	canal_id / tarefa_id	bigint	FK	8 bytes	Relaciona monetização ao canal e à coleta
arxivs	id	bigint	PK	8 bytes	Identificador interno da série histórica experimental
arxivs	ts	datetime	UK comp.	8 bytes	Timestamp da observação histórica
arxivs	url	string	–	255	URL observada no Web Archive
arxivs	obs	text	–	longo	Observações da coleta arquivística
arxivs	canal_id / tarefa_id	bigint	FK	8 bytes	Relação com canal e coleta

A.1.1 Diagrama de fluxo de dados

A Figura A.6 apresenta o fluxo geral de funcionamento do *SocialDash*. O processo é orientado por conjecturas analíticas relacionadas à polarização, monetização e toxicidade, permitindo diferentes perspectivas de investigação sobre o ecossistema do YouTube.

O fluxo inicia com a entrada de consultas, URLs ou IDs de vídeos e canais, seguida pelas etapas de coleta, hidratação e processamento dos dados. A partir dessas informações,

o sistema aciona módulos externos de análise, realiza inferências analíticas e armazena snapshots para fins de rastreabilidade e reprocessamento.

Por fim, os resultados são apresentados em dashboards, gráficos e widgets analíticos, integrando sinais discursivos, métricas de engajamento, monetização e toxicidade em uma interface exploratória unificada.

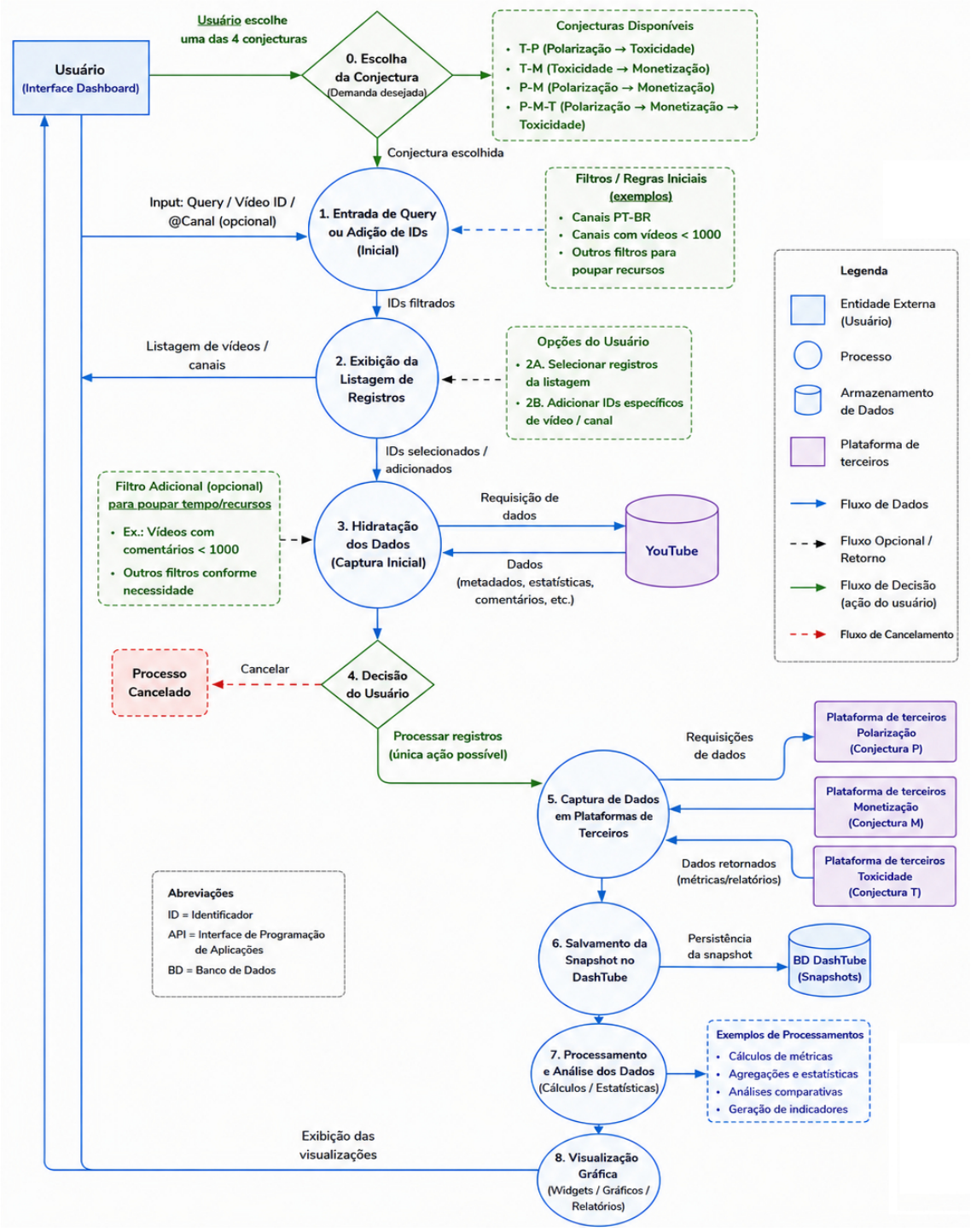


Figura A.6: Diagrama de fluxo de dados do SocialDash

O diagrama de fluxo de dados apresentado anteriormente representa uma abstração conceitual do funcionamento do artefato, descrevendo os principais fluxos de entrada, processamento, integração externa, persistência e visualização de dados.

A implementação prática dessa arquitetura materializa-se no *SocialDash*, desenvolvido como instanciação computacional do artefato proposto. Nesse contexto, o DFD representa o modelo lógico e operacional da solução, enquanto o *SocialDash* corresponde à sua realização concreta, integrando módulos de coleta, armazenamento, processamento analítico e visualização exploratória.

A.2 Tecnologias de implementação

A seleção das tecnologias foi orientada pelos requisitos do artefato, incluindo integração de múltiplas fontes de dados, automação de coleta, persistência de dados heterogêneos e suporte a visualizações interativas. A Tabela A.2 apresenta o ambiente tecnológico adotado e as justificativas funcionais de cada componente.

É importante destacar que essas escolhas constituem decisões de engenharia do artefato, e não prescrições diretas da literatura.

A aplicação foi desenvolvida em **PHP 8.3**, responsável pela lógica de processamento, integração com APIs e geração de saídas em `HTML` e `JSON`. Sobre essa base, foi utilizado o **Laravel 12**, estruturado em arquitetura *MVC*, responsável pela organização da aplicação, prototipação, autenticação, segurança e gerenciamento de rotas.

A persistência dos dados foi realizada em **MySQL 8.0**, adequado à modelagem relacional das entidades (buscas, canais, vídeos e comentários). O ambiente de execução utiliza **Linux Ubuntu 24.04**, permitindo controle direto via *shell* e maior flexibilidade operacional.

Na camada de interface, foi adotado o **Tailwind CSS 4**, responsável pela estilização da aplicação, e a biblioteca **Chart.js**, utilizada na construção de visualizações interativas.

O gerenciamento de dependências foi realizado com **Composer 2.4**, possibilitando a integração de bibliotecas externas, incluindo ferramentas de automação de navegação e acesso a conteúdos Web.

O sistema integra múltiplas APIs externas, como **Perspective API** (toxicidade), **Google NLP** (análise textual), **OpenAI** (inferência via LLM) e **GotIt** (NLP em português - durante o desenvolvimento da tese foi descontinuada) e **TranscriptAPI**. As credenciais

são armazenadas no arquivo `.env`, garantindo segurança e separação entre código e configuração.

O código-fonte está disponível em `https://github.com/lga37/dr2.git`, e o sistema encontra-se implantado em `www.tesedoutorado.com.br`.

Em síntese, o ambiente adotado atende aos requisitos de modularidade, integração e visualização necessários ao artefato e esta sumarizado na Tabela A.2.

Tabela A.2: Requisitos de ambiente e tecnologias utilizadas no SocialDash

Camada	**Tecnologia**	**Versão**	**Função no artefato**
Linguagem de programação	PHP	8.3	Camada principal de aplicação; integração com APIs, processamento no servidor, geração de saídas `HTML`/`JSON`
Framework Web	Laravel	12	Estrutura MVC, prototipação, rotas, autenticação, segurança, sessões, validação, organização modular
Banco de dados	MySQL	8.0	Persistência relacional de queries, canais, vídeos, comentários, snapshots e variáveis derivadas
Sistema operacional	Ubuntu Linux	24.04	Ambiente de execução do servidor, automação, controle via *shell*, integração com dependências externas
Estilização	Tailwind CSS	4	Estilo visual de páginas, tabelas, textos e componentes de interface
Biblioteca de gráficos	Chart.js	–	Construção de gráficos interativos e amigáveis para análise exploratória
Gerenciador de dependências	Composer	2.4	Instalação e gerenciamento de pacotes PHP externos
API de toxicidade	Perspective API	–	Cálculo de escores probabilísticos de toxicidade em comentários
API de NLP	Google NLP	–	Processamento de linguagem natural e análise textual
API de LLM	OpenAI API	–	Classificação contextual, inferência textual e apoio analítico
API de NLP em português	GotIt	–	Apoio à análise textual com foco em conteúdo em português
Configuração sensível	Arquivo `.env` (Laravel)	–	Armazenamento de chaves de API, senhas e variáveis de ambiente
Versionamento	GitHub	–	Repositório do código-fonte: `https://-github.com/lga37/dr2.git`
Hospedagem provisória	X	–	Ambiente de nuvem com acesso ao servidor Linux e implantação da instância `tesedoutorado.com.br`

A.3 Resumo das principais funções do *backend* do artefato

A.3.1 Conjectura — polarização e toxicidade

A conjectura C1 foi implementada no *SocialDash* por meio do *Widget 1 — Toxicidade da Audiência*, desenvolvido em Laravel, Livewire e Blade. O fluxo inicia com a seleção de vídeos via *queries*, URLs ou IDs específicos do YouTube.

Após a seleção, a função `avaliarVideos()` coordena o pipeline de coleta e análise. Os comentários são obtidos pela *YouTube Data API*, principalmente via endpoint `commentThreads`, utilizando a função `getCommentsByBucketsRelevance()` para distribuição temporal dos comentários em *buckets*.

A análise de toxicidade utiliza a Perspective API, responsável pela geração de escores contínuos de toxicidade. Esses valores são agregados por funções internas, como `toxMedia()`, produzindo métricas comparativas entre vídeos e séries temporais de evolução discursiva.

Os resultados são apresentados em tabelas e gráficos interativos desenvolvidos com Chart.js, integrando métricas de toxicidade, evolução temporal e classificações ideológicas inferidas a partir do conteúdo analisado.

A implementação também incorpora mecanismos de amostragem e limitação de processamento, visando reduzir custos computacionais e dependências externas durante a execução das análises.

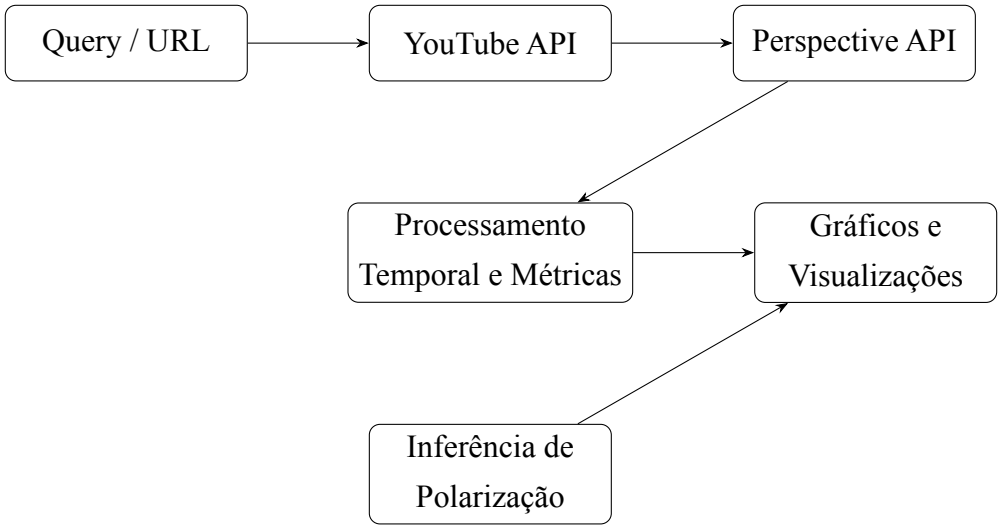


Figura A.7: Fluxo simplificado de processamento da conjectura C1 no SocialDash.

A.3.2 Conjectura — monetização e toxicidade

A conjectura C3 foi implementada na Tarefa 2 do *SocialDash*, integrando métricas de toxicidade e *proxies* de monetização em análises comparativas entre canais do YouTube. O fluxo inicia com a seleção de canais via *query*, URL ou ID, utilizando os componentes `Tarefa2.php` e `tarefa2.blade.php`.

A função `avaliarMonetizacaoToxicidade()` coordena o pipeline analítico, organizando os canais em unidades comparáveis. Os vídeos são coletados pela função

`getAllVideos()`, hidratados com metadados adicionais e distribuídos em janelas temporais por meio de `pmt_bucket_periods()`.

A análise por bucket é executada pela função `pmt_analisar_bucket_mt()`, responsável pelo cálculo de métricas de toxicidade derivadas da Perspective API e pela integração de sinais estimados de monetização. Entre os indicadores utilizados estão toxicidade média, toxicidade máxima, taxa de alta toxicidade, estimativas do VidIQ e quantidade de URLs externas associadas ao canal.

Os resultados são consolidados em estruturas comparativas exibidas na interface do sistema, permitindo observar relações temporais entre discurso, engajamento e monetização. A implementação também utiliza mecanismos de amostragem e limitação de coleta, reduzindo custos computacionais e dependências externas durante a execução das análises.

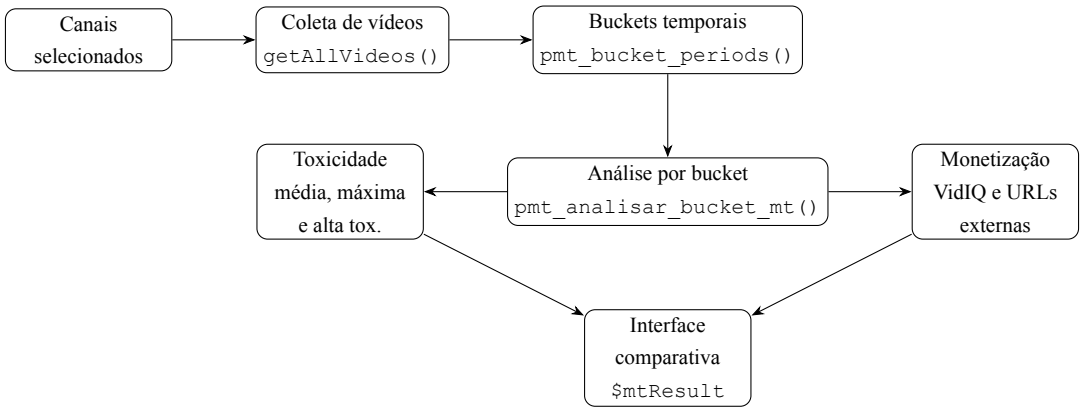


Figura A.8: Fluxo simplificado de processamento da conjectura C3 no SocialDash.

A.3.3 Conjectura — polarização e monetização

A conjectura C2 foi implementada na Tarefa 3 do *SocialDash*, voltada à análise integrada entre polarização discursiva e sinais estimados de monetização em canais do YouTube. O fluxo inicia com a seleção de canais por *query*, URL ou ID, utilizando os componentes `Tarefa3.php` e `tarefa3.blade.php`.

Após a validação da seleção, cada canal é tratado como unidade analítica. O sistema coleta vídeos, organiza os conteúdos em *buckets* temporais e processa sinais textuais associados a títulos, descrições, transcrições e tags.

A monetização é operacionalizada por indicadores aproximados, como estimativas do VidIQ, quantidade de URLs externas e média de URLs por vídeo. A polarização é inferida a partir de sinais textuais e classificações discursivas, apoiadas por rankings lexicais e

nuvens de palavras.

Os resultados são apresentados em cartões comparativos por canal, combinando indicadores temporais, sinais econômicos e estruturas textuais exploratórias. As nuvens de palavras atuam como apoio interpretativo, permitindo identificar temas recorrentes sem substituir análises semânticas mais aprofundadas.

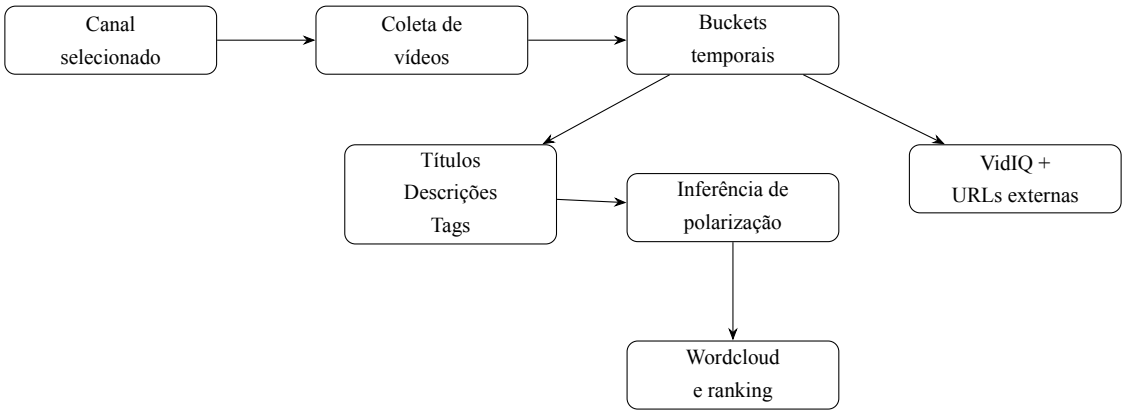


Figura A.9: Fluxo simplificado de processamento da conjectura C2 no SocialDash.

A.3.4 Conjectura — integração entre polarização, toxicidade e monetização

A conjectura C4 foi implementada no *Widget 4* do *SocialDash*, integrando em uma única rotina analítica sinais de polarização, toxicidade e monetização. O fluxo inicia com a seleção de vídeos por *query*, URL ou ID, utilizando os componentes `Tarefa4.php` e `tarefa4.blade.php`.

A função `avaliarVideos()` coordena o pipeline principal, persistindo buscas, canais e vídeos selecionados, além de executar as rotinas de toxicidade, monetização e inferência ideológica. Para viabilidade computacional, a implementação limita inicialmente a análise a pequenos conjuntos de vídeos.

A toxicidade é calculada a partir de comentários coletados pela YouTube Data API via `commentThreads`, utilizando funções de amostragem temporal em *buckets*. Os comentários recebem escores contínuos da Perspective API, posteriormente agregados em métricas comparativas e séries temporais renderizadas com Chart.js.

A monetização é operacionalizada por *proxies* inferenciais, principalmente estimativas do VidIQ e extração de URLs externas presentes nas descrições dos vídeos. Já a polarização é inferida por modelos de linguagem a partir de títulos, descrições e tags dos vídeos analisados.

Os resultados são integrados em uma estrutura única de visualização, permitindo observar associações entre discurso, toxicidade e sinais econômicos em vídeos selecionados. A proposta possui caráter exploratório, não buscando estabelecer causalidade direta entre as dimensões analisadas.

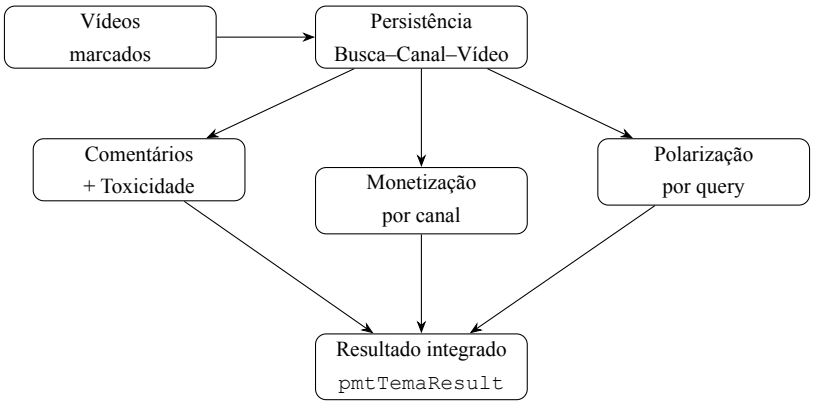


Figura A.10: Fluxo simplificado de processamento do Widget 4 no SocialDash.

A.4 Prompts utilizados

Listing A.1: Prompt utilizado para classificação discursiva dos vídeos

```
{
    ”videoId”: ”-wtrejN7MFw”,
    ”labs”: [
        ”base”
    ],
    ”labs_key”: ”base”,
    ”seed_name”: ”Padre Paulo Ricardo”,
    ”expected_label”: ”cristao”,
    ”videoTitle”: ”Homilia Diária.298: Sexta-feira da 19.ª Semana do Tempo Comum (P) - O
     Evangelho da família”,
    ”prompt”: ”Você é um classificador acadêmico de orientação discursiva em vídeos do
    YouTube.\n\n
    ...
}
```

Listing A.2: Prompt estruturado utilizado nos experimentos

```
{
    ”videoId”: ”-wtrejN7MFw”,
    ”labs”: [
        ”base”
    ],
    ”labs_key”: ”base”,
    ”seed_name”: ”Padre Paulo Ricardo”,
    ”expected_label”: ”cristao”,
    ”videoTitle”: ”Homilia Diária.298: Sexta-feira da 19.ª Semana do Tempo Comum (P) - O
     Evangelho da família”,
```

```
"prompt": "Você é um classificador acadêmico de orientação discursiva em vídeos do
YouTube.\n\n
A dimensão analítica esperada é: religiao\n\n
Configuração de dados usada neste teste: base\n\n
Classifique o VÍDEO, não o canal, apenas entre estas opções:\n
cristao, afro-brasileiro, islamico, ateu-secular, indefinido\n\n
----------------------------------------\n
REGRAS METODOLÓGICAS\n        ----------------------------------------\n\n
1. Use exclusivamente os dados fornecidos.\n
2. Não utilize conhecimento externo sobre o canal, o criador ou o veículo.\n
3. Não utilize o rótulo esperado do benchmark.\n
4. Classifique o enquadramento discursivo do vídeo, não apenas o tema.\n
5. Não classifique com base na reputação geral do canal.\n
6. A categoria do YouTube é apenas informação auxiliar, nunca decisiva.\n\n
----------------------------------------\n
HIERARQUIA DE EVIDÊNCIA\n        ----------------------------------------\n\n
Priorize nesta ordem:\n\n
1. Transcrição do vídeo, quando disponível.\n
2. Descrição do vídeo, quando disponível.\n
3. Título do vídeo.\n
4. Tags, hashtags e categoria, quando disponível.\n\n
Se houver conflito entre título\/descrição e transcrição, a transcrição deve
prevalecer.\n\n        ----------------------------------------\n
REGRAS DE DECISÃO\n        ----------------------------------------\n\n
- Classifique apenas se houver evidência discursiva explícita.\n
- Se o conteúdo for apenas informativo, descritivo, histórico, jornalístico ou
neutro, use \"indefinido\".\n
- Se houver sinais conflitantes, escolha um polo apenas quando houver predominância
clara.\n
- Se os sinais forem equilibrados ou insuficientes, use \"indefinido\".\n
- Não infira intenção ideológica implícita sem evidência textual.\n
- Não confunda crítica, descrição histórica ou discussão acadêmica com
posicionamento polarizado.\n\n        ----------------------------------------\n
DEFINIÇÕES ANALÍTICAS PARA CIÊNCIA\n        ----------------------------------------\
n\n
Quando a dimensão for \"ciencia\":\n\n
- \"cientifico\": conteúdo alinhado com evidências, consenso científico, método
científico, explicação educacional ou análise acadêmica.\n
- \"anti-ciencia\": conteúdo que rejeita, desacredita ou nega consenso científico
sem base robusta.\n
- \"conspiratorio\": conteúdo que sustenta narrativa conspiratória, causalidade
oculta, agentes secretos, manipulação deliberada ou explicação alternativa não
verificável.\n
- \"indefinido\": conteúdo descritivo, histórico, jornalístico, cultural ou sem
posição clara sobre ciência\/conspiração.\n\n
Importante:\n        - Só use \"conspiratorio\" se o vídeo defender ou promover uma
narrativa conspiratória como explicação central.\n\n
----------------------------------------\n
DEFINIÇÕES ANALÍTICAS PARA POLÍTICA\n
----------------------------------------\n\n
Quando a dimensão for \"politica\":\n\n
- \"esquerda\": enquadramento favorável a pautas progressistas, redistributivas,
trabalhistas, igualitárias, críticas ao conservadorismo ou à direita.\n
- \"direita\": enquadramento favorável a pautas conservadoras, liberais,
nacionalistas, religiosas conservadoras, anticomunistas, antipetistas ou críticas à
esquerda.\n
```

```
- \"centro\": enquadramento moderado, institucional, tecnocrático ou sem polarização
 clara.\n
- \"indefinido\": conteúdo factual ou jornalístico sem posição dominante clara.\n\n
      ----------------------------------------\n
DEFINIÇÕES ANALÍTICAS PARA RELIGIÃO\n
----------------------------------------\n\n
Quando a dimensão for \"religiao\":\n\n
- \"cristao\": enquadramento associado a doutrina, prática, defesa ou visão cristã.\
n
- \"afro-brasileiro\": enquadramento associado a religiões de matriz africana ou
afro-brasileira.\n
- \"islamico\": enquadramento associado ao Islã, cultura islâmica ou prática
muçulmana.\n
- \"ateu-secular\": enquadramento crítico à religião, secularista, ateísta ou laico
.\n
- \"indefinido\": conteúdo religioso descritivo, histórico ou sem posição clara.\n\n
       ----------------------------------------\n
AMBIGUIDADE\n        ----------------------------------------\n\n
Use is_ambiguous = true quando:\n\n
- houver sinais mistos relevantes;\n
- houver evidência insuficiente;\n
- o vídeo tratar do tema sem assumir posição clara;\n
- título, descrição e transcrição apontarem em direções diferentes.\n\n
Use is_ambiguous = false quando houver evidência consistente para a classificação
escolhida.\n\n       ----------------------------------------\n
EVIDÊNCIA OBRIGATÓRIA\n        ----------------------------------------\n\n
No campo evidence:\n\n
- liste 1 ou 2 evidências curtas;\n
- use trechos literais ou quase literais dos dados fornecidos;\n
- não invente evidência;\n        - não use conhecimento externo;\n
- se não houver evidência suficiente, explique isso no short_reason.\n\n
----------------------------------------\n
INTERPRETAÇÃO DOS CAMPOS\n        ----------------------------------------\n\n
confidence:\n
- 0.80 a 1.00 = evidência clara e consistente.\n
- 0.60 a 0.79 = evidência moderada.\n
- abaixo de 0.60 = baixa segurança.\n\n\n
sentiment_hint.valence:\n
- \"positivo\" quando o tom predominante for favorável, celebratório ou afirmativo.\
n
- \"negativo\" quando o tom predominante for crítico, hostil, alarmista ou
acusatório.\n
- \"neutro\" quando o tom for informativo\/descritivo.\n
- \"indefinido\" quando não houver evidência suficiente.\n\n
sentiment_hint.emotional_intensity:\n
- 0.00 a 0.20 = baixa carga emocional.\n
- 0.21 a 0.50 = carga emocional leve.\n
- 0.51 a 0.75 = carga emocional moderada.\n
- 0.76 a 1.00 = carga emocional alta.\n\n
----------------------------------------\n
DADOS DISPONÍVEIS\n        ----------------------------------------\n\n
[BASE]\nTítulo do vídeo: Homilia Diária.298: Sexta-feira da 19.ª Semana do Tempo
Comum (P) - O Evangelho da família\nIdioma esperado: BR\n\n
----------------------------------------\n
FORMATO DE RESPOSTA\n        ----------------------------------------\n\n
Responda exclusivamente em JSON válido, sem markdown, sem comentários e sem texto
```

```
    fora do JSON:\n\n        {\n
    \"dimension\": \"religiao\",\n        \"predicted_label\": \"\",\n        \"
    confidence\": 0.0,\n        \"is_ambiguous\": false,\n
    \"evidence\": [],\n        \"short_reason\": \"\",\n        \"sentiment_hint\": {\n
    \"valence\": \"positivo|neutro|negativo|indefinido\",\n            \"
    emotional_intensity\": 0.0\n        }\n
    }",
    "created_at": "2026-04-28 23:23:38"
}
```


A.5 Interfaces do artefato



Figura A.11: Página inicial do protótipo *SocialDash*, apresentando a proposta geral da pesquisa, os objetivos das tarefas experimentais e o fluxo inicial de interação do usuário com o sistema. A interface introduz os conceitos de monetização, toxicidade e polarização no contexto do YouTube e direciona o participante para as tarefas associadas às conjecturas investigadas.

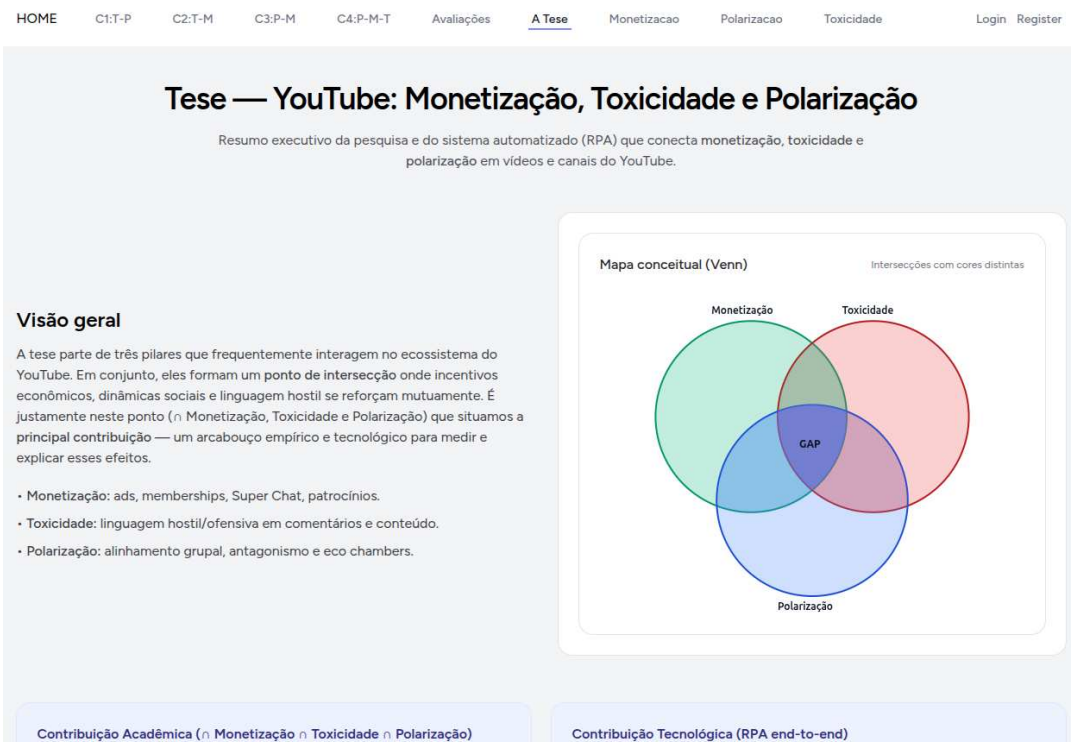


Figura A.12: Interface explicativa da tese e da arquitetura conceitual do *SocialDash*, apresentando a interseção entre monetização, toxicidade e polarização como núcleo analítico central da pesquisa. A tela resume visualmente os principais pilares teóricos e tecnológicos utilizados no artefato.

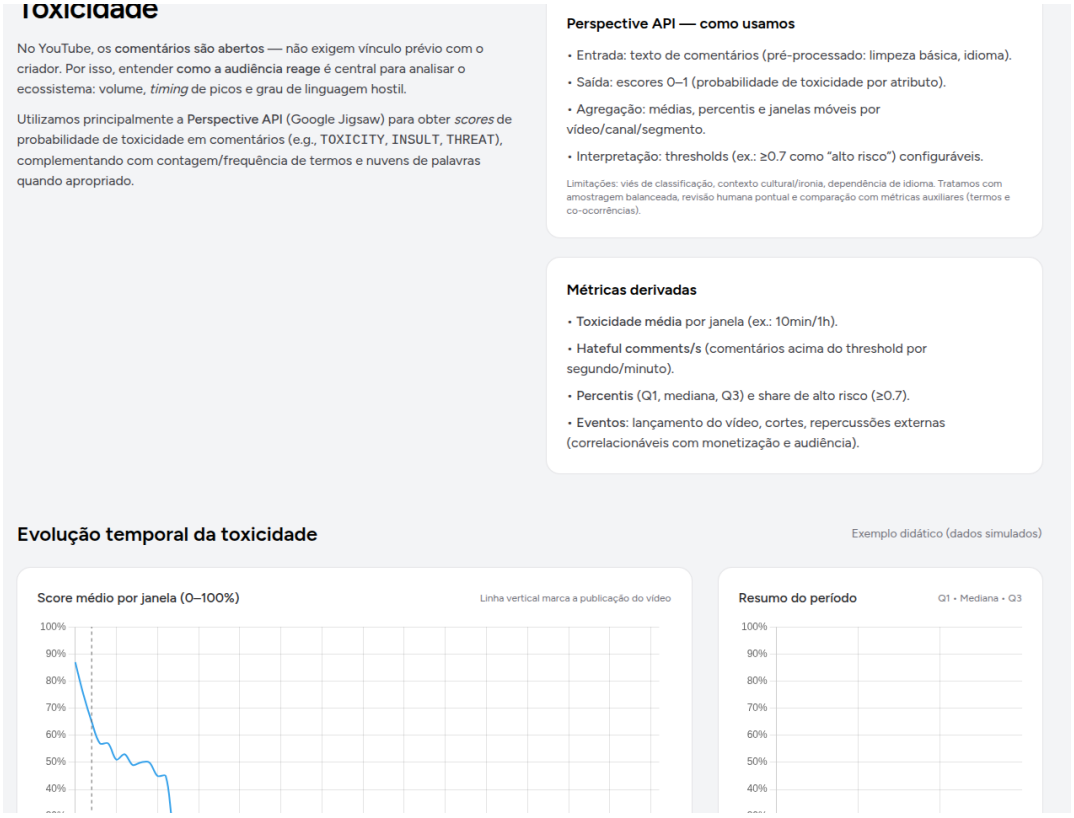


Figura A.13: Página explicativa sobre toxicidade, descrevendo o uso da Perspective API, mecanismos de agregação temporal dos comentários, métricas derivadas e exemplos de visualização da evolução da toxicidade ao longo do tempo.

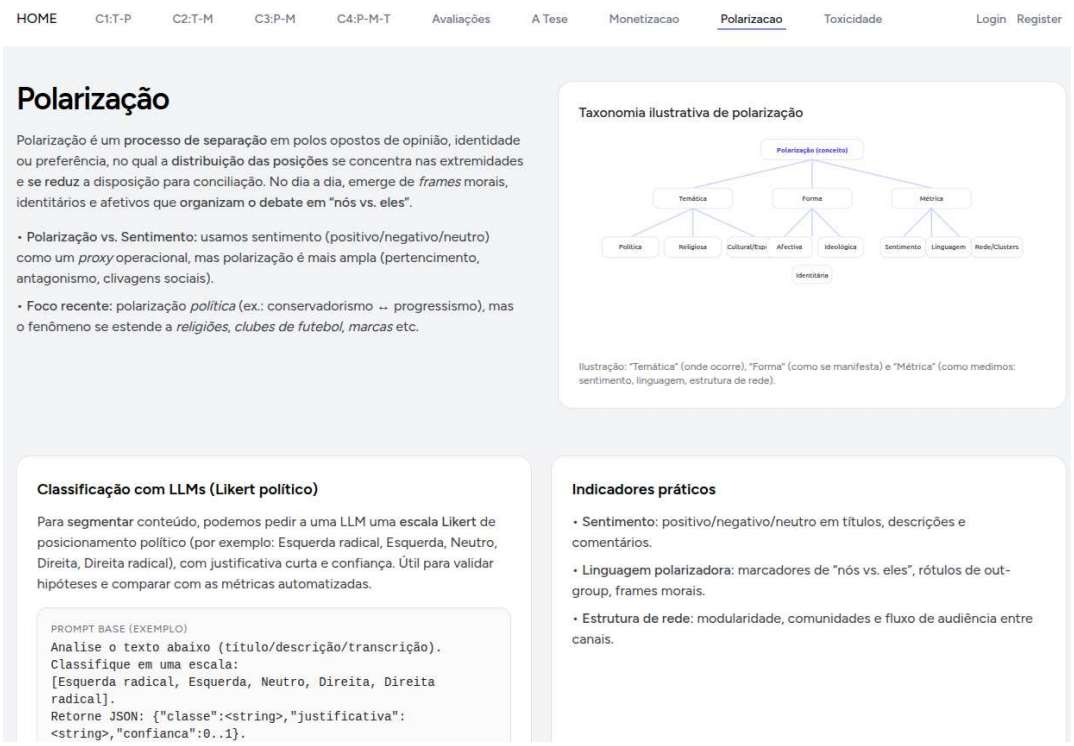


Figura A.14: Página explicativa sobre polarização discursiva, apresentando taxonomias ilustrativas, classificação ideológica assistida por LLMs e exemplos de indicadores utilizados pelo sistema para inferência de polarização em conteúdos do YouTube.

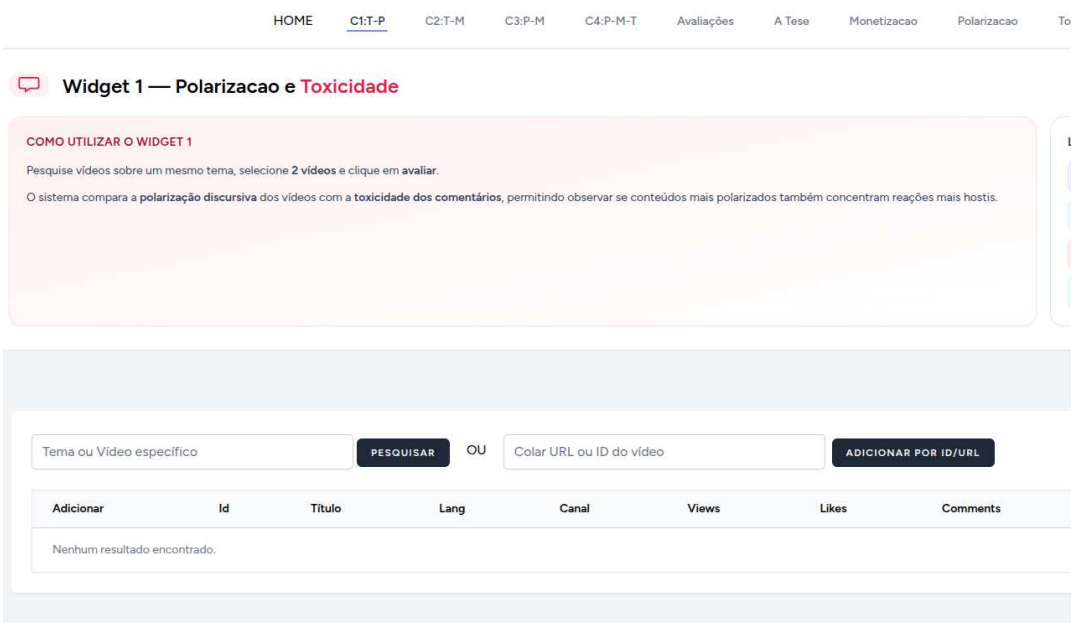


Figura A.15: Interface inicial de entrada da conjectura C1 (Polarização ↔ Toxicidade), permitindo ao usuário pesquisar vídeos, selecionar conteúdos específicos e iniciar o processo de avaliação comparativa entre sinais discursivos e toxicidade dos comentários.

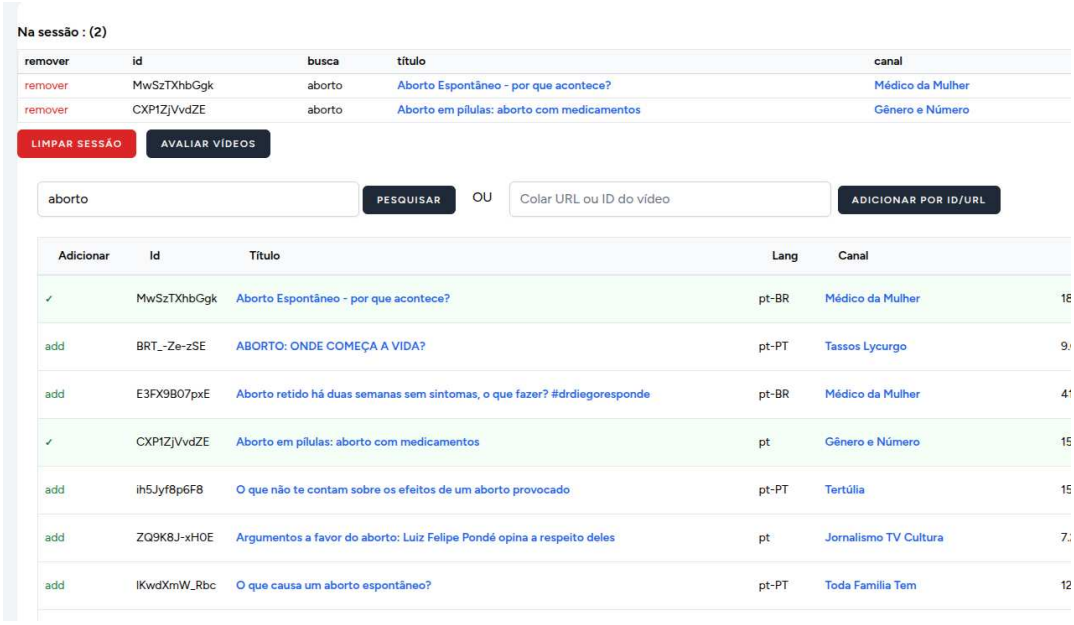


Figura A.16: Interface de seleção de vídeos utilizada nas tarefas experimentais do *SocialDash*. O sistema permite adicionar vídeos por busca textual, URL ou ID, formando uma sessão analítica temporária posteriormente utilizada nos widgets de avaliação das conjecturas.

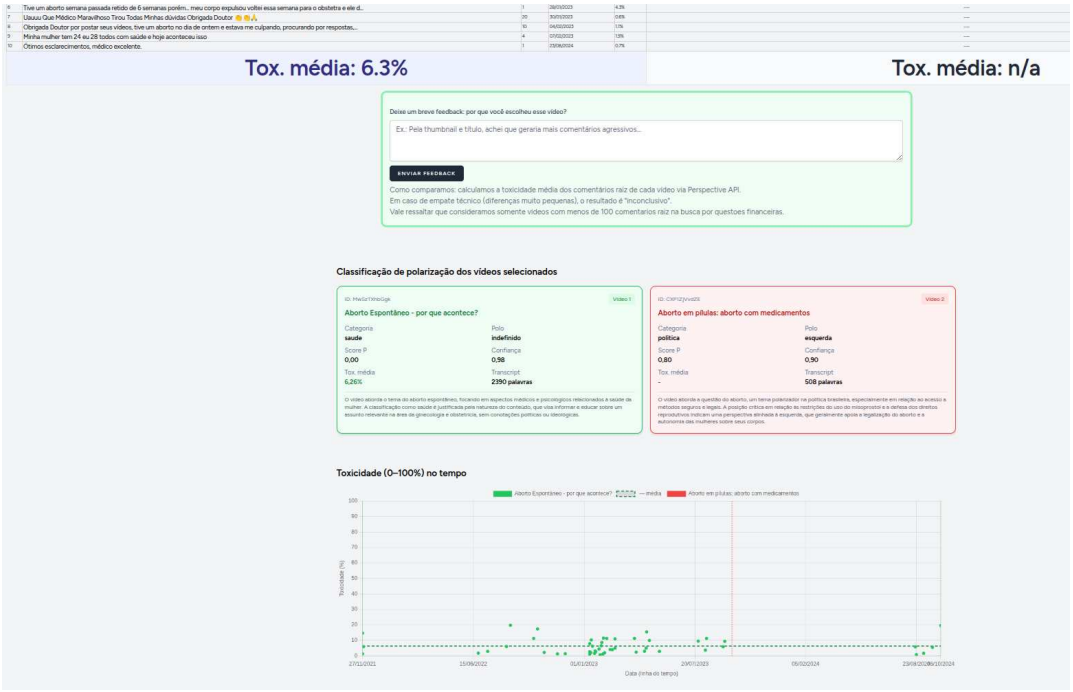


Figura A.17: Resultado da conjectura C1 (Polarização ↔ Toxicidade), integrando classificação ideológica inferencial dos vídeos, métricas derivadas da Perspective API e visualização temporal da evolução dos comentários tóxicos.

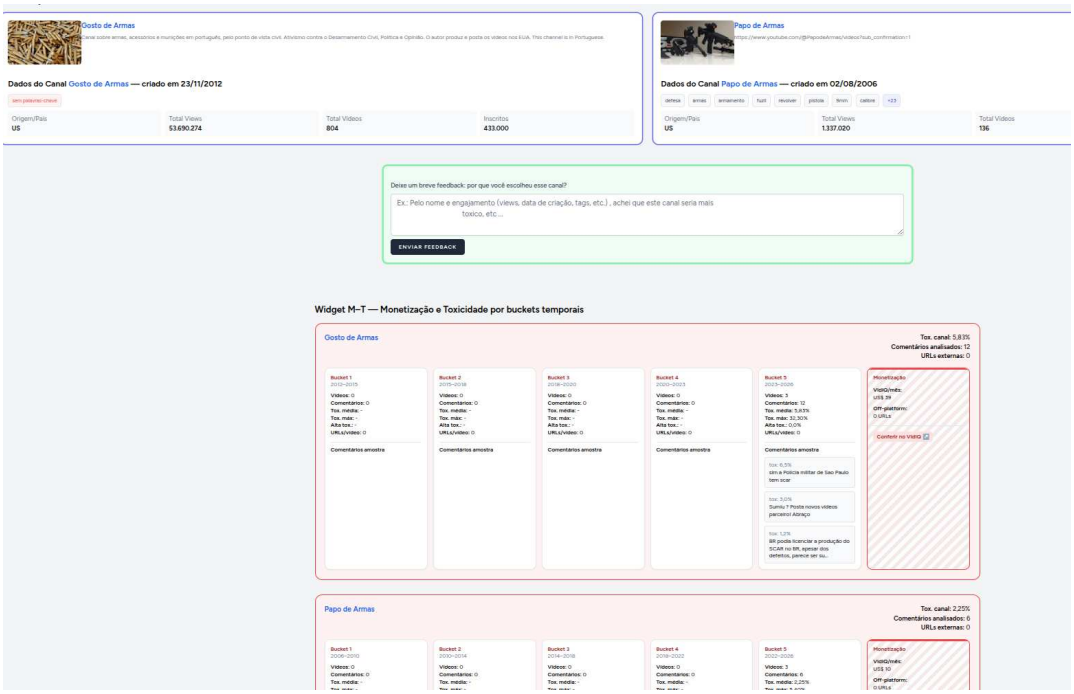


Figura A.18: Resultado da conjectura C2 (Toxicidade ↔ Monetização), apresentando buckets temporais, comentários amostrados, estimativas de monetização, URLs externas e métricas agregadas de toxicidade associadas aos canais analisados.

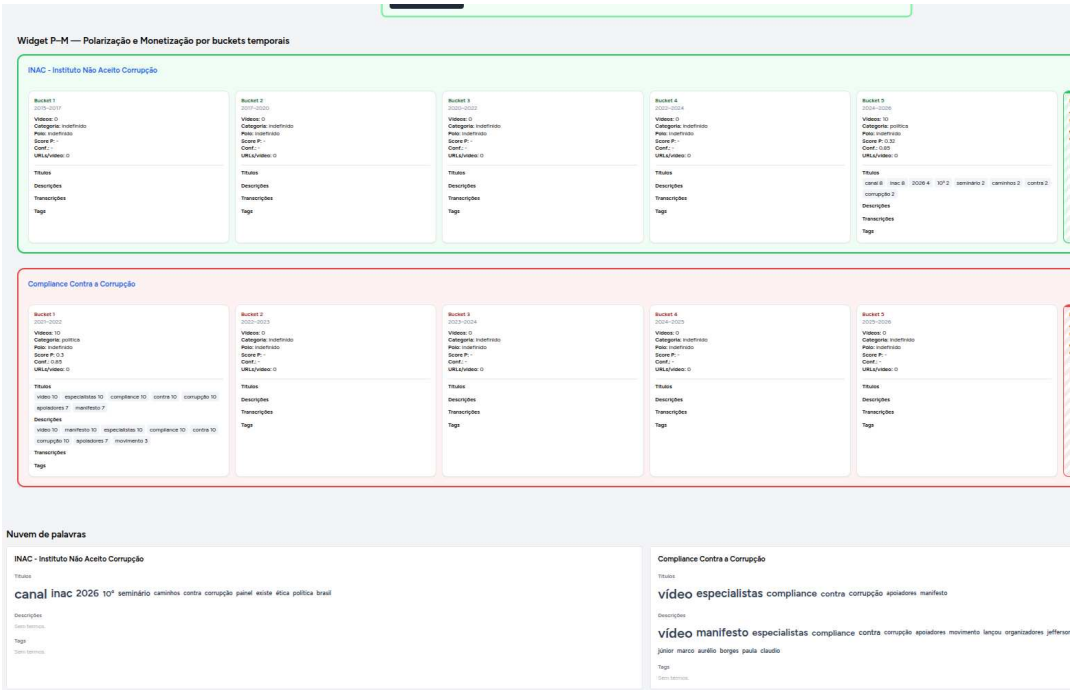


Figura A.19: Resultado da conjectura C3 (Polarização ↔ Monetização), utilizando buckets temporais e nuvens de palavras para representar padrões discursivos, classificação ideológica e possíveis relações entre alinhamento temático e indicadores aproximados de monetização.



Figura A.20: Resultado integrado da conjectura C4 (Polarização ↔ Monetização ↔ Toxicidade), apresentando em uma única interface analítica a classificação ideológica inferencial dos vídeos, estimativas de monetização, URLs externas, métricas derivadas de toxicidade e a evolução temporal dos comentários.