Análise de Sentimentos no YouTube para Conteúdo Infantil: IA para Reclassificar Resultados de Busca

Rafael Vargas Mesquita Santos[1], **João Victor de Salles**[1], **Flávio Izo**[1], **Sabrina Vargas**[2]

[1]Instituto Federal do Espírito Santo (IFES)

rafaelv@ifes.edu.br, joaovictormpl@gmail.com, fizo@ifes.edu.br

[2]Universidade do Espírito Santo

brinabrug@gmail.com

Abstract. *We present a native approach to re-rank children's YouTube searches via Sentiment Analysis. With a domain-specific corpus (2,749 sentences, balanced across three classes) and fine-tuning of BERTimbau, the final classifier — an* ensemble *of five cross-validation models — achieved 79.84% macro-F1 on the held-out test set (79.64% accuracy) and 91.2% F1 for the Negative class — a critical metric for child safety.* Random Oversampling *is applied only to the training* folds*; the test set (20%) remains untouched for generalization. A web prototype integrates an interpretable* Safety Score *to reorder results prioritizing educational videos. The solution operates in Brazilian Portuguese without machine translation, preserving linguistic nuances.*

Resumo. *Apresentamos abordagem nativa para reclassificar buscas infantis no YouTube via Análise de Sentimentos. Com corpus específico (2.749 sentenças, balanceado entre três classes) e ajuste fino do BERTimbau, o classificador final —* ensemble *de cinco modelos da validação cruzada — obteve 79,84% de F1-macro no teste retido (79,64% acurácia) e F1 91,2% na classe Negativo — métrica crítica para segurança infantil. O* Random Oversampling *é aplicado apenas nos* folds *de treino; o teste (20%) permanece intacto para generalização. Protótipo web integra* Score de Segurança *interpretável para reordenar resultados priorizando vídeos educativos. A solução opera em português do Brasil, sem tradução automática, preservando nuances linguísticas.*

1. Introdução

A onipresença de plataformas digitais como o YouTube exige soluções robustas de filtragem para públicos vulneráveis, como o infantil. A Análise de Sentimentos (AS) é central para automatizar a classificação de conteúdos subjetivos [Benevenuto et al. 2015]. Apesar de avanços com o BERTimbau [Souza et al. 2020], o português segue com menor representatividade e escassez de *datasets* específicos [Nunes 2023, Reis et al. 2015]. Traduzir textos para inglês degrada nuances [Reis et al. 2015].

Este trabalho parte do problema de como projetar, desenvolver e validar um filtro nativo em português capaz de classificar, com eficácia, conteúdos de vídeos infantis no YouTube. O objetivo geral é construir uma solução de software baseada em Processamento de Linguagem Natural (PLN) que identifique e restrinja conteúdos com

conotação emocional negativa (tristeza, raiva, medo), promovendo um ambiente mais seguro para crianças. O percurso metodológico inclui: (i) coleta e curadoria de um *corpus* de transcrições de vídeos infantis; (ii) pré-processamento e estruturação dos dados; (iii) desenvolvimento e ajuste fino de um modelo BERTimbau para classificação de sentimentos; (iv) implementação de um protótipo web que reordena buscas pelo *Score de Segurança*; (v) validação com métricas consagradas [Santos and Comarela 2025].

Do ponto de vista de Sistemas de Informação, a contribuição é dupla: tecnológica e social. Tecnologicamente, a solução supera abordagens genéricas e pipelines dependentes de tradução, preservando nuances culturais e linguísticas. Socialmente, entrega um mecanismo prático de curadoria de conteúdo infantil com base em evidências quantitativas e qualitativas. As principais contribuições são: (i) corpus rotulado de domínio; (ii) *ensemble* de modelos BERTimbau ajustados; (iii) protótipo funcional; (iv) *Score* interpretável para transparência ao usuário; (v) resultados robustos com alto F1[1] na classe Negativo.

Além disso, fundamentamos o recorte infantil pela maior incidência de expressões emocionais diretas, sentenças simples e vocabulário específico, que diferem marcadamente do discurso adulto e exigem recursos nativos. Avaliamos alternativas como léxicos generalistas e tradução automática, mas os riscos de perda de significado, ambiguidade e ruído motivaram uma arquitetura centrada em português do Brasil, com validação cruzada e *oversampling* para garantir desempenho estável em todas as classes.

No plano prático, o sistema proposto operacionaliza a curadoria semântica ao integrar a classificação de sentimentos ao fluxo de busca e ordenação, aproximando a pesquisa acadêmica da utilidade social. O mecanismo de pontuação (*Score*) confere transparência, substituindo decisões opacas por um indicador interpretável. Avaliações quantitativas — acurácia, F1 e matriz de confusão — e qualitativas — cenários reais de busca — sustentam a eficácia da solução. Os resultados mostram alta sensibilidade para a classe Negativo (F1 = 91,2%, *recall* = 89,1%), requisito crítico para a segurança infantil, com cobertura adequada das classes Neutro e Positivo, ainda que com confusão residual entre elas em razão de sua proximidade semântica.

Em relação à construção social, o sistema busca equilibrar bem-estar social com uso ético e sustentável da tecnologia: respeita privacidade ao operar sobre transcrições públicas sem dados pessoais; mitiga desigualdades e vieses por meio de curadoria balanceada e avaliação por classe; reduz assimetrias de poder e riscos de manipulação com transparência do *Score* e controle verificável por pais e educadores; e dialoga com abordagens interdisciplinares para orientar critérios de segurança e impacto social.

2. Fundamentação Teórica

PLN busca permitir que sistemas processem e compreendam linguagem natural [Gonzalez and Lima 2003, Firmino et al. 2025, Miranda and Rodrigues 2025]. Em síntese, etapas clássicas incluem: tokenização (segmentação em unidades mínimas), normalização (padronização de formas e remoção de ruídos), análise sintática (estrutura frasal), semântica (sentido literal) e pragmática (sentido contextual). Denomina-se *corpus* (plural *corpora*) um conjunto estruturado de textos representativos de um domínio

[1]F1 é a média harmônica entre precisão e *recall*, equilibrando ambas em uma única métrica.

linguístico, utilizado como base empírica para o treinamento, ajuste e avaliação de modelos de PLN. A eficácia do PLN depende criticamente de corpora representativos do domínio [Finatto et al. 2015].

A AS extrai polaridade (positivo/negativo/neutro) e pode estender-se à afetividade [Benevenuto et al. 2015, Rosa 2015]. Abordagens léxicas dependem de dicionários sensíveis a negação, intensificadores e n-gramas [Ramos and Freitas 2019, Rosa 2015], enquanto métodos supervisionados aprendem padrões de dados rotulados, com maior demanda de curadoria [Benevenuto et al. 2015]. No contexto infantil, em que léxico e sintaxe diferem do adulto, ferramentas genéricas e traduções automáticas tendem a perder nuances, reforçando a necessidade de soluções nativas de domínio [Benevenuto et al. 2015, Souza and Café 2018, Medeiros and de Freitas Neto 2025, Candido et al. 2025, Reis et al. 2015, Nunes 2023, Baracho et al. 2025]. Nesse cenário, corpora específicos e *fine-tuning* do BERTimbau em PT-BR [Souza et al. 2020] são estratégias adequadas para recuperar nuances de sentenças curtas e expressões emocionais diretas.

2.1. Trabalhos Relacionados

O modelo *Samba* [Le et al. 2022] combina metadados e legendas de vídeos em uma rede neural de fusão para identificar conteúdo impróprio para crianças no YouTube, atingindo 95% de acurácia em 70 mil vídeos. A abordagem, porém, opera exclusivamente em inglês e produz classificação binária, sem mecanismo de *re-ranking* ou pontuação contínua de segurança. Em [Chalkias et al. 2023], métodos léxicos como VADER e TextBlob foram aplicados a 167 mil comentários de vídeos educacionais para caracterizar o sentimento do público, revelando os limites dos léxicos generalistas em plataformas de vídeo. Da Rosa Jr. *et al.* [da Rosa Jr. et al. 2024] propuseram um classificador de *stance* em PLN para mapear a promoção de apostas no YouTube brasileiro, reforçando a urgência de mecanismos independentes de curadoria para grupos vulneráveis.

Nosso trabalho difere ao focar o universo infantil em português, construindo corpus próprio e realizando ajuste fino do BERTimbau sobre transcrições de vídeos. Assim, endereçamos lacunas de abordagens restritas ao inglês [Le et al. 2022, Chalkias et al. 2023] ou a domínios adultos [da Rosa Jr. et al. 2024], propondo *re-ranking* ativo voltado à segurança infantil com alta sensibilidade para conteúdo negativo.

A Tabela 1 sintetiza as diferenças entre as abordagens citadas e a presente proposta, evidenciando o conjunto específico de escolhas que a posicionam como contribuição original: idioma nativo, domínio infantil, técnica supervisionada baseada em *transformer* ajustado e saída combinada (rótulo categórico e *Score* contínuo).

Em particular, diferentemente do *Samba* [Le et al. 2022], das abordagens léxicas de [Chalkias et al. 2023] e do classificador de *stance* de da Rosa Jr. *et al.* [da Rosa Jr. et al. 2024] (cujo RoBERTa foi ajustado sobre corpus jurídico), este trabalho adota o BERTimbau (BrWaC, PT-BR geral) ajustado a corpus infantil para captar nuances de sentenças curtas. A saída não se limita a rótulo categórico: integra um *Score de Segurança* interpretável para sustentar reordenação de resultados e curadoria parental.

Tabela 1. Comparativo entre trabalhos relacionados e a proposta.

Trabalho	Idioma	Domínio	Técnica	Saída
[Le et al. 2022]	Inglês	YouTube (infantil)	Fusão neural (metadados+legendas)	Classificação
[Chalkias et al. 2023]	Inglês	YouTube (educacional)	Léxico (VADER, TextBlob)	Polaridade
[da Rosa Jr. et al. 2024]	PT-BR	YouTube (apostas)	Transformers (RoBERTa, Llama)	Posicionamento
Este trabalho	PT-BR	YouTube (infantil)	Transformers (BERTimbau)	Polaridade + *Score*

3. Metodologia

Esta seção descreve o percurso metodológico para construir uma solução nativa em português voltada à segurança infantil. Apresentamos a coleta e curadoria do corpus, as decisões de pré-processamento, o balanceamento e a arquitetura do protótipo. O objetivo é garantir reprodutibilidade, evitar *data leakage* e sustentar desempenho estável nas classes de sentimento. A rotulagem em três classes foi feita por quatro anotadores independentes (três alunos e um professor) com decisão por voto majoritário.

3.1. Critérios de Seleção dos Vídeos

Para a composição do corpus, foram definidos critérios de seleção que visavam abranger um espectro diversificado de conteúdos. A coleta incluiu vídeos de canais destinados ao público infantojuvenil e da plataforma YouTube Kids, que representam a base de conteúdos considerados seguros e apropriados. Em contrapartida, e com o objetivo de treinar o modelo a identificar conteúdos nocivos, também foram selecionadas transcrições de animações voltadas ao público adulto, como “South Park” e “The Boondocks”. A inclusão destes últimos justifica-se por possuírem um apelo estético visual que pode atrair a atenção de crianças, embora veiculem temáticas e linguagens profundamente inadequadas. Essa abordagem contrastiva é essencial para que o modelo de classificação aprenda a discernir os limites entre o conteúdo positivo/neutro e o negativo (potencialmente nocivo).

3.2. Coleta e Curadoria do Corpus

Foram coletadas transcrições de vídeos infantis (YouTube/YouTube Kids) e amostras contrastivas de animações adultas com apelo visual. As ferramentas Clipto e NotebookLM auxiliaram transcrição e organização [Clipto.ai 2025]. O corpus final inclui 2.749 frases rotuladas manualmente em três classes (Negativo, Neutro, Positivo) [Liu 2022], com distribuição balanceada entre as classes após esforço de curadoria. Desafios de reconhecimento automático de fala em contexto infantil incluem variabilidade vocal e ruído [Basak et al. 2023]. Além da extração, conduziu-se verificação de coerência das legendas, removendo trechos redundantes ou com ruído acentuado, e priorizando frases com carga emocional explícita e trechos informativos relevantes ao cotidiano infantil. A diversidade temática (educação, entretenimento, cotidiano familiar) foi considerada para reduzir vieses e ampliar representatividade. A Tabela 2 apresenta uma amostra estruturada do corpus utilizado, com duas frases representativas de cada classe.

3.3. Critérios de Rotulagem e Seleção de Conteúdo

As 2.749 frases foram rotuladas manualmente segundo diretrizes de polaridade: Negativo (tristeza, raiva, medo, insultos), Neutro (informativo/explicativo) e Positivo (elogio, ale-

Tabela 2. Amostra estruturada do corpus utilizado.

ID	Título do vídeo	Frase	Rótulo
`ocHOzZvdS1Y`	A Cigarra e a Formiga	A cigarra não tinha comido nada durante dias.	Negativo
`32mlN52ihg0`	O Lobo e os Sete Cabritinhos	Quando viu a porta aberta, ela soube que algo ruim tinha acontecido.	Negativo
`Yy4D7cf0z_s`	Operações com Frações	Hoje vamos aprender a multiplicar frações?	Neutro
`Lquvg-WfF_0`	Pré-história para Crianças	Esta etapa da pré-história se divide em Idades do Cobre, do Bronze e do Ferro.	Neutro
`ocHOzZvdS1Y`	A Cigarra e a Formiga	Obrigado por tudo. Você salvou a minha vida e nunca vou me esquecer.	Positivo
`H0Kx7IW7-Rc`	Peppa Pig em Português	A Peppa gosta de cuidar do seu irmãozinho, George.	Positivo

gria, carinho). Para robustez na distinção de risco, incluímos transcrições de animações adultas com apelo visual (p.ex., *South Park*) que, embora atrativas para crianças, carregam linguagem imprópria e temas inadequados. Esse contraste aprimora a fronteira de decisão do modelo, preparando-o para cenários reais. A rotulagem considerou casos ambíguos (humor, ironia) com revisão manual para mitigar enviesamentos. Cada instância foi classificada individualmente pelos quatro rotuladores, sem acesso às decisões dos demais; o rótulo final foi definido por voto majoritário (moda) e, em casos de empate, o professor responsável pelo projeto atuou como árbitro.

A consistência da rotulagem foi avaliada pelo Kappa de Fleiss [Fleiss 1971], apropriado a múltiplos rotuladores e categorias nominais. Sobre as 2.749 instâncias rotuladas pelos quatro anotadores, obteve-se $\kappa = 0{,}72$, na faixa de *substantial agreement* (0,61–0,80) da escala de [Landis and Koch 1977], reforçando a confiabilidade dos rótulos finais. A concordância média entre rotuladores foi de 81,5%, com aproximadamente 63% das frases recebendo voto unânime e o restante com dissenso de apenas um rotulador.

3.4. Pré-processamento e Estruturação

Limpeza básica (remoção de ruído e espaços extras) precedeu a tokenização WordPiece do BERTimbau; não se aplicou remoção de *stopwords* nem *stemming* para preservar contexto [Souza et al. 2020, Jurafsky and Martin 2023]. Manter *stopwords* e flexões preserva intensidade e intenção em expressões curtas, relevante para AS. Aplicou-se também marcação de negação: após termos como *não* ou *sem*, a palavra seguinte recebeu o prefixo `NEG_`, no treino e na avaliação. Metadados (título, descrição, canal) foram combinados a trechos da transcrição no texto de entrada do *Score*, capturando o arco narrativo. Priorizaram-se frases curtas e representativas, alinhadas aos vídeos infantis e às emoções do público-alvo. A Tabela 3 resume a distribuição inicial.

3.5. Balanceamento e Validação

Aplicou-se *Random Oversampling* no treino para mitigar desbalanceamento [Batista et al. 2004, He and Garcia 2009]. A escolha em detrimento do SMOTE deveu-se às sentenças curtas, em que a interpolação sintética pode produzir construções semanticamente incoerentes. Validação cruzada 5-fold com *hold-out* de 20% para teste final [Kohavi et al. 1995]. As fases de validação e teste foram monitoradas por métricas

Tabela 3. Distribuição das classes no corpus final (2.749 frases, balanceado).

Classe	Quantidade	%
Negativo	868	31,6%
Neutro	999	36,3%
Positivo	882	32,1%
Total	2.749	100%

complementares (*loss*, acurácia, F1 macro e ponderado), assegurando que melhorias de hiperparâmetros se refletissem em ganhos de generalização e detectando flutuações entre *folds* e sinais de sobreajuste.

3.6. Arquitetura e Protótipo

Modelo BERTimbau Base (*cased*) [Souza et al. 2020] implementado com *Transformers*/*PyTorch*. Protótipo web cliente-servidor integra extração de transcrições, classificação de sentimento e exposição de resultados via *Score de Segurança*. O *backend* expõe *endpoints* de busca e análise, orquestrando integração com a API do YouTube, pré-processamento e inferência do modelo ajustado. O *frontend* apresenta resultados com indicador visual de adequação (cores e pontuação), facilitando o uso por pais e educadores. A arquitetura modularizada permite composição de filtros adicionais (toxicidade, linguagem imprópria) preservando o contrato da API.

3.7. Definição dos Hiperparâmetros e *Ensemble* Final

Antes do treinamento final, realizou-se uma busca em grade manual variando épocas (1, 3, 5), tamanho de lote (8, 16, 32), taxa de aprendizado (2e-5, 3e-5, 5e-5) e passos de aquecimento (100, 500). A configuração selecionada (5 épocas, *batch* 8, taxa 3e-5 e 100 passos de aquecimento) apresentou melhor equilíbrio na validação cruzada, com prioridade para F1 macro e F1 elevado na classe Negativo. O classificador final é um *ensemble* (*soft voting*) dos cinco modelos da validação cruzada.

3.8. Estratégia de Validação e Ambiente

Para evitar *data leakage*, o *oversampling* foi aplicado apenas no conjunto de treino de cada *fold*, preservando validação e teste nas distribuições originais [He and Garcia 2009]. Adotou-se validação cruzada estratificada (k=5) sobre 80% dos dados, com 20% reservados para teste final [Kohavi et al. 1995]. O desenvolvimento ocorreu em Python com suporte a CUDA; em GPU Tesla T4, a inferência por sentença apresentou tempo médio de 15ms por modelo (75ms para o *ensemble*), viável para uso em tempo real no protótipo.

4. Resultados

Apresentamos os resultados em três frentes: métricas quantitativas globais e por classe, análise visual por matriz de confusão e avaliação prática no protótipo com exemplos de busca segura e cenários de risco.[2] Buscamos conectar as evidências numéricas ao comportamento observado nas interfaces, discutindo implicações para a curadoria de conteúdo infantil.

[2]Repositório: `https://github.com/ravarmes/wics-2026`

4.1. Desempenho Quantitativo

Com 5 épocas, *batch* 8 e taxa de aprendizado 3e-5, a validação cruzada 5-fold apresentou F1-macro de 77,46% ± 1,41% por *fold*. O *ensemble* dos cinco *folds* atinge 79,84% de F1-macro (79,64% de acurácia) no teste final (550 amostras). A Tabela 4 consolida as métricas globais e por classe no conjunto de teste retido.

Tabela 4. Métricas globais e por classe no conjunto de teste final (550 amostras).

Classe / Agregação	Precisão	*Recall*	F1
Negativo	93,4%	89,1%	91,2%
Neutro	69,0%	84,5%	76,0%
Positivo	82,0%	64,8%	72,4%
Macro	81,46%	79,45%	79,84%
Weighted	80,87%	79,64%	79,63%
Acurácia: **79,64%**			

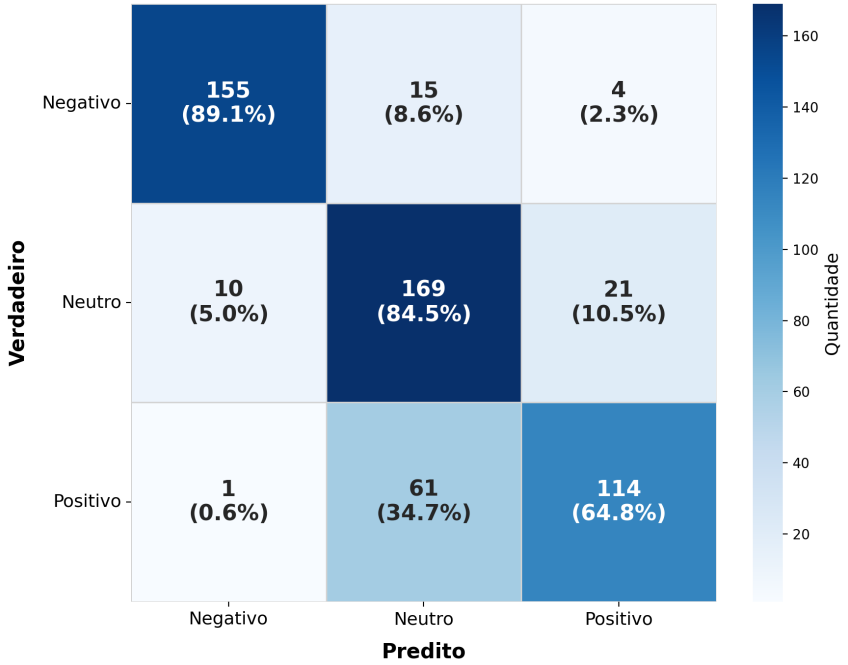

Figura 1. Matriz de confusão do *ensemble* no conjunto de teste (550 frases).

A Figura 1 evidencia a concentração de acertos na diagonal principal, com alto desempenho na classe Negativo (F1 = 91,2%) — métrica crítica para o objetivo de segurança infantil. Os erros mais comuns envolvem confusão entre Positivo e Neutro (61 instâncias positivas classificadas como neutras e 21 instâncias neutras classificadas como positivas), coerente com a proximidade semântica entre elogios sutis e declarações informativas de baixa carga afetiva.

A acurácia de 79,64% no teste retido supera a média de validação 5-fold por *fold* (F1-macro de 77,46% ± 1,41%). A alta sensibilidade na classe Negativo (*recall* de 89,1%) é o resultado mais importante para o caso de uso de filtragem infantil: o modelo raramente deixa passar conteúdo negativo, mesmo que ocasionalmente classifique conteúdo positivo como neutro — erro menos crítico do que o inverso.

4.2. Comparação com Baselines

Para situar o ganho do ajuste fino, comparamos a proposta a três *baselines* na mesma partição de teste (20%, estratificada, 550 amostras): (i) SentiLex-PT [Silva et al. 2012]; (ii) TF-IDF com Regressão Logística; e (iii) BERTimbau congelado como extrator de *embeddings* alimentando uma Regressão Logística — configuração próxima a *zero-shot*. A Tabela 5 sintetiza as métricas globais.

Tabela 5. Comparação com *baselines* no conjunto de teste retido (550 amostras, estratificado). F1 reportado em formato *macro* e por classe.

Método	Acurácia	F1 Macro	F1 Neg.	F1 Neu.	F1 Pos.
SentiLex-PT [Silva et al. 2012]	45,45%	44,68%	47,1%	48,2%	38,7%
TF-IDF + Reg. Logística	69,45%	69,66%	80,2%	65,3%	63,4%
BERTimbau congelado + LR	74,73%	74,81%	88,0%	69,6%	66,9%
BERTimbau *ensemble* (proposto)	**79,64%**	**79,84%**	**91,2%**	**76,0%**	**72,4%**

Quatro observações sustentam a escolha técnica. Primeiro, o SentiLex-PT [Silva et al. 2012] atinge 45,45% de acurácia e F1 macro de 44,68%, estabelecendo o piso léxico; a contagem de polaridade sem contexto distribucional falha em negações, ironia e idiomatismos, com F1 = 38,7% na classe Positivo.

Segundo, o TF-IDF eleva o desempenho para 69,45% ao capturar coocorrências; a diferença de $\sim$24pp em relação ao SentiLex-PT evidencia o ganho do aprendizado supervisionado sobre léxicos externos.

Terceiro, o BERTimbau congelado (*zero-shot*) atinge 74,73% sem ajuste fino, evidenciando o valor do pré-treinamento em larga escala em português; este *baseline* forte estabelece o piso contra o qual o ajuste fino deve provar valor.

Quarto, o ajuste fino e o *ensemble* acrescentam $+4{,}91$**pp de acurácia** e $+5{,}03$**pp de F1 macro** sobre o *zero-shot*. A **classe Negativo sobe de 88,0% para 91,2%** ($+3{,}2$**pp)**, mantendo o patamar elevado exigido pela aplicação de segurança infantil, enquanto os maiores avanços ocorrem em **Neutro** ($+6{,}4$**pp)** e **Positivo** ($+5{,}5$**pp)** — classes em que o *zero-shot* tinha mais dificuldade. O conjunto de resultados confirma que a combinação corpus-do-domínio + BERTimbau ajustado em *ensemble* é a escolha técnica mais adequada ao objetivo de segurança infantil.

4.3. Protótipo e Aplicabilidade

A interface web permite comparar um cenário de busca considerado seguro (p.ex., "Bob Esponja") com um cenário de risco (p.ex., "South Park"), reordenando os resultados quando o filtro está ativo. As Figuras 2, 3 e 4 ilustram esses dois casos.

A Figura 2 apresenta a interface inicial, onde o usuário insere consultas e habilita filtros de segurança. O retorno inclui o *Score* e a classificação predominante, informando a adequação do conteúdo, e o cabeçalho destaca a ativação dos filtros e o nível de confiança retornado pelo modelo, conferindo transparência à decisão de segurança. Os filtros de duração e engajamento já haviam sido implementados em trabalhos anteriores; este trabalho apresenta a integração do filtro de Análise de Sentimentos ao protótipo, com os demais (faixa etária, tópicos educacionais, toxicidade e linguagem imprópria) como trabalhos futuros a serem integrados ao fluxo de busca.

Figura 2. Tela principal do YouTube Safe Kids.

Figura 3. Busca segura

Na Figura 3 (parte superior), a busca por "Bob Esponja" exemplifica um cenário infantil, no qual a reordenação por *Score de Segurança* destaca conteúdos pedagógicos e inofensivos com pontuação elevada. Na Figura 3 (parte inferior), o ranqueamento apresenta vídeos com títulos e descrições alinhadas ao vocabulário infantil, com *scores* em verde para positivo/neutro. Os vídeos no topo correspondem, por construção, à classe Positivo com alta confiança; itens com *score* intermediário (Neutro) mantêm-se acessíveis, preservando a liberdade de escolha informada.

Esse desenho de interface materializa boas práticas de comunicação de risco, em que um indicador contínuo (*score*) complementa a classe categórica de sentimento, e cada item informa simultaneamente sentimento dominante e confiança do modelo.

A Figura 4 (parte superior) apresenta um cenário de estresse (p.ex., "South Park"), onde o sistema identifica predominância de sentimento negativo e penaliza os resultados. O sistema não filtra a busca: os vídeos retornados pelo YouTube permanecem listados, mas os de *score* mais negativo são empurrados para as últimas posições. Na Figura 4 (parte inferior), os vídeos são ordenados por *score* decrescente, com destaque em vermelho para conteúdos inadequados, auxiliando revisão parental ágil e reduzindo atrito com a experiência padrão da plataforma.

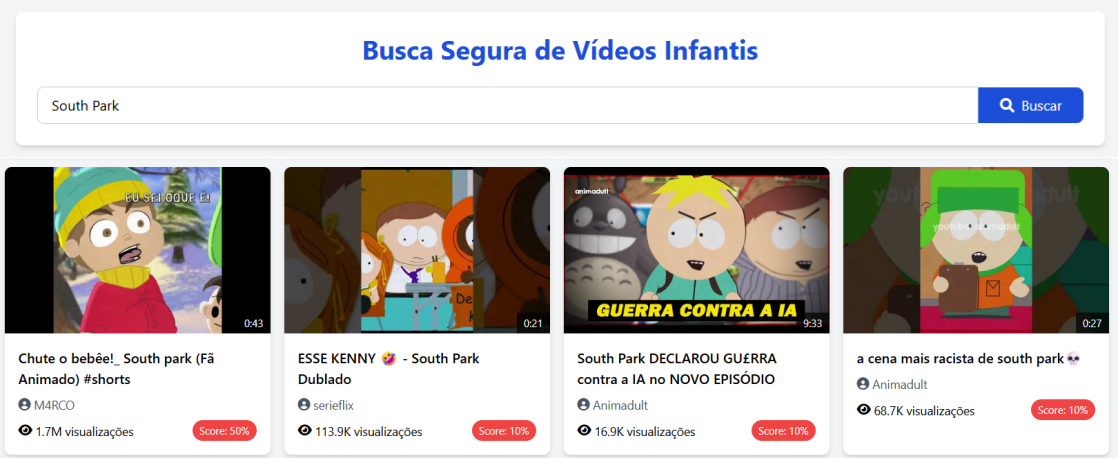

Figura 4. Busca de risco

4.4. Score de Segurança

O cálculo do ***Score* Global** processa os metadados e segmenta a transcrição em três momentos temporais para verificar o arco narrativo do conteúdo:

- **Início:** extração de sentença introdutória (primeiros 10% do vídeo);
- **Meio:** extração de sentença do desenvolvimento (intervalo entre 40% e 60%);
- **Fim:** extração de sentença de conclusão (últimos 10% do vídeo).

Esses trechos são concatenados ao Título e à Descrição, formando um único bloco textual unificado (T_{input}) que serve de entrada para o modelo:

$$T_{\text{input}} = \text{Título} + \text{Descrição} + \text{Frases}_{(\text{Início, Meio, Fim})} \tag{1}$$

O texto consolidado é submetido ao BERTimbau, que retorna a classe predominante (Negativo, Neutro ou Positivo) e o nível de confiança. Com base nesses valores, aplica-se o Algoritmo 1, que penaliza conteúdos Negativos e bonifica os seguros:

Algoritmo 1. Lógica de cálculo do *Score* baseado na classe e confiança.

```
# Logica de pontuacao refinada para seguranca infantil
if predicted_class == 0:  # Negativo
    # Penaliza fortemente conteudo negativo
    score = 0.1 + (0.2 * (1 - confidence))

elif predicted_class == 1:  # Neutro
    # Score base
    score = 0.5

else:  # Positivo
    # Bonifica conteudo positivo
    score = 0.7 + (0.3 * confidence)
```

Matematicamente, a função de *Score* (S) é definida por partes, dependendo da classe predita e da confiança (C) do modelo:

$$S = \begin{cases} 0{,}1 + 0{,}2 \cdot (1 - C) & \text{se Classe = Negativo} \\ 0{,}5 & \text{se Classe = Neutro} \\ 0{,}7 + 0{,}3 \cdot C & \text{se Classe = Positivo} \end{cases} \tag{2}$$

Essa abordagem assegura que vídeos classificados como **Negativos** recebam notas no intervalo $[0{,}10; 0{,}30]$, funcionando como um filtro de barreira. Já os conteúdos **Positivos** ocupam o intervalo superior $[0{,}70; 1{,}00]$, enquanto classificações **Neutras** mantêm uma pontuação intermediária fixa de $0{,}50$.

Os limites foram definidos para garantir separação visual perceptível entre as classes na interface (faixa vermelha para Negativo, amarela para Neutro e verde para Positivo), sem sobreposição de intervalos e mantendo uma diferença mínima de 0,40 entre as faixas extremas. Essa formulação por partes permite mapear o *score* a indicadores cromáticos sem ambiguidade e atribui margem proporcional à confiança apenas nos extremos, onde a decisão de bloqueio ou recomendação é mais sensível.

Para exemplificar a penalização, utiliza-se o cenário do vídeo "Debate familiar — Rick & Morty", cujo conteúdo encontra-se disponível publicamente no YouTube.[3] A Figura 5 ilustra o resultado na interface.

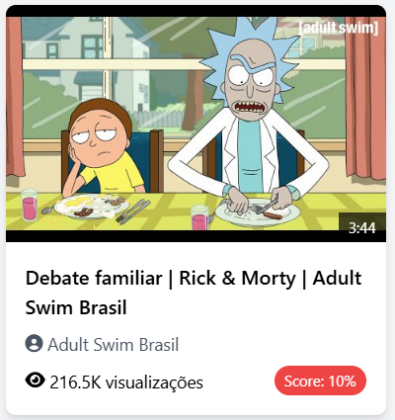

Figura 5. Score negativo

Neste exemplo, o modelo atribuiu a classe **Negativo** com confiança de 99,97% ($C \approx 0{,}9997$). Substituindo-se os valores na Equação 2:

$$S_{\text{final}} = 0{,}1 + 0{,}2 \cdot (1 - 0{,}9997) = 0{,}1 + 0{,}00006 \approx \mathbf{0{,}10}\ (10\%) \tag{3}$$

O cálculo acima demonstra a sensibilidade do algoritmo: com confiança próxima da certeza, o termo de ajuste $(1-C)$ torna-se insignificante, fazendo com que o *score* convirja para o valor base de $0{,}1$. Isso assegura que conteúdos inequivocamente inadequados sejam penalizados com a pontuação mínima.

Ressalta-se que a arquitetura foi projetada para suportar a aplicação simultânea de múltiplos filtros (por exemplo, combinando Análise de Sentimentos e Detecção de Toxicidade). Nesse cenário, o *Score de Segurança Global* (S_{global}) pode ser calculado pela média aritmética das pontuações individuais de cada filtro ativo, conforme a Equação 4:

$$S_{global} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{4}$$

[3]Endereço do vídeo analisado: `https://www.youtube.com/watch?v=QzPLj7B71ck`.

onde N é o número de filtros habilitados e S_i é a pontuação obtida no filtro i. Dessa forma, um vídeo precisa apresentar consistência nos critérios selecionados para receber uma recomendação mais segura. A média aritmética foi adotada como agregador inicial pela simplicidade e simetria; esquemas de ponderação que atribuam maior peso a filtros críticos (p.ex., toxicidade e linguagem imprópria), conforme seu impacto potencial na segurança e adequação do conteúdo, ficam como trabalho futuro.

5. Conclusões

O *ensemble* de modelos BERTimbau ajustados alcança 79,84% de F1-macro (79,64% de acurácia) no teste retido e F1 de 91,2% para a classe Negativo, com Fleiss Kappa de 0,72 na rotulagem (Seção 3.3), atendendo ao objetivo de segurança infantil. A validação cruzada 5-fold apresentou F1-macro de 77,46% $\pm$ 1,41% por *fold*; a agregação em *ensemble* sustenta a robustez metodológica. O protótipo web comprova aplicabilidade com reclassificação e ordenação por *Score*, sustentado por corpus de domínio e validação robusta. A arquitetura modular e o *Score* interpretável favorecem adoção por pais, escolas e plataformas; limitações incluem o tamanho moderado do corpus (2.749 frases) e a confusão semântica residual entre classes Positivo e Neutro (proximidade de elogios sutis e declarações informativas afetivamente leves), mitigadas por expansão futura e filtros complementares.

Como impacto prático, a integração do filtro de sentimentos ao protótipo valida o uso em tempo real e estabelece base para incorporar filtros adicionais (duração, faixa etária, tópicos educacionais, toxicidade e linguagem imprópria) com o mesmo padrão de transparência e interpretabilidade. O filtro de Análise de Sentimentos apresentado constitui um dos componentes em desenvolvimento na ferramenta; após a integração dos demais filtros previstos, está planejada uma fase de avaliação com usuários reais (pais, educadores e crianças), de modo a validar a percepção de utilidade, a usabilidade e o impacto socioeducacional do sistema em cenários cotidianos.

A pesquisa contribui para a curadoria responsável de conteúdo infantil em ambientes digitais, favorecendo o bem-estar social e a transparência nos processos automatizados. Operando apenas com dados públicos e explicitando as classificações via *Score* interpretável, a proposta reforça práticas de responsabilidade algorítmica e reduz riscos associados à opacidade dos sistemas de IA, alinhando-se a princípios de ciência aberta e às dimensões éticas, culturais e educacionais do desenvolvimento de tecnologias baseadas em inteligência artificial [Valença and Santos 2025].

Como trabalhos futuros, pretende-se ampliar a arquitetura com novos filtros (Linguagem Imprópria, Valor Educacional, Diversidade, Interatividade e Conteúdo Sensível), combinando PLN para análise textual e Visão Computacional para elementos visuais em uma abordagem híbrida, escalável e sensível a diferentes contextos de uso, cenários educacionais e demandas de curadoria parental. A combinação corpus-de-domínio, BERTimbau ajustado e *Score* interpretável posiciona-se como contribuição madura para a curadoria de conteúdo infantil em português brasileiro, conciliando rigor metodológico e impacto socioeducacional.

Referências

Baracho, J. K. d. C. M., Lisboa, L. A., and Lopes, R. V. V. (2025). Levantamento e análise qualitativa de bases de dados de fake news em português. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 169–180. SBC.

Basak, S., Agrawal, H., Jena, S., Gite, S., Bachute, M., Pradhan, B., and Assiri, M. (2023). Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. *Computer Modeling in Engineering & Sciences*, 135(2):899–929.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

Benevenuto, F., Ribeiro, F., and Araújo, M. (2015). Métodos para análise de sentimentos em mídias sociais. In *Short course in the Brazilian Symposium on Multimedia and the Web (Webmedia)*, pages 1–30.

Candido, L. S., Barbosa, C. A. d. M., Martins, L. G., and Costa, E. J. H. (2025). Análise de ferramentas de detecção de ia para textos científicos em português gerados por chatgpt, gemini e deepseek. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 78–91. SBC.

Chalkias, I. et al. (2023). Learning analytics on YouTube educational videos: Exploring sentiment analysis methods and topic clustering. *Electronics*, 12(18):3949.

Clipto.ai (2025). Clipto: Ai-powered video to text & content repurposing. Acesso em: 21 set. 2025.

da Rosa Jr., J. M. et al. (2024). Characterizing YouTube's role in online gambling promotion: A case study of Fortune Tiger in Brazil. In *Proceedings of the ACM Web Science Conference*. ACM.

Finatto, M. J. B., Lopes, L., and Silva, A. C. (2015). Processamento de linguagem natural, linguística de corpus e estudos linguísticos: uma parceria bem-sucedida. *Domínios de lingu@ gem. Uberlândia, MG. Vol. 9, n. 5 (dez. 2015), p.[41]-59*.

Firmino, V. P., Lopes, J. N. d. S., and Reis, V. Q. (2025). Identificando padrões de sexismo na música brasileira através do processamento de linguagem natural. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 59–69. SBC.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gonzalez, M. and Lima, V. L. S. (2003). Recuperação de informação e processamento da linguagem natural. In *XXIII Congresso da Sociedade Brasileira de Computação*, volume 3, pages 347–395. sn.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*. Stanford University, 3rd edition draft edition. Acesso em: 21 set. 2025.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Le, B., Tandon, R., Oinar, C., Liu, J., Durairaj, U., Guo, J., Zahabizadeh, S., Ilango, S., Tang, J., Morstatter, F., Woo, S., and Mirkovic, J. (2022). Samba: Identifying inappropriate videos for young children on YouTube. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4173–4177. ACM.

Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

Medeiros, M. C. R. and de Freitas Neto, F. P. (2025). Um estudo sobre vieses de gênero em modelos de pln aplicado em histórias geradas pelo gpt-3.5 e gemini. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 41–52. SBC.

Miranda, A. L. d. A. and Rodrigues, C. M. d. O. (2025). Uma abordagem integrada para detecção de discurso de ódio em mídias sociais utilizando vetorização de textos e emojis. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 247–255. SBC.

Nunes, M. d. G. V. (2023). E agora, pln? In Pardo, T. A. S. et al., editors, *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em português*, chapter 25. Brasileiras de PLN (Brapaln), São Carlos.

Ramos, B. and Freitas, C. (2019). Sentimento de quê?¨ uma lista de sentimentos para a análise de sentimentos. *STIL*, pages 15–18.

Reis, J., Gonçalves, P., Araújo, M., Pereira, A. C., and Benevenuto, F. (2015). Uma abordagem multilíngue para análise de sentimentos. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. SBC.

Rosa, R. L. (2015). *Análise de sentimentos e afetividade de textos extraídos das redes sociais.* PhD thesis, Universidade de São Paulo.

Santos, R. V. M. and Comarela, G. V. (2025). Development of an equity strategy for recommendation systems. In *Workshop on the Implications of Computing in Society (WICS)*, pages 24–35. SBC.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). SentiLex-PT: Principais características e evolução. In *Linguamática*, volume 4, pages 21–33.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Souza, R. R. and Café, L. M. A. (2018). Análise de sentimento aplicada ao estudo de letras de música. *Informação & Sociedade*, 28(3).

Valença, L. R. and Santos, R. d. S. (2025). Justiça algorítmica: Instrumentalização, limites conceituais e desafios na engenharia de software. In *Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 225–234. SBC.