**Catalogação na publicação**
**Seção de Catalogação e Classificação**

L864a Lopes, Arthur Ricardo Ribeiro.
Análise da presença de discurso de ódio em vídeos de recomendações do youtube / Arthur Ricardo Ribeiro Lopes. - João Pessoa, 2024.
12 f. : il.

Orientação: Yuri Malheiros.
TCC (Graduação) - UFPB/CI.

1. Discurso de ódio. 2. Análise de vídeo. 3. Youtube. 4. Reconhecimento de discurso. I. Malheiros, Yuri. II. Título.

UFPB/CI CDU 004.738.1

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

# Análise da presença de discurso de ódio em vídeos de recomendações do YouTube

**Arthur Ricardo Ribeiro Lopes**[1], **Yuri de Almeida Malheiros Barbosa**[1],

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

arthur.ricardo17@gmail.com, yuri@ci.ufpb.br

8 de setembro de 2024

## Resumo

A análise de vídeos com intuito de identificar discurso de ódio tem motivado o desenvolvimento de muitos trabalhos nos últimos anos, porém, em sua maioria, com apenas o propósito de reconhecê-los. Esse trabalho, no entanto, tem o intuito de não apenas identificar se esse discurso acontece ou não no vídeo, mas também de verificar qual o nível de contato de um usuário com esse tipo de discurso durante a utilização da plataforma YouTube. Para isso, foi simulada uma navegação automática entre os vídeos recomendados, a partir do primeiro deles e a partir de um deles escolhido de forma aleatória. Foi observado que, mesmo com todos os filtros que a plataforma possui para os seus conteúdos, ainda existe a presença de conteúdo principalmente do tipo Ofensivo, (5%-10% de cada vídeo), independente da escolha aleatório ou a partir do primeiro vídeo da lista de recomendados. Por fim, não se destaca nessa análise nenhum nível de profundidade de navegação específica.

## Abstract

The analysis of videos aimed at identifying hate speech has motivated the development of many studies in recent years, but mostly with the sole purpose of recognizing them. However, this work aims not only to identify whether this speech occurs in the video or not, but also to assess the level of user exposure to this type of speech while using the YouTube platform. To achieve this, an automatic navigation between recommended videos was simulated, starting from the first one and from one randomly chosen. It was observed that, despite all the filters the platform has for its content, there is still a presence of mainly Offensive content (5%-10% of each video), regardless of the random choice or starting from the first video in the recommended list. Finally, this analysis does not highlight any specific depth of navigation.

## 1 Introdução

Nos últimos anos, os trabalhos desenvolvidos relacionados a detecção de discurso de ódio, em sua maioria, tem como objetivo principal o treinamento e teste de modelos com este intuito. Outro grande foco nessas análises é na detecção desses conteúdos em texto de *tweets*. Porém, são poucos os trabalhos que se concentram na experiência dos usuários que interagem com mídias dessa natureza.

Em [1] e [2], foca-se na interação do usuário com o sistema de recomendação do YouTube. Em [1], tentou-se investigar se os algoritmos colocam os usuários em um "buraco de coelho", em que o usuário começa a pesquisar sobre um tema, indo cada vez mais fundo sobre ele, e se existem maneiras de mitigar o efeito desses algoritmos. Nesse trabalho o foco maior está relacionado ao alinhamento político dos usuários e como isso influência os conteúdos recomendados. Outro exemplo acontece em [2], onde o foco é testar algumas hipóteses relacionadas a quanto o algoritmo de recomendação do YouTube coloca seus usuários em bolhas de informações falsas e tendem a mantê-los ali, medindo o quanto isso pode ser prejudicial.

A importância de tentar identificar o quanto esses usuários podem ser expostos a conteúdos não "saudáveis", sejam eles violentos, ofensivos ou inapropriados, vem do fato de que, com cada vez mais frequência, passamos mais tempo utilizando essas plataformas, e cada vez mais cedo. Além disso, esses algoritmos de recomendação são treinados para manter o usuário na plataforma e, mesmo com filtros, pode acontecer o consumo de mídias de conteúdo dessa natureza e que, devido aos próprios sistemas de recomendação, podem ser reforçados.

Esse trabalho tem como intuito realizar a verificação da presença de vídeos inapropriados no YouTube. Além disso, analisaremos se a visualização de conteúdos dessa natureza sugerem, pelo sistema de recomendação da plataforma, outros vídeos semelhantes e com que profundidade, considerando a profundidade a quantidade de vídeos necessários interagir com até atingir o dado vídeo. Por fim, será feita uma análise dos discursos presentes nesses videos.

## 2 Trabalhos Relacionados

Para a pesquisa dos trabalhos relacionados foi levado em consideração principalmente dois fatores, o primeiro era entender as diferenças entre utilizar uma abordagem multimodal e uma abordagem unimodal, com apenas o texto, para análise do discurso de ódio em um vídeo. O segundo fator levado em consideração foi procurar por trabalhos que têm como foco o entendimento da experiência do usuário ao utilizar o sistema de recomendação do *YouTube*. Para o primeiro caso, as palavras-chave utilizadas para a pesquisa no Google Scholar foram "*Hate Speech in Videos*", "*Classification*", "*Multimodal Classification*" e "*Hate Speech in tweets*". Para o segundo caso, na pesquisa foram utilizadas as palavras-chave "*YouTube Recommendation System*" e "*YouTube Videos*".

No artigo "Detection of Hate Speech in Videos Using Machine Learning"[3], Ching Seh Wu e Unnathi Bhandary (2020) utilizam a base de dados coletada manualmente a partir de vídeos do YouTube, que continham discurso normal e termos ofensivos, com foco em vídeos com discurso racista e sexista. O estudo utilizou os modelos Naïve Bayes, Random Forest, Máquina de Vetor de Suporte (do inglês, *Suport Vector Machine* - SVM) e *Recurrent Neural Network* (RNN – Rede neural recorrente) para classificar e detectar discurso de ódio a partir da extração de áudio para análise de texto. O estudo obteve sucesso na tarefa com o modelo Random Forrest Classifier alcançando uma precisão de 96% na classificação dos vídeos. Isso indica que a abordagem de apresendizagem máquina utilizada foi eficaz na identificação de conteúdo ofensivo nos vídeos analisados. Além da precisão, foram utilizadas as métricas de *Recall* e *F1 Score*.

Por sua vez, no trabalho "HateMM: A Multi-Modal Dataset for Hate Video Classification"[4], Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta e Animesh Mukherjee (2023) utilizam um conjunto de dados próprio que consiste em vídeos coletados da plataforma BitChute, que é uma plataforma com baixa moderação em relação aos vídeos publicados. A metodologia empregada envolveu uma abordagem multimodal para detectar conteúdo de ódio nesses vídeos, combinando características de texto, áudio e visão, comparando modelos. Alguns exemplos são combinações do BERT, para texto, ViT, para os frames dos vídeos e do Mel-frequency cepstrum-MFCC, para os áudios. Como principais resultados obtidos, os autores identificaram as diferenças significativas entre essas combinações, em que os modelos multi-modais apresentaram um melhor desempenho, tendo como melhor resultado o modelo que utiliza as três formas de conteúdo. Como métricas para avaliação dos modelos foram utilizadas a acurácia, *F1-score*, *Recall* e precisão.

Em "Multi-modal Hate Speech Detection using Machine Learning"[5], Fariha Tahosin, Ponkoj Chandra e Golam Rabiul (2021) também utilizam uma abordagem multimodal no mesmo problema de classificação do discurso de ódio em vídeos. Nesse caso também foi criada uma base de dados própria, composta por vídeos que foram coletados de séries e filmes. Esse estudo utilizou uma abordagem multimodal para para essa detecção, analisando o áudio dessas cenas, o texto que é a transcrição deste áudio e vídeo em formato dos frames da cena. Para a classificação foram utilizados modelos de aprendizado de máquina clássicos tais como Random Forrest, Regressão logística, Adaboost, K-NN, Naive Bayes e Árvores de Decisão. De maneira geral, o modelo multimodal alcançou uma performance pior que os modelos individuais, tanto de imagem, áudio e texto. Isso se deve a natureza dos modelos que são mais simples para lidar com os tipos de dados específicos. As métricas empregadas para essa avaliação foram precisão, *Recall* e *F1-Score*.

Em "Multimodal Hate Speech Detection from Videos and Texts"[6], Nishchal Prasad, Sriparna Saha and Pushpak Bhattacharyya (2023) utilizaram uma base de dados de natureza multimodal, que consiste de vídeos da rede social Vine, juntamente com as legendas desses vídeos e alguns comentários associados a eles. A metodologia utilizada consiste em utilizar uma série de modelos e comparar, tanto de maneira

individual, quanto combinando esses modelos. Para lidar com texto foram utilizados os seguintes modelos: GloVe-text, BERT-text, XLNet-text, ResBiLSTM; para os vídeos foi utilizado a EfficientNet; e combinações desses modelos com algumas modificações, para lidar com a abordagem multimodal. Os principais resultados obtidos incluem o melhor desempenho das arquiteturas BERT-text e XLNet-text nas abordagens unimodal e um desempenho superior da abordagem multimodal. As métricas empregadas para avaliação foram acurácia, área sob a curva ROC (AUC) e *F1-Score*.

Raul Gomez, Jaume Gibert, Lluis Gomez (2019), no trabalho intitulado "Exploring Hate Speech Detection in Multimodal Publications"[7], utilizam a abordagem multimodal realizada com apenas a fonte visual e a textual. A base de dados utilizada é de tweets, utilizando o texto e as imagens vinculadas a eles. Caso existissem textos presentes nessas imagens, os mesmos eram extraídos utilizando ferramentas de OCR. Em relação às classificações, foram utilizados os modelos: Feature Concatenation Model (FCM), Spatial Concatenation Model (SCM) , Textual Kernels Model (TKM) e Long Short-Term Memory (LSTM). Os melhores resultados foram quase os mesmos utilizado de LSTM, com apenas o texto do tweet, e pelos FCM , SCM e TKM, utilizando os três tipos de entrada (o texto do tweet, a imagem e o texto presente na imagem), mostrando a boa eficácia na abordagem com apenas o texto. Para a avaliação dos modelos foi utilizado precisão, *Recall*, *F1-Score* e curva ROC

Em 2018, no trabalho "Indonesia Hate Speech Detection using Deep Learning"[8], Taufic Leonardo Sutejo e Dessi Puji Lestari utilizaram uma abordagem também multimodal, porém utilizando apenas o áudio e texto. A base foi criada pelos pesquisadores, que coletaram dados manualmente de várias plataformas de mídia social, como Facebook, Twitter, Line Today, comentários do YouTube e transcrições de vídeos do YouTube. Em relação aos áudios utilizados, esses consistem de gravações de homens e mulheres lendo textos coletados no primeiro momento. Foi utilizada uma abordagem de aprendizado profundo baseada em LSTM para detectar discurso de ódio. Para o modelo textual, foi comparada uma série de *embeddings*, para a abordagem multimodal utiliza-se de um técnica de combinação para agrupar os atributos do modelo de áudio e o de texto. Os resultados mais promissores nesse trabalho veio do modelo textual, utilizando CBOW e do modelo multimodal, utilizando CBOW e INTERSPEECH 2009, demonstrando a eficácia do uso de características textuais nessa detecção. A métrica utilizada para a avaliação dos modelos foi a *F1-Score*.

"Deep Learning for Hate Speech Detection in Tweets"[9] dos autores Pinkesh Badjatiya, Shashank Gupta, Manish Gupta e Vasudeva Varma (2017) utilizou de uma abordagem apenas unimodal, com texto, para comparar múltiplas arquiteturas de aprendizado profundo, como Convolutional Neural Network (CNNs), LSTMs e FastText, na detecção de discurso de ódio. Foi utilizado um conjunto de dados de 16 mil tweets rotulados como racista, sexista ou nem racista nem sexista. Os melhores resultados foram obtidos utilizando "LSTM + Random Embedding + GBDT", mostrando a eficácia nesse tipo de abordagem para essa detecção. As métricas empregadas para avaliação desses modelos foram a precisão, o *Recall* e o *F1-Score* ponderado.

No trabalho "Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning technique"[10], Akib Mohi Ud Din Khandaya, Syed Tanzeel Rabani, Qamar Rayees Khana e Showkat Hassan Malik (2022) utilizaram uma base de dados própria composta de tweets relacionados ao COVID-19 pré-processados para remover ruídos. Esse estudo visou detectar discursos de ódio nesses tweets utilizando várias técnicas diferentes, tais como Regressão Logística, Naïve Bayes , Support Vector Machine, Decision Tree e Stochastic Gradient Boosting e também várias *embeddings* diferentes. Os melhores resultados foram obtidos utilizando a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico (do inglês, *Term Frequency — Inverse Document Frequency* (TF/IDF), *Bag of Words* com o tamanho do tweet, em conjunto com Stochastic Gradient Boosting. Para comparar esses modelos foram utilizados Acurácia, *Recall*, Precisão e o *F1-Score*

Já em "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users"[11], Megan A. Brown, James Bisbee, Angela Lai, Ri-chard Bonneau, Jonathan Nagler e Joshua A. Tucker(2022) utilizam dados dos usuários reais do youtube para tentar identificar suas tendências políticas baseados na análise dos vídeos recomendados para eles e nos vídeos que clicam, traçando um perfil para esse usuário. Diferente dos demais vídeos, o foco não é identificar o discurso de ódio na mídia consumida, mas entender uma tendência no padrão de consumo.

Os trabalhos relacionados apresentados nessa seção estão sumarizados na Tabela 1. Apesar do presente trabalho não ser focado no treinamento de um classificador, podemos observar que em "Detection of Hate Speech in Videos Using Machine Learning","HateMM: A Multi-Modal Dataset for Hate Video Classification","Multi-modal Hate Speech Detection using Machine Learning","Multimodal Hate Speech Detection from Videos and Texts","Multimodal Hate Speech Detection from Videos and Texts" e "Exploring Hate Speech Detection in Multimodal Publications" que, apesar da abordagem multimodal

ter mostrado maior precisão, em relação a classificação de discurso de ódio em vídeos, nesses mesmos trabalhos é possível ver que a abordagem utilizando texto tem uma eficácia considerável.

Tabela 1: Trabalhos Relacionados

| **Título do Trabalho** | **Base Utilizada** | **Metodologia** | **Métricas Utilizadas** |
|---|---|---|---|
| Ching Seh Wu e Unnathi Bhandary (2020) | Vídeos do YouTube | Naïve Bayes, Random Forest, SVM e RNN | Precisão, *Recall* e *F1-Score* |
| Raul Gomez, Jaume Gibert, Lluis Gomez (2019) | Vídeos e áudios de atores simulando emoções. | BERT, ALBERT e multitask learning (MTL) | Precisão, *Recall*, *F1-score* e acurácia |
| Mithun Das,Rohit Raj, Punyajoy Saha, Binny Mathew , Manish Gupta e Animesh Mukherjee (2023) | Vídeos da plataforma BitChute | BERT, ViT e MFCC | *F1-Score* |
| Akib Mohi Ud Din Khandaya, Syed Tanzeel Rabani, Qamar Rayees Khana e Showkat Hassan Malik (2022) | Texto de tweets | Regressão Logística, Naïve Bayes, SVM, Árvore de Decisão e *Stochastic Gradient Boosting* | Precisão, *Recall*, *F1-Score*, Acurácia e Matriz de confusão. |
| Fariha Tahosin, Ponkoj Chandra e Golam Rabiul (2021) | Vídeos de séries e filmes | Regressão Logística, Naïve Bayes, SVM, Árvore de decisão e *Stochastic Gradient Boosting* | Precisão, *Recall* e *F1-Score* |
| Nishchal Prasad, Sriparna Saha e Pushpak Bhattacharyya (2023) | Vídeos e suas transcrições em conjuntos com seus comentários de várias fontes online | GloVe-text, BERT-text, XLNet-text, ResBiLSTM e EffecientNet-B0 | Acurácia, Área sob a curva ROC (AUC) e *F1-score* |
| Taufic Leonardo Sutejo e Dessi Puji Lestari (2018) | Vídeos retirados do Facebook, Twitter, Line Today, comentários do YouTube e transcrições de vídeos do YouTube | LSTM | *F1-score* |
| Pinkesh Badjatiya, Shashank Gupta, Manish Gupta e Vasudeva Varma (2017) | Texto de tweets | CNNs, LSTMs e FastText | Precisão, *Recall* e o *F1-score* ponderado. |
| Megan A. Brown,James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler e Joshua A. Tucker (2022) | Videos visitados por usuários reais. | Estudo do comportamento desses usuários | Gráficos para a análise do comportamento |

Em "Indonesia Hate Speech Detection using Deep Learning" é possível observar que a abordagem por meio de texto apresenta resultados melhores, quando comparado com o multimodal texto e áudio. "Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning technique" e "Deep Learning for Hate Speech Detection in Tweets", por sua vez, apresentam bons resultados quando utilizam apenas texto para a classificação de discurso de ódio. Esse trabalho, assim como em "Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users", foca no entendimento do padrão de consumo dos usuários. No caso do trabalho relacionado a esse, os autores focam na experiência de um usuário consumindo mídias, dentro do sistema de recomendação do Youtube, no entanto, nesse trabalho, simulamos a experiência de um novo usuário sem uma conta logada, portanto sem o seu histórico de consumo, mas ainda assim, dentro do sistema de recomendação do Youtube.

# 3 Metodologia

O experimento consiste em utilizar uma ferramenta de *web scraping* para simular um usuário novo do YouTube, que não está logado em uma conta. De maneira geral, o processo envolve visitar a página principal de notícias do YouTube e escolher aleatoriamente um dos vídeos para consumo. Depois que esse vídeo é selecionado, acessamos o vídeo e visitamos um dos vídeos recomendados, como descrito na próxima seção. Em seguida, é utilizada a transcrição desses vídeos para classificar os trechos dos vídeos e construir os resultados. Esse processo acontece como na fluxograma detalhado na Figura 1.

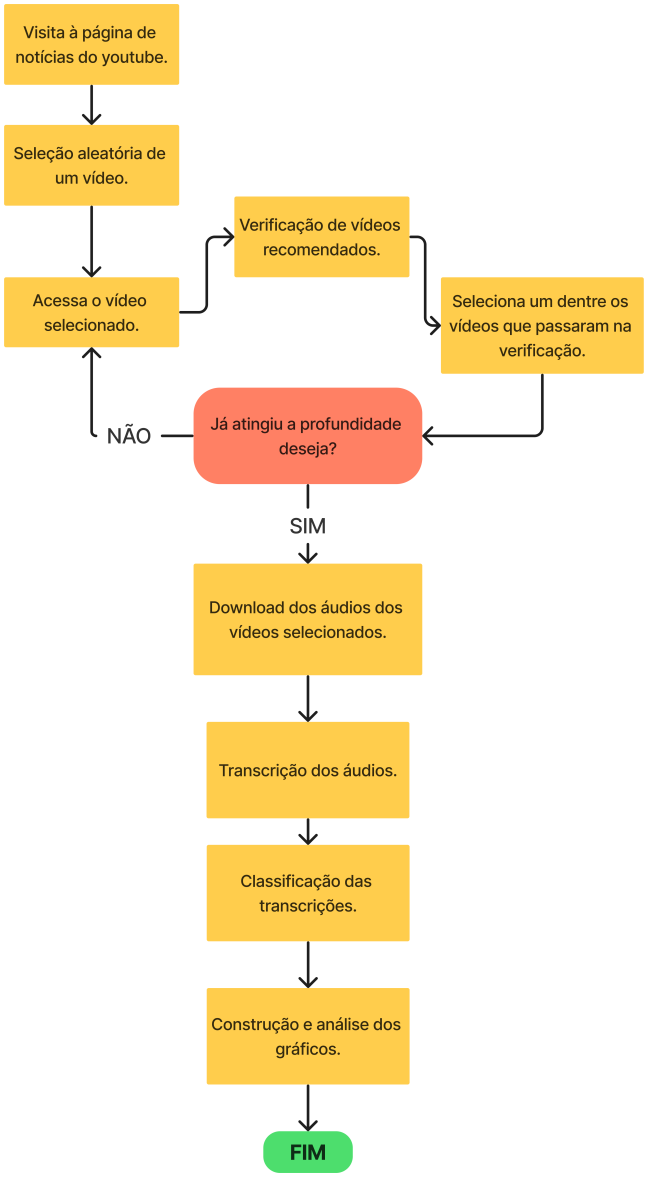


Figura 1: Fluxograma da metodologia.

## 3.1 Coleta de Dados

Para a coleta de dados, foi utilizada a biblioteca *Selenium* do Python para realizar o *web scraping* no site do YouTube. Essa coleta começa na página de notícias do YouTube, verificando todos os vídeos presentes nessa página e selecionando aleatoriamente um deles. Depois de selecionado o vídeo inicial, a navegação para os próximos $N$ vídeos, que é feita da seguinte maneira: verifica-se quais os vídeos recomendados a partir deste vídeo e, dentre esses recomendados, quais deles têm a duração menor ou

igual a 10 minutos. Depois dos vídeos serem verificados, existem duas maneiras de ir para o próximo vídeo: (1) selecionando um vídeo aleatório dentre os verificados ou (2) selecionando o primeiro vídeo dessa lista. Este processo é repetido utilizando o vídeo selecionado até atingir os $N$ vídeos desejados. Neste trabalho $N$ foi definido como 10. É considerado uma nova profundidade depois que selecionamos o próximo video e o acessamos.

Após essa seleção, é utilizada a biblioteca yt_dlp do Python, que, utilizando o *link* desses vídeos, faz requisições para baixar os áudios correspondentes. Depois disso, utilizamos a biblioteca *Whisper* (versão 1.1.10) para, através da ferramenta *Mel Spectrogram*, verificar em qual idioma esse áudio foi produzido. Depois de verificado o idioma, utilizamos a mesma biblioteca para realizar a transcrição desses áudios. É retornado por essa biblioteca o áudio dividido em segmentos. O tamanho dele é definido pelo próprio modelo da biblioteca, encaixando esses segmentos em pausas do áudio ou no final das frases, vírgulas ou momentos de silêncio. Para cada um desses segmentos, temos o texto presente naquele intervalo, além do tempo inicial e final dele, em relação ao áudio completo. Essa transcrição é salva em formato JSON para ser utilizada na próxima etapa de classificação do tipo de discurso, descrita a seguir.

Utilizando esse processo foram realizados um total de 40 experimentos, sendo 20 para a seleção aleatória dos vídeos recomendados e 20 para a seleção do primeiro video daqueles recomendados, sendo analisados um total de 1644 vídeos.

### 3.2 Classificação dos Dados

Para a classificação dos dados, foi utilizado um modelo da versão básica do BERT para inglês, o qual está disponível em um repositório no *HuggingFace* [1]. Este modelo classifica os fragmentos do texto em quatro categorias: "aceitável", quando não apresenta elementos inapropriados, ofensivos ou violentos; "inapropriado", quando contém linguagem obscena ou vulgar, porém não direcionada para um alvo específico; "ofensivo", quando inclui generalizações ofensivas, desprezo, desumanização ou comentários ofensivos indiretos; e "violento", quando apresenta ameaças, condescendência, desejo ou apelo para violência física contra um alvo, incluindo também apelar, negar ou glorificar crimes de guerra e crimes contra a humanidade [12].

### 3.3 Análise dos Dados

Após a realização do experimento, são utilizados gráficos que mostram de maneira proporcional as classificações por vídeo e como esses trechos se distribuem ao longo do vídeo para uma análise mais clara, tanto do experimento completo, quanto de um vídeo específico.

## 4 Resultados e discussão

Nessa seção serão analisados os resultados obtidos após a realização dos experimentos. Essa análise será realizada através dos gráficos construídos com os dados obtidos durante a execução da metodologia. Além dessa análise, será realizado um comparativo entre os experimentos aleatórios, quando é escolhido um vídeo aleatoriamente dentre os que são recomendados e não aleatórios, sempre que se escolhe o primeiro vídeo recomendado. Por fim, é feita uma análise geral do sistema de recomendação do YouTube, em relação a conteúdos com esse tipo de discurso.

### 4.1 Análise da classificação dos vídeos

O gráfico da Figura 1 mostra, em proporção, a presença de cada classe relacionada à análise do conteúdo (Aceitável, Inapropriado, Violento e Ofensivo) de todos os vídeos que foram recomendados em cada profundidade (1-10), a partir de um vídeo escolhido aleatoriamente dos recomendados. A mesma análise pode ser observada, mas a partir do primeiro vídeo sugerido entre os recomendados, na Figura 3. Por profundidade, para cada experimento, podem ser analisados de 1 a 20 vídeos. Esse número varia, pois são removidos aqueles com duração maior que 10 minutos e não estão em inglês. Os valores dos gráficos representam a média de 20 execuções dos experimentos.

[1] https://huggingface.co/IMSyPP/hate_speech_en

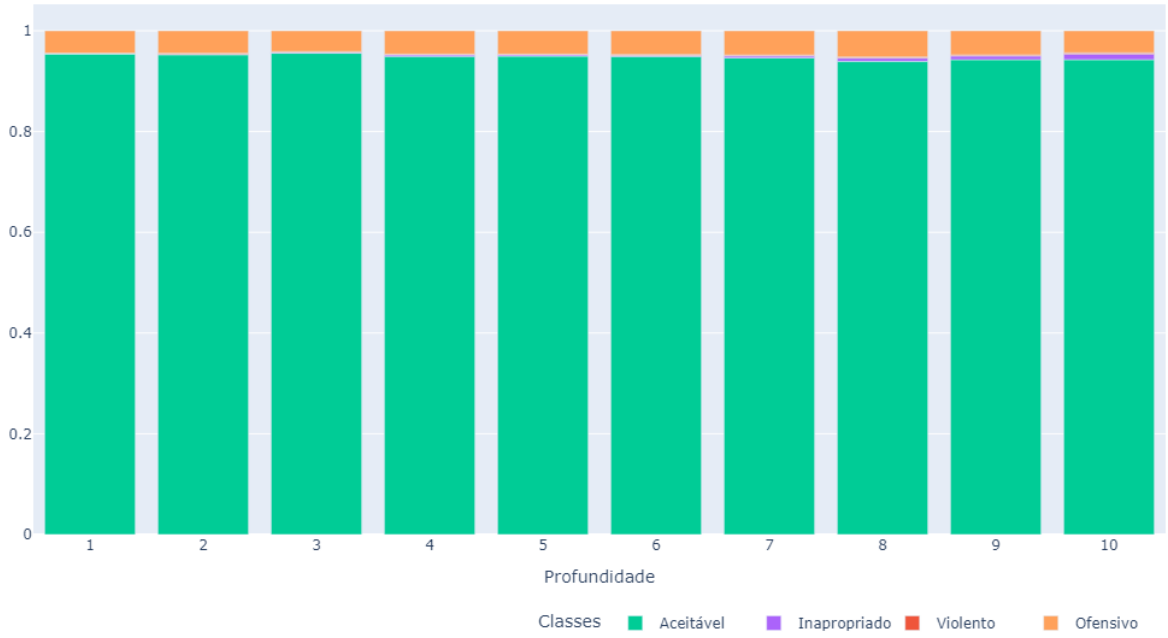


Figura 2: Proporção das classes dos recomendados por profundidade no experimento aleatório.

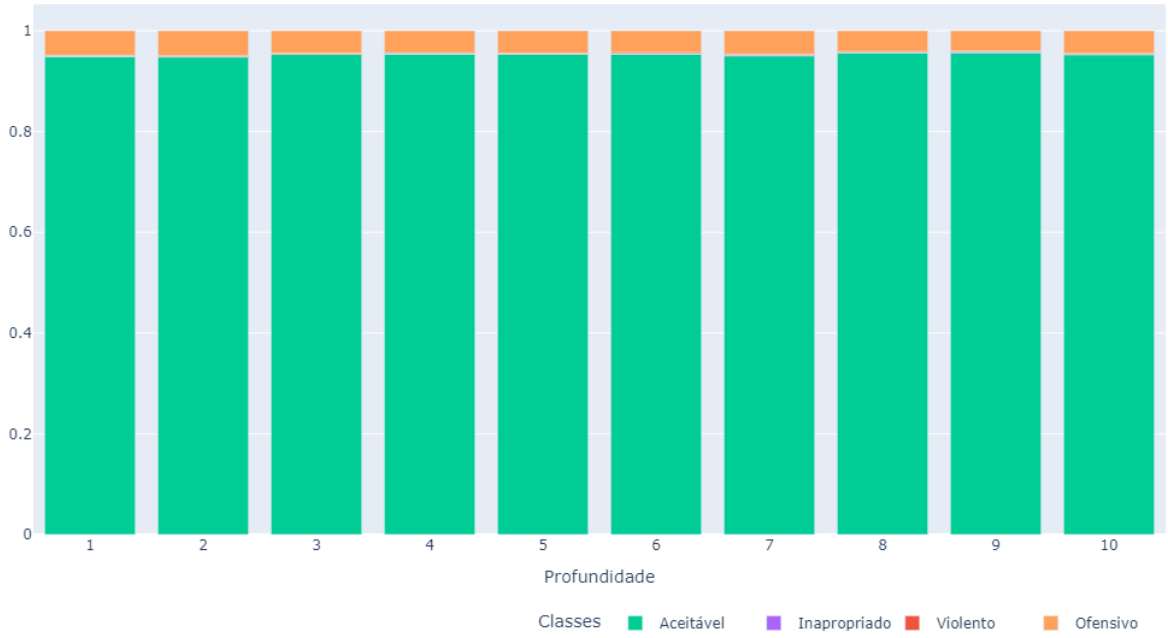


Figura 3: Proporção das classes dos recomendados por profundidade no experimento não aleatório.

Como pode ser observado nos gráficos, todas as profundidades apresentaram praticamente a mesma proporção das classes consideradas na classificação do vídeo. Destaca-se em todas as profundidades a Classe Aceitável, com mais de 95%, sendo a Classe Ofensivo a outra presente em maior quantidade. Foram ainda encontrados segmentos classificados como das Classes Inapropriado e Violento, mas em uma pequena proporção (Tabelas 2 e 3). Esse padrão para todas as as observações pode demonstrar o resultado da execução de filtros de moderação do YouTube, que tentam restringir o acesso a conteúdos

inapropriados e violentos de sua plataforma.

Podemos também analisar a média e o desvio padrão do número de segmentos por classe e por profundidade na Tabela 2. Nela observa-se o valor referente a todos os vídeos selecionados a partir de um, definido aleatoriamente, daqueles recomendados. A mesma análise foi feita, na tabela 3, com todos os vídeos recomendados a partir do primeiro listado na lista de recomendações. Os valores das tabelas representam a média de 20 execuções dos experimentos.

| **Profundidade** | **Aceitável** | | **Inapropriado** | | **Violento** | | **Ofensivo** | |
|---|---|---|---|---|---|---|---|---|
| | **Média** | **Desvio** | **Média** | **Desvio** | **Média** | **Desvio** | **Média** | **Desvio** |
| 0 | 50,8 | 38,54 | 0,0 | 0,0 | 0,0 | 0,0 | 2,0 | 2,37 |
| 1 | 73,35 | 44,04 | 0,0 | 0,0 | 0,05 | 0,22 | 3,5 | 3,69 |
| 2 | 83,35 | 38,63 | 0,3 | 0,95 | 0,25 | 0,7 | 3,9 | 4,32 |
| 3 | 81,6 | 42,22 | 1,1 | 4,79 | 0,1 | 0,3 | 4,1 | 5,39 |
| 4 | 82,1 | 48,6 | 0,0 | 0,0 | 0,15 | 0,36 | 3,4 | 3,93 |
| 5 | 106,0 | 50,27 | 0,05 | 0,22 | 0,1 | 0,44 | 4,0 | 3,69 |
| 6 | 78,5 | 40,27 | 0,2 | 0,68 | 0,3 | 0,95 | 3,8 | 5,18 |
| 7 | 100,1 | 58,43 | 0,6 | 2,03 | 0,45 | 0,67 | 4,15 | 3,75 |
| 8 | 84,55 | 50,54 | 0,2 | 0,87 | 0,05 | 0,22 | 4,0 | 5,44 |
| 9 | 102,45 | 59,21 | 0,85 | 2,52 | 0,2 | 0,87 | 4,75 | 5,4 |
| 10 | 100,85 | 48,31 | 0,75 | 2,45 | 0,35 | 0,65 | 5,25 | 5,13 |

Tabela 2: Média e desvio de segmentos por profundidade dos vídeos visitados de forma aleatória.

| **Profundidade** | **Aceitável** | | **Inapropriado** | | **Violento** | | **Ofensivo** | |
|---|---|---|---|---|---|---|---|---|
| | **Média** | **Desvio** | **Média** | **Desvio** | **Média** | **Desvio** | **Média** | **Desvio** |
| 0 | 43,45 | 26,62 | 0,0 | 0,0 | 0,05 | 0,22 | 2,0 | 2,26 |
| 1 | 75,95 | 35,93 | 0,05 | 0,22 | 0,2 | 0,51 | 4,25 | 5,77 |
| 2 | 70,05 | 56,81 | 0,1 | 0,3 | 0,15 | 0,36 | 4,85 | 6,29 |
| 3 | 69,0 | 43,35 | 0,1 | 0,3 | 0,25 | 0,54 | 4,55 | 6,09 |
| 4 | 73,45 | 56,23 | 0,05 | 0,22 | 0,15 | 0,48 | 3,7 | 6,17 |
| 5 | 64,4 | 46,04 | 0,1 | 0,3 | 0,25 | 0,54 | 4,0 | 5,91 |
| 6 | 70,05 | 47,18 | 0,05 | 0,22 | 0,1 | 0,3 | 3,9 | 5,66 |
| 7 | 70,25 | 44,92 | 0,15 | 0,36 | 0,2 | 0,51 | 4,3 | 5,91 |
| 8 | 77,45 | 45,17 | 0,15 | 0,36 | 0,15 | 0,48 | 3,65 | 5,77 |
| 9 | 72,7 | 40,61 | 0,1 | 0,3 | 0,25 | 0,54 | 4,15 | 5,95 |
| 10 | 82,6 | 38,28 | 0,05 | 0,22 | 0,2 | 0,51 | 3,75 | 4,84 |

Tabela 3: Média e desvio de segmentos por profundidade dos vídeos visitados de forma não aleatória.

Nas Tabelas 2 e 3, na média de segmentos analisados ao longo das 10 profundidades, as classes de discursos presentes tendem a manter a mesma média da quantidade de segmentos ao longo das profundidades. Nas mesmas tabelas 2 e 3, podemos observar um alto desvio padrão da Classe Aceitável, provavelmente devido à quantidade diferente de segmentos por vídeo. Como os vídeos têm diferentes tempos de duração, sempre menor ou igual a 10 minutos, gera-se diferentes quantidades de segmentos para cada um deles, refletido principalmente nas classes Aceitável e Ofensivo, majoritárias na amostra.

Na Figura 3 foi realizada a análise de um vídeo escolhido aleatoriamente dos recomendados em cada profundidade. Em todas as profundidades destaca-se a Classe Aceitável, sendo a Classe Ofensivo a outra presente em maior frequência. No entanto, diferente da primeira, observamos uma variação ao longo das profundidades, onde essas proporções variam entre mais de 95% para classe majoritária (Aceitável) e menos de 90% para essa mesma classe.

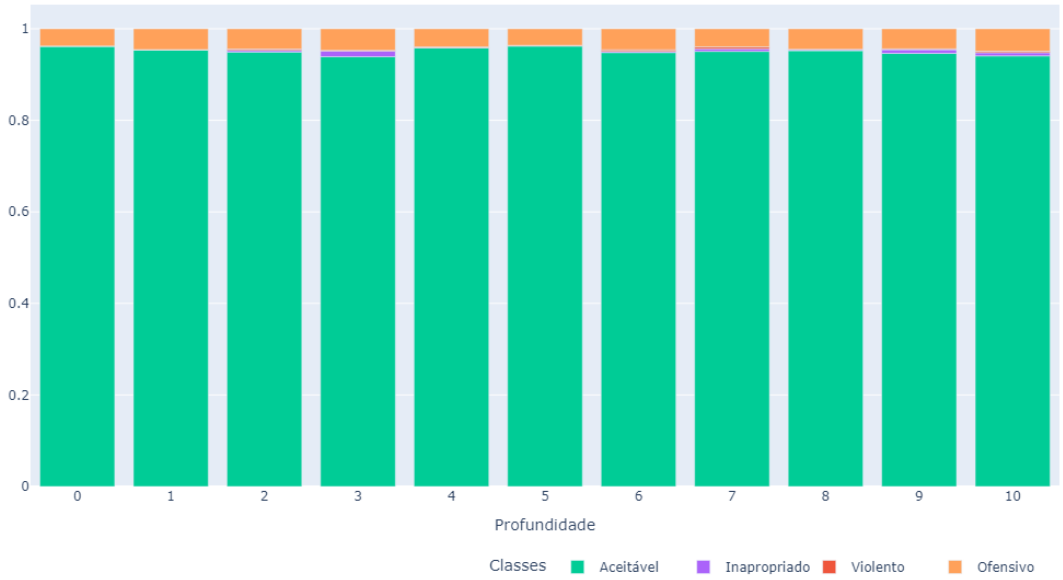


Figura 4: Proporção das classes dos vídeos selecionados por profundidade no experimento aleatório

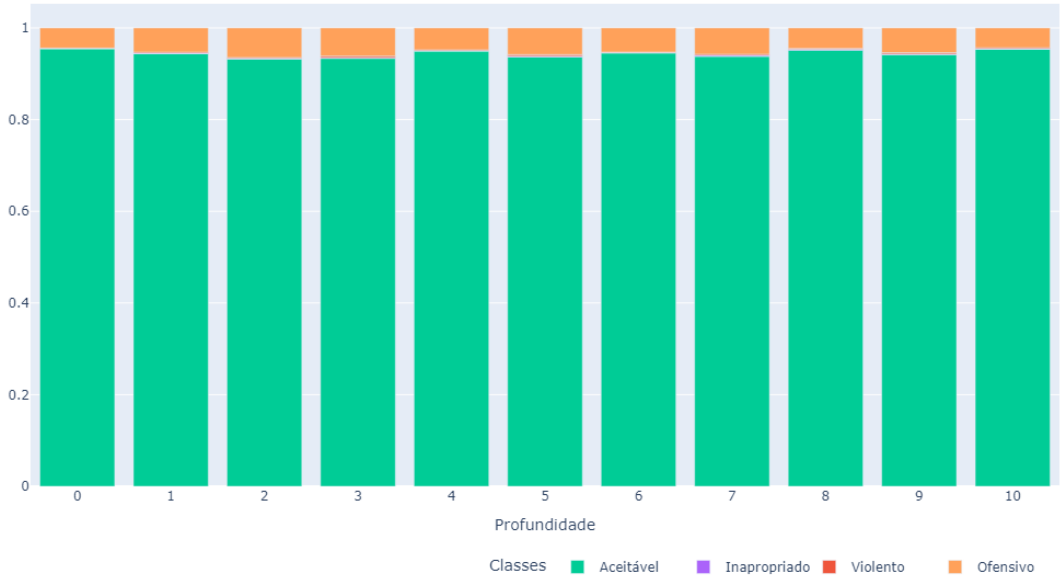


Figura 5: Proporção das classes dos vídeos selecionados por profundidade no experimento não aleatório

## 4.2 Distribuição de classes ao longo dos vídeos

O gráfico da Figura 3 mostra a análise das mesmas classes (Aceitável, Inapropriado, Violento e Ofensivo) ao longo de um vídeo específico para cada uma das profundidades (0-10). A Profundidade 0 corresponde ao vídeo que foi selecionado direto da página de notícias e não dos recomendados. É possível observar que, frequentemente, os segmentos classificados como ofensivos acontecem próximos uns dos outros, como esperado, já que pode ser um trecho único segmentado.


Figura 6: Distribuição das classes para cada vídeo por segundo do video.

Pelos dados demonstrados através dos gráficos e tabelas mostradas anteriormente, podemos ver que mesmo sendo em quantidade reduzida devido aos filtros que o YouTube possui para a seleção de seu conteúdo, esse tipo de discurso ainda está presente nos vídeos postados nessa plataforma.

É possível perceber que uma vez que esses discursos aparecem, o conteúdo tende a se manter presente nos demais vídeos recomendados. Também é possível observar, como mostrado nos trabalhos relacionados descritos anteriormente, que utilizar a transcrição do áudio do vídeo é uma maneira eficiente para a classificação desse tipo de conteúdo.

## 4.3 Análise dos conteúdos classificados como violentos, ofensivos ou inapropriados

Para avaliar se os conteúdos classificados como Violentos, Ofensivos ou Inapropriados, realmente pertencem a essas categorias, nessa seção foi selecionado um vídeo para podermos analisar os segmentos que foram classificados em uma dessas três classes e verificar se o conteúdo realmente possui esse teor os se apenas é uma limitação do modelo utilizado.

Para essa análise, o vídeo selecionado foi o vídeo com o maior número de segmentos rotulados como Inapropriado. Esse vídeo em questão é uma entrevista com o produtor musical, rapper e ator Fredro Starr falando sobre o assassinato do também rapper e ator Merlin Santana. O vídeo[2] foi publicado em 16/03/2024, conta com 1 milhão de visualizações e foi acessado em 06/04/2024.

Para esse vídeo, todos os segmentos que não foram classificados como aceitáveis estão presentes na Tabela 4, entretanto, serão analisados como mais detalhes alguns desses segmentos. Para os nove primeiro segmentos presentes na tabela, temos todos como inapropriados, neles podemos observar que todos têm ou a presença da palavra *fuck* ou *shit*. Esse parece ser o principal motivo para o modelo ter os classificados desta maneira. É importante ressaltar que, por exemplo, no sexto segmento, que é a frase: "So that's why I didn't really give a f***", o áudio do vídeo tem um sinal sonoro (*bip*), o que significa que o modelo recebeu o texto como está mostrado, com a censura, e mesmo assim o classificou com inapropriado.

Em relação aos segmentos classificados como ofensivos, o primeiro classificado dessa maneira foi a frase "N**** was paying.". Nela temos a presença da palavra n****s, que é uma gíria do idioma inglês, usada principalmente entre os afro-americanos e que significa "preto" ou "negro", só que é uma gíria depreciativa.

[2]https://www.youtube.com/watch?v=LUDiD1m8p7Y

Outro segmento classificado como ofensivo foi o “So he’s like, all right, you’re a rat for me.”. Nesta frase, traduzida como “Então ele disse, tudo bem, você é um rato para mim.”, observamos um teor ofensivo, apesar de não possuir palavras inapropriadas.

Outro segmento que merece a atenção é: o “Like telling n****s I hear a little sh**.”, que foi classificado como inapropriado, onde observamos novamente a presença da palavra n****s, porém acompanhada da palavra sh**, fazendo com que a classificação fosse inapropriada.

Para a Classe Violento, temos a presença de apenas um segmento que é o “Cut it out” traduzido como “Pare com isso”, porém, a presença da palavra *cut* pode ter levado o modelo a classificá-lo como violento, levando em consideração que essa palavra pode ser traduzida como cortar, ignorando seu contexto.

Levando em consideração os segmentos analisados para este vídeo, o modelo utilizado acertou a classificação da maioria dos segmentos, apesar de não avaliar o contexto dos segmentos próximos.

| Classificação | Tempo(s) | Texto |
|---|---|---|
| Inapropriado | 53 | Saturday night, Nick, this shit looked like the movie |
| Inapropriado | 97 | I never forget this shit. |
| Inapropriado | 118 | You know, I’m gonna give a f***, you know what I mean? |
| Inapropriado | 147 | getting ready for this audition for some Moisha shit. |
| Inapropriado | 150 | I didn’t even want to do this shit. |
| Inapropriado | 153 | So that’s why I didn’t really give a f***, |
| Inapropriado | 158 | Don’t do that shit. |
| Inapropriado | 162 | You don’t want to do that shit. |
| Inapropriado | 163 | Onyx is too hardcore for that shit. |
| Ofensivo | 167 | N**** was paying. |
| Inapropriado | 175 | So I’m like, oh, sh**. |
| Ofensivo | 177 | So he’s like, all right, you’re a rat for me. |
| Inapropriado | 193 | And he was like, oh, sh**. |
| Inapropriado | 222 | Like telling n****s I hear a little sh**. |
| Inapropriado | 283 | You talking about me, telling Nick I’m soft and all this shit. |
| Ofensivo | 345 | Like we was like, yo man, that’s stupid. |
| Ofensivo | 347 | It was dumb, you know, and, you know, |
| Inapropriado | 359 | There’s an episode of Moesha after all that beef sh** episode |
| Inapropriado | 380 | And this is after all the beef sh**. |
| Ofensivo | 395 | before I have to Usha your ass up out of here |
| Inapropriado | 397 | and I was like, what’s some sh** like that? |
| Inapropriado | 411 | That sh** was lit. |
| Ofensivo | 421 | Yo, they did Merlin dirty, man. |
| Ofensivo | 425 | They did him dirty, man. |
| Ofensivo | 434 | because you automatically relate to that sh** |
| Inapropriado | 455 | I don’t want to stay in f**king Beverly Hills, n***a. |
| Inapropriado | 460 | After that day, you ain’t never seen my black ass in the hood ever again. |
| Ofensivo | 492 | Hood, sh**t, ghetto, every f**king way. |
| Ofensivo | 497 | You in L.A. to do one f**king thing, n***a. |
| Ofensivo | 502 | Stop f**king bull sh**t. |
| Inapropriado | 504 | Stay your ass in f**king Beverly Hills. |
| Inapropriado | 507 | Stay your ass in... |
| Violento | 511 | Cut it out. |

Tabela 4: Segmentos que foram classificados diferente de Aceitável.

# 5 Conclusões

Como foi possível observar nos resultados obtidos, mesmo uma plataforma como o YouTube, que apresenta filtros em relação aos conteúdos que são publicados, foi possível encontrar discursos inapropriados, violentos ou ofensivos. No entanto, é importante destacar que alguns dos segmentos considerados em uma dessas classes são falsos positivos, sendo importante avaliar outros modelos de classificação.

Em relação às limitações do trabalho, um dos problemas encontrados foi a falta de modelos que lidam

com mídias com idiomas diferentes do inglês, como o português. Além disso, foi necessário limitar o tempo dos vídeos em 10 minutos ou menos, devido a limitação de hardware para lidar com um processamento maior.

Como trabalhos futuros pretende-se utilizar modelos mais complexos ou vídeos de maior duração. No entanto, é importante considerar um limiar mínimo para o tempo de vídeo, mantendo uma quantidade próxima de segmentos para cada mídia analisada. É importante treinar ainda um modelo que faça a classificação para a língua português e que levem em consideração o contexto, podendo melhorar as previsões das classes previstas.

Finalmente, destaca-se a necessidade de realizar mais experimentos, observando se o padrão de recomendação se mantém a partir de um vídeo com uma maior proporção de segmentos inapropriados, violentos e ofensivos, daqueles que não os tem.

# Referências


[1] Haroon, Muhammad, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq e Magdalena Wojcieszak: *YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations*, 2022.

[2] Srba, Ivan, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik e Maria Bielikova: *Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles*. ACM Transactions on Recommender Systems, 1(1):1–33, janeiro 2023, ISSN 2770-6699. http://dx.doi.org/10.1145/3568392.

[3] Wu, Ching Seh e Unnathi Bhandary: *Detection of Hate Speech in Videos Using Machine Learning*. Em *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, páginas 585–590, 2020.

[4] Das, Mithun, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta e Animesh Mukherjee: *HateMM: A Multi-Modal Dataset for Hate Video Classification*, 2023.

[5] Boishakhi, Fariha Tahosin, Ponkoj Chandra Shill e Md. Golam Rabiul Alam: *Multi-modal Hate Speech Detection using Machine Learning*. Em *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, dezembro 2021. http://dx.doi.org/10.1109/BigData52589.2021.9671955.

[6] Prasad, Nishchal, Sriparna Saha e Pushpak Bhattacharyya: *Multimodal Hate Speech Detection from Videos and Texts*. EasyChair Preprint no. 10743, EasyChair, 2023.

[7] Gomez, Raul, Jaume Gibert, Lluis Gomez e Dimosthenis Karatzas: *Exploring Hate Speech Detection in Multimodal Publications*, 2019.

[8] Sutejo, Taufic Leonardo e Dessi Puji Lestari: *Indonesia Hate Speech Detection Using Deep Learning*. Em *2018 International Conference on Asian Language Processing (IALP)*, páginas 39–43, 2018.

[9] Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta e Vasudeva Varma: *Deep Learning for Hate Speech Detection in Tweets*. Em *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, WWW '17 Companion. ACM Press, 2017. http://dx.doi.org/10.1145/3041021.3054223.

[10] Khanday, Akib Mohi Ud Din, Syed Tanzeel Rabani, Qamar Rayees Khan e Showkat Hassan Malik: *Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques*. International Journal of Information Management Data Insights, 2(2):100120, 2022, ISSN 2667-0968. https://www.sciencedirect.com/science/article/pii/S2667096822000635.

[11] Brown, Megan, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler e Joshua Aaron Tucker: *Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users*. 2022. https://ssrn.com/abstract=4114905, Available at SSRN: https://ssrn.com/abstract=4114905 or http://dx.doi.org/10.2139/ssrn.4114905.

[12] Kralj Novak, Petra, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič e Fabiana Zollo: *Handling Disagreement in Hate Speech Modelling*. Em Ciucci, Davide, Inés Couso, Jesús Medina, Dominik Ślezak, Davide Petturiti, Bernadette Bouchon-Meunier e Ronald R. Yager (editores): *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, páginas 681–695, Cham, 2022. Springer International Publishing, ISBN 978-3-031-08974-9.