Análise do Discurso de Ódio em Comentários de Vídeos no YouTube: Um Estudo de Caso da CPI da COVID-19 no Brasil

Marcela C. Teixeira, Julio C. S. Reis

[1]Departamento de Informática, Universidade Federal de Viçosa (UFV)

marcelacuperteixeira@gmail.com, jreis@ufv.br

Abstract. *Social platforms are part of most Brazilians' life. In different proportions, everyone consumes and produces content, are influenced, and influencers. In this context, regardless of the subject discussed, whether politics, religion, or sports, it is increasingly common to find speeches that are harmful to specific groups of people, manifesting themselves with a series of variations, such as intimidation, prejudice, offensive content, and incitement to hatred. Given this also destructive potential of social platforms, the present work aims to analyze these hate speeches in Portuguese, made on videos published on YouTube during a specific period, namely - April 27 to July 14, 2021, the date of installation of the COVID-19 CPI until the day it was decided it's an extension. Our results reveal interesting patterns of this hateful discourse that can be useful in the future to direct potential evolutions of the policies for using this platform and propose effective tools against the problem.*

Resumo. *As plataformas sociais são parte da vida da maioria dos brasileiros. Em diferentes proporções, todos consomem e produzem conteúdos, são influenciados e influenciadores. Nesse contexto, independente do assunto discutido, seja política, religião ou esportes, é cada vez mais comum encontrar discursos que são prejudiciais a certos grupos de pessoas, manifestando-se em uma série de variações, como intimidação, preconceitos, conteúdos ofensivos e incitação ao ódio. Diante desse potencial também destrutivo das plataformas sociais, o presente trabalho tem como objetivo analisar discursos de ódio, a partir de comentários em português feitos em vídeos publicados no YouTube durante um período específico, a saber - dia 27 de abril até 14 de julho de 2021, data da instalação da CPI do COVID-19 até o dia que foi decidido a prorrogação da mesma. Nossos resultados revelam padrões interessantes desse discurso odioso que podem ser úteis, futuramente, para direcionar potenciais evoluções das políticas de uso da plataforma bem como para a proposição de ferramentas que sejam efetivas contra o problema.*

1. Introdução

As plataformas sociais popularizaram-se muito nos últimos anos, e, durante o contexto da pandemia causada pelo coronavírus (COVID-19), tornaram-se essenciais para a manutenção das diversas relações devido, por exemplo, à necessidade de isolamento social. Nestes ambientes, os usuários consomem e compartilham uma grande variedade de conteúdos e se expressam livremente; entretanto, muitas vezes essa liberdade é confundida, abrindo espaço para opiniões pessoais carregadas de preconceitos e discriminação. Infelizmente,

o fato é que esses sistemas se tornaram palco de inúmeros casos de discurso de ódio online [Gagliardone et al. 2015], que tendem a se agravar em períodos envolvendo temas controversos e/ou divisíveis. Como resultado, este fenômeno tem sido reconhecido como um problema premente por muitos segmentos da sociedade atual e autoridades de muitos países, como a Alemanha [Knight 2018].

Frente a esse desafio, é necessário, antes de propor soluções, entendermos padrões do discurso de ódio disseminado no ambiente online, considerando diferentes plataformas e cenários. Isso é importante pois cada plataforma é proposta para fins específicos, e isso pode delinear e/ou influenciar aspectos da interação do usuário neste contexto com potenciais impactos em diversas áreas da nossa sociedade, como política (i.e., democracia) e saúde. Logo, é sobre esta lacuna que se enquadra o objetivo deste trabalho.

Particularmente, exploramos dados oriundos de comentários feitos em vídeos do YouTube para entender características do conteúdo odioso disseminado no contexto brasileiro considerando um evento específico, a CPI da COVID-19. Para isso, aplicamos uma abordagem para identificação de ódio que, embora tenha sido utilizado em trabalhos anteriores [Mondal et al. 2017], foi pouco explorado para o português brasileiro. Mais especificamente coletamos e analisamos cerca de 40 mil comentários feitos em vídeos publicados na plataforma durante 27 de abril a 14 de julho de 2021 (i.e., data de instauração da referida CPI). De forma geral, nossos resultados revelam um alto percentual de comentários tóxicos (cerca de 26%) realizados em vídeos do YouTube, uma plataforma moderada. As análises de padrões textuais também destacam características interessantes do conteúdo odioso postado nesta plataforma. Esperamos que esses resultados possam ser usados, futuramente, como insumo para a proposição de alternativas (e.g., ferramentas e políticas) que minimizem o impacto do discurso de ódio em plataformas digitais como YouTube.

2. Metodologia

Nesta seção é apresentada a metodologia adotada neste projeto, incluindo detalhes relativos ao processo de coleta de dados e a estratégia adotada para identificação de ódio em um comentário.

Coleta de Dados. Primeiramente, foi necessário definir os canais do YouTube que seriam explorados durante o estudo. Para isso, os seguintes critérios foram adotados: *(i)* canal que tenha publicado vídeo relacionado à CPI da COVID-19 durante a período de interesse (i.e., 27 de abril e 14 de julho de 2021); *(ii)* popularidade do canal (e.g., em termos de número de inscritos) e por fim; *(iii)* representatividade política. É importante mencionar que, para *(iii)*, foram selecionados canais representativos de ambos os espectros políticos (i.e., esquerda e direita), conforme estudo apresentado em [Oliveira et al. 2019].

A Tabela 1 apresenta uma visão geral dos dados coletados por canal, incluindo os identificadores (IDs) dos vídeos e o número de comentários (por vídeo e canal). No total, a partir da API oficial do YouTube[1], foram coletados 39.516 comentários[2], realizados pelos usuários da plataforma nos top-4 vídeos mais populares de cada um dos canais selecionados, a saber: "TV247", "Conversa Afiada" e "TV Folha" (como representantes

[1] `https://developers.google.com/youtube/v3/docs/?apix=true`

[2] Coleta realizada em Fevereiro de 2022 e limitada aos 5000 comentários mais populares de cada vídeo.

Tabela 1. Relação de vídeos por canal e número de comentários coletados.

Canal	ID do vídeo (# comentários coletados)	# por canal
TV 247	DcJGYFMxVjU (139), NwLmFva54tw (117), wkh-4zZ6w8g (136), f_7t5WoHST4 (93)	485
Conversa Afiada	civj0lt-bTs (629), mcnFYCkPnfE (1269) AaSHfj69-0k (2427), z8aY9HY4ok0 (1333)	5658
TV Folha	76OaMkk9yCU (128), VVIjdBltHp4 (1333) 7zrgJxEIhSg (233), y1scdSEQ2LA (44)	1738
Nando Moura	JgsSsjwETu0 (2149), bhnM_InKYCE (4005) 7PfBatVHNk4 (2795), 5dsNPXuZ-kI (2212)	11161
Folha Política	vUugOTc39Jw (4110), RsWkyj4mhy0 (2304) 24pbuysNUpY (3291), QzTqL2dShtU (4976)	14681
Mamãe Falei	PQ6iTA8nCFM (1349), 8AStCClubCA (947) QydzV6ISslo (1244), pHAFiTsGExE (2253)	5793

do jornalismo alternativo de esquerda) e "Nando Moura", "Folha Política" e "Mamãe Falei" (como representantes de canais com viés político de direita). Em suma, já podemos observar que o número de comentários postados em nos vídeos veiculados por canais representativos da direita (i.e., "Nando Moura", "Folha Política" e "Mamãe Falei") é mais expressivo em comparação a comentários postados em vídeos publicados por canais representativos da esquerda (cerca de 80% do número total de comentários).

Estratégia para Identificação de Ódio. Após a coleta dos comentários, foi necessário definir uma estratégia que nos permita distinguir um comentário odioso (ou tóxico) dos demais. Neste contexto, exploramos a Perspective API[3], uma plataforma que se baseia em aprendizado de máquina, para identificação deste tipo de conteúdo. Em resumo, o modelo fornece um resultado entre 0 e 1 que indica a probabilidade de que um dado conteúdo, fornecido como entrada, seja percebido como ofensivo/tóxico. Em outras palavras, quanto mais próximo de 1, maior a probabilidade de que o conteúdo seja percebido como tóxico. Esta ferramenta tem sido bastante explorada em trabalhos anteriores [Lima et al. 2020], no entanto, seu uso em dados em português, ainda é limitado a algumas plataformas, como Facebook [Guimarães et al. 2020].

Fundamentada a estratégia, e com base no limiar definido no estudo apresentado em [ElSherief et al. 2018], que também faz uso da Perspective API para identificação do discurso de ódio, porém em inglês, foi delimitado, para o presente trabalho, que comentários de ódio (ou tóxicos) seriam aqueles nos quais a probabilidade de toxicidade retornada pela API fosse maior que 0,8. Esse limiar tem sido utilizado como base em diversos trabalhos anteriores neste contexto [Lima et al. 2020]. Na próxima seção são apresentados os resultados das análises realizadas no trabalho.

3. Resultados

3.1. Toxicidade

A Figura 1 apresenta a distribuição cumulativa (CDF) de toxicidade para os comentários coletados. De forma geral, percebe-se uma tendência: cerca de 26% dos comentários analisados foram considerados tóxicos, independentemente do canal ou vídeo. Em outras palavras, 10.235 dos 39.516 comentários coletados e analisados possuem probabilidade de toxicidade maior que 0,8, o que sugere um volume significativo de conteúdo com alta probabilidade de conter discurso odioso. Ademais, embora esse percentual seja menor em comparação ao restante, ele é bastante expressivo, principalmente considerando que

[3] www.perspectiveapi.com

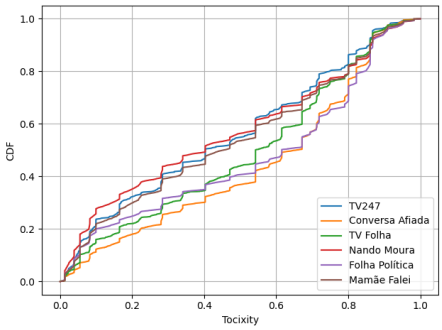

Figura 1. Distribuição Cumulativa (CDF) da toxicidade associada aos comentários coletados.

o YouTube, uma plataforma amplamente utilizada pelos brasileiros[4], é um ambiente moderado. Isso reforça a necessidade de realização de estudos como esse, que podem ser úteis para criação de subsídios para evolução das políticas de uso dessas plataformas bem como a proposição de ferramentas que possam ser efetivas na contenção da disseminação deste tipo de conteúdo nestes ambientes.

Na Tabela 2 apresentamos exemplos destes comentários com limiar maior que 0,8 de probabilidade de toxicidade, onde podemos observar qualitativamente o conteúdo considerado odioso. Neste contexto é válido ressaltar que nomes reais foram anonimizados (substituídos pela palavra NOME) para garantia da integridade das informações pessoais dos envolvidos.

3.2. Nuvem de Palavras

A Figura 2 apresenta nuvens de palavras geradas para ambos os grupos de comentários (i.e., tóxicos e não tóxicos). Observamos a presença de palavras que remetem diretamente a disseminação de ódio (ver Figura 2a), como por exemplo, “vagabundo”, “canalha” e “merda”; e também palavras que, associadas ao contexto, apontam um discurso carregado de animosidade, como “corrupto” e “bandido”. Enquanto na Figura 2b, as palavras também são sensíveis ao contexto de pandemia e política, porém menos, ou não (explicitamente) relacionadas ao ódio, como em “cpi”, “brasil”e “povo”.

3.3. Atributos Psicolinguísticos

Por fim, com objetivo de entender de maneira mais aprofundada características que possam ser úteis para distinguir comentários tóxicos dos demais, analisamos atributos psicolinguísticos dos comentários coletados. Para isso foi exploramos a versão mais recente do LIWC (*Linguistic Inquiry Word Count*) disponível para o idioma português do Brasil

[4] `www.similarweb.com/top-websites/`

Tabela 2. Exemplos de comentários e sua respectiva toxicidade.

Comentário	Valor de Toxicidade
“Cala a boca nojento falso, ladrão da saúde !”	0,98
“Não existe vacina pra vírus rotativo e mutável vcs são simplesmente cobaias idiota.”	0,95
“Esse mulher é ridicula, falando, pode me dizer ao menos sua religião, ridiculaaa, o que tem haver a religião ai sua pilantra, [...] porque é sem pé e nem cabeça uma pessoa questionar isso em uma CPI.”	0,90
“Eu tenho NOJO desses lixos bolsonaristas.”	0,87
“Não pode ficar impune, esse canalha, do *NOME*, que roubou o dinheiro da saúde [...], foi VCS canalhas com suas roubalheiras.”	0,86

(a) Toxicidade $>$ 0,8 **(b) Toxicidade $\leq$ 0,8**

Figura 2. Nuvem de palavras associada aos comentários coletados.

[Pennebaker et al. 2015]. O LIWC é um dicionário (léxico) que refere-se à uma coleção de palavras que definem uma determinada categoria, com objetivo de analisar emoções, componentes cognitivos e estruturais de textos, já utilizado em diversos esforços anteriores [Gonçalves et al. 2015]. Uma vez que o objetivo é investigar características mais intrínsecas ao texto, mensuramos a razão entre as porcentagens para cada uma das categorias. Por exemplo, para a categoria *money*, a porcentagem em textos com toxicidade $>$ 0,8 foi de 2.14% e para toxicidade $\leq$ 0,8 foi de 2.24. Nesse caso, a razão entre os valores (2.14/2.24 $\approx$ 0.96) indica que a ocorrência da categoria nos textos é aproximadamente a mesma. Logo, apresentamos na Tabela 3 apenas os resultados de diferenças significativas (i.e., razão $\geq$ 1,5 ou $\leq$ a 0,5).

Podemos notar que as categorias que apresentam diferenças mais significativas estão diretamente ligadas ao contexto de ódio, sendo elas relacionadas a emoção negativa (*negemo*) - (e.g., feio, nojento) - pontuando pelo menos 50% a mais nos textos considerados tóxicos; e a palavras consideradas de raiva (*anger*) - (e.g., ódio, matar) - que ocorrem quase duas vezes mais. Na Tabela 4 são apresentados comentários que contém exemplos de palavras pertencentes a tais categorias significativas. Pelos mesmos motivos apresentados anteriormente, os nomes reais foram anonimizados (substituídos pela palavra NOME).

De forma geral, embora as análises tenham sido feitas considerando uma amostra de vídeos e comentários populares, elas reforçam a necessidade de ferramentas efetivas para contenção de discursos de ódio, que ocorrem, inclusive, em plataformas moderadas como YouTube.

4. Conclusão

Neste trabalho apresentamos uma caracterização do discurso de ódio contido em comentários de vídeos do YouTube, publicados no intervalo de 27 de abril a 14 de julho de 2021, data da instalação da CPI do COVID-19 até o dia que foi decidido a prorrogação da mesma, um importante evento no contexto (político) brasileiro.

Em suma, os resultados apresentam evidências de que as políticas e mecanismos atuais ainda não são totalmente eficazes para conter comentários odiosos em plataformas como o YouTube. Além disso, questiona-se os impactos políticos e sociais da não restrição de discursos que ferem o respeito pelo outro, que podem culminar em danos

Tabela 3. Categorias do LIWC com razão significativa.

Categoria	Toxicidade $>$ 0,8	Toxicidade $\leq$ 0,8	Razão
negemo	7,2%	4,39%	1,64
anger	3,44%	1,75%	1,97

Tabela 4. Exemplos de comentários com palavras das categorias *negemo* e *anger*.

Comentário
"Corruptos saqueadores vocês mataram mais que Covid."
"Eu tô com tanto odio de *NOME*, Que queria ver ele perdendo na urna para o canalha do *NOME*. Podem me julgar!"
"Essa cpi é um nojo."
"Esse *NOME* é um babaca nojento. Deixa de ser imbecil. Qdo é pra fuder o presidente é certíssimo né. N vale merda esse cara. O *NOME* mentiu. Outro, q n tem postura e sem caráter."
"Que palhaçada! Esse *NOME* está plantando feio ! Deus está vendo tudo."

irreversíveis. Destacamos ainda, que tal fenômeno independe do canal e do vídeo. Além disso, as análises de conteúdo e atributos psicolinguísticos dos comentários evidenciaram características que podem ser úteis, futuramente, para a construção de abordagens que sejam capazes de distinguir um conteúdo tóxico dos demais, como por exemplo, padrões associados a presença de palavras relacionadas a emoção negativa e raiva, que incluem vocábulos como "nojento","ódio" e "matar".

Acreditamos que os resultados obtidos podem ser usados como insumo para a proposição de medidas (e.g., ferramentas e eventualmente adequação de políticas) que sejam efetivas no combate ao discurso de ódio no ambiente online. Como trabalhos futuros pretendemos explorar outras plataformas (e.g., Telegram) e investigar a eficácia de abordagens baseadas em aprendizado de máquina, por exemplo, na identificação do discurso de ódio disseminado em plataformas digitais.

Agradecimentos. Este trabalho foi parcialmente financiado pelo CNPq e FAPEMIG.

Referências

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Proc. of the ICWSM*, pages 430–434.

Gagliardone, I., Gal, D., Alves, T., and Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Gonçalves, P., Dalip, D., Reis, J., Messias, J., Ribeiro, F., Melo, P., Araújo, L., Gonçalves, M., and Benevenuto, F. (2015). Bazinga! caracterizando e detectando sarcasmo e ironia no twitter. In *Proc. of the BrasNAM*.

Guimarães, S. S., Reis, J. C., Ribeiro, F. N., and Benevenuto, F. (2020). Characterizing toxicity on facebook comments in brazil. In *Proc. of the WebMedia*, pages 253–260.

Knight, B. (2018). Germany implements new internet hate speech crackdown. `https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590`. (Acessado em 26/06/2023).

Lima, L., Reis, J. C., Melo, P., Murai, F., and Benevenuto, F. (2020). Characterizing (un) moderated textual data in social systems. In *Proc. of the ASONAM*, pages 430–434.

Mondal, M., Araújo, L. S., and Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proc. of the HYPERTEXT*, pages 85–94.

Oliveira, R., Ribeiro, M., and Ortellado, P. (2019). Uma descrição dos canais políticos no youtube. Technical report, Monitor do Debate Político no meio Digital.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.