Breve Análise do Sistema do Recomendação do YouTube

Dhruv Babani
d.babani0012edu.pucrs.br
PUC-RS
Porto Alegre, Rio Grande do Sul, Brasil

ACM Reference Format:
Dhruv Babani. 2022. Breve Análise do Sistema do Recomendação do YouTube. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

Buscou-se, neste trabalho, realizar uma breve análise do sistema de recomendação do YouTube, mediante estudo de dados coletados sobre o comportamento de seu algoritmo em relação às ações do usuário na plataforma num período de 10 dias.

2 INTRODUÇÃO

O YouTube é uma das mais populares, mas menos estudadas, plataformas de mídia social, e leva conteúdo a 95% da população mundial.[1] Tendo em vista sua popularidade e a facilidade em sua utilização, escolheu-se o YouTube para a realização deste trabalho, que tem o objetivo de estudar o algoritmo de seu sistema de recomendação a partir da análise de 2 contas de teste criadas na rede. O tema escolhido para o perfil de atuação foi o futebol e a outra conta atuará como perfil de controle. Para basear o estudo, definiu-se as seguintes hipóteses:

(1) Na página principal da conta de controle, haverá uma presença de pelo menos 25% de vídeos de futebol (tema mais popular), enquanto na seguinte conta, pelo menos 66% dos vídeos da página principal serão a respeito do respectivo tema. Ao assitir um vídeo, 2 entre os 3 primeiros vídeos recomenadados (66%) serão do mesmo tema do vídeo que está sendo assitido.
(2) Na página principal da conta de controle, não haverá um tema dominante, nenhum tema passando dos 25% de presença, todavia na conta alternativa, pelo menos 40% dos vídeos da página principal serão a respeito do tema escolhido. Ao assitir um vídeo, no máximo, 1 entre os 3 primeiros vídeos recomendados (33%) serão do mesmo tema do vídeo que está sendo assistido.

Nossas hipóteses, em suma, sugerem duas possibilidades a respeito do algoritmo de recomendação do YouTube. A primeira sugere que o YouTube, para a conta de controle, irá recomendar mais vídeos de futebol, uma vez que é o tema mais popular dentre os dois, com o intuito de manter o usuário na plataforma o máximo de tempo possível, sem arriscar mostrar algum outro assunto que possa desinteressá-lo. Para as demais contas, aconteceria o mesmo processo da conta de controle, no entanto, de maneira mais agressiva, tendo no mínimo dois terços dos vídeos recomendados com a mesma temática da respectiva conta. Para os 3 primeiros vídeos recomendados logo após assitir a um vídeo, espera-se que pelo menos 2 a cada 3 sejam da mesma temática do vídeo assistido, então, se o usuário estiver assistido a um vídeo de programação, 2 dos 3 primeiros vídeos recomendados serão sobre programação. Em contrapartida, a segunda hipótese sugere que, para a conta de controle, o usuário estaria exposto a diversos conteúdos sem que qualquer um destas temáticas de conteúdos ultrapassasse os 25% de presença. Para as demais contas, o algoritmo de recomendação apresentaria um comportamento similar ao esperado na conta de controle, expondo o usuário a diferentes temas, porém, mantendo uma percentagem de presença de 40% do tema escolhido para a conta. Para os 3 primeiros vídeos recomendados logo após assitir a um vídeo, espera-se que no máximo 1 desses 3 vídeos seja do mesmo conteúdo do vídeo assistido.

3 METODOLOGIA

Para a realização deste trabalho, criou-se 2 contas de teste na plataforma: 1 delas operariam apenas em seus respectivo tema:Futebol, e a seguinte foi definida como conta de controle, que assistiria vídeos do tema anterior e vídeos de temas aleatórios, alternando o tipo de tema entre os dias de coleta.

Após isso, definiram-se as seguintes variáveis independentes:

(1) Dias de coleta: dias úteis (mesmo com a ocorrência de feriados), com início no dia 01/11/22 e fim no dia 11/11/22, totalizando 10 dias;
(2) Horário de coleta: início exatamente às 19:20;
(3) Nº de vídeos assistidos: na conta normal assistiria apenas 3 vídeos sobre o tema por dia de coleta;
(4) Vídeos recomendados: após o fim de cada vídeo assistido, seriam coletados o tema dos 3 primeiros conteúdos recomendados.

Sobre as variáveis dependentes, definiu-se o seguinte:

(1) Duração dos vídeos: sofreria variação em cada análise, nas quais os vídeos deveriam durar entre 10 e 15 minutos, mas sempre com a duração mais próxima possível entre as 2 contas a cada análise;
(2) Operação da conta de controle: o conjunto de temas pesquisados era mudado a cada 2 dias de análise, com alternância entre o tema pré-definido (3 videos) e temas aleatórios (1 vídeo por tema). A organização dos dias foi defenida pelo seguinte:
    - Nos dias 3º, 4º, 7º, 8º e 10º, a conta operaria em vídeos de temas aleatórios;

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- Nos dias 1º, 2º, 5º, 6º e 9º, a conta operaria em vídeos do tema escolhido.

Em todos os dias de coleta e em todas as contas, realizou-se, inicialmente, a anotação dos temas dos 8 vídeos recomendados na página principal e, em seguida, a pesquisa dos temas e a busca por vídeos dentro do intervalo de tempo pré-definido, que só eram assistidos quando todas as contas tivessem encontrado vídeos que respeitassem essa condição. Após cada vídeo assistido, anotava-se os 3 primeiros vídeos recomendados logo após seu fim, então repetia-se a pesquisa do tema e buscavam-se novos vídeos.

Para todos os vídeos assistidos, foi anotado o seu título e tema, com o fim de evitar repetições do mesmo vídeo em outros dias de coleta. Após o fim da coleta de dados, construiu-se gráficos para cada conta que representassem, separadamente, os temas de vídeos da página principal e os temas de vídeos recomendados após cada vídeo. A partir das análises desses gráficos, calculou-se a porcentagem da presença de vídeos do mesmo tema na página principal, e a porcentagem da presença de vídeos do mesmo tema nos vídeos recomendados logo após assitir a um vídeo.

As porcentagens foram calculadas com a seguinte fórmula:

$$\frac{TT}{TV} * 100$$

Onde:

- TT = Total de vídeos de um tema - tema escolhido ou temas aleatórios (apenas conta controle);
- TV = Total de vídeos de uma recomendação - (80 para a página principal; 90 para os recomendados após os vídeos; 50 para recomendados após os vídeos na conta de controle).

4 TRABALHOS RELACIONADOS

4.1 Examining algorithmic biases in YouTube's recommendations of vaccine videos [1]

Neste trabalho, analisou-se como o Youtube recomenda vídeos relacionados à vacinação, a partir da coleta de duas listas de dados:

(1) 250 vídeos mais relevantes na busca de termos relacionados à vacinação;
(2) Vídeos sobre vacinação recomendados por cada vídeo da lista 1 (Aproximadamente 50 vídeos coletados/vídeo).

A base de dados formada, que possuía 2122 vídeos, incluia também os números de visualizações, de comentários, de likes e de deslikes e a data de publicação de cada vídeo, além dos IDs e nomes dos canais que os publicaram. Posteriormente, os vídeos foram divididos nos seguintes temas: pró-vacina; neutro e antivacina, e a partir dessa ordenação, usou-se um modelo de probabilidade para calcular a tendência de cada vídeo a recomendar outros ou de ser recomendado, baseado na sua relação com a vacinação.

Após os testes, concluiu-se os seguintes resultados:

(1) Atributos como likes, dislikes, quantia de comentários e de visualizações não tem efeitos significativos nas recomendações;
(2) Os vídeos pró-vacina foram quase 3 vezes mais recomendados que os antivacina;
(3) Vídeos pró-vacina recomendam mais conteúdo do mesmo tema em comparação aos antivacina.
(4) Vídeos pró-vacina tendem a recomendar uma quantidade maior de conteúdo pró-vacina ou neutro, enquanto os antivacina tendem a recomendar vídeos do mesmo tema em grande quantia, formando bolhas de conteúdo na rede.

Com base nessas conclusões, sugeriu-se que o YouTube possui um filtro em seu sistema de recomendação, que opera relacionado às opiniões que os vídeos transmitem, o que gerou preocupações de que os vídeos antivacinas sejam menos propensos a levar os usuários a vídeos pró-vacinas devido às bolhas observadas na rede de recomendação. Por fim, realizou-se uma breve discussão geral sobre a importância da transparência em algoritmos de recomendação e sobre a forma como plataformas de mídia social, como o YouTube, poderiam decidir qual conteúdo apresentar aos seus usuários.

4.2 "Down the Rabbit Hole" of Vaccine Misinformation on YouTube: Network Exposure Study [3]

Esse trabalho nos introduz uma discussão a respeito do algoritmo de recomendação do YouTube como um potencial propagador de desinformação a respeito de vacinas e a área da saúde. Para realizar essa pesquisa foram usados mais de 538 vídeos, tendo registrado mais 6 vídeos dentre os recomendados, e usando palavras-chave para pesquisa direta. Os temas dos vídeos variavam entre vídeos relacionados a vacinas, sendo pró ou contra; vídeos de saúde, sem envolver vacinas; e vídeos sobre autismo (apenas). Tendo esses vídeos, foi calculada a taxa de exposição a desinformação, o número 0 representa risco nulo de exposição.

Os resultados obtidos indicaram que a maioria dos vídeos pró-vacinas eram de fontes confiáveis, como hospitais e instituições governamentais, porém, o usuário fica sujeito a um alto risco de exposição a conteúdos anti-vacina. Vídeos não relacionados a vacinas tinham poucas chances de serem expostos a conteúdo anti-vacina, assim como vídeos relacionados a autismo e a saúde. No entanto, vídeos anti-vacina apresentaram grandes probabilidades de seresm expostos a outros vídeos de conteúdos similares.

Com essa pesquisa descobriu-se que ao usar a barra de pesquisa do YouTube, o usuário consegue fácil acesso a vídeos pró-vacina de fontes confiáveis. Além disso, mesmo que assistam vídeos pró-vacina, existe uma grande chance de serem recomendados vídeos anti-vacina. Por fim, calculou-se que cerca de apenas de 2 a 6 porcento dos vídeos sobre autismo, não estão relacionados a vacinas. Concluiu-se que, ao pesquisar diretamente pela barra de pesquisa do YouTube, há menores chances de encontrar vídeos que levem a desinformação, no entanto, dependendo do vídeo assistido, este pode levar o usuário desinformação.

4.3 Examining Political Bias within YouTube Search and Recommendation Algorithms [2]

Neste artigo, examinou-se o viés político do sistema de recomendação do YouTube por meio de dois experimentos. No primeiro, explorou-se as recomendações do sistema de busca do site, no qual foram testados 30 termos com viés políticos, coletando-se 200 vídeos para cada termo (24000 no total). Após isso, quatro contas com diferentes vieses foram criadas: esquerda; direita; ambas; nulo (não assistirá conteúdo político). As contas assistiram 20 vídeos de seus

vieses, e então pesquisaram os 30 termos separados inicialmente. Os vídeos recomendados nas pesquisas foram coletados, transcritos e analisados por uma IA especializada em calcular o viés político de um texto. Após análise, percebeu-se que a distribuição do viés na recomendação era muito próxima, mas com uma pequena tendência a recomendar conteúdos de esquerda.

No segundo experimento, analisou-se a recomendação de conteúdo que surge logo após finalizar a visualização de um vídeo na plataforma. Para isso, 1200 vídeos foram selecionados dentre o total usado no primeiro experimento, e foram divididos nas categorias "mínimo","moderado"e "extremo", tanto para esquerda quanto para direita. Os vídeos foram assistidos e selecionou-se o primeiro recomendado após o fim do vídeo base, num ciclo repetido 10 vezes. No fim, não houveram valores estatisticamente significantes para definir o viés geral na recomendação. A partir dos experimentos, concluiu-se que o sistema de recomendação em buscas do YouTube possui um viés levemente inclinado a conteúdeo de esquerda, mas para a recomendação após vídeos, a plataforma evita formar bolhas de conteúdo e tenta levar o usuário a outros conteúdos.

5 RESULTADOS

Após a análise dos dados e construção dos seguintes gráficos, obteu-se os seguintes resultados:

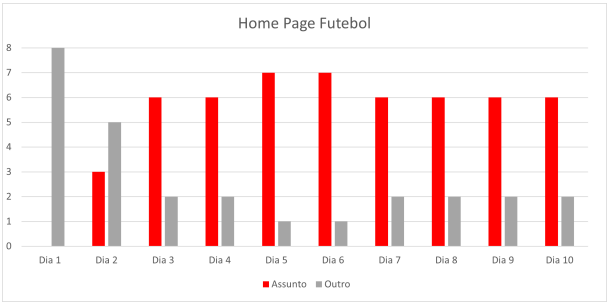

Figura 1: Vídeos recomendados na página principal para a conta de Futebol.

Figura 2: Vídeos recomendados após assistir a um vídeo na conta de Futebol.

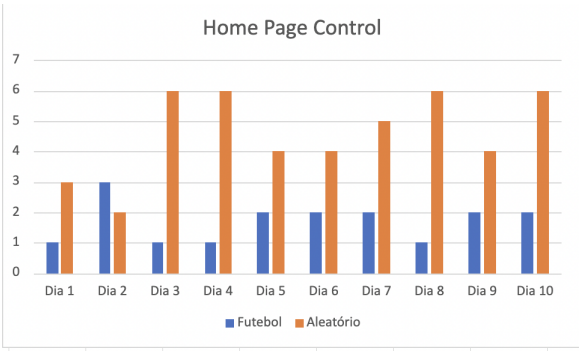

Figura 3: Vídeos recomendados na página principal para a conta de Controle.

(1) Futebol: Em média, 66% dos vídeos da home page eram sobre futebol (Figura 1), enquanto 95% dos vídeos recomendados após assistir a um vídeo também eram sobre futebol
(2) Controle: Na home page, em média de 20% dos vídeos eram de futebol; 80% de vídeos totalmente aleatórios Ao assistir a um vídeo de tema aleatório, 64% dos vídeos recomendados, em média, eram do mesmo assunto do vídeo assistido Quanto aos vídeos recomendados: Ao assitir a um vídeo de um dos temas escolhidos, 73% desses vídeos eram do mesmo assunto do vídeo assistido, em média.

6 CONSIDERAÇÕES FINAIS

Ao analisar os resultados, fizeram-se algumas constatações:

(1) Todas as contas, tiveram vídeos de futebol sendo recomendados;
(2) As conta de futebol obteve quase o mesmo resultado de suas páginas iniciais e nos mesmos dias.

Ao fim da pesquisa, pudemos analisar de forma clara e sucinta nossos dados e percebemos que nossas hipóteses estavam erradas, já que tanto a primeira quanto a segunda falharam em certo ponto. A primeira hipótese provou-se errada, já que na conta de controle a presença de vídeos de futebol na página principal do YouTube não ultrapassou os 20%.

Para os vídeos recomendados, ambas as hipóteses falharam, já que ultrapassou os 33% de presença para todas as contas, e, na conta de controle, os vídeos recomendandos do mesmo assunto chegaram a apenas 64%, não atingindo a meta esperada. Pode-se dizer que, apesar de ambas as hipóteses não terem sido confirmadas, a hipótese de número dois foi a que mais perto chegou de ser verdadeira, tendo sua única falha na afirmação acerca dos vídeos recomendados.

O que pode-se constatar a partir dos resultados que: na home page, o algoritmo de recomendação do YouTube, por mais que dê preferencia ao conteúdo mais assistido, ainda mantém uma boa margem de conteúdos desconhecidos ao usuário, no entanto, se o usuário decidir assistir vídeos de determinado tema, ele dificilmente irá assistir a outro vídeo de um conteúdo diferente.

REFERÊNCIAS

[1] Deena Abul-Fottouh, Melodie Yunju Song, and Anatoliy Gruzd. 2020. Examining algorithmic biases in YouTube's recommendations of vaccine videos. *International Journal of Medical Informatics* 140 (2020), 104175. https://doi.org/10.1016/j.

ijmedinf.2020.104175
[2] Michael Lutz, Sanjana Gadaginmath, Natraj Vairavan, and Phil Mui. 2021. Examining Political Bias within YouTube Search and Recommendation Algorithms. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1–7. https://doi.org/10.1109/SSCI50451.2021.9660012

[3] Lu Tang, Kayo Fujimoto, Muhammad Tuan Amith, Rachel Cunningham, Rebecca A Costantini, Felicia York, Grace Xiong, Julie A Boom, Cui Tao, et al. 2021. "Down the rabbit hole" of vaccine misinformation on YouTube: Network exposure study. *Journal of Medical Internet Research* 23, 1 (2021), e23262.