Characterizing Toxicity on Facebook Comments in Brazil

Samuel S. Guimarães
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
samuelsg@ufmg.br

Julio C. S. Reis
Universidade Federal de Minas Gerais
Universidade FUMEC
Belo Horizonte, Minas Gerais, Brasil
julio.reis@dcc.ufmg.br

Filipe N. Ribeiro
Universidade Federal de Ouro Preto
João Monlevade, Minas Gerais, Brasil
filipe.ribeiro@ufop.edu.br

Fabrício Benevenuto
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil
fabricio@dcc.ufmg.br

ABSTRACT

On social media platforms, comments associated with news pieces are usually filled with negativity and toxicity, many times promoting flamed discussions and insults among users. Although designed to encourage conversations and interactions, the high toxicity might end up contributing to create a hostile environment in the online space, which is detrimental to both social media platforms and their users. In this work, we provide a large-scale diagnostic about the toxicity in comments associated with news shared on Facebook. To do that, we collected all posts and comments from relevant pages during a major political event in Brazil, the release of Former President Lula from prison. We then used the Perspective API from Google to measure the toxicity of the comments and posts. Our analysis of the toxicity unveils features that influence toxicity associated with the news, especially in relation to public figures. We hope our findings may affect the design of better content policies able to mitigate the problem.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Applied computing** → *Sociology*.

KEYWORDS

Facebook, Social Media, Hate, Comments

1 INTRODUÇÃO

Mídias sociais como o Facebook e o Twitter se tornaram ambientes propícios para usuários encontrarem, compartilharem e discutirem notícias em tempo real sobre assuntos diversos. Quantificando essa tendência, trabalhos recentes estimaram que 68% dos americanos [26] e 66% dos brasileiros [20] consomem notícias principalmente a partir de mídias sociais. Essas plataformas excederam jornais impressos como uma fonte de notícias para os americanos desde 2018 [25], e no Brasil desde 2014 [20].

Esse novo ambiente proporcionado pelas mídias sociais trouxe consigo novos recursos que alteraram o ecossistema de notícias, modificando a maneira como elas são produzidas, disseminadas e discutidas. Primeiramente, nessas plataformas qualquer pessoa pode se registrar como um jornalista ou grupo de mídia, por exemplo, apenas criando uma página no Facebook alegando ser um. De fato, foram descobertas mais de 20.000 páginas no Facebook que se auto-declaram como uma fonte de notícias apenas nos Estados Unidos [23]. Além disso, as redes sociais online oferecem uma estratégia de disseminação na qual usuários auxiliam no compartilhamento de notícias para influenciar seus amigos [3, 31]. Ao mesmo tempo, sites de mídias sociais permitem que os usuários interajam e discutam entre si através das seções de comentários de cada postagem. Entre essas pessoas que normalmente comentam notícias, cerca de 77,9% deles comentam em redes sociais [28].

Esses comentários permitem que usuários se engajem em discussões com os autores, abrindo um espaço para diferentes perspectivas e criticismo construtivo. Os comentários também permitem transformação dos usuários de leitores passivos para participativos, permitindo que fontes de notícias recebam sugestões de melhorias [12]. De forma complementar, o motivo mais comum para essas pessoas expressarem uma opinião online, entre as que tendem a comentar, é externar um sentimento ou desejo [28], e apesar das vantagens na interação, o sentimento expressado pode ser extremamente desagradável. Dessa forma, esses comentários rudes frequentemente surgem em notícias online, independentemente da manchete da notícia [22]. Isso pode transformar as seções de comentários em terrenos férteis para toxicidade e hostilidade, similar à algo que ocorreu com as mensagens em sites de notícias [9], o que levou à remoção das seções de comentários nesses sites.

Essa toxicidade pode ser observada em vários outros contextos e dentro de diferentes redes sociais como Twitter [18, 27], Youtube [24], Gab [17] e Facebook [11, 16]. Todavia, quando focamos em comentários tóxicos, especialmente em notícias e conteúdo semelhante, o Facebook se destaca. Ele é a maior rede social do mundo, com 2.5 bilhões de usuários ativos em dezembro de 2019 [4], sendo que 67% dos americanos buscam suas notícias na plataforma [26]. Para o Brasil, essa estatística é de 54% [20]. Contudo, as postagens contendo as notícias são sobrecarregadas de comentários rudes [13, 29]. E dentro dessa parte do ecossistema

WebMedia '20, November 30-December 4, 2020, São Luís, Brazil

ACM ISBN 978-1-4503-8196-3/20/11…$15.00
https://doi.org/10.1145/3428658.3430974

de notícias no Brasil, pouco se sabe sobre a distribuição dessas mensagens tóxicas.

Este artigo procura preencher essa lacuna provendo um extenso diagnóstico sobre a toxicidade nos comentários associados com notícias compartilhadas no Facebook. Mais especificamente, as *questões de pesquisa* que este trabalho busca responder são: (P1) *Que tipos de páginas recebem mais comentários tóxicos?*; (P2) *Quais fatores influenciam a proporção desses comentários?* e; (P3) *Quais são as caraterísticas comuns das mensagens tóxicas?*

Para responder a essas perguntas, foram coletadas e analisadas postagens de páginas relevantes no Facebook durante um evento político brasileiro de destaque: a saída do Ex-Presidente Luiz Inácio Lula da Silva da prisão em 8 de novembro de 2019. Mais especificamente, foram coletadas postagens e comentários de uma semana antes e uma semana depois deste evento de interesse, com foco em páginas políticas e páginas de notícias que possuem seções políticas. Em seguida, foi medida a toxicidade dos comentários e postagens usando a Perspective API do Google. A análise dessas medições em várias páginas e publicações constitui a principal contribuição deste artigo. Dentre os principais resultados verificou-se que páginas de mídia recebem mais toxicidade que páginas de figuras públicas, com a política dessas figuras afetando pouco a toxicidade. No entanto, os comentários tóxicos são mais prevalentes quando as figuras públicas se tornam o tópico de uma postagem. Em geral, os comentários que respondem a outros comentários são mais tóxicos do que as respostas gerais, e há mais mensagens tóxicas nas páginas com publicações mais tóxicas. Essa toxicidade também se mostra concentrada, com cerca de 20% das páginas responsáveis por 60% dos comentários tóxicos e 56% de todas as postagens tóxicas. Por fim, foi descoberto que a saída de Lula da prisão mostra que um determinado evento político pode aumentar a toxicidade, mas o efeito pode não ser significativo. Com esses resultados espera-se que os principais fatores que influenciam os comentários tóxicos possam ajudar plataformas de redes sociais no desenvolvimento de políticas de conteúdo, especialmente no Brasil.

O restante do artigo está organizado da seguinte forma. A segunda seção apresenta uma revisão bibliográfica. Em seguida, é discutida a metodologia experimental. Posteriormente, são descritos os resultados que respondem as perguntas de pesquisa. E, finalmente, são discutidas as conclusões e possíveis direções futuras.

2 TRABALHOS RELACIONADOS

O discurso de ódio, a polarização política, e a análise de interações em redes sociais têm sido tópicos de interesse nos últimos anos, especialmente depois de alguns países aprovarem leis com intenção de reduzir a quantidade de discurso de ódio online, sem contar com o aparente pico de polarização política [10]. Buscando por trabalhos recentes que como este analisam comentários de notícias, especialmente trabalhos que lidam com dados obtidos do Facebook, observa-se que eles se dividem entre os que usaram uma gradação de quanto ódio existe em uma mensagem, alguns que discutiram como detectar o ódio, e poucos outros que analisaram o quadro geral. Entre os trabalhos que medem ódio nos comentários, Su *et al.* [29] descobriu que grosseria e incivilidade foram mais prevalentes nos comentários nas páginas do Facebook de sites de notícias que eram conservadoras ou locais, com cerca de 20% a 40% dos comentários nessas notícias consistindo de comentários rudes. Da mesma forma, Reis *et al.* [22] mostraram que as seções de comentários de jornais influentes estão se tornando um local de hostilidade propício para ação de trolls, coletando mensagens de sites de notícias e usando análise de sentimento para confirmar essa tendência. A análise de comentários em notícias feita por esses trabalhos também é presente neste artigo, também incluindo uma comparação política através das figuras públicas, mas sendo feita uma análise agora focada no Brasil e utilizando o conceito toxicidade da Perspective API[1].

Explorando diferentes métodos de detecção de ódio, Almeida *et al.* [1] usaram os dados produzidos por Davidson *et al.* [6], divididos em mensagens com ódio, ofensivas e regulares, para explorar quantificadores da área de teoria da informação para aprimorar o uso de TF-IDF na classificação do texto. Também procurando aprimorar a classificação, Pelle *et al.*[21], com uma definição de linguagem ofensiva similar à presente na Perspective API, usou a combinação de classificadores para capturar diferentes aspectos de algoritmos distintos, buscando reduzir a variância nos resultados. Como a detecção de ódio é um problema aberto, e as subdivisões do ódio ainda não são claras [30], neste trabalho foi utilizado um dos modelos que são considerados estado da arte, a Perspective API.

Paralelamente, poucos artigos coletam ou usam dados dos comentários de notícias no Facebook, não só medindo o ódio. Entre os que criaram um conjunto de dados do Facebook, Khan e Chang [14] coletaram as postagens da página da Amazon durante cinco anos utilizando a Graph API do Facebook que permite a coleta de diversos dados das páginas e, após geração de atributos, utilizaram diferentes redes neurais para prever o número de interações distintas que a postagem receberá, com base no conteúdo, a forma do conteúdo (texto, imagem, etc.) e algumas informações temporais. Kolhatkar e Taboada [15] também agregaram 1.121 comentários, sendo focados na noção de construtividade dos comentários das notícias, avaliando o resultado com uma abordagem de deep learning. Nesse caso, comentários construtivos seriam comentários que não apenas geram uma resposta emocional, mas criam um dialogo civilizado com os envolvidos. A construtividade foi comparada com a toxicidade da Perspective API. Para dados brasileiros, Pelle e Moreira [7] criaram o conjunto de dados *OffComBr*[2] usando um dos maiores sites de notícias do Brasil, o G1. Seu foco era a criação de um banco de dados que pudesse ser usado em aprendizado de máquina, rotulando 1.250 comentários da seção de comentários do site. Contrastando esses artigos observa-se que as análises tiveram um foco em conceitos diferentes da toxicidade ou tinham um escopo mais fechado em grupos específicos, analisando cerca de milhares de comentários.

Em resumo, medir a toxicidade de dados online está se tornando algo mais preciso e útil. No trabalho aqui apresentado foi realizada uma análise em uma escala maior que trabalhos anteriores no contexto brasileiro, concentrada no Facebook, a maior plataforma de mídia social utilizada para compartilhamento de notícias no Brasil.

3 METODOLOGIA

Nesta seção é apresentada a metodologia proposta para este trabalho, incluindo a estratégia para selecionar e agrupar veículos de

[1] https://www.perspectiveapi.com
[2] https://github.com/rogersdepelle/OffComBR

notícias e páginas políticas no Facebook, além de como foi inferida a toxicidade dos comentários e postagens associados à eles.

3.1 Encontrando Páginas do Facebook de Política e Notícias

O primeiro passo da metodologia adotada é a seleção das páginas do Facebook que seriam monitoradas. A seleção foi feita a partir de uma lista inicial de 22 páginas do Facebook introduzida por Moretto e Ortellado [19], que inclui veículos de notícias brasileiros da mídia tradicional e alternativa, bem como outras meta-informações, como sua inclinação política e alcance. Como apenas 14 páginas ainda eram ativas, usamos a ferramenta *Audience Insights* do Facebook[3] para expandir o conjunto inicial.

Essa ferramenta auxilia anunciantes a refinar o público para o qual desejam exibir um anúncio, definindo um conjunto de atributos como idade, localização, sexo e *interesses*. Esses *interesses* consistem em um grande conjunto de tópicos inferidos pelo Facebook que representam assuntos que provavelmente irão atrair o interesse do usuário, incluindo figuras públicas, políticos, partidos, tipos de comida, restaurantes ou atividades. Uma das funções do *Audience Insights* é ajudar a detalhar o público desejado para o anúncio, sugerindo tópicos relacionados. Dessa forma ao informar um interesse, a ferramenta sugere páginas do Facebook com um público semelhante (menu 'Curtidas na página').

A partir disso, foram escolhidos quatro interesses relacionados ao cenário político brasileiro como sementes e buscadas as páginas relacionadas manualmente. Os seguintes interesses foram usados: 1) **Jair Messias Bolsonaro**, o atual presidente (direita); 2) **Lula**, o ex-presidente (esquerda); 3) **o Partido Social Liberal (PSL)**, um partido de direita[4]; e 4) o **Partido dos Trabalhadores (PT)**, um partido de esquerda. Das páginas sugeridas correspondentes a cada um dos quatro interesses, incluímos as das seguintes categorias[5]: Figuras públicas, Políticos, Funcionários do governo , Autores, Organizações Políticas, Partidos Políticos, Mídia, Sites de Notícias e Mídia, Empresas de Notícias/Mídia, Empresas de Transmissão e Produção de Mídia, Revistas, Jornalistas, Programas de TV (relacionados a Notícias) e Jornais.

Esse processo resultou em 63 páginas brasileiras no Facebook de várias categorias[6], mas que são agrupadas principalmente em duas categorias principais: Figuras Públicas e Mídia, subdivididas em cinco subgrupos. Esses cinco subgrupos são organizados em: (1) Figuras Públicas: **Figuras Públicas de Direita**, **Figuras Públicas de Centro**, e **Figuras Públicas de Esquerda**, e; (2) Mídia: **Mídias Tradicionais** e **Mídias Alternativas**.

Como a lista inclui não apenas políticos, mas também ativistas políticos, foi preferido o uso do termo 'figuras públicas' ao invés de 'políticos' neste estudo. Para essas figuras públicas, foi utilizada a posição política auto-declarada para designar os seus subgrupos, e foram ignoradas as páginas associadas a figuras públicas que não possuíam posição política clara. Para mídias, foram usados dados de organizações oficiais de imprensa brasileiras, como a Associação Nacional de Jornais (ANJ), a Associação Nacional de Editores de Revistas (ANER) e a Agência Nacional de Telecomunicações (ANATEL), para classificar as páginas de mídias como **alternativa** se não tiverem nenhum registro nessas instituições, e **tradicionais**, se tiverem. Essas mídias alternativas tendem a se apresentar apenas online ou, às vezes, tem apenas páginas do Facebook, onde publicam exclusivamente. Em seguida, são apresentados os detalhes dos dados coletados.

3.2 Coletando os dados

Com a lista de páginas criada, foi usada a Graph API[7] do Fabebook para coletar postagens e comentários associados às mesmas. Os dados de postagens e comentários incluem conteúdo textual, o número de curtidas, data de publicação e menções à outros usuários do Facebook. Também foram coletadas todas as respostas aos comentários. Usando esta API, as postagens do período de 27 de outubro a 16 de novembro de 2019 nas páginas selecionadas foram o foco para compor nosso conjunto de dados, cerca de uma semana antes e após a saída de Lula da prisão, capturando todos as publicações disponíveis na API.

3.3 Inferindo Toxicidade

Para inferir a toxicidade das postagens e comentários nas páginas do Facebook foi usada a Perspective API do Google. Existem vários modelos fornecidos pela API, que são descritos na Tabela 1.

Tabela 1: Visão geral das métricas da Perspective API.

Métrica	Descrição do Modelo
TOXICITY	Comentário rude, desrespeitoso ou irracional que provavelmente fará as pessoas abandonarem uma discussão
SEVERE_TOXICITY	Uma versão da *Toxicity* que é menos sensível a comentários que incluem o uso positivo de palavrões
IDENTITY_ATTACK	Comentários negativos ou odiosos dirigidos a alguém por causa de sua identidade
INSULT	Comentário insultuoso, inflamatório ou negativo focado em uma pessoa ou grupo
PROFANITY	Palavrões ou outra linguagem obscena ou profana
THREAT	Descreve a intenção de infligir dor, ferimento ou violência contra um indivíduo ou grupo

Em todos os casos, dado um texto, os modelos retornam uma métrica igual a probabilidade do mesmo ser da classe desejada. Quando o texto é confuso ou com erros ortográficos, o modelo pode não retornar uma pontuação. Para diminuir esse efeito uma limpeza de hashtags e links foi feita antes da medição, não excluindo emoticons já que a API lida bem com eles. Para 8,17% das postagens e 9,13% dos comentários em nosso conjunto de dados, a API não foi capaz de medir a toxicidade. Usamos todos os modelos para medir a toxicidade de postagens e comentários, mas descobrimos que a maioria dos resultados está altamente correlacionada entre si

[3] https://www.facebook.com/ads/audience-insights/.
[4] Partido pelo qual Bolsonaro concorreu à presidência.
[5] O administrador da página do Facebook é quem designa as categorias à página entre a lista predefinida de categorias do Facebook.
[6] Essas 63 páginas são das seguintes categorias - 1 Autor, 4 Empresas de Transmissão e Produção de Mídia, 5 Revistas, 1 Mídia, 16 Empresas de Notícias/Mídia, 9 Sites de Notícias e Mídia, 1 Organização Política, 15 Políticos e 11 Figuras Públicas.
[7] https://developers.facebook.com/docs/graph-api

Tabela 2: Visão geral dos dados coletados rotulados como figuras públicas de direita (FgPD), figuras públicas de centro (FgPC), figuras públicas de esquerda (FgPE), todas as figuras públicas (FgP) e também mídia tradicional (MTr) e alternativa (MAlt).

Tipo	Páginas	Posts	Coment.	Curtidas	Seguidores	Falando a Respeito	Coment. Tóxicos	Posts Tóxicos	Média	Desvio Padrão	Erro Quad. Médio	Entropia
FgPD	10	875	1.308.573	27.533.991	30.224.588	2.468.575	9,50%	2,29%	0,2855	0,1028	0,0111	0,3141
FgPC	4	125	65.793	4.646.812	4.618.715	55.567	9,09%	0,00%	0,3135	0,1273	0,0180	0,3045
FgPE	12	1.100	547.581	13.024.855	13.907.604	1.194.221	9.82%	1,00%	0,2929	0,1237	0,0169	0,3212
FgP	*26*	*2.100*	*1.921.947*	*45.205.658*	*48.750.907*	*3.718.363*	*9,58%*	*1,48%*	*0,2886*	*0,1168*	*0,0146*	*0,3158*
MTr	15	6.725	1.404.119	63.962.840	64.071.959	7.644.183	16,19%	0,22%	0,3903	0,1633	0,0323	0,4428
MAlt	22	7.000	769.853	12.443.604	12.484.364	3.151.038	17,15%	1,24%	0,3912	0,1945	0,0430	0,4583
Mídia	*37*	*13.725*	*2.173.972*	*76.406.444*	*76.556.323*	*10.795.221*	*16,53%*	*0,74%*	*0,3906*	*0,1799*	*0,0377*	*0,4483*
Total	**63**	**15.825**	**4.095.919**	**121.612.102**	**125.307.230**	**14.513.584**	**13,27%**	**0,84%**	**0,3427**	**0,1728**	**0,0346**	**0,3915**

(o coeficiente de correlação de Pearson maior que 0,8 ($p < 0,05$)). Assim, neste trabalho, são apresentados os resultados utilizando o modelo de toxicidade (*Toxicity*). A Tabela 3 mostra exemplos de comentários com suas pontuações de Toxicidade correspondentes.

Tabela 3: Exemplos de comentários e sua toxicidade.

Comentário	Toxicidade
F*P não ganha mais nada, é vi*do	0,905
Que tal ... matar-mos Lula e tds os bandidos do STF	0,884
Vão tudo pra cadeia	0,506
sabe ele não gosta de pobre	0,673
Genteeeee... ninguém faz nada pra cessar essa corja?	0,369
Não deixem soltar Barrabás novamente Não podem cometerem este erro de novo ...	0,199
claro que ta ...ele quer que continue sendo estatal pra meter a mao na grana que entra ..muito simples	0,128

3.4 Base de dados

A partir da lista de páginas citada, com a Graph API foram coletados 4.095.919 comentários de 15.825 postagens. A Tabela 2 mostra algumas estatísticas agregadas da base de dados dividida em diferentes categorias[8]. Pode-se observar, considerando a popularidade na divisão dos grupos, que mídias têm um público maior do que as figuras públicas em todas as métricas de popularidade coletadas. Na proporção de comentários tóxicos, a mídia também tem uma presença maior de mensagens tóxicas. Na próxima seção principal, mais detalhes sobre essa diferença serão apresentados. Essa tendência entre mídia e figuras públicas também ocorre com a toxicidade média de um comentário. Mas a toxicidade média dos comentários por publicação possui uma maior variação nas postagens de mídias do que nas postagens de figuras públicas, considerando as métricas de estabilidade da toxicidade. Diferentemente, a proporção de postagens tóxicas é 1,48% para todas as figuras públicas e 0,74% para ambas as mídias, como visto na Tabela 2, com a mídia tradicional tendo a menor proporção.

3.5 Potenciais Limitações

Existem algumas limitações nos dados coletados que são discutidas a seguir.

Acurácia dos modelos de toxicidade da Perspective API para Português. Medir a toxicidade de um texto ainda é um tópico de pesquisa em desenvolvimento. A Perspective API representa uma das primeiras "ferramentas de prateleira" disponíveis, e não há estudos atuais sobre sua precisão na língua portuguesa. Portanto, para estimar sua precisão, foi rotulada manualmente uma amostra dos dados e medido o coeficiente de concordância Kappa de Cohen [5] entre os rótulos, além da concordância com a Perspective API. Dois voluntários rotularam como tóxicos ou não tóxicos 2.000 comentários selecionados aleatoriamente, atingindo um coeficiente kappa de 0, 45, com um intervalo de confiança entre 0, 36 e 0, 55. Os voluntários então discutiram o conteúdo e chegaram a um veredicto final sobre os rótulos, que foi então comparada com a classificação dada pelo modelo da Perspective API usando o melhor ponto de corte encontrado, de 0, 8. O kappa entre o rótulo humano e a API foi 0, 44, com um intervalo de confiança entre 0, 36 e 0, 52. Da mesma forma, traduzir o texto e, em seguida, usar a Perspective API em inglês, retorna um kappa de 0, 43, com um intervalo de confiança entre 0, 34 e 0, 52, o que torna o uso da versão em português melhor do que o que é disponível em inglês, neste caso. Esses resultados também mostram que medir a toxicidade de textos é um teste difícil e o Perspective API pode ser tão bom quanto um humano. Um futuro benchmark entre a API e alternativas em português pode mostrar a melhor solução, porém isso extrapola o escopo deste trabalho.

Detecção de comentários tóxicos pelo próprio Facebook. Outra limitação do conjunto de dados está relacionada às remoções de comentários feitas pelo próprio Facebook a fim de amenizar o ambiente tóxico que possa ser criado a partir destes comentários. Como a coleta de todos os dados demorou alguns dias, os dados podem não representar a situação dos comentários tal como foram publicados, considerando que o Facebook pode ter excluído parte das mensagens mais tóxicas. Depois de medir quantos comentários foram apagados dois meses após nossa pesquisa inicial em uma amostra aleatória, descobrimos que cerca de 1% deles foram excluídos. Coletar os comentários das postagens assim que elas são publicadas seria a única forma de garantir obter todos os comentários, todavia, isso se torna inviável dependendo da escala, como neste trabalho.

Mesmo com essas limitações, esta base de dados pode fornecer uma visão interessante sobre a toxicidade nos comentários do Facebook. Na seção seguinte, serão apresentados e discutidos os principais resultados da caracterização da toxicidade.

3.6 Metodologia da análise dos dados

Para analisar os dados neste trabalho foram utilizadas as médias e desvios padrões da toxicidade dentro das categorias estabelecidas

[8] A tabela completa por página e outros materiais adicionais estão disponíveis em https://homepages.dcc.ufmg.br/~samuel.guimaraes/WebMedia2020

na seção 3.1, com as Funções de Distribuição Aculumada (Cummulative Distribution Function, em inglês, abreviada como CDF) sendo analisadas através do teste U de Mann-Whitney, que compara a probabilidade de pontos de uma curva serem maiores que pontos em outra curva. Em especial para avaliar o impacto do evento analisado, também são utilizados o teste qui-quadrado de Pearson, bem como calcular a razão de possibilidades dos dados.

4 ANÁLISE DA TOXICIDADE

Nesta seção, será analisado o nível de toxicidade de comentários e postagens para verificar até que ponto a toxicidade está correlacionada com diferentes categorias e inclinações ideológicas.

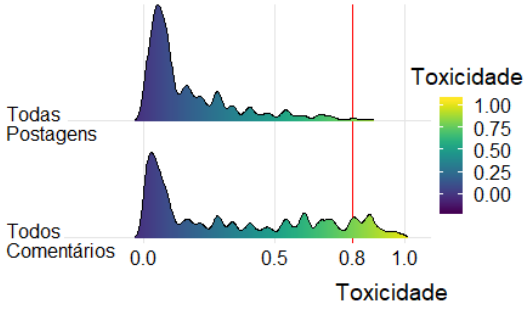


Figura 1: Distribuição de toxicidade para todos os comentários e postagens.

A Figura 1 mostra a distribuição das pontuações de toxicidade para todos os comentários e postagens. Neste trabalho, um comentário ou postagem foram considerados como sendo tóxicos se a toxicidade for superior a 0,8, seguindo o ponto de corte validado e o trabalho anterior de ElSherief et al [8]. Usando esse limiar, foram considerados tóxicos 13,27% dos comentários e 0,84% das postagens. Embora essas porcentagens de comentários e postagens tóxicas possam parecer baixas, o nível de toxicidade varia entre as páginas por um fator de dois a dez em comparação com os valores médios. A Figura 2 mostra a CDF da mesma distribuição da Figura 1. Pode-se observar que 50% de todas as postagens tem menos de 0.1 de toxicidade, e que na faixa entre 0.1 e 0.9 de toxicidade, a distribuição dos comentários está mais enviesada para maiores valores que as postagens. Ao verificar o quanto essa distribuição varia entre páginas, foi descoberto que 20% de todas as páginas são responsáveis por 60% dos comentários tóxicos e 56% de todas as postagens tóxicas. Um grande número de comentários tóxicos em uma página pode indicar que um grupo de usuários irritados está atacando a página ou que há uma briga ocorrendo entre os usuários, criada por discussão na seção de comentários. Mas isso também pode indicar que uma página está incitando esses comentários. Para testar isso foi usado o teste qui-quadrado de Pearson entre as postagens tóxicas e as postagens com a proporção de comentários tóxicos acima da média dos dados. Dessa forma, foi demonstrado que **a toxicidade da postagem e a proporção dos comentários tóxicos possuem dependência** (χ^2 = 36,867, $p < 0{,}005$). Além disso, usando o teste U de Mann-Whitney encontramos que **as repostas a um comentário tendem a ser mais tóxicas** (U > $3.6x10^{11}$, $p < 0{,}005$). Na Figura 3 mais detalhes dessa comparação mostram que por mais que no ponto de corte haja pouca diferença na porcentagem, como o teste mostra, existe uma maior probabilidade da toxicidade das respostas serem maiores. Essas análises já respondem alguns fatores que influenciam a proporção de comentários tóxicos, que é a pergunta feita pela **segunda questão de pesquisa** (P2).

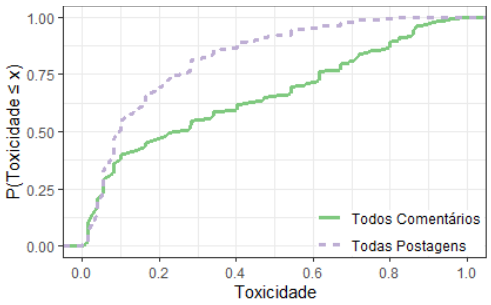


Figura 2: Função de Distribuição Acumulada (CDF) da quantidade de comentários e postagens tóxicas.

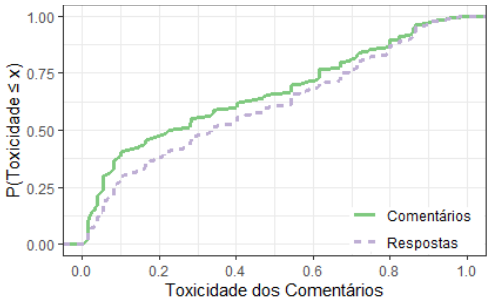


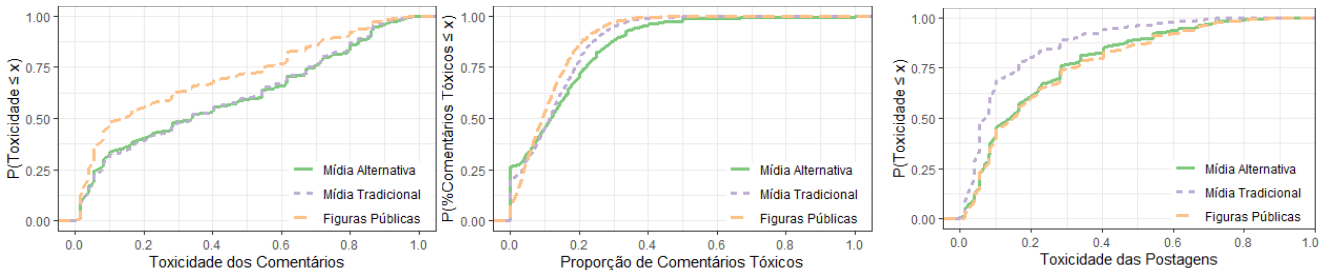
Figura 3: Função de Distribuição Acumulada (CDF) da toxicidade dos comentários em comparação com suas respostas.

4.1 Toxicidade em Páginas Brasileiras

Nesta seção, é examinado como várias características das páginas estão relacionadas ao nível de toxicidade nas seções de comentários, respondendo a **primeira questão de pequisa** (P1). Para analisar a toxicidade de várias páginas, inicialmente foi calculado as proporções das mensagens tóxicas em cada uma das postagens.

Primeiramente, a tendência política de figuras públicas (esquerda, centro ou direita) e o tipo de mídia (tradicional ou alternativa) das páginas foram analisadas quanto ao nível de toxicidade nos comentários. Para cada subgrupo, agregaram-se todos os comentários das páginas correspondentes. Observa-se pela Tabela 2 que a média da toxicidade dos comentários para as figuras públicas (0,2886) é ligeiramente menor do que para mídia (0,3906), mas que a posição política das figuras públicas não afeta significativamente tal média.

A Figura 4a ilustra essa diferença, mostrando que isso se estende a toda a distribuição, com a toxicidade sendo estatisticamente menor para a figuras públicas, com significância segundo o teste U de Mann-Whitney (U > $3.52x10^{12}$, $p < 0{,}005$). A Figura 4b semelhantemente mostra que em termos de proporção de comentários tóxicos (ou seja, comentários cuja toxicidade é superior a 0,8), as médias das



(a) Toxicidade de comentários **(b) Comentários tóxicos.** **(c) Toxicidade de postagens.**

Figura 4: Função de Distribuição Acumulada (CDF) da toxicidade em comentários e postagens por tipo de página.

proporções de comentários tóxicos em publicações de figuras públicas são menores que de publicações na mídia com médias de 9,58% e 16,53%, respectivamente. Usando o teste U de Mann-Whitney para avaliar a diferença nessas proporções, foi visto um efeito significativo no tipo de páginas ($U > 1.17x10^7$, $p < 0{,}005$), mostrando que as **publicações na mídia recebem estatisticamente mais toxicidade do que publicações de figuras públicas**.

Uma possível razão é que as pessoas podem considerar as páginas de figuras públicas como espaços homogêneos de discussão política onde apoia-se a figura, enquanto consideram as páginas da mídia como espaços de discussão onde pessoas com várias inclinações políticas se reúnem [2]. No primeiro caso, quando se fala com pessoas que pensam da mesma maneira, pode haver menos toxicidade, ou segundo ambiente heterogêneo poderia desencadear mais toxicidade. Um sinal de que o ambiente em si pode ser responsável pela diferença pode ser visto na Figura 4c. Nela, é visto que a mídia tradicional possui uma distribuição da toxicidade das postagens com menor toxicidade, mostrando que a proporção maior de comentários tóxicos não vem de muitas postagens com texto tóxico. Dentre as páginas da mídia, todas exceto uma têm mais de 10% de comentários tóxicos, enquanto que metade das páginas de figuras públicas não apresenta tal nível de toxicidade em seus comentários. As páginas de Lula, Michel Temer e Jair Bolsonaro, que foram presidentes do Brasil, são a primeira, segunda, e quinta páginas de proporções mais baixas de comentários tóxicos.

Como Lula e Bolsonaro são figuras altamente polarizantes, isso é inesperado. Além da possibilidade da homogeneidade política, outra possível explicação é que, como políticos, suas contas podem ser mantidas por profissionais. Como a presença online é extremamente importante para eles, seus assistentes podem tentar remover comentários tóxicos. Em contraste, certas mídias podem não ter o mesmo esforço e, em alguns casos, podem até se beneficiar se as pessoas visitarem a página e brigarem, gerando audiência. Portanto, seu esforço para moderação do conteúdo pode ser menor.

4.2 Toxicidade em Postagens

Em seguida, a proporção de comentários tóxicos foi examinada, selecionando as piores mensagens e também analisando se elas são majoritariamente políticas. Para filtrar as mensagens com poucos comentários, calculou-se o número médio de comentários, ignorando as mensagens abaixo da média, uma vez que podem ter apenas um comentário tóxico, e essa proporção de 100% seria irrelevante. Os resultados são apresentados na Tabela 4.

Ao considerar postagens com um número de comentários acima da média, páginas de figuras de direita são mais prevalentes. Isso pode indicar que enquanto a esquerda política pode receber mensagens tóxicas, páginas de inclinação à direita atraem maiores audiências e maiores quantidades de comentários tóxicos. Contabilizando o conteúdo, pode-se confirmar a prevalência da política como um tópico. Mas, como o maior evento do período usado foi política, não é possível uma generalização dos resultados. Entre as postagens com mais comentários tóxicos apenas duas foram criadas por figuras públicas, escritas pelo Deputado *Eduardo Bolsonaro*, e pelo vereador *Carlos Bolsonaro*, ambos filhos do atual presidente. A página de Eduardo Bolsonaro também aparece na próxima análise.

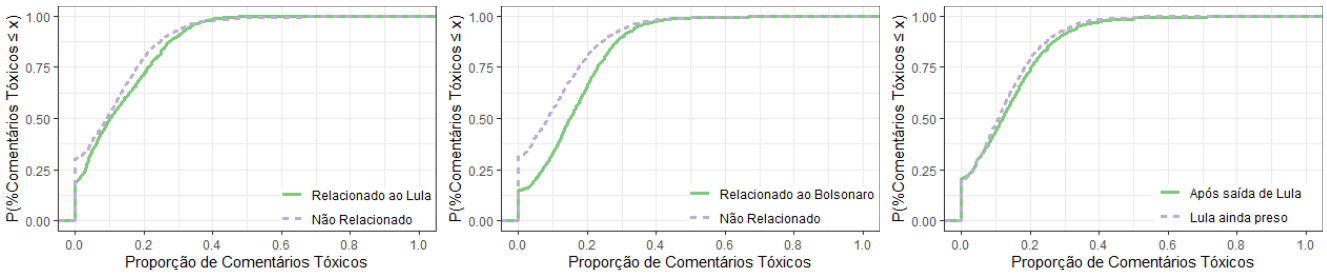
Tabela 4: As dez postagens acima da média em número de comentários com maior proporção de comentários tóxicos.

Página	Conteúdo	Toxicidade Post.	Coment. Total	% Tóxica
Jornal da Cidade Online	Cusparada em Bolsonaro	0,8652	1730	49,94%
Jornal da Cidade Online	Política e atentado ao Bolsonaro	0,8004	1505	47,97%
Jornal da Cidade Online	Gasto de dinheiro público	0,8641	3238	47,25%
O Antagonista	Frase de Lula	0,2841	1844	46,31%
Eduardo Bolsonaro	Retorno do PT	0,4028	6219	45,57%
Jornal da Cidade Online	Briga política	0,8224	1257	44,87%
Diário do Brasil	Gasto de dinheiro público	0,1620	866	44,46%
Jornal da Cidade Online	Frase de Lula	0,1342	1042	44,34%
Caneta Desesquerdizadora	Política e ambientalismo	0,8004	2720	44,12%
Carlos Bolsonaro	Frase de Lula	0,7899	1666	43,88%

4.3 Toxicidade nos Comentários

Aprofundando a análise da relação entre o que faz um usuário publicar comentários tóxicos, todos os comentários cuja pontuação de toxicidade é 1 (a pontuação máxima) foram analisados. Primeiro, verificou-se de onde esses comentários vieram dentre os cinco subgrupos (Figura 6). E, dessa forma, nota-se que as figuras de direita têm a maior proporção dos comentários mais tóxicos, e as mídias alternativas têm um número ligeiramente maior destes comentários do que a mídia tradicional.

Dois padrões interessantes são encontrados por exame manual. Em primeiro lugar, como dito anteriormente, Jair Bolsonaro tem



(a) Conteúdo relacionado a Lula. **(b) Conteúdo relacionado a Bolsonaro.** **(c) Conteúdo antes e depois da soltura de Lula**

Figura 5: Função de Distribuição Acumulada (CDF) de comentários tóxicos em relação à Lula e Jair Bolsonaro.

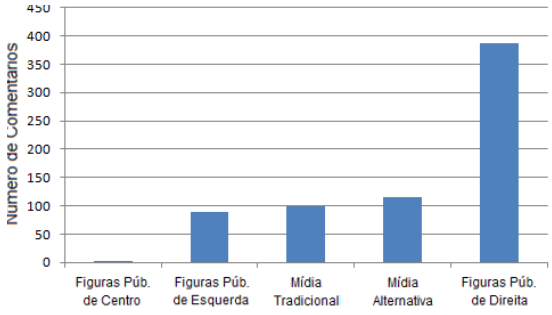


Figura 6: Distribuição dos comentários com toxicidade igual a 1 por página.

uma baixa proporção de comentários tóxicos, mas ele ainda recebe muitas mensagens extremamente tóxicas (ou seja, de toxicidade igual a 1). Isso pode implicar que ele não remove ativamente os comentários tóxicos, mas devido ao seu grande público, a proporção é relativamente baixa. Ou, esses comentários podem não ter como alvo o Bolsonaro e, portanto, não são sinalizados. Em segundo lugar, também descobrimos que *Eduardo Bolsonaro* e o *Jornal da Cidade Online*, uma página de mídia alternativa, tiveram mais comentários com toxicidade máxima do que os recebidos pelo Presidente. A página de Eduardo tem um quarto do total de mensagens de seu pai, mas 3 vezes mais comentários tóxicos, mostrando que é possível que mais seguidores concordando com o atual presidente diminua seu percentual.

O conteúdo dos dez piores comentários com toxicidade menor que 1 também foi avaliado, e é mostrado na Tabela 5. Nota-se que o tópico político é onipresente, com Bolsonaro e Lula sendo tópicos em sete dos dez comentários. Isso novamente contrasta o fato de que a página de ambos tem pouquíssimos comentários tóxicos. Por esse motivo, ele foi procurado como um tópico nas postagens tóxicas através de seu nome completo e apelido no texto da postagem. 8,24% das postagens de todas as páginas são sobre ele. A figura 5a mostra que a distribuição da proporção de comentários tóxicos para postagens sobre Lula é um pouco maior, com uma média de 13,07% contra uma média de 11,35% para postagens não relacionadas. Isso indica que as pessoas deixam **comentários tóxicos com muito mais frequência em postagens sobre ele**, e ele se torna um tópico que atrai muito ódio. A figura 5b mostra que Bolsonaro tem um efeito semelhante. Com 9,66% de todas as postagens o citando, a proporção média de comentários tóxicos nas postagens que o mencionam é de 15,92%, o que é superior a 11,02% das postagens sem menções. Assim encontramos algumas características de comentários tóxicos, respondendo a **terceira questão de pesquisa** (P3), sendo possível notar que **política é um assunto comum**, que apesar disso **comentários sobre figuras públicas tendem a ser mais tóxicas**, e tais mensagens não costumam ser enviadas nas páginas destas figuras, mas em **páginas de notícias falando sobre eles**.

4.3.1 Como a saída de Lula da prisão afetou os comentários. Por fim, busca-se voltar a analisar a **segunda questão**, verificando como um incidente político pode alterar a proporção de comentários tóxicos. Como citado anteriormente, esta coleta de dados é um recurso valioso para ver como as discussões online foram feitas em torno da libertação de Lula. A partir destes dados, foi comparada a toxicidade das postagens de uma semana antes e uma semana depois que sua saída foi relatada como eminente, dia 07 de novembro. A Figura 5c apresenta os resultados. Verificou-se então que as postagens tinham, em média, 12,32% de comentários tóxicos antes do evento, e que aumentaram para 13,60%, após a decisão. Mas ao fazer o teste qui-quadrado de Pearson, bem como calcular a razão de possibilidades dos dados descobriu-se que não foi encontrada uma correlação significativa. Isso mostra que um único incidente político pode alterar a toxicidade dos comentários, mas **não necessariamente de maneira significativa**.

5 CONCLUSÃO

Neste artigo, uma caracterização detalhada da toxicidade de 4.095.919 comentários online em 15.825 postagens de 63 páginas do Facebook, coletadas entre 27 de outubro e 16 de novembro de 2019, foi apresentada. Após dividir as páginas em Figuras Públicas, Mídias tradicionais, ou alternativas, sua toxicidade foi investigada. Esses resultados revelaram uma série de tendências interessantes. Verificou-se que as postagens que possuem notícias com textos tóxicos e os comentários tóxicos são ambos minoria, sendo 0,84% e 13,27% do total, respectivamente. Mas esse resultado é concentrado, com cerca de 20% das páginas responsáveis por cerca de 60% dos comentários tóxicos e 56% de todas as postagens tóxicas. Em geral, comentários que respondem a outros comentários são mais tóxicos, e há mensagens mais tóxicas comentadas em páginas com publicações mais tóxicas, mesmo que a toxicidade da postagem e dos comentários não esteja diretamente correlacionada. Com a subdivisão utilizada, as páginas de mídia recebem mais toxicidade

Tabela 5: Os dez comentários mais tóxicos dentre o conjunto de dados com toxicidade abaixo de 1.

Página	Commentário	Toxicidade
Magno Malta	Bandido FDP*ta! Pastor facista ! Vai te f*der !	0,99998
Jair Messias Bolsonaro	Bial babaca, você só dá fora babacão, bial trouxa comunista de merda	0,99998
Jornal da Cidade Online	Vossa excelência hiena filho da p*ta vagabundo bandido lixo rato de esgoto	0,99996
Lula	Velho maldito gente maldita quero que vocês tudo se f*dam	0,99996
Ivan Valente	Morra logo e vá para o inferno, velhinho safado, comunista asqueroso!!!	0,99995
O Globo	seu ridiculo nojento péste va pro inferno canalha!!!!!	0,99995
Jair Messias Bolsonaro	F*da se seu bosta presidente de merda vai faze alguma coisa pelo país seu merda só sabe aparece na TV	0,99994
Jornal da Cidade Online	Vai cantar no banheiro seu idiota, te f*de pra deixar de defender bandido	0,99994
O Globo	Esse imbecil so fala merda.	0,99993
Eduardo Bolsonaro	Esse imbecil só fala merda.	0,99993

do que as páginas de figuras públicas, mas quando figuras públicas são citadas em uma postagem, a proporção de comentários tóxicos aumenta. A afiliação política das figuras públicas não afetou a proporção de comentários ou postagens tóxicas em si, mas as mídias tradicionais recebem mais comentários tóxicos do que as alternativas. Por fim, constatou-se que a saída de Lula da prisão mostra que um determinado evento político pode aumentar momentaneamente a toxicidade, mas o efeito pode não ser significativo.

Com esses resultados espera-se que os principais fatores que influenciam os comentários tóxicos associados às notícias possam ajudar plataformas de redes sociais no design de políticas de conteúdo capazes de minimizar esse problema, especialmente no Brasil. Em especial, a metodologia apresentada pode facilmente ser replicada em outros países, já que a divisão das páginas nas cinco subcategorias apresentadas e o uso do *Audience Insights* não dependem do país analisado ser o Brasil.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo projeto Capacidades Analíticas do Ministério Público de Minas Gerais e por CNPq, CAPES e Fapemig.

REFERENCES

[1] Thais G Almeida, Bruno À Souza, Fabíola G Nakamura, and Eduardo F Nakamura. 2017. Detecting hate, offensive, and regular speech in short comments. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*. 225–228.

[2] Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 68–79.

[3] Samuel Barros and Rodrigo Carreiro. 2015. A discussão pública e as redes sociais online: o comentário de notícias no Facebook. *Fronteiras-estudos midiáticos* 17, 2 (2015), 174–185.

[4] J Clement. 2019. Number of social network users worldwide from 2010 to 2021 (in billions). (2019).

[5] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[6] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[7] Rogers Prates de Pelle and Viviane P Moreira. 2017. Offensive Comments in the Brazilian Web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

[8] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.

[9] Klint Finley. 2015. A brief history of the end of the comments. *Wired, October* 8 (2015).

[10] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online Hate Speech*. UNESCO.

[11] Jan Hanzelka and Ina Schmidt. 2017. Dynamics of cyber hate in social media: A comparative analysis of anti-Muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology* 11, 1 (2017), 143–160.

[12] Rachel L. Harris and Lisa Tarchak. 2019. Yes, Our Writers Do Read Your Comments. (December 2019). https://www.nytimes.com/2019/12/30/opinion/nytimes-columnists-comments.html The New York Times.

[13] Sanne Hille and Piet Bakker. 2014. Engaging the social news user: Comments on news sites and Facebook. *Journalism Practice* 8, 5 (2014), 563–572.

[14] Sabih Ahmad Khan and Hsien-Tsung Chang. 2019. Comparative analysis on Facebook post interaction using DNN, ELM and LSTM. *PloS one* 14, 11 (2019).

[15] Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*. 11–17.

[16] Ioanna K Lekea and Panagiotis Karampelas. 2018. Detecting Hate Speech Within the Terrorist Argument: A Greek Case. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 1084–1091.

[17] Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 515–522.

[18] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, 85–94.

[19] Marcio Moretto and Pablo Ortellado. 2018. *Quanto mais velhos, mais polarizados: Perfil dos usuários que interagem com páginas de notícias no Facebook*. Technical Report 1. Monitor do Debate Político no Meio Digital.

[20] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Nielsen. 2019. *Reuters Institute digital news report 2019*. Vol. 2019. Reuters Institute for the Study of Journalism.

[21] Rogers Pelle, Cleber Alcântara, and Viviane P Moreira. 2018. A classifier ensemble for offensive text detection. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. 237–243.

[22] Julio Reis, Fabricio Benevenuto, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the News: First Impressions Matter on Online News. In *ICWSM*. Oxford, UK.

[23] Filipe N Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth International AAAI Conference on Web and Social Media*.

[24] Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2019. Online hate ratings vary by extremes: a statistical analysis. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 213–217.

[25] Elisa Shearer. 2018. Social media outpaces print newspapers in the U.S. as a news source. https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/. (December 2018). Pew Research Center.

[26] Elisa Shearer and Katerina Eva Matsa. 2018. News Use Across Social Media Platforms 2018. https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/. (September 2018).

[27] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media.. In *ICWSM*. 687–690.

[28] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. 2016. News commenters and news comment readers. *Microsoft Word-Egaging News Projet* (2016), 1–21.

[29] Leona Yi-Fan Su, Michael A Xenos, Kathleen M Rose, Christopher Wirz, Dietram A Scheufele, and Dominique Brossard. 2018. Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society* 20, 10 (2018), 3678–3699.

[30] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899* (2017).

[31] Brian E Weeks and R Lance Holbert. 2013. Predicting dissemination of news content in social media: A focus on reception, friending, and partisanship. *Journalism & Mass Communication Quarterly* 90, 2 (2013), 212–232.