RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

A gestão do discurso de ódio nas plataformas de redes sociais digitais: um comparativo entre Facebook, Twitter e Youtube

Luiz Rogério Lopes Silva
Universidade Federal do Paraná, Doutorado em Gestão da Informação, Curitiba, PR, Brasil
luizlopescomunicacao@gmail.com

Rodrigo Eduardo Botelho-Francisco
Universidade Federal do Paraná, Departamento de Ciência e Gestão da Informação, Curitiba, PR, Brasil
rodrigobotelho@ufpr.br

Alisson Augusto de Oliveira
Universidade Federal do Paraná, Graduação em Gestão da Informação, Curitiba, PR, Brasil
alissonaug@icloud.com

Vinicius Ramos Pontes
Universidade Federal do Paraná, Graduação em Gestão da Informação, Curitiba, PR, Brasil
ramosvinicius83@gmail.com

ARTIGOS

DOI: https://doi.org/10.26512/rici.v12.n2.2019.22025

Recebido/Recibido/Received: 2018-12-12
Aceitado/Aceptado/Accepted: 2019-01-21

Resumo: O impasse social sobre discurso de ódio e liberdade de expressão na Internet tem impulsionado os Sites de Redes Sociais (SRS) a intensificarem suas políticas de moderação de conteúdo. A gestão do conteúdo de ódio em plataformas como Facebook, Twitter e Youtube é tão complexa, haja vista seu caráter multifacetado e o grande número de interagentes, que executivos destas empresas assumem a ineficiência de seus recursos (humanos e tecnológicos) na tentativa de controlar o escalonamento, duração, difusão e circunspecção de crimes e discursos odientos. O problema tem despertado a atenção de governos e organizações civis, que por sua vez aumentam a pressão sobre as plataformas no intuito de melhorarem suas escolhas editoriais e sua logística de monitoramento e remoção deste tipo de interação. Neste contexto, este trabalho tem como objetivo comparar as ações realizadas pelo Facebook, Twitter e Youtube no que tange a formulação e ampliação de políticas e decisões sobre conteúdo individual de ódio. Para isto, foi realizado uma análise histórica (2015-2018) de dados secundários das políticas e termos de comunidade específicos de cada SRS numa análise das tomadas de decisão com os cinco tópicos do termo de compromisso que as empresas assumiram com a Liga Anti-Difamação, em 2013, no combate ao discurso de ódio online. Os resultados apontam o Facebook como o SRS que mais investiu em estratégias de combate a intolerância e incivilidade online, apesar da empresa não deixar claro os métodos empregados para tal fim. De modo geral, todas as plataformas evoluíram na estrutura operacional de denúncia de conteúdo odiento, mas se mostraram ineficientes em moderação, remoção e contenção da propagação do *cyberhate.*

Palavras-chave: Discurso de ódio. Facebook. Redes Sociais. Twitter. Youtube.

The management of hate speech on the platforms of digital social networks: a comparison between Facebook, Twitter and YouTube

Abstract: The social stalemate over hate speech and free speech on the Internet has pushed Social Networking Sites (SNS) to intensify their content moderation policies. The management of hateful content on platforms such as Facebook, Twitter and Youtube is so complex, given its multifaceted character and the large number of interactors, that executives of these companies assume the inefficiency of their resources (human and technological) in an attempt to control the scheduling, duration, diffusion, and circumspection of crimes and hate speeches. The problem has awakened the attention of governments and civil organizations, which in turn increase the pressure on the platforms to improve their editorial choices and their logistics of monitoring and removing this type of interaction. In this context, this work aims to compare the actions carried out by Facebook, Twitter, and YouTube in the formulation and expansion of policies and decisions on individual hateful content. For this, a historical analysis (2015-2018) of secondary data of the specific policies and community terms of each SRS was carried out in a review of the decision-making with the five topics of the commitment term that the companies assumed with the Anti-Defamation League, in 2013, in the fight against hate speech online. The results point to Facebook as the SRS that most invested in strategies to combat intolerance and incivility online, although the company did not make clear the methods used for this purpose. Overall, all platforms evolved in the operational structure of denouncing cyberhate but were inefficient in moderation, removal, and containment of cyberhate propagation.

Keywords: Hate speech. Facebook. Social Network. Twitter. YouTube.

La gestión del discurso de odio en las plataformas de redes sociales digitales: un comparativo entre Facebook, Twitter y Youtube

Resumen: El impasse social sobre discurso de odio y libertad de expresión en Internet ha impulsado a los Sitios de Redes Sociales (SRS) a intensificar sus políticas de moderación de contenido. La gestión del contenido de odio en plataformas como Facebook, Twitter y Youtube implica un carácter multifacético y un gran número de interagentes. Además, es tan compleja que ejecutivos de estas empresas asumen la ineficiencia de sus recursos (humanos y tecnológicos) en el intento de controlar el escalonamiento, duración, difusión y circunspección de crímenes y discursos odientos. El problema ha despertado la atención de gobiernos y organizaciones civiles, que a su vez aumentan la presión sobre las plataformas con el fin de mejorar sus elecciones editoriales y su logística de monitoreo y remoción de este tipo de interacción. En este contexto, este trabajo tiene como objetivo comparar las acciones realizadas por Facebook, Twitter e Youtube en lo que se refiere a la formulación y ampliación de políticas y decisiones sobre contenido individual de odio. En ese sentido, se realizó un análisis histórico (2015-2018) de datos secundarios de las políticas y términos de comunidad específicos de cada SRS en un análisis de las tomas de decisión con los cincotópicos del término de compromiso que las empresas asumieron con la Liga Anti-Difamación, en 2013, en el combate al discurso de odio online. Los resultados apuntan a Facebook como el SRS que más ha invertido en estrategias de combate a la intolerancia e incivilidad en línea, a pesar de que la empresa no deja claro los métodos empleados para tal fin. En general, todas las plataformas evolucionaron en la estructura operativa de denuncia de contenido odiante, pero se mostraron ineficientes en moderación, remoción y contención de la propagación del cyberhate.

Palabras-clave: Discurso de odio. Facebook. Redes sociales. Twitter. Youtube.

1 Introdução

Os *Sites* de Redes Sociais (SRS) são exemplos de espaços na Internet onde diferentes grupos de interesses reproduzem atitudes e condutas de diferentes naturezas, entre elas o ódio. Nestes espaços, os contextos normativos e as ordenações discursivas não são evidentes (RECUERO, 2014) e as ideologias podem ser interpretadas de diferentes modos, produzindo efeitos distintos entre os que detiveram a mesma informação. Sobre isso, estudos apontam que SRS como o Facebook, Twitter e Youtube desempenham um papel instrumental na

propagação do ódio e na tradução do discurso em ação (BULINGE, 2014; COHEN-ALMAGOR, 2009; KNOBEL, 2012; KUDLACEK *et al*. 2017; LI *et al.* 2017; ROST *et al.*, 2016; SCHÄFER, *et al.* 2015; TYNES, *et al.* 2013; XIANG *et al.* 2012). O conteúdo odiento externalizado nestes espaços, por sua vez, pode assustar, intimidar ou silenciar usuários da plataforma, sendo que alguns deles podem inspirar outros usuários a cometerem violência (SALEEM *et al*. 2017).

A literatura também sugere que os padrões da comunidade e a política de discurso de ódio dos SRS são orientados pela motivação administrativa e para rentabilizar interações (BEN-DAVID, MATAMOROS-FERNÁNDEZ, 2016; SHEPHERD *et al*. 2015) e que a estrutura física e de mão-de-obra das empresas não suportam a alta demanda de conteúdo a ser moderado. Ben-Davi e Matamoros (2016) apontam que os recursos tecnológicos e a lógica corporativa dos *Sites* de Redes Sociais interferem na dinâmica das performances de ódio e contribuem para a percepção da retórica do ódio como informação legítima. Para as autoras, a lógica do algoritmo não é neutra e pode discriminar de acordo com os interesses da empresa que administra o SRS. O ódio, portanto, não surge fortuitamente de eventuais discordâncias, mas é também o resultado inevitável do funcionamento das plataformas.

Bem-David e Matamoros-Fernandez (2016) também apontam que a lógica das plataformas ainda permite que pessoas ou grupos continuem alcançando novas audiências, recrutando novos membros e criando comunidades de ódio que possibilitam que a violência saia das telas e alcance as ruas. As ações que as empresas têm feito ainda é incipiente diante de uma demanda tão grande de denúncias. Visando atender esta demanda, a detecção automatizada tem sido utilizada por seu potencial de examinar, filtrar e categorizar grandes conjuntos de dados em um curto período de tempo e fornecer análises complexas que não podem ser feitas de maneira tradicional.

Por outro lado, Facebook, Twitter e Youtube têm declarado o fenômeno do ódio online como prioridade em suas políticas de conteúdo. Executivos das três empresas confirmam um aumento de rigor na remoção de conteúdo com discurso de ódio, sem ferir os princípios de liberdade de expressão. Sheryl Sandberg, diretora de operações do Facebook, afirma que *"não há espaço para ódio ou violência no Facebook"* e reforça que *"a rede social usa tecnologia como inteligência artificial para encontrar e remover propaganda terrorista, equipes com especialistas em contraterrorismo e revisores em todo o mundo para manter conteúdos extremistas fora da nossa plataforma"*. (FACEBOOK lança...)

Mark Zuckerberg, CEO do SRS, em testemunho na Suprema Corte Americana, reconheceu as limitações do *site* e se comprometeu de maneira pessoal a encontrar alternativas para o problema: *"o mundo se sente ansioso e dividido, e o Facebook tem muito trabalho a fazer - seja protegendo a nossa comunidade de abusos e ódio, defendendo a*

interferência de estados-nação ou assegurando que o tempo gasto no Facebook seja tempo bem gasto". Zuckerberg disse estar otimista com as ferramentas de Inteligência Artificial que identificam conteúdo linguístico de ódio, mas assumiu que *"até automatizarmos mais o processo, há um índice de erros maior do que eu gostaria"* (CNBC, 2018).

O Twitter, por sua vez, disponibilizou novos termos de segurança em 2018, acentuando a austeridade quanto a questão do discurso de ódio online: *"Não é permitido promover violência, ameaçar ou assediar outras pessoas com base em raça, etnia, nacionalidade, orientação sexual, sexo, identidade de gênero, religião, idade, deficiência ou doença grave"* (TWITTER, 2018). O SRS norte-americano também faz uso de Inteligência Artificial para remover *tweets* ofensivos ou notificar seu usuário sobre a exclusão da postagem. O CEO do Twitter, Jack Dorsey, em sua própria página, perguntou aos seus seguidores quais seriam as melhorias que a plataforma poderia fazer para o ano de 2017 (DORSEY). A maioria dos participantes pedia que o SRS fosse mais rigoroso quanto aos casos de assédio e que disponibilizasse mais ferramentas para que os usuários pudessem denunciar *tuítes* ofensivos. Além disso, pediam uma transparência em razão de como eram tratadas as infrações cometidas por usuários do *microblog*.

Figura 1. Tweet do CEO do Twitter Jack Dorsey em 01 de março de 2018

Fonte: Twitter @jack

Já para o Youtube, um dos SRS da Google, "*nem tudo que é maldoso ou ofensivo é considerado incitação ao ódio*" (YOUTUBE, Padrões..., 2018). A empresa reconhece que existe uma linha tênue entre discurso de ódio e liberdade de expressão, o que exige cautela na elaboração de políticas de segurança e na remoção de conteúdo. Em nota oficial, o SRS afirmou que tem expandido o trabalho contra abusos que ferem as diretrizes da comunidade. Desde junho de 2017, os moderadores revisaram quase 2 milhões de vídeos para conteúdo extremista violento, ajudando a treinar tecnologia de aprendizado de máquina para identificar vídeos

semelhantes no futuro (YOUTUBE, 2017). A combinação de moderadores e tecnologia na detecção de *spam*, nudez e discurso de ódio permitiu que a plataforma removesse mais de 150.000 vídeos por extremismo violento.

Para o Youtube, o conteúdo terrorista e discurso de ódio representam uma proporção muito pequena do conteúdo que viola as diretrizes de comunidade, cerca de 1% do que foi sinalizado em 2015, por exemplo. Mesmo assim, a empresa sofreu com o fim de anúncios publicitários de empresas como AT&T, Verizon e Johnson & Johnson, que descobriram que seus anúncios foram veiculados junto a vídeos que promoviam o terrorismo e a homofobia (GLOBO).

Apesar dos esforços em impor iniciativas de controle e cerceamento das práticas de discurso de ódio, os SRSs têm sido pressionados por governos e organizações de direitos humanos quanto a ineficiência de suas ações. Na Alemanha, por exemplo, o governo aprovou a lei *Network Enforcement Act*, conhecida como *NetzDG*, que exige que as corporações que administram os SRS excluam o conteúdo "claramente ilegal" (conteúdo difamatório, propaganda neonazista e chamadas à violência) de suas plataformas em até 24 horas, ou enfrentarão multas que podem chegar a € 50 milhões (US$ 60 milhões) (ALEMANHA, 2017). A lei aplica-se à maioria dos SRS na Alemanha, alegando que se tratam de empresas privadas, e é contrária à legislação dos EUA, onde os SRSs não são responsáveis pelo conteúdo de seus interagentes, compartilhado em seus serviços.

A *NetzDG* provocou mudanças emergenciais no gestão de discurso de ódio destes *sites*, ao mesmo tempo em que enfrenta dificuldades de aplicação e aceitação no contexto europeu. Ativistas dos direitos da Internet e políticos da oposição na Alemanha, contrários à nova lei, argumentam que a legislação dificultou a liberdade de expressão, deixando nas mãos das corporações dos EUA decidir o que é ou não discurso de ódio.

No Brasil o tratamento conferido à liberdade de expressão nas redes sociais digitais compete ao *Marco Civil da Internet* (MCI) desde 2014 (BRASIL, 2014). A lei tem por intuito preservar a livre expressão e evitar a censura na rede, garantindo que qualquer pessoa possa se expressar livremente *online*, a fim de promover equilíbrio entre as garantias constitucionais de proteção da liberdade de expressão e da intimidade, da honra e da imagem das pessoas. O MCI também assegura um ambiente aberto, democrático e livre, onde a remoção de conteúdo precisa passar por ordem judicial, justamente para evitar conflito entre liberdade de expressão e o direito à privacidade, sendo mais adequado que o julgamento seja do juiz e não dos provedores. Com isto, no caso do discurso de ódio *online* brasileiro, as plataformas podem remover conteúdo com base em seus padrões de comunidade, já que o usuário concordou

com os termos de uso, mas não podem remover conteúdos denunciados de seus usuários sem que haja uma ordem judicial específica.

Neste contexto geral e para os fins desta pesquisa, Facebook, Twitter e Youtube foram identificados como SRS que: (a) possuem alta demanda de conteúdo a ser moderado, (b) declararam publicamente sua preocupação com o discurso de ódio em suas plataformas e (c) assinaram o termo de compromisso da ADL Rede Anti-difamação numa cooperação à causa da tolerância e o combate ao *cyberhate*. A partir destas premissas, o objetivo é comparar as ações realizadas pelas empresas no que tange à formulação e ampliação de políticas e decisões sobre conteúdo individual de ódio. Para isto, foi realizada uma análise histórica (2015-2018) das políticas, termos de comunidade e avanço dos recursos tecnológico de cada SRS na moderação de conteúdo odiento. Para tanto fez-se uso das notícias oficiais publicadas no *blog* de cada uma das empresas para verificar o avanço das ações dentro da perspectiva dos cinco tópicos do termo de compromisso assumido com a ADL - Liga Anti-difamação, em 2013, no combate ao *cyberhate*.

O resultado desta pesquisa poderá ser conferido a partir dos tópicos abaixo. O texto está dividido em quatro partes: o acordo assinado pelos SRS com a Liga Anti-difamação, as políticas de conteúdo e remoção de conteúdo odiento dos SRS, a estrutura de coleta das informações e a descrição das ações desenvolvidos pelas empresas e, por fim, um comparativo entre os avanços do Facebook, Twitter e Youtube em relação aos termos do acordo.

2 O termo de acordo com a Liga Anti-Difamação

A Liga Anti-Difamação (ADL) é uma organização norte-americana sem fins lucrativos, não partidária e de direitos humanos do país. Dentre suas ações, destacam-se as iniciativas contra o anti-semitismo, a gestão de informações que auxiliam autoridades policiais e comunidades a se protegerem de ameaças extremistas de todo tipo, consultorias na elaboração de leis de defesa dos direitos humanos, promoção do respeito a comunidade israelense e o desenvolvimento de atividades nas escolas que combatem o preconceito, valorizam as diferenças e criam ambientes acolhedores.

A ADL recebe o apoio de parceiros e apoiadores públicos e privados para "impedir a difamação contra o povo judeu e garantir justiça e tratamento digno à todos" (ADL, 2018). Em 2013, a organização firmou um esforço colaborativo com Facebook, Google, Microsoft e Twitter, para resolver o problema do ódio *online* e, ao mesmo tempo, respeitar a liberdade de expressão. Representantes de algumas das principais corporações de tecnologia, líderes da sociedade civil, membros da comunidade jurídica e acadêmicos participaram de uma reunião da Força-Tarefa Inter-parlamentar de Coalizão pelo Combate ao Anti-Semitismo (ICCA) sobre o

ódio na Internet, na Universidade de Stanford. O encontro resultou no *Best Practices for Challenging Cyberhate (BPCC)*, um documento que recomenda diretrizes aos provedores de mídia social e à comunidade da Internet.

Dentre as 10 melhores práticas do BPCC, cinco são posturas que os SRS devem praticar: (1) considerar os relatos sobre o discurso de ódio online de forma comprometida, atento aos princípios fundamentais de liberdade de expressão, dignidade humana, segurança pessoal e respeito pelo estado de direito; (2) os provedores que apresentam conteúdo gerado pelo usuário devem oferecer uma explicação clara de sua abordagem para avaliar e resolver os relatórios de conteúdo odioso, destacando seus termos de serviço relevantes junto ao usuário; (3) oferecer mecanismos e procedimentos de fácil manuseio para denúncia de conteúdo odioso; (4) responder aos relatórios do usuário em tempo hábil; e (5) aplicar as sanções que seus termos de serviço contemplarem de maneira consistente e justa (CITRON, 2014).

As práticas aconselhadas à comunidade da Internet são (1) trabalhar em conjunto para lidar com as consequências prejudiciais do ódio *online*; (2) identificar, implementar e/ou encorajar estratégias eficazes de contra-fala - incluindo resposta direta, comédia e sátira quando apropriado ou simplesmente definindo o registro diretamente; (3) compartilhar conhecimento e ajudar a desenvolver materiais e programas educacionais que estimulem o pensamento crítico tanto na atividade *online* proativa quanto na reativa; (4) encorajar outras partes interessadas a ajudar a aumentar a conscientização sobre o problema dos *haters* e a necessidade urgente de abordá-los; e (5) acolher novas ideias e novas iniciativas para promover um ambiente online civil.

Cinco anos após o acordo, as três empresas afirmam ter aprimorado suas políticas de discurso de ódio, políticas de dados e padrões de comunidade, além de ter investido em recursos tecnológicos e na contratação de mais moderadores de conteúdo exclusivos para conter o fenômeno. Tanto Facebook quanto Twitter e Youtube publicaram em seus *sites* políticas de discurso de ódio *online* (FACEBOOK, 2018; TWITTER, 2018) e criaram procedimentos para "educar" algoritmos, que com o tempo repetem as respostas automaticamente, por meio de recursos avançados de identificação de imagens ou textos ofensivos.

Os recursos humanos empregados na moderação de possíveis conteúdo de ódio têm sido criticados pela Imprensa e por entidades que lidam com a problemática. Em entrevista à BBC Brasil, por exemplo, um brasileiro que trabalhou como moderador de discurso de ódio no Facebook afirmou que a rotina de trabalho é estressante, com metas que chegavam até 3.500 monitoramentos diários, o que representaria uma análise a cada oito segundos (SENRA, 2017). O ex-funcionário afirmou que os moderadores só visualizam o nome do autor das publicações

sem acesso a seus perfis completos, que a missão é apagar, ignorar ou encaminhar a publicação para a avaliação superior - o que ocorre especialmente em casos de suicídio ou pedofilia, que por sua vez são encaminhados a autoridades.

Verifica-se, com esta descrição, que os SRS mantém o apoio e a parceria com a ADL em ações de combate a intolerância e violência online. A Liga tem colaborado com as políticas de discurso de ódio dos sites e conquistado grandes feitos em seus próprios objetivos. É possível observar, neste caso, a presença da ADL como suporte estratégico e operacional das tomadas de decisão das empresas. As ações de cada um dos SRS, por sua vez, são descritas nas etapas seguintes deste artigo.

3 Política de discurso de ódio dos SRS

A compreensão do discurso de ódio pelos SRS não é consensual. Cada uma das empresas abordadas neste artigo trata o fenômeno a partir de um conjunto de documentos próprios, onde sinalizam regras ou práticas aceitas em suas comunidades. Facebook, Twitter e Youtube fazem a gestão de conteúdos odientos por meio de estratégias que envolvem moderados humanos e Inteligência Artificial. Apesar de desenvolverem ações colaborativas entre si, cada empresa trabalha com este tipo de informação a seu modo, exercendo o poder de controlar, bloquear, filtrar e remover qualquer conteúdo agressivo que viole seus termos.

O Facebook possui um documento chamado “Padrões da Comunidade” (FACEBOOK, *Padrões*), no qual lista o tipo de conteúdo que pode ser publicado e compartilhado na plataforma. Segundo o documento, estas políticas foram criadas para proteger os usuários e delimitar que tipo de conteúdo que pode ser removido por meio de denúncias. Em março de 2015, o documento foi atualizado e acrescentou-se o discurso de ódio na lista dos conteúdos passíveis de remoção. Os termos proíbem ataque direto a pessoas com base no que o SRS chama de características protegidas: raça, etnia, nacionalidade, filiação religiosa, orientação sexual, casta, sexo, gênero, identidade de gênero, *status* migratório e doença ou deficiência grave. Também oferece proteções para o *status* migratório.

O Twitter, por sua vez, disponibiliza o documento “Imposição de Nossas Regras” (TWITTER, 2018), onde lista o que é considerado ofensivo e passível de remoção ou suspensão de contas: (a) comportamento abusivo, (b) mídias íntimas, (c) conduta de propagação de ódio, (d) exaltação da violência, (e) contas afiliadas a grupos extremistas violentos, (f) *spam*, (g) promoção ou incentivo a suicídio e automutilação, (h) conteúdo sensível, (i) falsa identidade e (j) divulgação de informações privadas. A empresa afirma que seus termos *"englobam as tendências mais recentes de comportamento online, levando em considerações as diferenças culturais e os contextos sociais a definir expectativas a respeito do que é permitido na*

plataforma" (TWITTER, 2018). A política contra propagação do ódio inclui a proibição de conteúdo que promove a violência, ataque diretamente ou ameace outras pessoas com base em raça, etnia, nacionalidade, orientação sexual, sexo, identidade de gênero, religião, idade, deficiência ou doença grave.

O Youtube incentiva a liberdade de expressão e defende o direito de expressar pontos de vista divergentes, o que não significa que o SRS tolere conteúdo odiento. A organização reconhece que existe uma linha tênue entre o que é ou não é considerado discurso de ódio e em suas *Diretrizes de Comunidade* (YOUTUBE, 2018) não permite conteúdos que promovem a violência ou tenham como objetivo principal incitar o ódio contra indivíduos ou grupos, com base em raça ou etnia, religião, deficiência, sexo, idade, *status* de reservista militar, orientação/ identidade sexual.

As três empresas afirmam trabalhar de maneira intensa para estabelecer um ambiente seguro para livre expressão de seus interagentes, mas encontram dificuldades operacionais diante do desafio de atuar simultaneamente em diversas jurisdições, além do grande fluxo de dados e crescente volume de denúncias a serem analisadas. O Facebook, apesar de não deixar claro os seus mecanismos de classificação e remoção de discurso de ódio, promulga que o procedimento acontece em três níveis: *(1) interações de forma escrita ou visual que promova a violência contra grupos ou indivíduos contemplados pela política de discurso de ódio da empresa* - neste nível, os discursos degradantes que comparam pessoas ou grupos a animais ou sub-humanos, tratando-os de maneira inferior física ou mentalmente, também são passíveis de remoção, assim como debochar do conceito, de eventos ou de vítimas de crimes de ódio, mesmo que nenhuma pessoa real apareça na imagem; *(2) ataques que visam diminuir ou macular a integridade física, mental e moral de grupos e indivíduos; expressões de desprezo ou seu equivalente visual; expressões de repulsa ou seu equivalente visual e xingamentos a um indivíduo ou grupo de pessoas que compartilhem características protegidas*; e (3) *apelos pela exclusão ou segregação de um indivíduo ou grupo de pessoas com base nas características citadas na política de discurso de ódio, ou que descreva ou vise negativamente pessoas por meio de difamação* (FACEBOOK, 2018**).** O SRS reforça que o processo de remoção de conteúdo leva em consideração quão ativa é a ameaça, quão sofisticados são os atores, quanto mal está sendo feito e como a ameaça vai impactar de maneira global.

O Twitter, por sua vez, também remove conteúdos de ódio e faz uso de uma combinação de ações para "punir" os usuários que violam as regras da plataforma. O SRS trabalha com uma lista chamada "medidas corretivas", o que envolve a análise de *tweets*, mensagens diretas e contas, aplicando restrições tais como: (1) alteração da conta para o modo somente-leitura (a pessoa só poderá ler o conteúdo de sua timeline e enviar mensagens diretas

aos seus seguidores); (2) verificação de propriedade da conta usando um número de telefone ou endereço de *e-mail*; e (3) suspensão permanente da conta do infrator, que passa a não ter mais permissão para criar novas contas.

Quanto aos *tweets*, as providências de remoção restringem-se à visibilidade do *tweet* nos resultados de busca, nas respostas e nas timelines. Uma notificação é enviada por *e-mail* identificando o *tweet* e quais políticas foram violadas. O interagente-violador pode excluir o *tweet* ou recorrer à análise se acreditar que a moderação é um equívoco. Já em relação a mensagens diretas, a plataforma interrompe as conversas entre o infrator denunciado e a conta do denunciante. A conversa também é removida da caixa de entrada do denunciante, exceto em casos em que a própria vítima opte por continuar enviando mensagens diretas ao infrator.

Além disso, se o perfil ou conteúdo de mídia de uma conta não estiver em conformidade com as políticas do Twitter, ele pode ser suspenso temporariamente até que o infrator edite a mídia ou as informações em seu perfil em conformidade com a solicitação. Os violadores podem recorrer das medidas corretivas por meio da interface da plataforma ou registrando uma denúncia. Caso seja concluído que a suspensão é válida, a empresa responde à contestação com as informações sobre a política que foi violada pela conta. Caso contrário, a conta é reativada.

Por fim, o Youtube oferece recursos de sinalização de conteúdo inapropriado, estejam eles na forma de vídeo, miniatura, comentário, mensagem de bate-papo ou canal. A "sinalização de conteúdo" é aconselhada para denunciar uma possível violação às diretrizes da comunidade, embora o SRS pontue que *"nem tudo que é maldoso ou ofensivo é considerado incitação ao ódio*" (YOUTUBE,2018) e aconselha que os interagentes bloqueiem as pessoas e/ou conteúdos que de alguma maneira provoque um desconforto.

A plataforma de vídeos também dispõe do Programa de Revisor Confiável - uma ferramenta de sinalização em massa para denunciar diversos vídeos ao mesmo tempo, além de oferecer suporte a um fórum privado para tirar dúvidas sobre o processo de aplicação de políticas, transparência em decisões sobre conteúdo sinalizado e análises prioritárias de sinalizações para ação mais rápida.

Na interface do Youtube também há possibilidade de denunciar vários vídeos, comentários ou até mesmo toda a conta de um usuário de uma só vez. Com a "ferramenta de denúncia" é possível detalhar a violação dentro das seguintes categorias: assédio e *bullying* virtual, falsificação de identidade, ameaças violentas, risco para crianças, incitação ao ódio contra uma minoria, *spams* e golpes.

4 Análise histórica e documental do discurso de ódio dos SRS

A fim de realizar uma análise histórica e documental dos procedimentos estratégicos em relação ao discurso de ódio das empresas observadas neste trabalho entre os anos de 2015 e 2018, a investigação utilizou-se de dados secundários das políticas e termos de comunidade específicos de cada SRS e das informações dos blogs oficiais do Facebook[1], Twitter[2] e Youtube[3], abrangendo o período determinado. Para pesquisa, foram utilizadas as seguintes palavras-chaves no sistema interno de busca de cada um dos *blogs*: "HATE SPEECH", "HATEFUL", "CYBERHATE".

Como resultado, foram encontradas, no Facebook, 33 notícias utilizando o termo *Hate Speech*, quatro com *Hateful* e uma com *Cyberhate.* No Twitter, as informações encontradas para os mesmos termos, respectivamente, foram 35, 14 e uma. Já no *blog* do Youtube foram encontradas 40 notícias sobre *Hate Speech*, 17 sobre *Hateful* e nenhuma sobre *Cyberhate.*

Estas postagens, por sua vez, foram lidas em sua completude possibilitando construir uma cronologia das ações desenvolvidas por cada uma das empresas ao longo dos anos de 2015-2018. Além disso, também foram realizadas denúncias em cada uma das plataformas para compreender o fluxo de relatórios e as providências tomadas pelos SRS. As publicações denunciadas foram escolhidas de modo aleatório, sem compromisso com a possibilidade de ser ou não um discurso de ódio. O intuito era acompanhar as etapas do processo de denúncia para descrevê-las fielmente ao exercido pelas empresas. As denúncias foram realizadas por meio da conta pessoal de dois dos autores deste artigo no mês de setembro de 2018. Todas as informações foram analisadas com base nos tópicos do *Best Practices for Challenging Cyberhate (BPCC),* no que compete aos provedores.

É importante salientar que muitas das estratégias usadas pelos SRS não são divulgadas por uma questão de segurança. Pessoas e organizações interessadas em prejudicar a saúde das relações nos espaços digitais podem usar das informações para criar soluções tecnológicas que atrapalham o processo de moderação. Desta forma, não é possível afirmar com precisão os aspectos tecnológicos usados pelas empresas para conter o fenômeno, mas é sabido pelos documentos analisados quais as direções principais que os SRS resolveram seguir e até onde já avançaram.

Como forma de apresentar os resultados, optou-se por construir uma narrativa descritiva a partir das informações encontradas, organizadas de acordo com cada plataforma,

1 https://newsroom.fb.com

2 https://blog.twitter.com

3 https://youtube.googleblog.com

conforme poderá ser conferido nos próximos tópicos. Outros termos de sistematização, especialmente em relação às categorias de análise, também poderão ser conferidos na discussão e nas considerações finais.

4.1 Facebook

Desde 2015, o Facebook tem afirmado melhorias no índice de remoção de conteúdo odiento em até 24 horas. O SRS utiliza da combinação de IA (Inteligência Artificial) e moderadores treinados para encontrar e remover propaganda terrorista, xenofóbica, apoio ao nazismo e outras formas de discurso de ódio. A empresa publicou em seu *blog* oficial, em 15 de março de 2015, que pode remover ou restringir o acesso ao conteúdo que possa violar seus padrões de comunidade ou alguma lei em um país específico (FACEBOOK, 2015). Para isso, tem disponibilizado o *Relatório Global de Solicitações do Governo*, que inclui os conteúdos removidos e os dados de contas, conforme a Lei de Vigilância de Inteligência Estrangeira dos EUA e as Cartas de Segurança Nacional. O SRS afirma remover com cautela os conteúdos solicitados por governos, na tentativa de não ceder a uma possível censura ou interesses políticos, mas assume remover ou limitar o acesso a tais solicitações quando ferem alguma lei local mesmo que não viole os seus padrões de comunidade.

A empresa de Mark Zuckemberg confirma que (1) atualiza constantemente a equipe de Operações do Usuário para avaliar os relatórios de violações de Padrões Comunitários em torno do discurso de ódio; (2) capacita as equipes que analisam e avaliam relatórios de discurso odiento ou conteúdo prejudicial; (3) aumenta a responsabilidade dos criadores de conteúdos cruel ou insensível, insistindo para que os autores sustentem o conteúdo que criam; (4) estabelece linhas de comunicação mais formais e diretas com representantes de grupos que trabalham na área do discurso de ódio, incluindo grupos de mulheres, para garantir o tratamento acelerado de conteúdo que eles acreditam que violam os padrões; (5) incentiva o grupo de trabalho *anti-cyberdate* da Liga anti-difamação e outros grupos de trabalho internacionais para melhor identificar as considerações de livre expressão e o efeito de discurso de ódio online sobre as experiências de minorias sociais; e (6) avalia o progresso das ações implementadas e dos objetivos coletivos (FACEBOOK, 2015).

Em parceria com a Edventure Partners[4], o Facebook criou o *P2P (Peer to Peer: Facebook Global Digital Challenge)* que envolve estudantes universitários de todo o mundo em competições por meio das quais os alunos criam campanhas de mídia social e estratégias

4 Empresa americana voltada a desenvolver oportunidades de aprendizado experiencial para estudantes e educadores. Site da empresa: https://edventurepartners.com/

offline para conter narrativas extremistas e odiosas. A empresa afirma que o P2P alcançou mais de 56 milhões de pessoas em todo o mundo por meio de mais de 500 campanhas anti-ódio e extremismo criadas por mais de 5.500 estudantes entre 2015 e 2017(FACEBOOK, 2017).

Juntamente com Twitter, Google e Microsoft, o SRS criou o *Cyberhate Problem-Solving Lab*, um laboratório com profissionais focados em novas soluções técnicas e estratégicas para identificação e abordagem do ódio nas plataformas. Desde 2016, o laboratório é administrado pelo Centro de Tecnologia e Sociedade da ADL no Vale do Silício e tem servido de suporte para iniciativas individuais dos SRS à criação de tecnologia *anti-hate*. No mesmo ano, o Facebook instituiu uma equipe de especialistas contra o terrorismo formada por funcionários próprios e terceirizados para revisar os comentários denunciados por usuários ou identificados pelo algoritmo. O número de moderadores chegou a 7.500, sendo que deste total 1.200 trabalhavam em Berlim na tentativa de responder a lei local *NetzDG*, que exige remoção de conteúdo de ódio em até 24 horas.

Em 2016, a empresa investiu em iniciativas acadêmicas para avaliar o impacto das mídias sociais nas eleições americanas e criou o serviço de suporte OCCI (*Online Civil Courage Initiative*), que fornece apoio a ONGs e ativistas europeus que trabalham para combater o extremismo online e o discurso de ódio, desenvolve os melhores métodos para responder ao extremismo online e ao discurso de ódio e auxilia a pesquisa sobre a relação entre o discurso de ódio online e a violência offline (FACEBOOK, 2016). O OCCI teve boa aceitação e se estendeu à Alemanha, França e Reino Unido. O serviço é composto por um *help desk*, relatórios de *insights* mensais e grupos regionais fechados do Facebook, dando suporte a 100 organizações anti-ódio e anti-extremismo e atingindo até setembro de 2018, 3,5 milhões de pessoas por meio de sua página no SRS.

No ano subsequente, o Facebook realizou uma série de oficinas de aprendizado em parceria com a CTED (Comitê de Contra-Terrorismo das Nações Unidas) e a Fundação Suíça ICT4Peace. Juntamente com Twitter, Microsoft e Google, a empresa criou o Banco de Dados de Hash da Indústria Compartilhada - um repositório para trocar as melhores práticas de desenvolvimento de técnicas de detecção e classificação de conteúdo odiento usando aprendizado de máquina e as impressões deixadas por outros interagentes violadores (FACEBOOK, 2017).

Ainda em 2017, o SRS definiu métodos padronizados de relatório de transparência para remoção de conteúdo terrorista e tornou mais eficiente sua política de anúncios, comprometendo-se a não permitir que a publicidade seja usada para ódio ou discriminação (FACEBOOK, 2017). Desde então, os anúncios são analisados por ferramentas automatizadas e moderadores na tentativa de responder quando as pessoas ocultam, bloqueiam ou marcam

anúncios como ofensivos. A empresa reforçou seu compromisso com a atualização de determinados recursos, como formatos, métricas e controles de anúncios.

Em 2018, a empresa declarou que a métrica mais importante usada para moderação e remoção de conteúdo odiento é o impacto total da publicação (FACEBOOK, 2018), ou seja, é mensurada a frequência com que o conteúdo é visto e quão grave é o impacto de cada visualização nas pessoas e comunidades que o visualizam. O impacto total é, portanto, a quantidade de visualizações vezes o impacto da violação de conteúdo.

O Facebook afirma ter aprimorado o fluxo informacional que envolve o processo de denúncia e remoção do conteúdo de discurso de ódio. Em 2015, ao denunciar um comentário, página, *post* ou conversa privada na plataforma, o interagente recebia um teste de verificação de identidade para comprovar que não se tratava de um "trote" ou de um "bot". A partir de 2016, os casos de denúncia de ofensa racista, homofóbica, xenofóbica, tidos como mais graves, passaram a ser revistos por moderadores que analisam e encaminham um alerta ao infrator. Em casos extremos, a plataforma oferece um recurso de contato para intervenção direta com as autoridades.

O procedimento evoluiu e em 2018 se tornou mais direto. Ao enviar o conteúdo para análise, o interagente recebe uma mensagem em sua própria conta informando que a equipe responsável tomará as devidas providências. O denunciante, por sua vez, pode acompanhar o *status* da denúncia (em análise ou encerrado) ou até mesmo cancelar o procedimento em sua Caixa de Entrada de Suporte. Após a análise, o Facebook oferece um parecer agradecendo pela iniciativa do interagente e informa a remoção ou não do conteúdo.

O SRS pede o motivo pelo qual o interagente deseja realizar o procedimento de denúncia e o tipo de conteúdo impróprio, oferecendo uma lista que inclui diretamente o discurso de ódio. A interface oferece uma série de ações que o próprio interagente pode tomar tais como bloquear o violador, deixar de seguí-lo, enviar o conteúdo para análise, desfazer a amizade e enviar uma mensagem para o violador pedindo que remova o conteúdo. Todo o processo busca preservar a identidade do denunciante e em todas as etapas o interagente tem a opção de bloquear ou desfazer a amizade com o violador.

Para responder a alta demanda de denúncias, sem apelar para uma censura generalizada e inconsequente, o Facebook tem aplicado o *Cross Check,* ou seja, uma segunda camada de revisão para garantir a aplicação correta das políticas de remoção de conteúdo. Desde 2016, o SRS aplica um modelo de aprendizado de máquina que usa vários sinais de engajamento, incluindo o *feedback* de pessoas no Facebook, para identificar conteúdo potencialmente falso, odioso ou abusivo. Os conteúdos identificados são encaminhados para

verificadores (terceirizados ou funcionários) para revisão, independente da fonte da violação ou a posição social do violador.

A empresa aumentou o número de pessoas trabalhando em equipes de segurança para 20.000, incluindo mais de 7.500 revisores de conteúdo (FACEBOOK, 2018). Também investiu em novas tecnologias para auxiliar no envio de relatórios a revisores com os conhecimentos adequados, para eliminar relatórios duplicados e para ajudar a detectar e remover imagens de propaganda terrorista e de abuso sexual infantil antes que elas sejam denunciadas. A própria chefe operacional, Sheryl Sandberg, passou a solicitar resumos semanais das moderação de conteúdo e reuniões com os líderes da equipe de análise de conteúdo para discutir as melhores maneiras de lidar com as escalações do conteúdo de ódio e como responder a problemas de moderação de conteúdo em tempo real (FACEBOOK, 2018).

Também em 2018, a empresa passou a permitir que os interagentes que tiverem conteúdo removido solicitem uma revisão sobre a violação dos Padrões da Comunidade, para comprovar se realmente o conteúdo feria algum dos termos (FACEBOOK, 2018). No serviço *Messenger,* o Facebook ampliou o trabalho das equipes de Operações da Comunidade para agilizar o processo de relatórios. A empresa se propõe analisar o conteúdo denunciado, em mais de 50 idiomas, e emitir relatório ao denunciante em até 48 horas.

Quanto às sanções, o Facebook pode remover as contas e publicações de violadores. Em agosto de 2018, o SRS derrubou várias contas no Brasil e em Myammar de indivíduos e organizações que cometeram ou permitiram graves abusos aos direitos humanos e propagavam inverdades na plataforma, sobretudo em contextos políticos.

4.2 Twitter

Além das ações em conjunto com os demais SRS, conforme já descrito anteriormente, o Twitter operou algumas iniciativas próprias. Em 2015, atualizou suas políticas e produtos relacionados a conteúdos proibidos e intensificou medidas corretivas. A empresa ampliou o termo de conteúdo abusivo para "ameaças ou promoção de ameaças ao próximo" e foi pioneira nos relatórios de transparência solicitados por governos.

O SRS aumentou o engajamento com organizações da sociedade civil no combate a conteúdos odiosos ou prejudiciais, trabalhando com mais de 250 organizações e ativistas, participando de mais de 35 eventos dedicados a combater o extremismo violento (CVE) e realizando sessões de treinamento em 15 países da EMEA[5]. A empresa também investiu em

5 Designação geográfica de países da Europa, Oriente Médio e África.

campanhas de conscientização à tolerância pós-ataque em Paris[6], de igualdade de gênero na Irlanda e de *anti-bullying* ao redor do mundo.

Já em 2016, o Twitter aderiu à Política de Remoção de Conteúdo Odiento em até 24 horas proposta pelo código de conduta da União Europeia e tornou-se membro do Grupo de Alto Nível da União Europeia para combater o racismo, a xenofobia e outras formas de intolerância (EUROPEAN COMMISSION, 2015). O CEO da companhia, Jack Dorsey, anunciou a luta contra o discurso de ódio na plataforma como uma de suas prioridades. A empresa criou o próprio Conselho de Confiança e Segurança para garantir que as pessoas se sintam seguras ao se expressarem na rede social. Também apoiou campanhas como *#exithate* e *#KlickItOut,* além de trabalhar em parceria com o Instituto de Diversidade de Mídia na criação de um guia, em quatro idiomas, contra o preconceito, distribuído de forma gratuita a professores e estudantes europeus.

Em 2017, o Twitter prometeu um endurecimento de regras voltadas para o discurso de ódio, principalmente o racista e xenofóbico. Por meio do Conselho de Segurança e Confiança estabeleceu novas políticas de remoção de conteúdo abusivo e odioso que usam ou promovem violência contra civis ou advogam a favor da intolerância (TWITTER, 2018). As novas regras também determinam remoção de "conteúdo que glorifica a violência ou os perpetradores de um ato violento" , de "qualquer conta que abuse ou ameace outras pessoas por meio de suas informações de perfil , incluindo seu nome de usuário, nome de exibição ou perfil bio", e "imagens odiosas como logos, símbolos ou imagens cuja finalidade é promover hostilidade e malícia contra outros com base em sua raça, religião, deficiência, orientação sexual ou etnia/ nacionalidade" (TWITTER, 2018).

Também em 2017, o Twitter participou do Fórum Global da Internet para combater o Terrorismo. Na ocasião, a empresa apresentou as iniciativas em conjunto com outros SRS, organizações civis e pesquisadores acadêmicos no combate ao Discurso de Ódio Online. As propostas foram semelhantes às do Facebook, citadas anteriormente. Em novembro do mesmo ano, o SRS fez mais alterações e passou a interromper a criação de novas contas abusivas, apresentar resultados de pesquisa mais seguros e recolher *tweets* potencialmente abusivos ou de baixa qualidade. Por outro lado, no mesmo ano, Dick Costolo, ex-CEO do Twitter, admitiu falhas da plataforma e se diz envergonhado em como a empresa lida com o *cyberbullying*. Ele afirmou que o Twitter não consegue ter controle do que é postado na rede social e por isso tem perdido grande número de usuários (GUARDIAN (THE), 2018).

6 Tiroteios e explosões são registrados em Paris. Disponível em: http://g1.globo.com/mundo/noticia/2015/11/tiroteios-e-explosoes-sao-registrados-em-paris-diz-imprensa.html Acessado em 21 de Setembro de 2018

Em 2018, o Twitter expandiu ainda mais sua política de conduta odienta, incluindo conteúdo que desumaniza os outros com base em sua participação em um grupo identificável, mesmo quando o material não inclui um alvo direto. Para isto, solicitou aos seus interagentes que respondessem a um questionário que valida as perspectivas globais do combate ao discurso de ódio e ajuda a clarear como a nova política da empresa poderia afetar diferentes comunidades e culturas. Mais de 8.000 interações foram coletadas e serão analisadas pelo Centro de Segurança do SRS até o fim do ano quando pretende implantar as alterações (TWITTER, 2018).

Quanto ao procedimento de denúncias no Twitter, na primeira etapa o interagente pode escolher entre quatro opções: (1) não tenho interesse neste *Tweet*; (2) é *spam*; (3) mostra uma imagem sensível ou imprópria; e (4) é abusivo ou nocivo. A interface oferece também um *link* para saber mais sobre “denunciar violações”.

Ao iniciar uma denúncia de conteúdo abusivo ou nocivo, a plataforma indica, entre as categorias prejudiciais, a de "ódio contra uma categoria protegida (por exemplo, raça, religião, gênero, orientação sexual ou deficiência)". A interface subsequente sugere três possíveis violadores - uma conta, uma outra pessoa ou um grupo de pessoas. E a partir desta escolha, a plataforma sugere outros *tweets* que também podem ser identificados como semelhantes ao denunciado. O procedimento tem como objetivo o aprendizado de máquina e ajuda em futuras identificações de conteúdo de ódio, de forma mais rápida.

Durante o processo de denúncia é possível bloquear ou silenciar a conta do violador ou as mensagens diretas. Além disso, o Twitter também indica procedimentos que o interagente pode tomar enquanto aguarda a decisão da denúncia e recomendações de outras providências para melhorar a experiência na plataforma. O SRS envia um *e-mail* com os textos dos *tweets* denunciados e uma mensagem de agradecimento.

Desde 2015, o Twitter usa recursos que levam em consideração sinais e o contexto na análise de conteúdo denunciado, incluindo a idade da conta e a similaridade de um *tweet* a outro conteúdo identificado anteriormente pela equipe de segurança como abusivo. De lá para cá, o SRS diz aprimorar relatórios com maior probabilidade de violar as regras, impondo sanções que vão desde o bloqueio até a suspensão definitiva da conta. Após análise por um de seus agentes, depois de feita a verificação da conta, se confirmado que possui violações de seus termos, a empresa determina uma de suas punições, isso de acordo com a regra que foi violada.

4.3 Youtube

O Youtube também atualizou as políticas de conteúdo em 2015 e tornou mais rígidas as políticas relacionadas a conteúdos proibidos e abusivos, ampliando a moderação de interações com ameaças ou promoção ao ódio e ao terrorismo. No primeiro semestre de 2018, a empresa intensificou os esforços para cumprir a lei alemã *NetzDG* e adicionou opções para os anunciantes evitarem patrocinar vídeos com conteúdo de ódio - problema que em 2017 gerou ônus significativos devido a debandada de anunciantes que tiveram suas marcas atreladas a vídeos terroristas e de ódio (YOUTUBE, 2018).

Por outro lado, um ano antes, em 2017, a empresa anunciou que "conteúdo inflamatório religioso ou de supremacia" que não violasse suas políticas seria permitido com rótulos de aviso e uma restrição que os tornasse inelegíveis para a receita de anúncios (YOUTUBE, 2018).

No Youtube, as categorias de conteúdo para denúncia são: conteúdo sexual, conteúdo violento ou repulsivo, conteúdo de incitação ao ódio ou abusivo, comportamentos perigosos e nocivos, abuso infantil, promoção ao terrorismo, *spam* ou enganoso, violação de direitos e problema com as legendas. Na opção de conteúdo de ódio são apresentadas especificações como "estimula ódio ou violência", "abuso de indivíduos vulneráveis", "*bullying*", "título ou descrição abusivos".

Após a delimitação do tipo de conteúdo de ódio, a plataforma pede mais detalhes sobre a denúncia e registra o horário e a data da solicitação. Uma mensagem de agradecimento e um aviso de que o conteúdo foi enviado para análise aparecem na sequência. Durante todo o procedimento o Youtube oferece a possibilidade de denunciar o canal.

A plataforma usa de sistemas automatizados, de membros do programa Revisor confiável (ONGs, agências governamentais e indivíduos) ou de equipe própria para identificar, revisar e remover o conteúdo tido como discurso de ódio. Depois que um conteúdo potencialmente problemático é detectado pelos sistemas automatizados ou denunciado por algum interagente, a análise humana verifica se ele realmente viola as políticas do Youtube. Como as denúncias dos Revisores Confiáveis têm uma taxa de ação mais alta do que as de um interagente normal, o Youtube prioriza estes dados para revisão.

Equipes em várias partes do mundo revisam vídeos denunciados e removem conteúdos que violam os termos. A restrição de vídeos, comentários, miniaturas e contas também depende da análise de um dos membros especializados, mas com o aumento do aprendizado e precisão das máquinas alguns conteúdos de ódio já são removidos automaticamente. Em junho de 2017, 40% dos vídeos removidos por extremismo violento foram excluídos antes de serem denunciados por pessoas. Este número aumentou

rapidamente para 76% em agosto de 2017 e para 83% em outubro do mesmo ano. O relatório divulgado pelo Youtube também reforça que em dezembro de 2017, 98% dos vídeos removidos por extremismo violento foram identificados pelos algoritmos de aprendizado de máquina, sendo a maioria antes mesmo de obter alguma visualização. (YOUTUBE, 2017).

Além do trabalho com moderadores e aprendizado de máquina, a Google anunciou em 2018 um fundo de inovação de US$ 5 milhões para combater o ódio e o extremismo. Este financiamento apoiará soluções orientadas para a tecnologia, bem como esforços de base, como projetos comunitários de jovens que ajudem a construir comunidades e promovam resistência à radicalização.

5 Discussão

A análise documental deste trabalho revelou uma preocupação e um progresso dos SRS em relação ao fenômeno do discurso de ódio em suas plataformas. Após a assinatura do termo de compromisso com a Liga Anti-difamação, em 2013, tanto Facebook quanto Twitter e Youtube estabeleceram regras mais severas em suas políticas de comunidade, visando a manutenção de um ambiente seguro, onde seus interagentes possam se expressar livremente. A partir de 2015, os SRS trataram da problemática do ódio *online* com ainda mais seriedade, pautando suas políticas nos princípios fundamentais de liberdade de expressão, dignidade humana, segurança pessoal e respeito pelo estado de direito.

Por outro lado, a pesquisa revelou que os SRS foram mais ágeis em suas ações após a interferência de governos, escândalos de vazamento de dados, ataques terroristas e a vigência de novas leis sobre conteúdo odiento na Internet. Naturalmente, tal constatação possibilita questionar o quanto os sites aqui analisados estariam preocupados com a manutenção da civilidade em suas plataformas ou se agiram pontualmente para conter os prejuízos financeiros e as sanções governamentais e da sociedade.

As iniciativas coletivas, como o banco de dados Hash e o Fórum Global da Internet, parecem depor favoravelmente em relação ao primeiro questionamento, haja vista que o interesse em conter o problema do ódio superou os objetivos individuais comerciais de cada empresa. Porém, apesar destas atitudes, que teoricamente significam que os SRS não toleram o discurso de ódio e estão unidos pela causa, a prática tem sido diferente. Neste sentido, verifica-se que os procedimentos de remoção de conteúdo ainda são obscuros aos interagentes, sobretudo no Twitter e no Youtube. Os documentos analisados não oferecem uma clara explicação do que é considerado discurso de ódio e os relatórios encaminhados não trazem explicações detalhadas do porquê o conteúdo viola ou não a política de propagação do ódio.

Nos anos de 2015 e 2016, os principais esforços das empresas parecem estar voltados para uma abordagem legal (alteração de política da plataforma, criação de conselhos etc) e abordagem educativa, investindo em parcerias, programas e iniciativas com foco na educação do interagente à tolerância, civilidade e respeito. Já em 2017 e 2018, os documentos oferecem informações de abordagem mais tecnológica, com o desenvolvimento e aperfeiçoamento de recursos de Inteligência Artifical e Aprendizagem de Máquina

A revisão dos procedimentos legais que cada um dos SRS estabeleceu para manter a segurança em seus *sites* e adaptar suas políticas à realidade dos países onde estão presentes foi constante. O Twitter merece destaque ao levar em consideração a opinião de seus interagentes neste processo, disponibilizando espaço para que eles se manifestassem quanto à política de moderação de conteúdo. Isto ocorreu tanto por meio da conta pessoal do CEO quanto na conta oficial @twittersafety, numa consulta que resultou em mais de 8.000 considerações.

Quanto aos mecanismos e procedimentos para denúncia, todos os SRS oferecem maneiras intuitivas e de fácil manuseio. O Facebook prefere o nome "*feedback* a publicação" ao invés do termo "denunciar" usado pelo Twitter e Youtube, o que pode causar um estranhamento da parte dos interagentes. As demais etapas do processo direcionam para denúncia de discurso de ódio de forma indubitável.

Nos testes realizados pela pesquisa, as respostas às denúncias aconteceram dentro do prazo máximo estabelecido de 48 horas, sendo que o Facebook foi o mais ágil no parecer, respondendo antes de 24 horas do início do procedimento. Não podemos afirmar, no entanto, que isto aconteça com frequência, e a depender da complexidade do conteúdo é bem possível que o cenário mude. A intenção, neste caso, não foi comparar o tempo dentro das 48 horas, mas se algum dos SRSs não conseguiria cumprir o acordado, o que não aconteceu.

As sanções aplicadas ainda estão sendo aprimoradas pelos SRS. O que ficou evidente foi que a exclusão da conta e a proibição de criar uma nova é o máximo que o interagente violador pode sofrer, exceto em casos que envolve as autoridades locais. Desta forma, a externalização do discurso de ódio se paga, ou seja, os violadores não recebem grandes penalizações, o que pode indicar uma propensão a reincidência de comportamento.

6 Considerações finais

O trabalho aqui proposto trouxe colaborações aos estudos de discurso de ódio online ao abordar o progresso que Facebook, Twitter e Youtube tiveram ao longo dos anos de 2015 a 2018 em suas políticas e ações de combate ao fenômeno do ódio online. Os documentos disponíveis nos *blogs* oficiais dos três SRS revelaram que as políticas de comunidade foram

aprimorados com o tempo e que os procedimentos de moderação de conteúdo odiento cresceram em tecnologia, artifícios legais, transparência e em recursos humanos.

As informações oficiais disponíveis e o procedimento de denúncia testados pelos autores deste artigo em cada uma das plataformas permitiram uma observação mais apurada sobre as tomadas de decisão das empresas em relação ao escalonamento, duração, difusão e escalada do discurso de ódio. Os termos de compromisso assinado pelos provedores com a Liga Anti-difamação em 2013 parece ter impulsionado a agilidade dos SRS com a problemática. Facebook, Twitter e Youtube se mostraram abertas ao diálogo com especialistas e organizações locais e reconhecem que muito ainda precisa ser realizado.

Por fim, ressalta-se que este estudo faz parte de um esforço de compreender o ódio em espaços digitais na perspectiva da criação de uma ontologia que auxilie os SRS na moderação do discurso de ódio brasileiro, baseado em evidências existentes. Mais pesquisas são necessárias para analisar como os fatos e informações de ódio estão sendo geridos pelas grandes corporações, tanto no que compete às características de uma empresa privada quanto no entendimento dos SRSs como plataformas públicas.

Referências

ANTI-DEFAMATION LEAGUE (ADL). *Who we are.* Dísponível em: <https://www.adl.org/who-we-are>. Acessado em 20 de agosto de 2018.

ALEMANHA. *Lei NetzDG, de 01 de Setembro de 2017.* Disponível em: <https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html> Acessado em 20 de agosto de 2018

BEN-DAVID, A.; MATAMOROS-FERNANDEZ, A. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, v. 10, p. 1167-1193, Feb. 2016.

BRASIL. *Lei nº 12.965, de 23 de abril de 2014.* Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Brasília. 2014. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm>.

BULINGE, F. Radicalisation sur Internet: Méthodes et techniques de manipulation. *Cahiers de la sécurité et de la justice*, v. 30, p. 32–42, 2014. Disponível em: <https://archivesic.ccsd.cnrs.fr/sic_01804903> Acessado em: 15 out. 2018.

CITRON, Danielle Keats. *Hate crimes in cyberspace*. Boston: Harvard University Press, 2014.

CNBC. *Mark Zuckerberg said an independent 'Supreme Court' could fix Facebook's content problems*. Disponível em: <https://www.cnbc.com/2018/04/02/facebook-ceo-mark-zuckerberg-on-a-supreme-court-for-content.html> Acessado em 23 de Agosto de 2018.

COHEN-ALMAGOR, Raphael. *Holocaust denial is a form of hate speech.* Amsterdam LF, v. 2, p. 33, 2009.

DORSEY, J. Disponível em: https://twitter.com/jack. Acessado em 21 de Setembro de 2018.

EUROPEAN COMMISSION. *EU High Level Group on combating racism, xenophobia and other forms of intolerance.* Disponível em: <https://ec.europa.eu/knowledge4policy/organisation/eu-

high-level-group-combating-racism-xenophobia-other-forms-intolerance_en> Acessado em 23 de agosto de 2018.

FACEBOOK. *Explaining Our Community Standards and Approach to Government Requests*. Disponível em: <https://newsroom.fb.com/news/2015/03/explaining-our-community-standards-and-approach-to-government-requests/> Acessado em 20 de agosto de 2018.

FACEBOOK. *Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism.* June 26, 2017. Disponível em: <https://newsroom.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/> Acessado em 21 de agosto de 2018.

FACEBOOK lança campanha para combater o discurso de ódio no Reino Unido. Disponível em: <http://tecnologia.ig.com.br/2017-06-23/facebook-discurso-odio.html> Acessado em 16 de Setembro de 2017.

FACEBOOK. *Padrões de Comunidade do Facebook*. 23 junho de 2017. Disponível em: <https://www.facebook.com/communitystandards/>. Acessado em 21 de Setembro de 2018.

FACEBOOK. *Trabalhamos para manter o facebook um lugar seguro*. 17 julho de 2018. Disponível em: <https://br.newsroom.fb.com/news/2018/07/trabalhamos-para-manter-o-facebook-um-lugar-seguro/> Acessado em 22 de agosto de 2018.

FACEBOOK. *Questões complexas: como combatemos o terrorismo.* 15 de junho de 2017. Disponível em: <https://br.newsroom.fb.com/news/2017/06/questoes-complexas-como-combatemos-o-terrorismo/> Acessado em 22 de agosto de 2018.

FACEBOOK. *Trabalhamos para manter o facebook um lugar seguro.* 17 de julho de 2018. Disponível em: <https://br.newsroom.fb.com/news/2018/07/trabalhamos-para-manter-o-facebook-um-lugar-seguro/> Acessado em 22 de agosto de 2018.

GLOBO (O). *Verizon, AT&T e Johnson & Johnson removem seus anúncios do YouTube.* **Globo**, Rio de Janeiro, Caderno Economia, 23 março 2017. Disponível em: <https://oglobo.globo.com/economia/verizon-att-johnson-johnson-removem-seus-anuncios-do-youtube-21105710> Acessado em 18 de agosto de 2018.

KNOBEL, M. *L'internet de la haine. Racistes, antisémites, néonazis, intégristes, isla- mistes, terroristes et homophobes à l'assaut du web.* Paris: Berg International Editeurs, 2012.

KUDLACEK, D.; TRESKOW, L.; MARSH, B.; FLEISCHER, S., PHELPS, M.; HALILOVIC PASTUOVIC, M. *Gap analysis on counter radicalisation measures.* Hannover: Kriminologisches Forschungsinstitut Niedersachsen, 2017. Disponível em: <https://f. hypotheses.org/wp-content/blogs.dir/2725/files/2017/11/Pericles-D1.2-Gap- Analysis-Report.pdf> Acessado em: 15 out. 2018.

RECUERO, R. *A conversação em rede:* Comunicação Mediada Pelo Computador e Redes Sociais na Internet. Porto Alegre*:* Sulina**,** 2014.

ROST, K.; STAHEL, L.; FREY, B. S. Digital social norm enforcement: Online firestorms in social media, *PLoS ONE,* v. 11, n. 6, 2016. Disponível em: <doi:10.1371/journal.pone.0155923> Acessado em: 15 out. 2018.

SANTOS, M. A. M dos. *O discurso de ódio em Redes Sociais.* São Paulo: Lura Editorial, 2016.

SENRA, R. 'Checava se alguém se mataria ao vivo': a rotina do brasileiro que moderava posts denunciados no Facebook. *BBC Brasil*, 8 nov. 2017. Disponível em: <https://www.bbc.com/portuguese/geral-41912670>. Acessado em 22 de Setembro de 2018.

SHEPHERD, T.; HARVEY, A.; JORDAN, T.; SRAUY, S.; MILTNER, K. Histories of hating. *Social Media + Society*, v. 1, *n.* 2, 2015. Disponível em: <doi:2056305115603997> Acessado em: 15 out. 2018.

GUARDIAN (THE). *What are four of the top social media networks doing to protect children?* Disponível em: <https://www.theguardian.com/sustainable-business/2016/feb/09/social-media-networks-child-protection-policies-facebook-twitter-instagram-snapchat> Acessado em 22 de set. de 2018.

TWITTER. *Imposição de nossa regras.* Disponível em: <https://about.twitter.com/pt/safety/enforcing-our-rules.html> Acessado em: 21 de set. de 2018.

TWITTER. *Twitter anuncia atualizações nas regras de segurança.* 18 dez. 2017. Disponível em: <https://blog.twitter.com/official/pt_br/topics/company/2017/twitter-anuncia-atualizacoes-nas-regras-de-seguranca.html> Acessado em: 21 de set. de 2018.

TWITTER. *Twitter detalha medidas contra uso indevido de robôs e desinformação*. 28 setembro de 2017. Disponível em: <https://blog.twitter.com/official/pt_br/topics/company/2017/twitter-detalha-medidas-contra-uso-indevido-de-robos-e-desinformacao.html> Acessado em: 21 de set. de 2018.

TYNES, B. M.; ROSE, C. A.; MARKOE, S. L. Extending campus life to the Internet: Social media, discrimination, and perceptions of racial climate. *Journal of Diversity in Higher Education,* v. 6, n. 2, p. 102–114, 2014. Disponível em: <http://dx.doi.org/10.1037/a0033267> Acessado em: 15 out. 2018.

WANG, B.; WANG, T.; WANG, A.; NIKA, H.; ZHENG, B.; ZHAO, B. Y. Whispers in the dark: Analyzing an anonymous social network. In: ACM CONFERENCE ON INTERNET MEASUREMENT CONFERENCE, 2014. *Proceedings.* Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C1C6471A097391527B0337E5C6B82967?doi=10.1.1.697.312&rep=rep1&type=pdf> Acessado em: 15 out. 2018.

XIANG, G.; FAN, B.; WANG, L.; HONG, J.; ROSE, C. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 21., 2012. *Proceedings.* p. 1980–1984, 2012.

YOUTUBE. *About policies.* Disponível em: <https://www.youtube.com/intl/pt-BR/yt/about/policies/#community-guidelines>. Acessado em: 21 de set. de 2018.

YOUTUBE. *Expanding our work against abuse of our platform.* December 4, 2017. Disponível em: <https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html> Acessado em: 21 de set. de 2018.