Desinformação e plataformas digitais:

elas estão enfrentando as fake news?

Helena Martins (Intervozes)

intervozes
coletivo brasil de comunicação social

Caminhos da pesquisa

A pesquisa analisa a desinformação, considerada uma estratégia para obtenção de ganhos políticos e econômicos.

Foram mapeadas as medidas adotadas a partir de 2018 pelas plataformas de redes sociais: Facebook, Instagram, WhatsApp, YouTube e Twitter

Para reunião e análise, definimos quatro categorias:
1. Abordagem do fenômeno;
2. Moderação de conteúdo;
3. Promoção de informações e transparência;
4. Medidas correlatas.

Além disso, estudadas recomendações de órgãos de direitos humanos e apresentadas propostas já formuladas pelo Intervozes.

1. Abordagem do fenômeno

Facebook	Instagram	WhatsApp	YouTube	Twitter
Não possui uma política específica nem trabalha com uma definição própria. **Apresenta as estratégias contra o fenômeno de maneira resumida nos Padrões da Comunidade.** A elaboração de medidas a serem adotadas é realizada pela equipe de "Políticas Globais de Conteúdo". Em 2020, criou o Conselho de Supervisão de Conteúdo, com membros externos.	Os Termos de Uso determinam que o usuário não pode "fazer algo ilícito, enganoso, fraudulento ou com finalidade ilegal ou não autorizada". **Não possui estrutura institucional ou processos específicos para tratar de desinformação.** Usa fluxos e procedimentos adotados pelo Facebook, incluindo as equipes de revisão de conteúdo e o Conselho de Supervisão.	Apresenta-se como plataforma de mensagens privadas criptografadas quenão modera conteúdo. **Contradição com a existência de grupos e listas de transmissão.** Não assume o uso frequente da plataforma para desinformação. Seguindo tal entendimento, não possui política, processos e/ou estrutura institucional nem conceito de desinformação definidos.	Também não trabalha com uma única definição nem possui uma política específica. Restrições à disseminação de desinformação aparecem em diferentes políticas, como sobre danos e mídias manipuladas. **Conteúdos que violam as Diretrizes da Comunidade são removidos e notificados.** Casos reincidentes sofrem restrições de publicação e monetização e o canal pode ser excluído.	Não possui política para validar a autenticidade de conteúdos, tampouco trabalha com definição. **Oferece contexto relacionado a post contestado, agindo diretamente em casos de possíveis danos causados por mídias manipuladas, informações enganosas sobre processos eleitorais e, recentemente, relacionadas à Covid-19.** Proíbe o uso de robôs manipuladores da rede. Tem um Conselho de Confiança e Segurança.

2. Moderação de conteúdo

Facebook	Instagram	WhatsApp	YouTube	Twitter
Encaminha conteúdos para verificação por agências. **Em caso de classificação como "falso" ou "parcialmente falso", usa rótulo específico e passa a ser acompanhado de "artigos relacionados".** Declarações de líderes políticos não passam por verificação. **Vídeos e imagens podem ser rotulados como "manipuladas", "tiradas de contexto" ou "falsas".** Não há remoção, mas a circulação pode ser reduzida. **Desinformação que cause violência ou dano ou que comprometa processos eleitorais é removida.** Anúncios com mensagens falsas não podem ser veiculados e anunciantes podem ser sancionados.	Usa agências de verificação. Conteúdos são marcados quando são verificados como desinformativos e podem ter circulação reduzida. **Conteúdos relacionados a falsos tratamentos sobre a Covid-19, teorias da conspiração ou falsas alegações registradas como danosas pelas autoridades de saúde são removidos, assim como anúncios e # desinformativos.** Na busca, prioriza informações de autoridades de saúde. **Nas eleições, se um conteúdo for classificado como falso, um filtro cinza é exibido sobre a imagem e usuários são alertados.**	Informa que não modera conteúdo. No campo da moderação, apontamos apenas que, desde 2018, tem **atuado para reduzir a circulação, por meio do estabelecimento de limites para encaminhamento de mensagens.** Primeiro, o limite de envio de uma vez foi fixado em 20 chats. Depois, 5. **Desde a pandemia, as mensagens altamente encaminhadas só podem ser encaminhadas para um contato de cada vez.** Mesmo essa política não é, em geral, associada ao combate à desinformação. Tal vinculação só foi expressa no contexto da pandemia.	**Tira do ar conteúdos que violam suas diretrizes e reduz o alcance de "desinformação danosa" e "conteúdos limítrofes", deixando de recomendá-los.** Vídeos editados e adulterados são proibidos pela política de "mídia manipulada". **Remove conteúdos e canais que desrespeitam a política para eleições e passou a remover informações enganosas sobre a Covid-19, sobretudo desinformação médica, medicamentos perigosos e origem do vírus.** Nem toda a publicidade em torno da pandemia foi restrita. Possui um Grupo de Análise de Ameaças para identificar desinformação por governos.	Mídias manipuladas, como deepfakes, são marcadas, podem ter visibilidade reduzida, receber link com explicações ou serem removidas. **Conteúdos que possam enganar as pessoas sobre voto são proibidos, dentro da política de integridade nas eleições, que nos EUA sinaliza as postagens de candidatos.** No Brasil, não. **Anúncios políticos e de veículos de comunicação estatais não são mais permitidos.** Postagens de líderes globais podem seguir no ar. **Pós Covid-19, ampliou a definição de "dano" e remove ou rotula posts danosos.**

3. Promoção de informações e transparência

Facebook	Instagram	WhatsApp	YouTube	Twitter
Oferece ícone para buscar mais informações sobre fonte. **Sobre Covid-19, apresenta informações de autoridades de saúde nos resultados da busca (direcionando usuários para o site da OMS). Lançou o Centro de Informações sobre Coronavírus. Pessoas que interagiram com conteúdos "falsos nocivos" recebem mensagens.** Financia iniciativas de educação midiática, desenvolveu materiais e campanhas sobre o tema e tem parceria com agências de checagem. Apoia experiências jornalísticas visando ampliar a circulação de notícias profissionais. Concede dados para 60 pesquisadores no mundo.	Sistemas informatizados buscam "correspondência de imagens" para rotular conteúdos já classificados como falsos, inclusive quando eles inicialmente publicados no Facebook. **Disponibiliza textos de contraponto elaborados por verificadores de fatos, com "desmentidos" ou informações oficiais sobre o tema.** Buscas sobre vacinas direcionam usuários para o site da OMS. **No contexto da Covid-19, passou a disponibilizar informações do órgão no topo do feed dos usuários**, além de oferecer imagens sobre medidas de prevenção para compartilhamento.	As informações são dispersas e pouco acessíveis a um usuário comum. Sobretudo a partir das eleições de 2018, foram feitas parcerias institucionais para verificação de conteúdos e realização de pesquisas, mas o alcance e a efetividade delas são questionáveis. Da mesma forma, as ações de sensibilização do público amplo têm sido restritas. **As medidas de promoção de informações apenas foram ampliadas na pandemia, como criação de hub.** Falta transparência quanto às práticas adotadas pelo WhastApp, como proibição de criação ou remoção de contas, além de relatórios sobre suas medidas.	**Prioriza o que chama de "vozes autorizadas", como canais de jornalismo, nas buscas por informações sobre eventos e política e nos "próximos vídeos" sobre esses temas.** Exibe painéis de informação com dados de contexto de fontes autorizadas sobre assuntos históricos, cientificamente comprovados e teorias da conspiração. Checagens de fatos por editores independentes podem ser exibidas. Não há checagem sobre cada vídeo. **A Covid-19 ganhou sessões especiais em todas essas ferramentas. Nas eleições do EUA, oferece dados adicionais sobre candidatos.** Apoia o jornalismo digital no combate à desinformação e ações de educação para a mídia.	**Possui a ferramenta #ConheçaosFatos, que é exibida em pesquisas associadas a vacinas e Covid-19. Informações não-críveis sobre saúde não aparecem nas buscas, que priorizam fontes oficiais .** Adicionou aba especial sobre a pandemia na função #Explorar. No recurso "Eventos", também seleciona informações de credibilidade e disponibiliza no topo das timelines. Aplica rótulo em posts sobre Covid e 5G, encorajando a checagem. **Abriu a plataforma para acompanhamento de postagens sobre a Covid por desenvolvedores e pesquisadores.** Apoia iniciativas de jornalismo, checagem de fatos e de educação para a mídia.

4. Medidas correlatas

Facebook	Instagram	WhatsApp	YouTube	Twitter
Adotou política de proibição do discurso de ódio em 2018 e melhorou sistemas automatizados para identificar vídeos violentos e restringir transmissões ao vivo. **A plataforma derruba contas falsas, propagadoras de spam ou consideradas de "comportamento não autêntico coordenado".** Também reduz a circulação de contas difusoras de conteúdo de baixa qualidade (caça-cliques, anúncios maliciosos e chocantes). Passou a restringir o acesso à coleta de dados por apps de terceiros. **Como medida de proteção à integridade eleitoral, disponibiliza um arquivo de anúncios políticos com informações, além de rotular essas peças como "propaganda eleitoral".**	Diretrizes da Comunidade condenam o discurso de ódio, mensagens indesejadas enviadas repetidamente e ataques a pessoas com intenção de constrangê-las. **Contas falsas ou redes de contas enquadradas na categoria de "comportamento inautêntico" são derrubadas. Contas que registrem determinados números de violações em um período podem ser desativadas, com direito à apelação.** O comércio de avaliações falsas de usuários é proibido; e curtidas e comentários não autênticos para "impulsionar a popularidade" são removidos. **Tem ampliado medidas contra *bullying* e assédio.**	**Define usos envolvendo declarações falsas, incorretas ou enganosas como violações em seus Termos de Serviço. Sinaliza mensagens altamente encaminhadas com uma ou duas setas,** indicando maior viralização. Foi retirado também o botão que permitia encaminhamento rápido de conteúdos e inserida uma lupa ao lado das mensagens assinaladas com duas setas. Além disso, **utiliza ferramentas de tratamento de spam e aprendizagem avançada de máquinas para retirar mensagens automatizadas em massa e banir contas de usuários com comportamentos inadequados,** a exemplo do envio de mensagens em massa e da criação de múltiplas contas.	**Possui uma política própria contra discurso de ódio e outra contra assédio e ameaças.** Vídeos supremacistas são proibidos e aqueles que não poder ser caracterizados como de ódio, mas se aproximam disso sofrem restrições . Qualquer pessoa ou canal intimidando, perseguindo, desumanizando e incentivando comportamento violento, inclusive via comentários. pode ser banido da rede. Possui política para tratar de temas como o privacidade e contra a difamação, para recebimento de denúncias. **Não autoriza distribuiçaõ de conteúdos perigosos ou nocivos e tem uma política contra falsificação de identidade e engajamento falso. Restringe o acesso por sistema automatizado.**	**Política contra propagação de ódio proíbe conteúdos com ameaças violentas, estereótipos, medo e assédio. Comportamentos abusivos também não são permitidos.** Infratores podem ter que excluir conteúdo, passar tempo sem publicar ou interagir ou ser suspensos totalmente. Combate o spam e a automação mal-intencionada, monitorando e desafiando atividades suspeitas das contas, como por meio da verificação de identidades. **O foco é barrar contas falsas, o engajamento não autêntico e atividades coordenadas para influenciar artificialmente conversas.** Divulga arquivos de posts de atividades coordenadas apoiadas por governos.

Análise comparada

As plataformas digitais não apresentam política e processos estruturados sobre o problema da desinformação e desenvolvem ações pontuais e reativas no combate ao fenômeno.

Nenhuma das empresas relatou ter conceito e estrutura específica para abordar a questão da desinformação, o que pode dificultar a coordenação das iniciativas.

Em relação à moderação de conteúdo desinformativo, a verificação de conteúdos, principalmente por agências externas, é prática presente em boa parte das plataformas, a partir de suas categorizações.

A complexidade de analisar os “tons de cinza” entre um e outro extremo enseja riscos de avaliações questionáveis, razão pela qual **a verificação deveria contar com mecanismos de devido processo** para a mitigação de abusos e erros.

A extensão das medidas sobre desinformação aos anúncios é fundamental.

Análise comparada

As plataformas resistiram em remover conteúdos desinformativos como o fazem em outras categorias, a partir do que é unilateralmente definido em suas diretrizes. Tal postura começou a mudar no contexto da pandemia do novo coronavírus.

Nesse contexto, a admissão de situações excepcionais de retirada em casos de risco evidente de danos graves parece uma possibilidade razoável, desde que conectada a regras de devido processo que permitam a contestação, a avaliação dos recursos por pessoas e a reparação em caso de erro na moderação aplicada, o que não é garantido em nenhuma das plataformas analisadas.

Do ponto de vista da informação sobre como lida com conteúdos desinformativos, chamou atenção a baixa transparência das plataformas. Não há também avaliação da efetividade do que tem sido implementado.

Análise comparada

No caso das eleições, os riscos são ainda maiores pela manutenção de práticas como impulsionamento, tratamento de dados sem transparência, ausência de acompanhamento do Judiciário etc. Dois anos depois de Bolsonaro, os riscos persistem.

O problema da desinformação precisa ser efetivamente reconhecido, comunicado e enfrentado pelas plataformas, o que passa pela revisão da estrutura e do modelo de negócios dessas empresas.

Obrigada!
helena.martins@ufc.br