21/05/2018 10:01 - Facebook remove 2,5 milhões de posts com discurso de ódio em 6 meses

O Facebook retirou do ar 2,5 milhões de publicações que foram identificadas como contendo discurso de ódio no primeiro semestre do ano. A informação foi divulgada no relatório de transparência da plataforma, publicado pela primeira vez na semana passada. O documento traz os resultados das ações de moderação de conteúdo praticadas pela empresa, como o monitoramento e a exclusão de mensagens publicadas.

A avaliação é feita com base em diretrizes estabelecidas pela companhia. Segundo elas, discurso de ódio é considerado "um ataque direto a pessoas com base no que chamamos de características protegidas: raça, etnia, nacionalidade, filiação religiosa, orientação sexual, sexo, gênero, identidade de gênero e doença ou deficiência grave", além do status migratório. "Ataques" são "discursos violentos ou degradantes, declarações de inferioridade ou incentivo à exclusão e segregação".

O Facebook também excluiu 3,5 milhões de conteúdos violentos. Estes são definidos nas diretrizes como uma mensagem "que exalte a violência ou celebre a humilhação ou o sofrimento de outras pessoas". São permitidas publicações com imagens explícitas em alguns casos mas, segundo a empresa, "para ajudar as pessoas a gerar conscientização sobre algumas questões".

O monitoramento de conteúdo do Facebook identificou e derrubou 21 milhões de conteúdos de nudez ou pornografia. A empresa estima que a cada 10 mil publicações, entre 7 e 9 traziam algum tipo de conteúdo que violava os padrões sobre nudez ou pornogafia.

A moderação também busca contas falsas. De acordo com o relatório, no primeiro semestre foram derrubados 583 milhões de perfis deste tipo. O número representa 26,5% do total de usuários que a plataforma tem (2,2 bilhões, segundo dados de abril). Contudo, não necessariamente as contas já existiam. De acordo com o documento, a maioria dos perfis considerados falsos é excluída minutos após a criação.

Automatização

Um dos pontos exaltados pelo Facebook em seu relatório é a atuação de seus sistemas para identificar os conteúdos violadores de suas regras para exclusão. No caso das publicações com nudez e pornografia, 96% foram marcadas pela tecnologia da plataforma. Nas mensagens com imagens de violência, o índice ficou em 86%. Já nos conteúdos com discurso de ódio, a proporção de mensagens sinalizadas pelo sistema da companhia cai bastante, ficando em 38%.

"Tecnologias como a inteligência artificial, que embora seja promissora, ainda está longe de ser efetiva para a maioria dos conteúdos de baixa qualidade, ja que uma análise do contexto também é muito importante. Por exemplo, a inteligência artificial não é boa o suficiente para determinar se alguém está proclamando ódio ou se está descrevendo uma situação ocorrida consigo mesma para gerar conscientização sobre o assunto", disse Guy Rosen, vice-presidente de gerenciamento de produto, em texto publicado no site oficial da empresa.

Fonte: Redação Notícias RO