Twitter tests Safety Mode to block internet trolls

September 1 2021

The feature will initially be tested by a small number of English-speaking users, Twitter said.

Twitter on Wednesday announced it is testing a new feature that automatically blocks hateful messages, as the US site comes under

increasing pressure to protect its users from online abuse.

Users who activate the new Safety Mode will see their "mentions" filtered for seven days so that they don't see messages flagged as likely to contain hate speech or insults.

The feature will initially be tested by a small number of English-speaking users, Twitter said, with priority given to "marginalized communities and female journalists" who often find themselves targets of abuse.

"We want to do more to reduce the burden on people dealing with unwelcome interactions," Twitter said in a statement, adding that the platform is committed to hosting "healthy conversations".

Like other social media giants, Twitter allows users to report posts they consider to be hateful, including racist, homophobic and sexist messages.

But campaigners have long complained that holes in Twitter's policy allow violent and discriminatory comments to stay online in many cases.

The platform is being sued in France by six anti-discrimination groups that accuse the company of "long-term and persistent" failures to block hateful comments.

Safety Mode is the latest in a series of features introduced to give Twitter users more control over who can interact with them. Previous measures have included the ability to limit who can reply to a tweet.

Twitter said the new feature was a work in progress, mindful that it might accidentally block messages that were not in fact abusive.

"We won't always get this right and may make mistakes, so Safety Mode

autoblocks can be seen and undone at any time in your Settings," the company said.

To assess whether a message should be auto-blocked, Twitter's software will take cues from the language as well as previous interactions between the author and recipient.

Twitter said it had consulted experts in online safety, mental health and human rights while building the tool.

ARTICLE 19, a UK-based digital rights group that took part in the talks, called the feature "another step in the right direction towards making Twitter a safe place to participate in the public conversation without fear of abuse".

The announcement came after Instagram last month unveiled new tools to curb abusive and racist content, following a slew of hateful comments directed at footballers after the Euro championship.