FRA Press Release
Vienna, 29 November 2023

EMBARGO: 29 November 2023 at 06:00 CET

Online hate: we need to improve content moderation to effectively tackle hate speech

Abusive comments, harassment and incitement to violence easily slip through online platforms' content moderation tools, finds a new report from the EU Agency for Fundamental Rights (FRA). It shows that most online hate targets women, but people of African descent, Roma and Jews are also affected. A lack of access to platforms' data and understanding of what constitutes hate speech hampers efforts to tackle online hate. FRA calls for more transparency and guidance to ensure a safer online space for all.

FRA's online content moderation report looks at the challenges of detecting and removing hate speech from social media.

It highlights that there is no commonly agreed definition of online hate speech. Online content moderation systems are also not open to researchers' scrutiny. This makes it difficult to get a full picture of the extent of online hate and hampers efforts to tackle it.

FRA's analysis of posts and comments published on social media platforms between January–June 2022 reveals:

- **Widespread online hate -** out of 1,500 posts already assessed by content moderation tools, more than half (53%) are still considered hateful by human coders.
- **Misogyny -** women are the main targets of online hate across all researched platforms and countries. Most hate speech towards women includes abusive language, harassment, and incitement to sexual violence.
- **Negative stereotyping -** people of African descent, Roma and Jews are most often targets of negative stereotyping.
- **Harassment –** almost half (47%) of all hateful posts are direct harassment.

To tackle online hate, the EU and online platforms should:

- **Provide safer online space for all** – to prevent online hate, platforms should pay particular attention to protected characteristics like gender and ethnicity in their content moderation and monitoring efforts. Very large online platforms, such as X (formerly Twitter) or YouTube, should include misogyny in their risk assessment and mitigation measures under the Digital Services Act (DSA). All EU member states should also ratify the Istanbul Convention to better protect women online.
- **Provide more guidance –** it is not always clear what is considered hate speech and what is protected under freedom of speech. The EU and national regulators should provide more guidance on identifying illegal online hate.

- **Capture all forms of online hate** – to ensure that different types of online hate are detected, the European Commission and national governments should create and fund a network of trusted flaggers, involving civil society. The police, content moderators and flaggers should be properly trained, to ensure that platforms do not miss or over-remove content.
- **Test technology for bias -** providers and users of automated content moderation tools should test their technology for bias to protect people from discrimination, as FRA's previous report on bias in AI also highlighted.
- **Ensure access to data for independent research** – the European Commission should ensure that platforms' own risk assessments under the DSA are complemented by independent research. Only a variety of approaches and tests will provide a full picture of what type of hate is not properly identified and taken down, and what the impact on people's fundamental rights is.

The report covers four social media platforms (Reddit, Telegram, X, and YouTube) in Bulgaria, Germany, Italy and Sweden. FRA was not able to access data from Facebook and Instagram for this research.

Between January–June 2022, FRA collected almost 350,000 posts and comments based on specific key words. Human coders assessed about 400 random posts from each country to determine if they were hateful. 40 random posts were then assessed in more detail by coders and legal experts. This report shows the different types of hate speech found across the countries, target groups and platforms covered.

Quote from FRA Director Michael O'Flaherty:

"The sheer volume of hate we identified on social media clearly shows that the EU, its Member States, and online platforms can step up their efforts to create a safer online space for all, in respect for human rights including freedom of expression. It is unacceptable to attack people online just because of their gender, skin colour or religion."

For more, please contact: media@fra.europa.eu / Tel.: +43 1 580 30 653