SAMARITANS

Implementing effective moderation for self-harm and suicide content

This information sheet provides guidance for sites and platforms hosting user-generated content on implementing effective moderation for self-harm and suicide content. All sites and platforms must moderate content to ensure that policies are upheld, and users are protected from material that could be potentially harmful.

Content moderation can be achieved through human moderation and using artificial intelligence (AI) approaches. AI approaches should be adopted if they are cost-effective and offer proportionate solutions to increase the speed and efficiency of moderation. However, companies should never rely solely on AI approaches. Instead, they should be used to prioritise content for human moderation.

Human moderation

Human moderation can be an effective way of detecting and responding to self-harm and suicide content online. All sites using human moderators must ensure that moderators receive high quality training and support. See **Guidance for supporting the wellbeing of moderators** for more information.

Benefits include:

- **More accurate than AI** at understanding slang, sarcasm, acronyms and the nuance around self-harm and suicide language.
- **Can understand the context of content and react to emerging trends.** For example, an image of a railway line may not initially seem harmful but could be captioned in a way that refers to suicide.

Potential limitations:

- **Inefficient for high volumes of content,** as it can be quite time intensive.
- **Can be inconsistent,** if moderators are not provided with adequate training.
- **Content can impact on moderator wellbeing,** if they are reviewing large volumes of content to relating to self-harm and suicide.

Artificial intelligence

Artificial intelligence can be an effective mechanism for enabling early detection of self-harm and suicide content at scale, preventing wider distribution and enabling prompt interventions. AI approaches are particularly important for sites that have large quantities of content uploaded and shared regularly.

Examples of AI include:

- **Text filters** to detect and remove content that contains particular words or themes, such as methods of suicide.
- **Username filters** to detect inappropriate usernames when users first register.
- **Triaging reported content** by prioritising and flagging reports for human moderation and closing false reports.
- **Image hashing** to detect and prevent the upload of images that are known to be harmful.

Implementing effective moderation for self-harm and suicide content

SAMARITANS

Benefits include:

- **Faster than human moderation,** at detecting harmful content at scale.
- **Can detect self-harm and suicide content during unsociable hours,** such as overnight.
- **Removes self-harm and suicide content before it is published,** limiting the number of users who may view it.
- **Flags potentially harmful content to human moderators for review,** preventing moderators from having to review all posted content.
- **Reveals the prevalence of self-harm and suicide content,** allowing sites to better understand the type of content being posted and ensure mechanisms are in place to support users and respond to content safely.
- **Decreases the amount of harmful content to be viewed by moderators.**

Potential limitations:

- **It can't always detect the nuance in language** around self-harm and suicide.
- **Can be 'tricked' by users** who may bypass filters through the use of emojis or change their language in order to post about topics that break the community guidelines.
- **False positives** can arise, by incorrectly identifying content as harmful self-harm and suicide content which can impact on user experience.
- **False negatives** can arise, by failing to detect content that would be considered harmful.

For guidance on using AI to keep online communities safe, see:

- **Cambridge Consultants for Ofcom (2019) Use of AI in online content moderation [accessed 24 March 2020].**
- **UK Government Office for Artificial Intelligence (2019) A guide to using artificial intelligence in the public sector [accessed 24 March 2020].**

How Samaritans can help you

For further support and advice on responding to self-harm and suicide content online, please see our website or contact the Samaritans' Online Harms Advisory Service at onlineharms@samaritans.org Email monitored Monday–Friday, 9am–5pm.