APR 10
Mack DeGeurin

AI-powered filter tools could help social audio apps address harmful content in real-time

Article

Intel is testing out a new AI-powered tool used to identify and automatically remove offensive words in an audio chat. Called “Bleep,” the gaming-focused software reportedly uses the AI

processing power on Intel-powered PCs to remove content deemed harmful within categories including LGBTQ+ hate, aggression, misogyny, name-calling, racism and xenophobia, sexually explicit language, swearing, and white nationalism, per PC Mag. Rather than remove harmful content entirely, Bleep appears to work on a sliding scale, with users able to select whether they want none, some, most, or all of any particular content. The early version also offers a single on-off switch for "N-Word."

Toxic speech is a persistent problem in gaming that alienates users and causes some to disengage entirely. According to a 2020 Anti-Defamation League survey of gamers identifying as LGBTQ+, Jewish, Muslim, African American, and Hispanic / Latinx, 28% who experienced online harassment said they avoided certain games based on their reputation for abuse, while 22% claimed to have stopped playing certain games altogether. Similar concerns over harassment occur in the esports field. 40% of respondents surveyed in a 2020 Foley and Lardner report said they thought cyberbullying within games posed the greatest risk to the esports industry, an 8% increase from the previous year.

The issue of audio content extends far beyond video games and will likely become an even greater issue thanks to the rise of live social audio apps like Clubhouse. Social audio apps started gathering steam when Discord launched in 2015, but are currently experiencing a watershed moment thanks to the sudden popularity of Clubhouse. Between January and February this year, monthly installs of Clubhouse worldwide increased from 2.4 million to 10.1 million, per Sensor Tower data. That rapid popularity wasn't lost on competitors: Twitter, Facebook, Spotify, and even LinkedIn are all currently developing or have released their own clearly Clubhouse-inspired social audio apps.

Real-time AI audio filters could help solve the hard problem of moderating audio content, but they carry the baggage of AI's own problems with bias and censorship. Real-time audio moderation tools are still in their infancy and lag behind text and image-based moderation solutions. So far, social audio apps have attempted to address moderation concerns by recording user audio and searching through them to adjudicate content violations. Twitter keeps an audio recording of Spaces rooms for 30 days or longer in order to review Twitter Rules violations. Clubhouse, on the other hand, reportedly deletes recordings if a session ends and there isn't an immediate complaint lodged. Recording social audio raises several problems though. This moderation method both runs counter to the supposedly ephemeral nature of social audio and is limited to addressing violations only afterthey've occurred. A real-time AI moderation filter like the one Intel is building could help solve both of these

problems but inevitably subjects itself to another set of problems regarding bias and censorship.