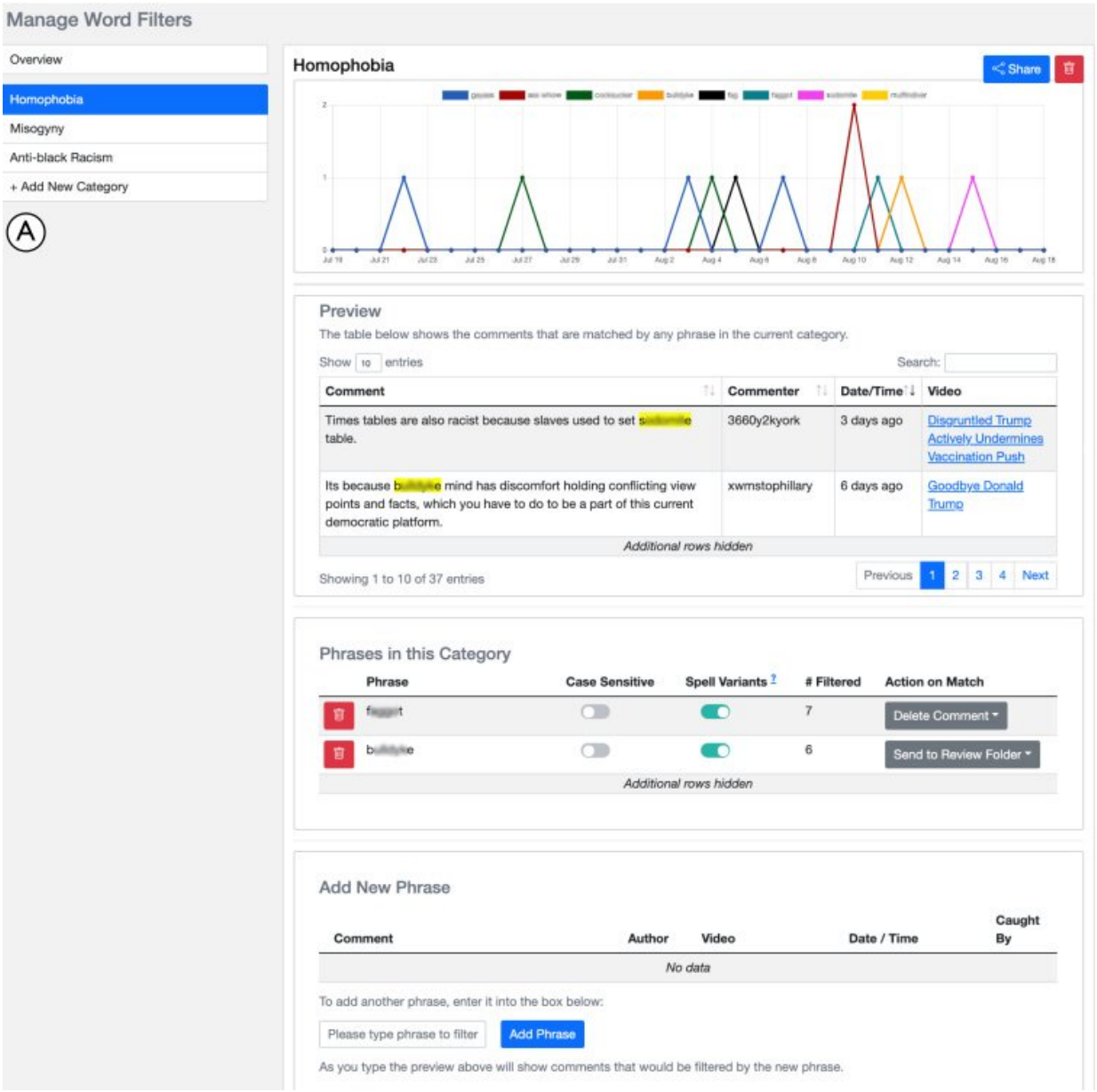
Researcher creates free comment moderation software for YouTube

May 3 2022



A screenshot of FilterBuddy's category page. Credit: Shagun Jhaver

As an expert in building a safer and fairer internet, Rutgers Assistant Professor Shagun Jhaver long suspected that digital content creators from minority groups suffered disproportionate online harassment. But when he heard about a 35-year-old Brazilian YouTuber who suffered a nervous breakdown and was hospitalized after a barrage of digital hate, he decided to do something about it.

Jhaver, who teaches in the Department of Library and Information Science at the Rutgers School of Communication and Information, built a comment moderation tool for YouTube to ensure that users like the video host in Brazil—who eventually shut down his channel—can work without fear of attack. Called FilterBuddy, the free, open-source tool is designed to give professional content creators—anyone who makes a living publishing on social media platforms—the power to protect themselves and their audiences.

To develop FilterBuddy, Jhaver investigated content moderation tools that come standard with platforms like YouTube, Facebook, Twitter and Twitch. In interviews with 19 creators from the Americas, Asia, Europe and the Middle East, Jhaver and colleagues from the University of Washington discovered that content creators, especially from underrepresented communities, are often attacked for their appearance, beliefs and background. "They face disproportionate abuse, so they have disproportionate need," he said.

Jhaver also learned the keyword filtering tools that platforms provide are hard to use and largely ineffective. "These tools are very rudimentary, very hard to scale up and on many platforms, you can't even search for the filtering rules that you have configured."

The findings were recently published in the *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Based on these data, Jhaver and colleagues created a system for YouTube creators to support better authoring, maintenance and auditing of word filters, which can tag and quarantine potentially offensive comments before they become public. FilterBuddy also includes an interactive filter capture preview (to test the filter's effectiveness before deployment), the ability to build atop filters created by others and time-series graphs and tables to understand what comments are caught by what filters over time.

Ultimately, Jhaver hopes FilterBuddy will inspire platforms to develop more powerful resources in house—and to make it easier for third-party apps to connect.

"We used standard development tools to create FilterBuddy, but the features are so appreciated by the community and so desperately needed," he said. "The fact that we built this service on a shoestring budget demonstrates that at the moment, the platforms aren't paying enough attention to creators' needs."

Most importantly, Jhaver said, he hopes his software will contribute to a safer online space—for users and for creators. "No one should have to abandon their work because of abusive comments," he said.

More information: Shagun Jhaver et al, Designing Word Filter Tools for Creator-led Comment Moderation, *CHI Conference on Human Factors in Computing Systems* (2022). DOI: 10.1145/3491102.3517505

Provided by Rutgers University