Hate speech in social media: How platforms can do better

February 18 2022, by Dory Knight-Ingram

ADL CENTER FOR TECHNOLOGY & SOCIETY

The Belfer Fellowship Series

VERY FINE PEOPLE

What Social Media Platforms Miss About White Supremacist Speech

Report by:
Libby Hemphill,
University of Michigan

Credit: University of Michigan

With all of the resources, power and influence they possess, social media platforms could and should do more to detect hate speech, says a University of Michigan researcher.

In a report from the Anti-Defamation League, Libby Hemphill, an associate research professor at U-M's Institute for Social Research and an ADL Belfer Fellow, explores social media platforms' shortcomings when it comes to white supremacist speech and how it differs from general or nonextremist speech, and recommends ways to improve automated hate speech identification methods.

"We also sought to determine whether and how white supremacists adapt their speech to avoid detection," said Hemphill, who is also a professor at U-M's School of Information. "We found that platforms often miss discussions of conspiracy theories about white genocide and Jewish power and malicious grievances against Jews and people of color. Platforms also let decorous but defamatory speech persist."

How platforms can do better

White supremacist speech is readily detectable, Hemphill says, detailing the ways it is distinguishable from commonplace speech in social media, including:

- Frequently referencing racial and ethnic groups using plural noun forms (whites, etc.)
- Appending "white" to otherwise unmarked terms (e.g., power)
- Using less profanity than is common in social media to elude detection based on "offensive" language
- Being congruent on both extremist and mainstream platforms
- Keeping complaints and messaging consistent from year to year

- Describing Jews in racial, rather than religious, terms

"Given the identifiable linguistic markers and consistency across platforms, social media companies should be able to recognize white supremacist speech and distinguish it from general, nontoxic speech," Hemphill said.

The research team used commonly available computing resources, existing algorithms from machine learning and dynamic topic modeling to conduct the study.

"We needed data from both extremist and mainstream platforms," said Hemphill, noting that mainstream user data comes from Reddit and extremist website user data comes from Stormfront.

What should happen next?

Even though the research team found that white supremacist speech is indentifiable and consistent—with more sophisticated computing capabilities and additional data—social media platforms still miss a lot and struggle to distinguish nonprofane, hateful speech from profane, innocuous speech.

"Leveraging more specific training datasets, and reducing their emphasis on profanity can improve platforms' performance," Hemphill said.

The report recommends that social media platforms: 1) enforce their own rules; 2) use data from extremist sites to create detection models; 3) look for specific linguistic markers; 4) deemphasize profanity in toxicity detection; and 5) train moderators and algorithms to recognize that white supremacists' conversations are dangerous and hateful.

"Social media platforms can enable social support, political dialog and productive collective action. But the companies behind them have civic responsibilities to combat abuse and prevent hateful users and groups from harming others," Hemphill said. "We hope these findings and recommendations help platforms fulfill these responsibilities now and in the future."

More information: Very Fine People: What Social Media Platforms Miss About White Supremacist Speech: www.adl.org/language-of-white-supremacy

Provided by University of Michigan

Citation: Hate speech in social media: How platforms can do better (2022, February 18) retrieved 2 May 2024 from https://phys.org/news/2022-02-speech-social-media-platforms.html