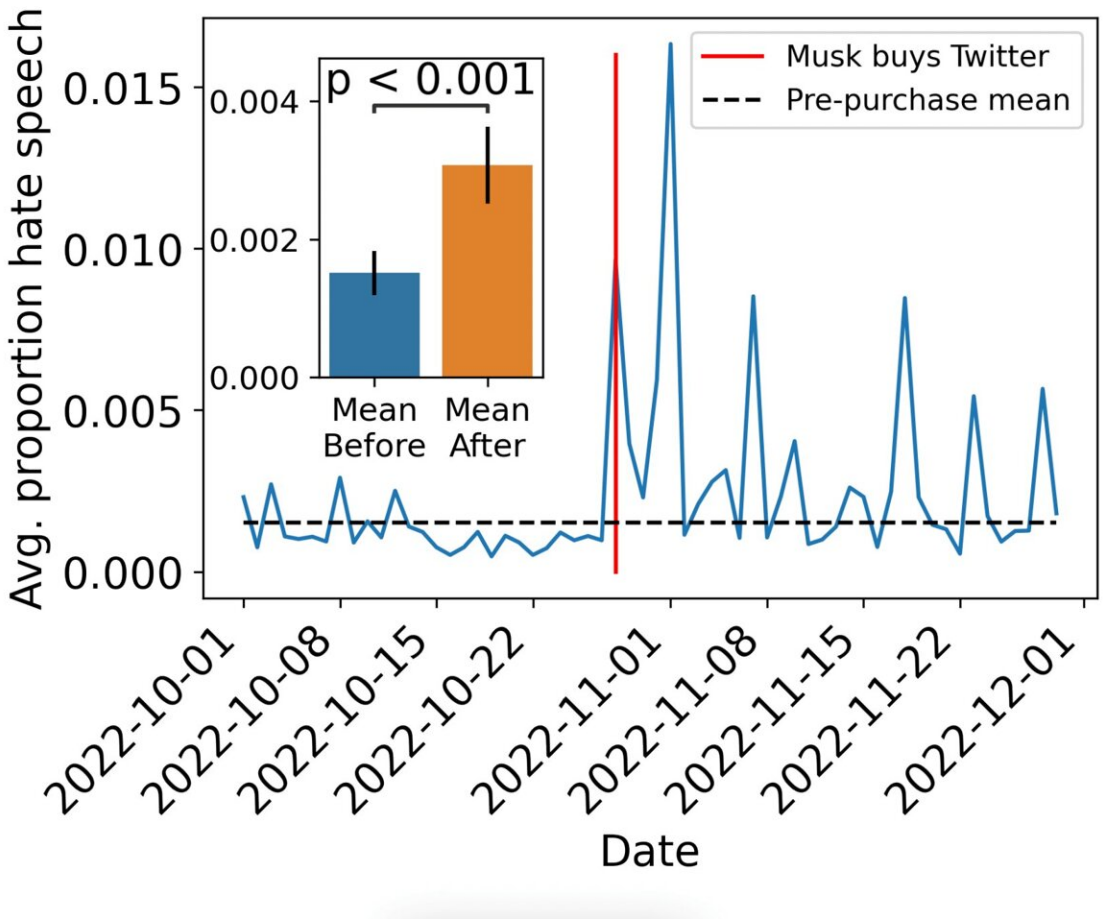
Analysis finds hate speech has significantly increased on Twitter

April 24 2023, by Julia Cohen



Hateful users increased their hate speech after Musk bought Twitter. Credit: University of Southern California

Computer scientist Keith Burghardt at the Information Sciences Institute (ISI), a research institute of the USC Viterbi School of Engineering, has been studying social media for five years, and specifically studying online hate for the last year.

Among other things, he has quantified how hateful online communities in Reddit increase the hate speech of new members; developed techniques to detect hateful subreddits and determine how subreddit members' early interactions affect their activity within the group; and found ways to understand how online extremism occurs and predict anti-vaccine users on Twitter.

His latest paper, "Auditing Elon Musk's Impact on Hate Speech and Bots," quantifies hate and bots on Twitter. It has been peer-reviewed and accepted as a poster paper in the 2023 International AAAI Conference on Web and Social Media (ICWSM), to be held June 5–8 in Limassol, Cypress.

When Elon Musk purchased Twitter on October 27th, 2022, two of his stated goals were to have less restrictive content moderation and to remove spam bots. In this paper, Burghardt and his fellow researchers set out to look at the impact of the former, and the success of the latter.

Less moderation means more hate

Previous research has shown that lighter moderation is associated with increased hate speech on social media platforms, therefore Burghardt's team hypothesized that hate speech on Twitter would increase following Musk's acquisition. The question was how to quantify this.

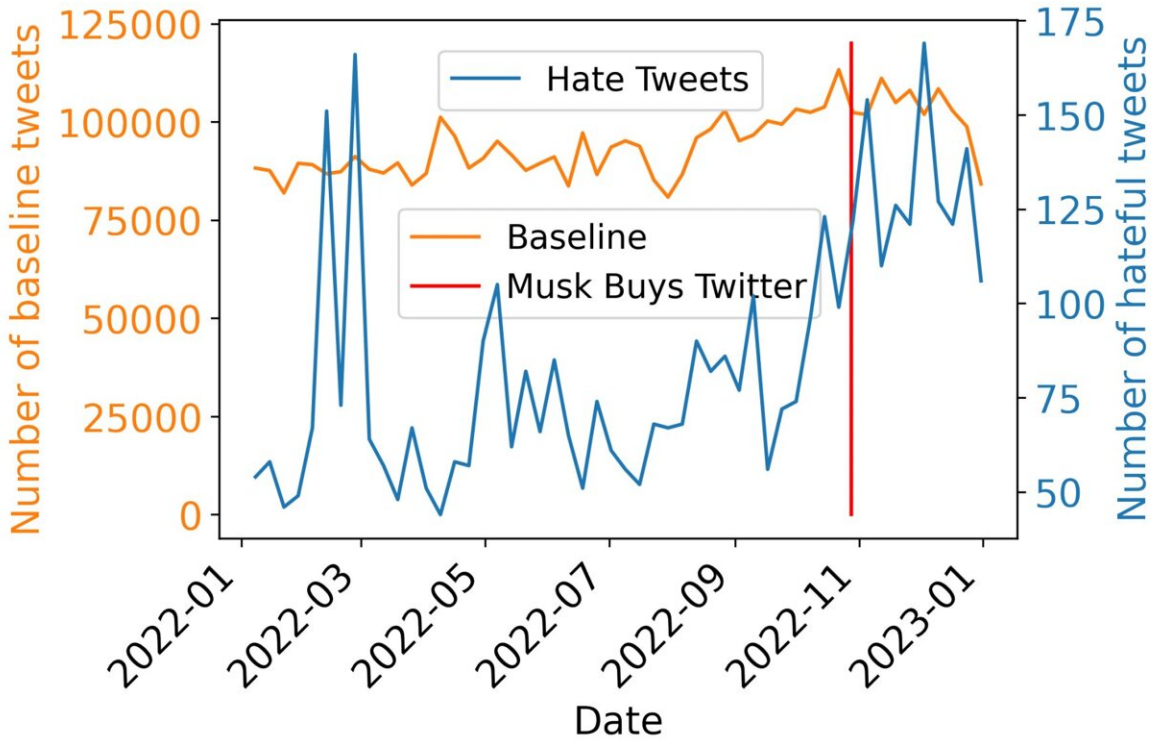
"We first had to create a set of words that we could determine as being hateful. Our aim was to find words that were relatively high precision, meaning that if people are using these words, it's unlikely they're being

used in a non hateful manner," said Burghardt. He and his team originally created this methodology to understand hate speech on Reddit. Here, they have applied the same methodology to Twitter. It provided them with 49 hate keywords (WARNING: Contains offensive terms).

"In addition, to weed out non-hateful or sexual uses of these words," Burghardt said, "we only considered tweets that an AI tool, Perspective API, judged were toxic and not sexual in nature."

Since the presence of a hate keyword doesn't necessarily mean a tweet is hateful, the team used Perspective API (application programming interface), a free and publicly available API that uses machine learning to identify toxic conversations. Though originally trained on New York Times data, it has been verified on a variety of social media platforms including Twitter. The team filtered the hateful users' tweets on Perspective API's toxicity metric, which defines a toxic comment as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion."

Using this methodology, the team extracted timelines of a sample of users who posted hateful tweets one month before and after Musk's purchase and measured their daily rates of hate speech during the same time period. This gave them a measure of the degree to which already hateful tweeters changed their level of hate.



Overall hate speech on Twitter increased after Musk bought the platform. (The spike in March, 2022 coincides with the Canada convoy protests.). Credit: University of Southern California

They found that the proportion of hate words in hateful users' tweets increased after Musk bought Twitter. And the average daily hate speech of hateful users nearly doubled.

Then, they measured the overall volume of hateful tweets throughout 2022. To ensure fluctuations in hate speech were not reflective of fluctuations in overall user activity, they also sampled a baseline set of tweets collected during the same time intervals using benign keywords (e.g., "thing").

They found that the daily average overall usage of hate keywords on Twitter nearly doubled after Musk bought Twitter.

Now, about those bots

Musk highlighted a reduction in bot accounts as one of his goals for the platform. As such, Burghardt and his fellow researchers hypothesized that the prevalence of bots would decrease after his purchase of Twitter.

While bots are not necessarily connected to hate speech, this was a continuation of the audit they were performing, assessing and quantifying the stated goals Musk had for the social media platform.

It should be noted however, that it has been shown that bots pose risks to social media platforms by spreading misinformation and hate.

The team used the Botometer API, developed, among others, by Emilio Ferrara, team leader at ISI and a professor of communication and computer science at USC Annenberg and at the USC Viterbi Department of Computer Science. The Botometer considers over 1,000 features to predict if a Twitter account is a bot and what type of bot (e.g., spam bot versus a fake follower).

The team collected Botometer scores of a sample of random accounts both before and after Musk bought Twitter. Burghardt said, "We found some types of bots became more prevalent, some types became a little less prevalent. But overall there was no significant change in the amount of bots."

"I've always been interested in understanding influence. That led to trying to understand if there are ways that people become influenced to do things that are harmful," said Burghardt. This audit of Twitter is a piece of that work, as are his other current projects: predicting if users

will join hate groups; and how external events can push people to have extremist views.

Provided by University of Southern California