FEBRUARY 2024

SOCIAL SCIENCE JOURNAL FOR TEENS

How does fear speech spread on social media?

Authors:
Punyajoy Saha, Kiran Garimella, and others

Associate Editors:
Miranda Wilson and Rachel Watson

Abstract

Are you on social media? There are lots of fun things on social media, but there are also some bad things. Social media platforms try to manage hurtful content like hate speech. But there are other types of hurtful content that are harder to manage. One of these is fear speech. This is when people use words to create fear about a group.

We wanted to know how common fear speech is on social media. So we created a model to identify fear speech and hate speech from 21 million social media posts. Then we examined things like how connected users were, how they interacted with people who didn't post hurtful things, and the characteristics of their posts. We found that users who post a lot of fear speech are more connected to others. Fear speech posts also have a lot more activity than hate speech posts. This could contribute to the spread of fear speech on social media.

Introduction

Have you ever been on **social media** and seen something that seems hurtful? No one wants to see that in their feed. Many social media platforms, like Facebook and Twitter/X, try to **moderate** harmful content. They have computer programs or use trained people to remove hurtful posts. **Hate speech** is one common type of harmful content that often gets removed. Hate speech targets an individual or a group of people. Toxic words are used and the intent is to hurt the target. Toxic words are usually based on things like race, religion, ethnicity, disability, sexual orientation, or gender.

Unfortunately, newer forms of harmful content are spreading on social media. One example is **fear speech**. This is when someone uses words to create fear about a group of people. Fear speech uses a series of logical arguments to create fear. It doesn't use toxic words like hate speech does. Sometimes fear speech can even encourage violence in the real world. Fear speech is much harder than hate speech to recognize. This makes it hard to moderate on social media.

We wanted to know how common fear speech is on social media. We also wanted to know more about how it spreads. Anything we can learn will help us moderate social media better. This will make social media a safer place for everyone.

Illustration of a Western dragon. **Image**: JL G from Pixabay

Fictional examples of hate and fear speech

Hate speech: Dragons are stupid, heartless creatures. They steal food and children. You can't trust them. Those worms will turn on you in a heartbeat!

Fear speech: Dragons have killed hundreds of our people. They won't stop coming after our food and our families. This is a fight for survival, and we have to protect our homes.

Methods

We created a dataset of posts from Gab.com from October 2016–July 2018. Gab.com is a social media platform that doesn't moderate harmful content very much. We used 4 experts and 103 **crowdsourced** workers to classify 10,000 posts. We labeled them as fear speech, hate speech, or normal speech. We then used these results to develop a model to classify the 21 million posts in our dataset.

We used the data to investigate:

→ **Extreme users**: We calculated how many users (top 10%) posted the most fear speech and hate speech. We called them extreme users. We then looked at the number of followers they had and how many people were following them.

→ Interactions between extreme and **normal users**: We looked at how many normal followers the extreme users had. We used a subset of normal users to look at how often they received and reposted material from extreme users.

→ Believability of posts: We used crowdsourced judges to read pairs of fear speech and hate speech posts. We then asked which post they believed more.

→ Activity on posts: We looked at hashtags, **URL** use, likes, replies, and reposts of fear and hate speech posts.

Finally, we used our model to look at more than 4 million tweets on Twitter between 2016 and 2022. We also used our model to look at almost 200,000 posts from Facebook over the same time.

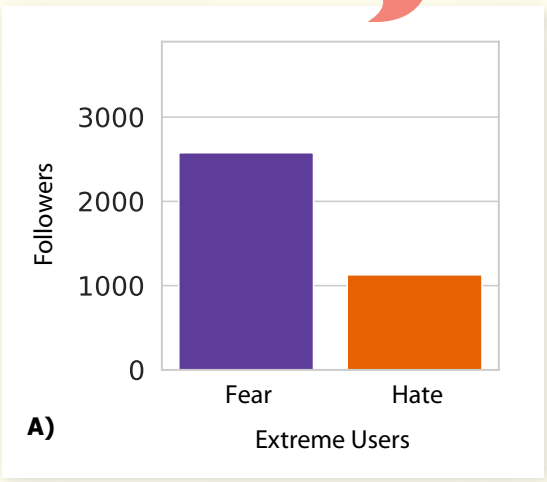
Results

We identified 476 extreme fear users and 478 extreme hate users. We found the extreme fear users have more total and normal users following them than extreme hate users do. They also follow more total people (Fig. 1). We saw that normal users received fewer posts from extreme fear users than extreme hate users. But they reposted more of what they received from extreme fear users.

We found that fear speech posts were more believable than hate speech posts. This is true even though fear speech posts often had unconfirmed information in them or fake URLs. We found that fear speech posts adopted hashtags from normal users faster than hate speech posts. We also saw that posts by extreme fear users got more likes, replies, and reposts than posts from extreme hate users.

Finally, we found that about 10% of Twitter posts and 4% of Facebook posts contained fear speech (Fig. 2). Both platforms had low levels of hate speech.

Compare the number of followers extreme users have. Based on this, do you think fear speech or hate speech is likely to spread more quickly?



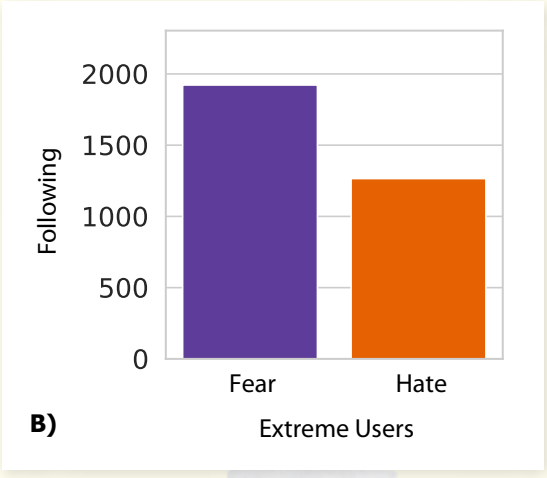


Figure 1:
(A) The average number of followers extreme users have and (B) the average number of people that extreme users follow.

Please see Figure 2 on Page 3

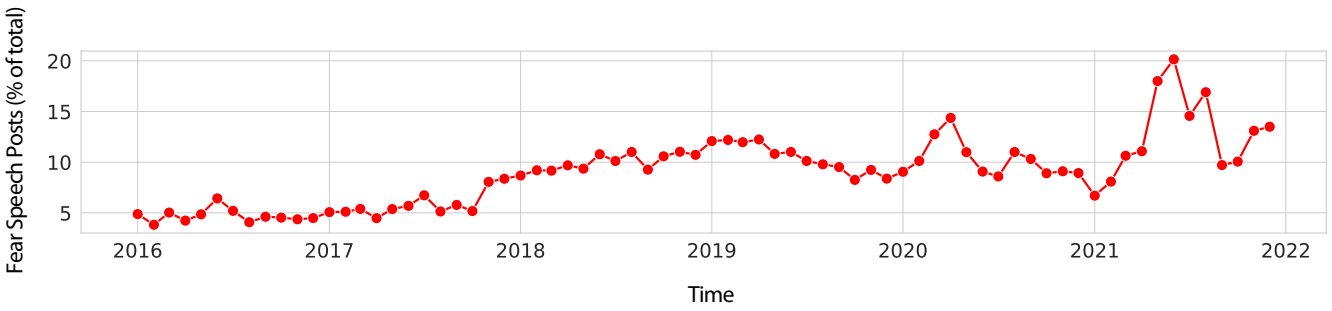


Figure 2:
Trend of fear speech on Twitter from 2016 to 2022.

Discussion

Fear speech is a common occurrence on social media. Extreme fear users are better connected than extreme hate users. They also engage with normal users more. Normal users are more likely to repost things from extreme fear users. This is because fear speech is more believable than hate speech. Fear speech uses arguments and misinformation in subtle ways that create fear. It is possible that these types of posts could cause violence in the real world. The danger is that fear speech seems to spread more quickly than hate speech through social media.

Unfortunately, fear speech will be hard to moderate or ban from social media. There are no toxic words for computer programs to recognize. And people need to be trained to recognize fear speech. We have several suggestions to help decrease the impact of fear speech on social media.

1. Educate social media users about what fear speech is and how to recognize it.
2. Check the information in fear speech posts. Learn how to recognize exaggerated arguments and false information.
3. Practice making arguments to oppose fear speech posts. These should focus on reducing potential violence.

Conclusion

We use social media for a lot of things. We use it to keep in touch with people, to get our news, to shop…and sometimes just to watch cat videos! Unfortunately, some people use social media to post hurtful things. You can help make social media a safer place for everyone by:

- Learning about the difference between hate speech and fear speech and how to recognize them. You can teach your friends, too!
- Reporting harmful or hurtful posts to social media moderators or telling trusted adults.
- Recognizing that not everything on social media is true. Make sure to check your sources and be aware of misinformation.

Acknowledgement: This article's adaptation was supported by the GM Foundation.

Glossary of Key Terms

Crowdsource - getting a large number of people to help with a task. The people could be paid or unpaid, depending on the job.

Extreme users - users who were in the top 10% of those who posted the most fear speech or hate speech posts. So, we called users who posted the most fear speech posts extreme fear users.

Fear speech - words that try to create fear of a target group based on things like race, religion, ethnicity, sexual orientation, disability, or gender. They use logical language to make their point. They don't use toxic or derogatory words.

Hate speech - words that show hatred toward an individual or group. It is intended to be humiliating or insulting to the members of the group. It is based on things like race, religion, ethnicity, sexual orientation, or gender. Hate speech can be easy to identify because it often uses toxic language like name calling.

Moderate - controlling the posting of wanted and unwanted posts on social media. This can be done by computer programs or by trained people.

Normal users - people who never post hate speech or fear speech.

Social media - virtual communities and networks where people can interact. People can share, create, or exchange information over social media. Some popular examples include: Twitter/X, Facebook, Instagram, and TikTok.

URL - the address of a webpage.

Check your understanding

1. What is the difference between fear speech and hate speech?
2. What can social media platforms do to prevent harmful content on their sites?
3. When you use social media, you probably hope that lots of people interact with your content (like, share, comment, etc.). For this study, what things did we look at to see how many people might be exposed to fear speech and hate speech?
4. Why is fear speech more common on social media than hate speech?
5. Pretend you have a younger sibling who wants to start using social media. Write them a letter. What can you tell them about how to recognize fear and hate speech? What should they do if they see dangerous content on social media? What other advice do you have for them about using social media safely?

REFERENCES

Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Matthew, and Animesh Mukherjee (2023) *On the rise of fear speech in online social media.* Proceedings of the National Academy of Sciences.
https://www.pnas.org/doi/suppl/10.1073/pnas.2212270120

Julia Angwin (2023) "Few are addressing one of social media's greatest perils". New York Times.
https://www.nytimes.com/2023/05/06/opinion/fear-speech-social-media.html

United Nations: Hate speech
https://www.un.org/en/hate-speech

American Psychological Association: Social media brings benefits and risks to teens
https://www.apa.org/monitor/2023/09/protecting-teens-on-social-media