PNAS RESEARCH ARTICLE | COMPUTER SCIENCES

On the rise of fear speech in online social media

Punyajoy Saha[a,1], Kiran Garimella[b,1], Narla Komal Kalyan[a], Saurabh Kumar Pandey[a], Pauras Mangesh Meher[a], Binny Mathew[a], and Animesh Mukherjee[a]

Edited by Jeffrey Ullman, Stanford University (Retired), Stanford, CA; received July 22, 2022; accepted November 29, 2022

Recently, social media platforms are heavily moderated to prevent the spread of online hate speech, which is usually fertile in toxic words and is directed toward an individual or a community. Owing to such heavy moderation, newer and more subtle techniques are being deployed. One of the most striking among these is fear speech. Fear speech, as the name suggests, attempts to incite fear about a target community. Although subtle, it might be highly effective, often pushing communities toward a physical conflict. Therefore, understanding their prevalence in social media is of paramount importance. This article presents a large-scale study to understand the prevalence of 400K fear speech and over 700K hate speech posts collected from Gab.com. Remarkably, users posting a large number of fear speech accrue more followers and occupy more central positions in social networks than users posting a large number of hate speech. They can also reach out to benign users more effectively than hate speech users through replies, reposts, and mentions. This connects to the fact that, unlike hate speech, fear speech has almost zero toxic content, making it look plausible. Moreover, while fear speech topics mostly portray a community as a perpetrator using a (fake) chain of argumentation, hate speech topics hurl direct multitarget insults, thus pointing to why general users could be more gullible to fear speech. Our findings transcend even to other platforms (Twitter and Facebook) and thus necessitate using sophisticated moderation policies and mass awareness to combat fear speech.

fear speech | hate speech | social media | Gab | prevalence

Significance

Existential fear has always been a concern across human history and even transcends to the rest of the animal world. This fear is so deeply ingrained that even the slightest "knock" to it could spark a violent conflict among different groups. Here, we demonstrate how social media platforms are used to extensively mediate elements of existential fear as fear speech posts. Their nontoxic and argumentative nature makes them appealing to even benign users who in turn contribute to their wide prevalence by resharing, liking, and replying to them. Remarkably, this prevalence is far stronger than the more well-known hate speech posts. Our work necessitates consolidated moderation efforts and awareness campaigns to mitigate the harmful effects of fear speech.

Content moderation plays an important role in removing harmful and irrelevant posts (spam), thereby keeping the platforms safe. Social media companies like Facebook* and Twitter† have detailed guidelines as what is considered hateful in their platforms. These companies use such guidelines to appoint manual and automatic moderators to delete hateful posts/suspend hateful users‡. Subsequently, the research community has started putting consolidated efforts to automate and, thereby, scale up this moderation, creating better datasets and machine learning models to accurately detect hate speech. The datasets span across different platforms including Twitter (1, 2), Gab (3), Reddit (4), etc. Further, the models also range from simple ones like mSVM (5) to complex AI architectures like transformers (6).

While these advances are indeed encouraging, newer and more subtle forms of harmful content are inflicting the online world, which most often go unnoticed. One such form of malicious content is fear speech, which involves spreading fear about one or more target communities online and, eventually, the physical world. In this context, we note that existential fear can bias peaceful people toward extremism. In a controlled experiment (7), a group of Iranian students were found to support doctrines related to the understanding of the value of human life as opposed to a jihadist call for suicide bombing. However, when they were frightened about death, they subscribed toward the bomber, even expressing a desire to become a martyr themselves. From time to time, mortality salience polarizes an individual or a group to stick firmly to their own beliefs while demonizing others with opposing beliefs. This arises from the fear of endangerment of their own clan. For instance, while the fear that was generated due to the 9/11 incident was real, it also made Americans more vulnerable to psychological manipulation. In this context, Florette Cohen notes that "fear tactics have been used by politicians for years to sway votes."§ In a survey (8) conducted by Cohen et al., the authors asked the participants to think about the fear of death and then gave them statements from three fictitious political personalities. One of them was a charismatic who stressed in-group favoritism, the second,

*https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/.

†https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

‡https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/.

§https://www.psychologicalscience.org/news/releases/the-political-effects-of-existential-fear.html.

Author affiliations: [a]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India; and [b]School of Communication and Information, Rutgers University, New Brunswick, NJ 08901

Author contributions: P.S., K.G., B.M., and A.M. designed research; P.S., N.K.K., and S.K.P. performed research; P.S., N.K.K., S.K.P., and P.M.M. analyzed data; and P.S., K.G., B.M., and A.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: punyajoys@iitkgp.ac.in or garimell@mit.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2212270120/-/DCSupplemental.

Published March 6, 2023.

a technocrat presenting practical solutions to realistic problems, and, the third, preaching democratic values. When primed with the fear of death, the support for the fictional charismatic leader went up by eightfolds. With the advent of social media, it has become easier to propel the prevalence of such fear tactics.

In real life, elements of fear are often found to be associated with events of violence. The posts of the alleged attacker who shot worshippers at the Pittsburgh synagogue in October 2018 portrayed the HIAS[¶] as an organization supporting refugee invasion (9). Similarly, the shooter of the Christchurch event in 2019 released a manifesto—"Great Replacement." This manifesto contained elements of fear in the form of "nonwhites" replacing "whites" in the future (10). A recent mass shooting in Buffalo shooting (11) also denotes another such racially motivated attack. Such association is also well grounded in the literature of intergroup conflicts (12).

Fear is also used by politicians and media figures. Politicians in the United States (13) and European nations (14) portray immigration as an invasion and asylum seekers as dangerous. A viral poster in The Brexit campaign—"Breaking point"—shows nonwhites as invaders and as a danger to the British resources (15). Media figures like Tucker Carlson often cite low birth rates among Americans as a threat to cultural identity. Previous work on "fear speech" (16) also found similar themes in public political WhatsApp groups in India during the general elections of 2019.

One of the representative messages from our study reproduced below shows the intricate structure of fear speech.

> Hundreds of South Americans are marching through Mexico, aiming to cross the US Border and demand asylum in the US. No one in Mexico is stopping them. This is a national security threat and should be dealt with by force if necessary. What else is our military good for if they can't stop an invading force?

Note that this message has no toxic words and is weaved into a series of arguments citing evidence, establishing a case of nationwide fear and finally inciting users to take an action. Such views often resonate with the opinions of the "common" audience, and they, in turn, contribute to spread the message deeper and farther into the network.

The central objective of this article is to investigate the prevalence of fear speech (See *User Characterisation* section) in a loosely moderated social media platform like Gab.com. Since no known dataset is available for such a study, we devise an algorithmic pipeline to first build a dataset of 400,000 fear speech posts to be contrasted with another 700,000 hate speech posts. Based on the analysis of this dataset, the central result that we arrive at demonstrates how users posting a large number of fear speech are successful in garnering significantly more followers compared to the users posting a large number of hate speech (*Position in the social network*). The former are also more effective in reaching out to the general users through reposts, replies, and mentions (See *User Characterisation* section). We elucidate that this is because of the nontoxic and argumentative nature of the posts that make them look more plausible and thus widely accepted. Some such prevailing arguments in fear speech correspond to violence by the Muslim community (10% of all fear speech posts), Jews controlling media and culture (10% of all fear speech posts), white genocide in South Africa (7% of all fear speech posts), etc. In contrast to this, the traditional hate speech posts mostly correspond to hurling insulting remarks or calling for deportation of the target community (see section *Topic modeling and Dataset* for the definitions for popular topics and this section for definitions). The seemingly benign nature makes fear speech more credulous to the users than hate speech, facilitating its increased prevalence in the network.

We stress that such forms of highly destructive speech should not go unnoticed and call for more sophisticated moderation mechanisms along with mass awareness. We believe that this article can lay the foundation stone for such an initiative.

Dataset

There are no data available in the literature that allow for the study of the prevalence of fear speech (See *Materials and Methods* section) in social media. Therefore, we had to set up an end-to-end pipeline to build our dataset. We make use of the Gab platform for data collection. Gab is a social media platform alternative to Twitter and was launched in May, 2016. It has 100,000 estimated active and 4 million total users.[#] Unlike Twitter, it has a "lax" moderation policy for harmful content and presents itself as a champion of "freedom of speech." It came under scrutiny in the Pittsburgh shooting case, where the sole suspect posted a message on Gab indicating an immediate intent to cause harm before the shooting and also had a history of antisemitic posts (17). Recently, Gab was one of the platforms used to plan the storming of the US Capitol on January 6, 2021 (18). Given these facts, we reasoned that Gab should be a breeding ground for the type of data we wanted for our investigation.

The site allows anyone to read and write posts up to 3,000 characters called "gabs." In Gab, posts can be reposted, quoted, and used as replies to other posts. Similar to Twitter, Gab also supports mentions and hashtags, and users can follow one another. We started off with a huge dump already crawled from Gab in a previous study (19). This contains all the posts and their metadata from October 2016 to July 2018. In total, there are 21 million posts. Further, it has repost and reply information for each post. In addition, the dump also hosts user bios and the follower/followee information per month. In total, there are ~280,000 users having at least one post[‖].

In order to prepare the dataset for our study, we annotate 10K posts from Gab using a hybrid set of annotators. A group of four expert annotators and 103 crowd workers from Amazon Mechanical Turk were chosen based on a rigorous test of their annotation performance (*Methods* for details). The task was to mark each post as a) fear speech, b) hate speech, or c) normal. Further, a post could have both fear and hate components, and, thus, these were annotated with multiple labels.

The annotators were asked to strictly adhere to the operational definitions of fear speech and hate speech as follows: Fear speech is an expression aimed at instilling (existential) fear of a target group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender (16), and hate speech is a language used to express hatred toward a targeted individual or group or is intended to be derogatory, to humiliate, or to insult the members of the group, based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender (22). In addition, the crowd workers were given a number

[¶] https://www.hias.org/, a nonprofit organization that provides aids to the refugees.

[#] https://en.wikipedia.org/wiki/Gab_(social_network) as of March 2021.

[‖] Out of these users, only 2% have a following/follower ratio higher than 10 (20) and around 1.2% users post more than 3 messages per day (21). Thus, the number of users that can be classified as bots based on the above two measures is negligible.

Table 1. Example of fear and hate speech in the Gab dataset

Fear speech	Hate speech
Germany is no longer German. German media celebrates school where 80% of class is non-German #GabFam #Politics #Europe #Merkel #Relocation #Muslims #BanIslam #Invasion #StopRelocation #WhiteGenocide	You are a camel piss drinking goat f**king imbecile now get off my timeline you disgusting piece of s*it
TILL White people won't protest for their SAFETY. Hell, it's not just Whites. Asian & Middle Eastern shopkeepers are frequent victims.Young Black Males are a DANGER to society. SOME are ok, but we don't know who is who. We need PROTECTION & the RIGHT NOT to race mix!	I hear Botswana is lovely in the spring. All ni**ers should go there. And stay
Jewish poison pouring out of our media and Hollywood is destroying Christianity	Because Jews are lying pigs. I'm really thinking this is a genetic thing

of examples and multiple rounds of training to enable them to perform the annotations as accurately as possible. The annotation went in 24 rounds with a small number of samples annotated in each round so as to reduce the overall mental toll faced by the annotators. In each round, the sample posts were chosen based on the presence of a set of manually prepared keywords, which increases the possibility of gathering such samples that are susceptible to be fear or hate speech. All rounds of annotation were closely monitored by the experts, and corrective steps were taken as and when necessary (*Methods* for full details of the annotation process).

After this elaborate process, we arrived at a dataset consisting of ∼10,000 annotated posts. Out of these, around 1,800 were fear speech and 4,000 were hate speech. The interannotator agreement values were as follows: Krippendorff's $\alpha = 0.30$ and Fleiss $\kappa = 0.34$. These agreement values are comparable with such complex tasks in a similar domain and settings (16, 23, 24). Some examples of fear vs hate speech are noted in Table 1. Note the use of various arguments in the fear speech posts such as the target community a) replacing indigenous population (first instance), b) being a physical danger to the society (second instance), and c) causing cultural threat (third instance).

Scaled-Up Dataset. Our objective was to study the large-scale prevalence of fear speech and compare the same with that of hate speech. Therefore, we needed to scale up the annotated data. One easy way to achieve this would be to use standard toxicity classifiers over the whole dataset. We verify whether this is possible using one of the state-of-the-art tools—the Perspective API (25). If we pass our base dataset through this classifier, we observe that the average toxicity score of the fear speech posts returned by the API is 0.51 as opposed to 0.69 for the hate speech posts. This difference is also statistically significant with $p < 1e^{-6}$ as per the Mann–Whitney U (M-W U) test (26). The normal posts have a toxicity score of 0.47 and is very close to that of the fear speech post. Thus, distinguishing fear speech from normal speech using such classifiers would be very difficult if not impossible. Hence, we develop a sophisticated BERT-based architecture to perform multilabel classification of an input post. We train and test the model using the base dataset and obtain a macro-F1 score of 0.63 (*Methods* for a detailed description of the model). We next ran this model to classify all the 2 million posts in our dataset. For fear speech, if we consider only those machine-generated labels as correct where the confidence of the classifier is >0.7, our results are >70% accurate (confirmed by a second round of expert annotation of a small number of samples). For hate speech, a similar accuracy is obtained if the decision confidence of the classifier is >0.9. We manually observe that increasing the threshold further did not improve the score further. Therefore, we empirically fix these two decision confidence levels to finally obtain a scaled-up dataset comprising ∼400K fear speech and ∼700K hate speech posts. (*Methods* for more details.) All our analyses that follow in this article are performed on this dataset.

Prevalence of Fear Speech

The prevalence of a particular entity in any social network can be directly attributed to its users, and fear speech is no exception. Therefore, the first task is to identify users who have a strong propensity to post fear speech. Similarly, for comparison, we also need to select users posting hate speech.

User Selection. Out of 280K users, we observe that as high as 9,200 users have posted at least 10 fear/hate speech posts. However, we were interested in the extreme behavior, i.e., we wanted to identify those users who have extreme propensity to post fear speech or hate speech. For this purpose, we find users falling in the top 10% percentile in terms of the number of fear speech or hate speech posted by them. We remove those common users that belong to both these sets[**]. We end up with 479 extreme fear speech (ExFear) and 483 extreme hate speech (ExHate) users[††]. The choice of these set of users is motivated by the fact that they would be the central actors responsible for the prevalence of fear/hate speech.

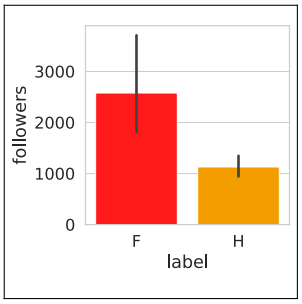
User Characterization. In this section, we compare the ExFear users with the ExHate users. In total, ExFear users posted 2.6 millon posts, out of which 104k were fear speech and 26k were hate speech. Similarly, ExHate users posted 2 million posts, out of which 184k were hate speech and 18k were fear speech. We consider three different aspects—their position in the social network, their overall reach of the normal users, and temporal trends.

Position in the social network. We construct the social network based on all the follower–followee relationships among the users till the end of the timeline (i.e., June 2018) (27). This network consists of 279,961 nodes and 1960,869 edges.

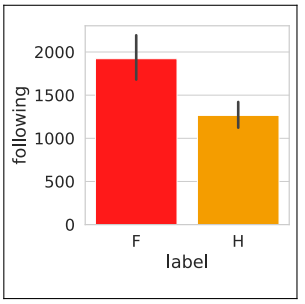
The first quantities that we compare are the number of followers and followings for each type of user. The plot in Fig. 1 shows that both the number of followers and the followings of ExFear users are larger than that of the ExHate users. The results are statistically significant with $P < 0.0001$ (M-W U test).

[**]We perform a separate analysis on this set of users in *SI Appendix*, Text in section 5.

[††]Out of this set, 476 users matched in terms of propensity score-based matching. *SI Appendix*, Text in section 1B for more details.

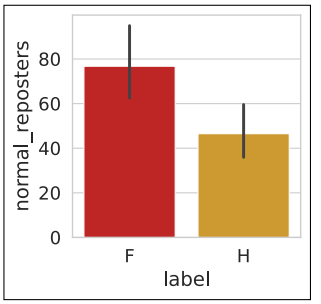


(A) Average number of followers for different groups of users.

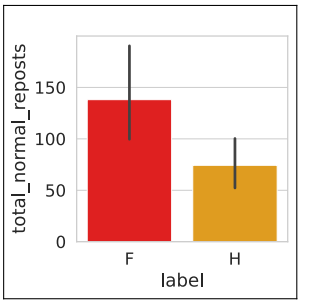


(B) Average number of followings for different groups of users.

Fig. 1. Plots denoting follower–following properties for ExFear (F) and ExHate (H) users. The results are significant at $P < 0.0001$ using the M-W U test.



(A) # of normal reposters per user.



(B) # of reposts from normal users per user.

Fig. 3. Distribution of reposts from normal users for ExFear (F) and ExHate (H) users. The results are significant at $P < 1e^{-6}$ using the M-W U test.

Next, we compute the betweenness and the eigen-vector centrality of the nodes from the undirected version of this network. These metrics are known to express the positional importance of the nodes; while eigenvector centrality indicates the influence of the nodes, betweenness centrality indicates the degree to which a node stands in between other nodes. From Fig. 2*A*, we observe that, in terms of eigenvector centrality, the ExFear users are more central compared to the ExHate users. Once again, the results are statistically significant with $P < 0.001$ (M-W U test). The observations remain the same for the betweenness centrality with the ExFear users far more central compared to the ExHate users (Fig. 2*B*).

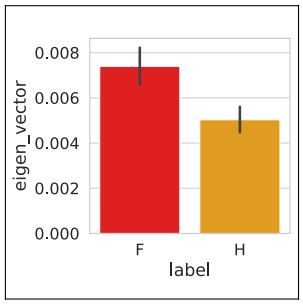
These results together show that the ExFear users are far more strategically placed in the network compared to the ExHate users. Such advantageous positions of the ExFear users are a natural source for higher prevalence of fear speech in the network.

Reach of the normal users. We label users who never post fear or hate speech (decision confidence of the model < 0.5) as "normal users." Here, we investigate how the ExFear and ExHate users interact with the normal users. First, we find that the average percentage of normal followers out of all followers for ExFear users (21%) is higher than for ExHate users (18%). This difference is statistically significant with $P < 1e^{-6}$ (M-W U test).

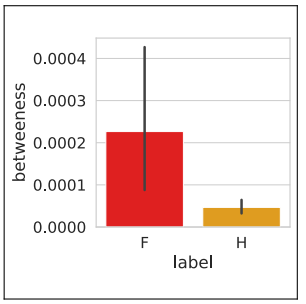
The number of posts made by ExFear and that by ExHate users are both of the tune of two million each, i.e., their posting activity is quite similar. Therefore, we plot the number of normal users reposting the posts of ExFear vs ExHate users. The results in Fig. 3*A* show that a larger number of normal users repost the posts of ExFear users compared to that of ExHate users ($P < 1e^{-6}$, M-W U test). Further, the total number of reposts by normal users to the posts made by ExFear users is larger than posts of ExHate users (Fig. 3*B*). The same trend persists for both mentions and replies. ExFear users mention more number of normal users in their posts (Fig. 4*A*) compared to ExHate users. Moreover, the total number of posts by the former having normal users mentioned is also higher (Fig. 4*B*). The number of normal users replying to the posts of ExFear users is higher than that of ExHate users (Fig. 5*A*). Finally, the number of replies obtained from the normal users by the ExFear users is larger than that of ExHate users (Fig. 5*B*). Hence, we show that ExFear users impact the normal users more.

We also analyze the impact on normal users based on the posts they receive. Since it is not possible to know whether someone received a post directly, we assumed that a user "A" would receive a post from a particular user "B" if she is following that user "B." We consider the top 500 normal users based on their number of posts. An additional constraint was that they should have at least one ExFear and one ExHate user in their following. This resulted in 179 users. We find that these users receive around 1.5% fear speech and 2% hate speech posts from their followings. Surprisingly, although the percentage of fear speech received by them is less, they end up reposting the fear speech almost four times more (average of 1.10 posts) compared to hate speech (average of 0.28 posts). The results are statistically significant ($P < 0.001$, M-W U test).

In our final experiment in this section, we go a step forward to assess the perception about fear vs. hate speech among in-the-wild users. We recruit human judges from Amazon Mechanical Turk for this experiment. We create a survey by posing pairs of fear and hate speech and ask human judges to select the post they believed

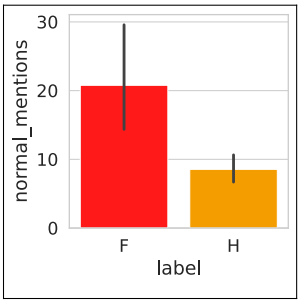


(A) Eigenvector centrality of users.

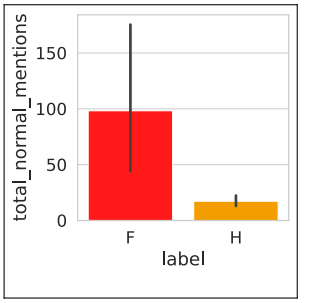


(B) Betweenness centrality of users.

Fig. 2. Centrality measures of ExFear (F) and ExHate (H) users. The results are significant at $P < 0.0001$ using the M-W U test.

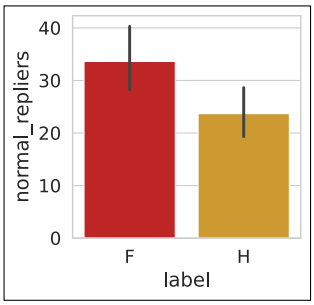


(A) # of normal mentioned users per user.

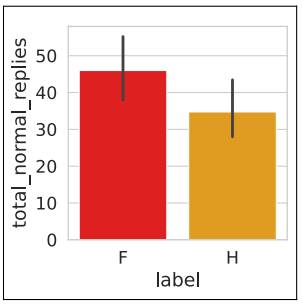


(B) # of normal mentions per user.

Fig. 4. Distribution of normal mentions for ExHate (H) and ExFear (F) users. The results are significant at $P < 1e^{-6}$ using the M-W U test.



(A) # of normal repliers per user.



(B) # of replies from normal users per user.

Fig. 5. Distributions of replies from normal users for ExFear (F) and ExHate (H) users. The results are significant at $P < 1e^{-6}$ using the M-W U test.

in more. Each pair of posts is judged by nine random judges. All these judges have high approval rates (>95%) and high approved hits (>1,000). We got the posts judged in three batches, with an incremental number of post pairs in each batch—precisely, 25, 30, and 45 pairs in the three successive batches. Each batch is further divided into pages of three pairs each. We make sure that the judges in successive batches do not overlap to ensure diversity of opinions. In total, 246 unique judges participated in the task (68 in the first batch, 82 in the second, and 96 in the third). For each pair, we select that post between fear and hate speech to be believable, which receives the majority of the votes. We find that out of 100 pairs, in as many as 69 pairs (i.e., 69%), fear speech posts were voted to be the more believable out of the two.

Overall, we observe that the ExFear users are far more well connected with the normal users compared to the ExHate users. Manual analysis reveals that the top reposted/replied/liked fear speech posts contain emotionally loaded language and/or urgent tone with the occasional usage of capital letters as shown in Table 2. Often, the posts pretend to narrate real incidents, foretell how bleak the future could be, and cite (fake) statistics to make the content look realistic and convincing. We also obtain the top 10 normal users mentioned by ExFear users and find that they usually have a large number of followers (~1,200) and followings (~1,700) but have less number of posts (~17). Manually analyzing their profiles from Gab, we find that their posts are generally on benign topics, but they repost a lot of controversial topics, which might be the reason why they get mentioned more by the ExFear users.

Temporal trends. In this section, we deep dive into the results obtained earlier to investigate the temporal evolution of different observables of interest.

As a first step, we investigate how the ExFear and ExHate users move in the follower–followee network over time. To this purpose, for each month, we construct an undirected follower–followee network and perform the standard k-core decomposition (28). Such a decomposition is known to segregate the network into "shells" with the innermost few shells containing the most influential nodes. We divide the nodes into 10 buckets in terms of the percentile ranks based on their k-core values, i.e., top 10% nodes in the first bucket, next 10% nodes in the second bucket, and so on. Note that therefore the first bucket consists of the most influential nodes, while the last contains the least influential ones. Next, we observe how the users move from one bucket to the other over time since they had joined the network. The temporal movement of the ExFear and ExHate users across the different buckets over time is shown in Fig. 6. Both the ExFear and ExHate users are predominantly in the outer shell of the network at the time of their joining. However, as time progresses, they accelerate steadily to the inner shells with the maximum influx happening in October 2016 for ExFear users and in August 2017 for the ExHate users[‡‡]. The maximum influx for ExHate users coincides with the Unite the Right rally, Charlottesville[§§], while for ExFear users, we see a jump toward the initial time period. This possibly indicates that a fraction of the ExFear users have remained all through in the core of the

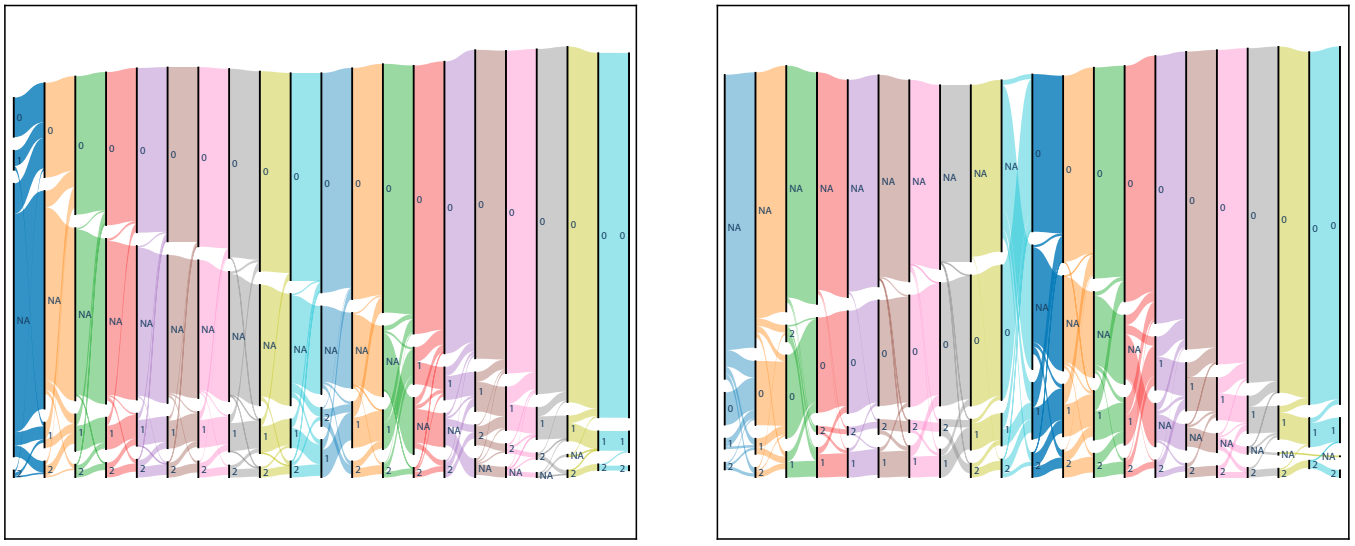
[‡‡]Here, the maximum influx is defined as the maximum users jumping to an inner core considering their current core.

[§§]https://time.com/charlottesville-white-nationalist-rally-clashes/.

Table 2. Examples of fear speech which are popular in terms of likes/replies/reposts

Post	Likes	Reposts	Replies
It's the future. I was promised flying cars and cured cancer. Instead I got "hate speech," a third world invasion, and an internet controlled by the ADL and SPLC. I'll be damned if I let this be the "future" my kids grow up in	**920**	**361**	41
PUBLIC SERVICE ANNOUNCEMENT FROM IDENTITY EVROPA San Francisco is a dangerous sanctuary city where the law does not apply to illegal invaders. Enter at your own risk!	**593**	205	16
This family lost a mother. She was killed by a Sudanese migrant in church yesterday in Antioch, Tennessee. Media silence is deafening.#MelanieSmith	**625**	**268**	0
80K whites dead in South Africa in an ongoing genocide = Silence. 30 dead in a highly suspicious unconfirmed G-S attack in the Middle East = World War 3	**588**	**282**	15
It's not too late. A Charlottesville 2 could feature a memorial for Heather Heyer blaming antifa for jostling a land whale with an explosive heart. That'd probably get attention, and it'd probably be hard to separate from the message that the Alt-Right were innocent victims just trying to speak before jewish domestic terrorists started killing Whites	0	0	**77**
I had an uber driver telling me this recently, after me going on about ni**ers (We in oz have a SMALL population of ni**ers per capita) he finally came clean they slashed his seats with knives and they SMELL particularly bad, Again this is in Australia where I see Africans maybe 10 a year. So far one tried to rob me and my uber driver had that happen! Imagine the US!!	9	1	**85**

The bold number per row shows the engagement factor based on which the specific post is cited.



(A) Extreme fear speech (ExFear) user movement

(B) Extreme hate speech (ExHate) user movement

Fig. 6. Alluvial diagram showing the core transition for the users. The stubs represent the dynamic graph state with the first stub indicating October 2016. A lower core value represents that a node is situated deeper in the network. "NA" denotes the set of users who are yet to join the networks each month from the total set of users. We show only the transitions among the three innermost cores for better visualization. The dark blue band shows the month with the maximum influx for each graph. Maximum influx means that during that month, the maximum number of users have jumped to an inner core with respect to their core.

network from the very beginning. On average, ExFear users take lesser time (2.83 mo) to reach the innermost core of the network compared to the ExHate users (3.32 mo).

Next, we investigate the temporal evolution of the engagement to the posts made by ExFear or ExHate users. When considering replies by normal users, we observe that while for the first 2 to 3 mo the trajectories are similar, after January 2017, the replies to ExFear users keep increasing while replies to ExHate users suffer a dip. The replies to ExFear users have a sudden peak around June 2017. After this, the replies to ExFear users dip below ExHate users, possibly due to the influence of an external event in the form of Unite the Right rally, Charlottesville. This might also suggest that many normal users started to subscribe to the hateful notions. If we consider the reposts by normal users[¶¶], we find that here the distributions are similar with the peaks occurring at a similar time (March 2018). Considering the normal users' mentions by the two groups of users temporally, we find a significant difference in the two curves. While ExHate users use very less mentions of normal users, ExFear users heavily use the same with the peak occurring (60 times per mentioned users) in December 2017. Manual analysis revealed that many of the fear speech posts had a comment about a target community followed by mentions of several users, social media influencers, news media sources, etc. For e.g., "Muslims want to double the number of mosques in France <link> @MichelOsef @Isleofcarl @Bill_Murray @SatanIsAllah @Brea @HEDGE @PigtownGrump @Zucotic @TaratheLeo @kingmack @Psnow @TwoPats @MaryJane @TupacZaday @Reef ".

Overall, this section demonstrates that fear speech has a significantly larger prevalence in the social network compared to hate speech. (More analysis with a larger set of users can be found in *SI Appendix*, Text.) In the next section, we investigate the content structure of fear speech, which undoubtedly plays the central role in its wider prevalence.

[¶¶] We have the reposts information from August 2017 in the dataset.

Content Structure

In this section, we investigate the differences in the content structure of the fear speech posts from those of the hate speech posts. These differences rooted in their content play a key role in shaping their prevalence.

Fear Speech Is Presented as Topical Arguments. We analyze the text present in the fear speech post and compare them with the hate speech posts using widely popular NLP tools as follows.

Topic modeling. We use the LDA model (29) to extract the topics in the fear and hate speech posts (More details in *SI Appendix*, Text section 6). Next, for each month, we plot their normalized distribution considering the total posts in that month. Overall, we notice one very important difference between fear speech (Fig. 7) and hate speech topics (Fig. 8). Topics in the fear speech mostly portrayed other communities as perpetrators in a subtle and argumentative style, while topics in the hate speech were dehumanizing or insulting the target communities.

Some of the illustrative examples of fear speech topics are "America needs to wake up" and "Ideology of Islam is dangerous," which are prevalent across all the months. Here, the topic "America needs to wake up" makes implicit calls to Americans to see the atrocities by other communities. The topic "violence by Muslim communities" notes the various unconfirmed violent activities by the Muslim communities. It had a tiny share initially (October & November 2016) but increased to a significant ratio afterward. On the other hand, the topic "immigrants manipulating elections" was prominent during the initial time periods but died out after April 2017. Another interesting topic was "jews controlling media"—which points out how Jews control media platforms. Apart from that, illegal immigration as a problem was portrayed in topics like "illegal immigration in Europe," "illegal immigration in the USA," etc.

Among the hate speech topics, three of the most consistent ones are "multitarget insults"—where a single hate post targeted multiple communities, "women being projected as prostitutes," and hate against voters from a different demography. Other topics

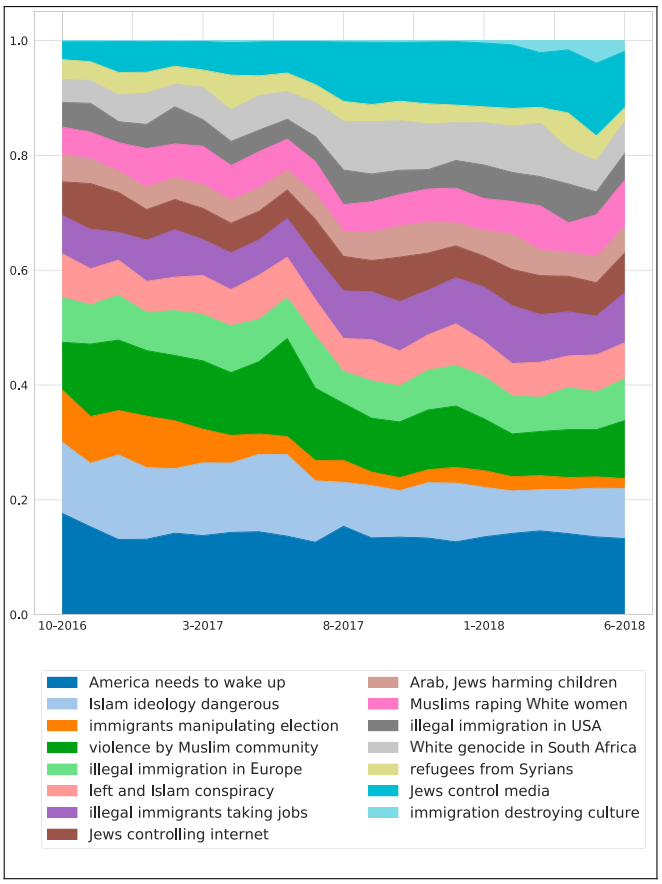


Fig. 7. Top 10 topics and their normalized distribution per month for fear speech posts.

like insults of Muslims and Canadians occur rarely and have smaller ratios. Insults of the Jewish community rose after August 2017. This might be an effect of the influx of a lot of new users during that time period. The topic which has posts targeting both homosexuals and Muslims reduced after March 2017 since it possibly merged with the multitarget insults. The ratio of posts under topics like "support for Nazi" and "insulting and blaming Africans" increased significantly after August 2017.

Reaction of normal users. A careful observation of the topics extracted from the fear speech posts shows that the arguments presented in these topics most often look quite acceptable and amenable to the normal users resulting in their direct involvement in reposting of and replying to the messages corresponding to these topics. The topics in the fear speech category receive around 1,000 reposts from normal users with the highest average reposts being received by the topic "violence by Muslim community" (∼2,500). On the other hand, for the topics extracted from the hate speech posts hurling direct attacks on different communities are usually found to be repulsive by normal users and are much less frequently reposted or replied to. The average number of reposts per topic is about 500 for hate speech topics with the highest average reposts being received by the topic "deport illegal immigrants" (∼1,100). Note that, in general, the average number of reposts for any post on the platform is around 2 per posts.

Hashtags and Web Domains.

Hashtags. Hashtags are an important component of the overall content of any social media post. We investigate how fast a hashtag originating from one form of speech is adopted to scribe another form of speech. A hashtag is considered to have originated in fear/hate/normal speech if a fear/hate/normal user uses it for the first time in one of their posts. One of the most surprising findings is how fast hashtags originating from normal speech get adopted to fear speech (∼83 d); this is significantly less compared to the time needed by hashtags originating from normal speech and getting adopted to hate speech (∼124 d) ($p < 1e^{-6}$, M-W U test, one-sided). This suggests that users posting fear speech carefully craft their messages to include hashtags mainly used by normal users. Consequently, the visibility of the corresponding fear speech post gets enhanced among the normal users. In addition, another observation is that the median time for a hate speech hashtag to get adopted into a fear speech post (∼73 d) is significantly ($P < 1e^{-6}$, M-W U test, one-sided) lower than a fear speech hashtag to get adopted into a hate speech post (∼88 d). This once again shows that fear speech users cleverly include hashtags used by hate speech users in their posts.

Web domains. We investigate the popular domains shared by the fear and hate speech users. Around ∼6,000 unique URLs were shared by each of these types of users. We manually inspected some of the most frequent domains (top 20) that were shared (Table 3). Many of the fear speech posts shared URLs of unconfirmed blogs on atrocities by the Muslim community—https://islamexposedblog.blogspot.com, https://thereligionofpeace.com, and https://counterjihad.com. Few domains were right biased media

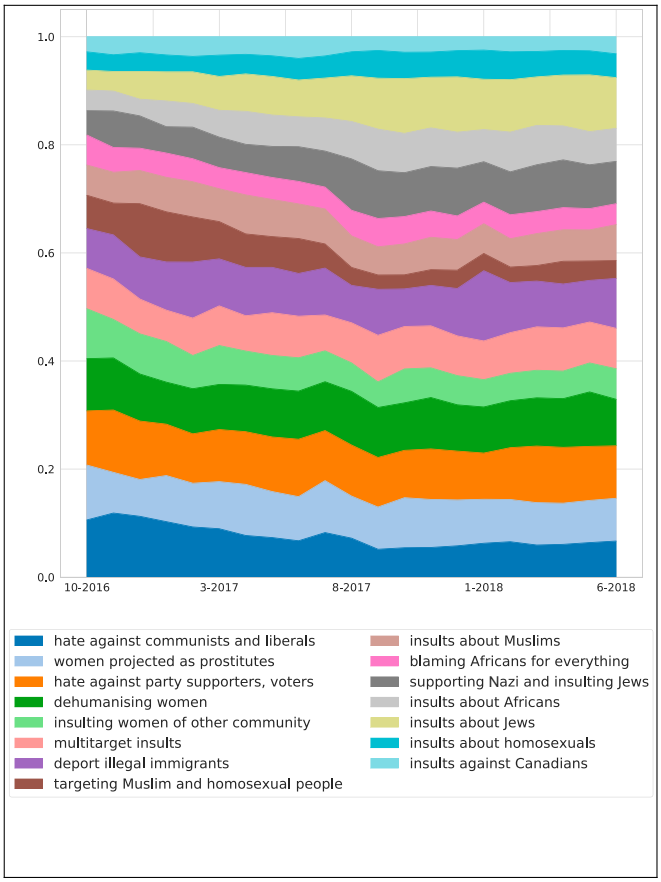


Fig. 8. Top 10 topics and their normalized distribution per month for hate speech posts.

Table 3. Some of the top relevant URLs along with the number of fear/hate speech posts

Fear speech	Hate speech
aclg (243)	pagesix (65)
whitenationnetwork (54)	towleroad (68)
islamexposedblog (72)	dailystormer (63)
thereligionofpeace (40)	weaselzippers (45)
sputniknews (37)	godhatesfags (28)
counterjihad (33)	thesmokinggun (20)

having low credibility like American Center for Law and Justice[##] and Sputnik news (https://sputnik.com/). Another website portrays the unconfirmed atrocities on the white community, whitenationnetwork (https://tinyurl.com/567n8rat). In fact, this website has been currently shut down. Other forms of conspiracy theories like coronavirus is a hoax also showed up on some of these websites. Overall, majority of the URLs shared by the fear speech users have fake/unconfirmed content which, most often, makes them highly believable to the benign social media users.

Popular domains in hate speech posts are quite different in nature. We find pagesix (https://pagesix.com/ accessed on March 10, 2022), an entertainment news website, and towleroad (https://www.towleroad.com/), an entertainment website for Gay and LGBTQ+ community, which are both authentic. Both these websites are benign in nature, but the hate speech posts referred to them to insult the celebrities mentioned on these platforms. We also find dailystormer (https://en.wikipedia.org/wiki/The_Daily_Stormer), godhatefags (https://en.wikipedia.org/wiki/Westboro_Baptist_Church), etc., which are popular far-right websites.

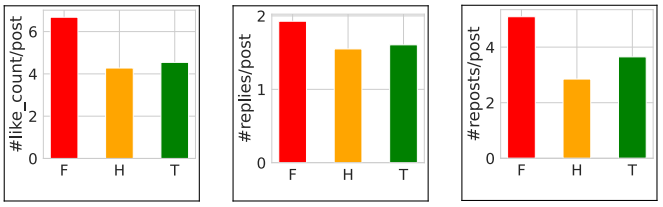
Interaction of the Users with the Content. Interaction with a post can be an essential indicator of how the audience engages with the post. We measure this using the reposts, replies, and likes frequency. Here, we compare these interactions for fear and hate speech posts. As a baseline, we also compare these with the overall level of interaction with all posts.

#Likes. Fear and hate speech posts taken together receive ∼65% of likes, while at the overall level, less than ∼60% posts receive one or more likes. As illustrated in Fig. 9*A*, we find that the average number of likes for fear speech is around ∼7 per post, which is significantly more ($P < 1e^{-6}$, M-W U test, one-sided) than that of hate speech. We have shown examples of the highly liked fear speech posts in Table 2.

#Replies. Fear and hate speech posts taken together receive one or more replies in ∼16% cases, while at the overall level, less than ∼10% posts receive one or more replies. Once again, as shown in Fig. 9*B*, the mean number of replies per post is higher for fear speech as compared to hate speech ($P < 1e^{-6}$, M-W U test, one-sided). We have shown examples of the highly replied fear speech posts in Table 2. Manual analysis revealed that interestingly, the post receiving higher reposts usually had fewer replies and likes. Further, around 0.3% of the replies of the fear speech are from normal users, whereas 0.2% of the replies of the hate speech are from normal users.

#Reposts. In terms of reposts, we observe that more number fear speech posts (∼18%) is reposted as compared to hate speech and overall posts (∼11 to 13%). The average number of reposts per post is significantly ($P < 1e^{-6}$, M-W U test, one-sided) higher for fear speech (5 per post) than for hate speech (3 per post

[##] http://www.aclj.org as accessed on Mar 7, 2022.



(A) Mean likes per post. (B) Mean replies per post. (C) Mean re-posts per post.

Fig. 9. Interaction of users with posts. Here, in the x-axis, we show the type of posts where F, H, and T denote fear speech, hate speech, and total (overall) posts, respectively.

(Fig. 9*C*). We have shown examples of the highly reposted fear speech posts in Table 2. Further, around 6% of the reposts of the fear speech are from normal users, whereas 3% of the reposts of the hate speech are from normal users.

In summary, we observe that the average level of engagement of users with fear speech posts is much higher than hate speech posts, which we believe is another reason for their prevalence.

Pervasive Impact of Fear Speech Transcending to Other Social Media Platforms. In this section, we demonstrate that the problem of fear speech is of significant general interest as it also prevails in other extensively moderated social media platforms, e.g., Twitter and Facebook. Note that the choice of these two platforms is motivated by the fact that both of them have their own strict hate speech policies in place and are constantly vigilant to remove harmful contents. We crawl large chunks of data from both these platforms and classify them as fear, hate, or normal speech using our prediction model discussed earlier. Once again, we use the same confidence value-based thresholding as used for the Gab dataset to designate a post to be fear/hate speech.

Twitter. For Twitter, we use the topical keywords (the exact list will be shared in the repository) from the topics in Fig. 7 and the academic research API to search through the history of tweets having those keywords. This way, we collect around 4,103,145 tweets over 6 y (2016 to 2022). We find that out of the entire dataset of around 4 million tweets, around 400k tweets (∼10%) were marked as fear speech by our model (examples in Table 4). We further plot the timeline of the posts and find that there is

Table 4. Examples of fear speech from the data collected from Twitter along with their dates

Text	Date
@AmosPosner Christians left tons of time for Jews to control media in th silence b/w the beat & when ppl yell "Santa Claus is comin to town"	5/12/2016
MIGRANT SCANDAL: 200 illegals a DAY caught sneaking into UK - and that's in just... https://t.co/avZNtrJyXk by #rvaidya2000 via @c0nvey	10/2/2017
@JudgeJeanine QUESTION PATRIOTS? ARE OUR OFFICIALS BREAKING THE LAW BY NOT UPHOLDING THE LAWS THEY WROTE, ALLOWING ILLEGAL IMMIGRATION TO OVERRUN OUR COUNTRY? IF YOU THINK SO.. SCREW BEING FIRED! HOW ABOUT CITIZENS ARREST?	10/4/2019
@FBI San Diego Antifa leader calling for the killing of white men and raping white women. https://t.co/rqHhL5pxD2	28/6/2020

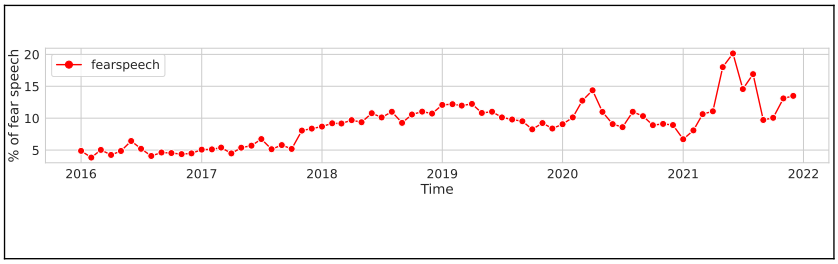


Fig. 10. Percentage of posts that were fear speech per month in the Twitter data.

an increasing trend in the number of fear speech posts (Fig. 10) over time. The presence of such a huge volume of fear speech and its increasing temporal trend is alarming and should be analyzed by moderation policy experts. Not surprisingly, our model could predict only around 31,000 posts as hate speech, which shows that Twitter is quite active in moderating such hateful content.

Facebook. For Facebook, we use the historical search of Crowd-Tangle (https://crowdtangle.com) and use 73 public white supremacist pages used in the previous literature. We collect all the posts from these pages (30). This way, we obtain around 191,666 posts over 6 y (2016 to 2022). Our model predicts that around 10k posts (around 4%) are marked as fear speech (examples in Table 5). We plot the timeline of the posts and find that there is a slightly decreasing trend in the number of fear speech posts (Fig. 11). This decrease could possibly be because of the overall moderation of the white supremacist pages and not specifically the individual fear speech posts. Once again, our model marked only 196 posts as hate speech, pointing toward the strict hateful content moderation on Facebook.

We believe that these results together point to the pervasive nature of the problem and the necessity for special all-round attention from the community.

Discussion

In this study, we aimed to understand what role fear plays in polarized conversations and how it differs from the traditional form of polarized content—hate speech. We find a significant difference in how extreme hate speech (ExHate) and fear speech users (ExFear) exist and interact with other users in the network.

Table 5. Some examples of fear speech from the data collected from Facebook along with their dates

Text	Date
Have you noticed Islam is growing stronger? The "girl next door" is even jumping on the jihad train	18/5/2017
#Turkey says its released 47,000 migrants into #Europe. That us 47,000 on a #hijrah most are men of military age. https://www.trtworld.com/turkey/number-of-migrants-leaving-turkey-reaches-47-113-minister-34211	29/2/2020
(Bangladesh: Muslims threaten to murder atheist blogger for criticizing political Islam, defending Buddhists) has been published on Jihad Watch	26/8/2020
Most attention goes to illegal aliens crossing by land, but data show rising numbers trying to come by water	14/1/2021

ExFear users have more followers and can effectively interact with the general audience than the ExHate users. This indicates out that even within a polarized, hateful context, fear has a different reach in the audience. In the correct context, such type of polarized content may act as tipping points (12, 31) during some event. This is especially so when there are groups of coordinated actors who are interested in propagating an agenda (16). Hence, it becomes important for the research community to understand how to moderate such different forms of extreme content. It is also interesting to think about some prioritization when moderating different forms of extreme content.

One of the main reasons why these differences exists is the language of the text used. While fear speech uses arguments and subtle ways to show some community as a threat, hate speech (32) uses slurs and insults to dehumanize the community. There is a huge body of work on how hate speech spreads in social media and can be analyzed (33–35), detected in monolingual (22, 36–38) and multilingual scenarios (6, 39–41), and mitigated using suspension (42) and counter speech (43–45). The presence of fear speech will create problems while deciding about the moderation policies because we might not be able to directly ban or suspend fear speech. The paper introducing the "fear speech" concept (12) suggests creation of alternative arguments to the arguments given in fear speech. These alternative arguments should aim at diffusing the violence potential of fear speech. Since many instances of fear speech may also contain misinformation to exaggerate their arguments, researchers in the misinformation (46) domain, news media, and fact-checking organizations can play an important role. However, even such measures might not be effective unless the end user is aware. Hence, awareness events, similar to the ones done for hate speech (47), should be conducted to make the users question the content they are receiving.

Past research in this community has focused on the role of social media in polarization (48, 49) and the role of user accounts in spreading such content (50). Our research takes a step back and tries to understand the types in which polarization happens and whether there exists a difference in the audience using/perceiving it.

One limitation of our study is that our detection model is trained on Gab data. Further, most of the prevalence analysis is also based on the Gab data. Our choice of Gab as a social media platform is motivated by its unmoderated nature, which makes studying hate speech easier. Further, obtaining certain nuanced data such as the time-varying structure of the followership network is easily possible for Gab. To complement this study further, we perform some basic prevalence analysis on Twitter and Facebook as well and find a significant amount of fear speech on these platforms. Our study renders hope that the investigation of fear speech can be easily extended to other platforms. Nevertheless, it seems that as these platforms continue moderating hate speech, actors spreading such content might shift to more

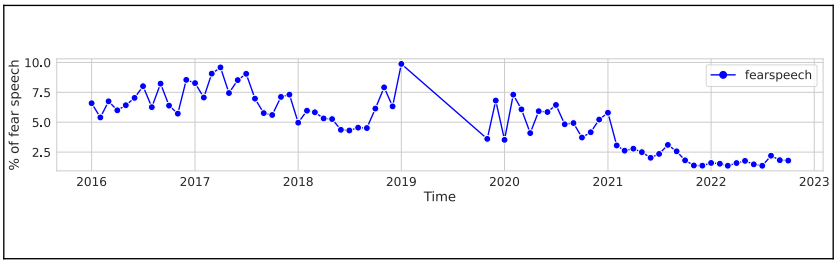


Fig. 11. Percentage of posts that were fear speech per month in the Facebook data.

subtle ways like fear speech. Moreover, the content on "fringe" platforms does not stay only on those platforms anymore, and we have seen instances of seemingly fringe platforms affecting mainstream conversations (51). Second, in an effort to scale up our findings to millions of posts, our study relies on the performance of automated classifiers on an inherently difficult task. While we have taken additional care while deciding the category of the posts, we might be missing some form of fear speech/hate speech. *SI Appendix*, Text section 3 provides robustness checks on our models.

Materials and Methods

This section provides details on data collection, annotation, and labeling and user-level classification. Statistical tools used in this analysis are noted in *SI Appendix*, Text section 1.

Data Annotation. There are different forms of toxic speech on social media. In this work, we primarily target hate speech and fear speech. For each post shown to the annotators, the annotator has to mark whether the speech is fear speech or hate speech, or normal. Further, they also need to mark the target communities toward which the particular posts are targeted. *SI Appendix*, Text in section 2 for more details.

To annotate the posts, we follow a hybrid strategy comprising both expert and crowd annotators. The expert annotators are a group of 4 undergraduate students who were trained using gold label annotations and detailed discussion sessions. The crowd workers were recruited from Amazon Mechanical Turk. We use a multilabel annotation framework, where a post can be assigned to both fear speech and hate speech.

To finalize the annotation guidelines and difficulty of the annotations, we first annotated a set of 1,000 posts using the expert annotators. The expert annotators achieve a set of 0.51 Krippendorff's alpha. Next, we created using a pilot study to select the crowd workers. The pilot study is a set of 15 gold annotated posts from these 1,000 data points used to test the Mechanical Turk workers who agreed to take part in our study. Out of 400 interested crowd workers, we selected 192 annotators. Of these annotators, 103 participated in the study.

The annotation process comprised 24 rounds. In each round, we gave a fixed number of posts to annotate. The number of posts per round was kept low, around 150 initially and finally increased to 500 as the annotators became more familiar with the task at hand. For sampling the posts, we employed different strategies. Initially, our strategy revolved around using community-based keywords. In each round, we removed the keywords that gave more normal samples in order to retrieve more fear speech/hate speech posts.

Post Classification. We develop a bunch of classification models for this task. As baselines, we use three different feature extraction techniques–bag of words vectors (BoW), GloVe word embeddings (WE), and TF-IDF features. We then use two one-vs-rest classifier–logistic regression (LR), support vector classifier (SVC) as well as XGBoost. Additional details of the baseline models are noted in *SI Appendix*, Text section 2 for more details.

Transformers are a recent NLP architecture formed using a stack of self-attention blocks having superior performance across a lot of benchmarks. We use several variations of the transformer models–i) pretrained models like bert-base-uncased, roberta-base, ii) models which are fine-tuned using data from hate speech-related tasks like HateXplain (22), Twitter-roberta-hate (52), etc., and iii) models which are pretrained using social media dataset–HateBERT (53). In the category iii), we also use a filtered-out version of the Gab dataset to pretrain a bert-base-uncased model further and name it Gab-BERT (We shall release this model upon acceptance of the paper). All these models are added with a classification head. Gab-BERT is the best model among all others with a macro F1 score of 0.62. (The full set of results for all the models are presented in *SI Appendix*, Text section 3.

We further hypothesize that hate speech and fear speech might show different forms of emotions. We use an emotion vector predicted using the model used in previous research work (54). This additional input vector increases the performance of the Gab-BERT model by 1 point for the F1 score and 4 points for accuracy.

User Analysis. To conduct the user analysis, we wanted to understand the characteristics of the extreme fear and extreme hate users. To do this, we find the users in the top 10% percentile in terms of the number of fear speech posts and hate speech posts separately. We remove the intersection of the users in both these sets. Finally, we end with 476 extreme fear speech (ExFear) and 478 extreme hate speech users (ExHate). We further perform the study on an extended set of users as well (noted in *SI Appendix*, Text section 4).

Temporal Movement of Users. To understand the temporal influence of the users over the entire timeline, we utilize the follower–followee network per month, which was referred to in (55). Then, for each month, we calculate the k-core or coreness metric (28) to identify the influential users in the undirected version of the follower–followee network. Next, we subdivide the nodes into 10 buckets based on their percentile ranks in terms of k-core value, i.e., the bottom 10% percentile to the top 10% percentile. Following this, we measure the time in months for a user to reach the inner core (core-0) in the network (further referred to as time-to-reach-core) from the time they join the network.

Data, Materials, and Software Availability. A repository of the data necessary to reproduce, analyze, and interpret all findings in this paper is available https://osf.io/dc7vu/?view_only=8144833546e54a399ab883f0b0e3e7f7. The code (including software information) for all studies and the analysis is available at https://github.com/punyajoy/Fearspeech-project.

1. V. Basile *et al.*, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter" in *13th International Workshop on Semantic Evaluation* (Association for Computational Linguistics, 2019), pp. 54–63.
2. Z. Waseem, D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter" in *Proceedings of NAACL Study Research Workshop* (2016), pp. 88–93.
3. B. Kennedy *et al.*, The gab hate corpus: A collection of 27k posts annotated for hate speech. PsyArXiv (2018).
4. R. Hada *et al.*, "Ruddit: Norms of offensiveness for English Reddit comments" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Association for Computational Linguistics, Online, 2021), pp. 2700–2717.
5. S. MacAvaney *et al.*, Hate speech detection: Challenges and solutions. *PloS One* **14**, e0221152 (2019).

6. S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, "A deep dive into multilingual hate speech classification" in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V* (Springer-Verlag, Berlin, Heidelberg, 2020), pp. 423–439.
7. T. Pyszczynski *et al*., Mortality salience, martyrdom, and military might: The great satan versus the axis of evil. *Pers. Soc. Psychol. Bull.* **32**, 525–537 (2006).
8. Fatal attraction: The effects of mortality salience on evaluations of charismatic, task-oriented, and relationship-oriented leaders. *Psychol. Sci.* **15**, 846–851 (2004).
9. "The radicalizing language of fear and threat | dangerous speech project." https://dangerousspeech.org/the-radicalizing-language-of-fear-and-threat/. Accessed 27 March 2022.
10. A. Iftikhar, "Christchurch anniversary: The islamophobic 'great replacement' theory - bridge initiative" https://bridge.georgetown.edu/research/christchurch-anniversary-the-islamophobic-great-replacement-theory/. Accessed 27 March 2022.
11. M. Morales, K. Sgueglia, "Buffalo mass shooting suspect to be arraigned thursday on 25 counts, including murder - CNN" (2022). https://edition.cnn.com/2022/06/02/us/buffalo-mass-shooting-suspect-indictment/index.html. Accessed 01 July 2022.
12. A. Buyse, Words of violence: Fear speech, or how violent conflict escalation relates to the freedom of expression. *Hum. Rts. Q.* **36**, 779 (2014).
13. "Trump and others stoke migrant caravan conspiracy theories." https://www.yahoo.com/news/trump-republican-lawmakers-stoke-migrant-caravan-conspiracy-theories-224959187.html. Accessed 06 April 2022.
14. "Wilders tells dutch parliament refugee crisis is 'islamic invasion' | reuters." https://www.reuters.com/article/us-europe-migrants-netherlands/wilders-tells-dutch-parliament-refugee-crisis-is-islamic-invasion-idUSKCN0RA0WY20150910. Accessed 06 April 2022.
15. A. Reid, Buses and breaking point: Freedom of expression and the 'brexit' campaign. *Ethical Theory Moral Practice* **22**, 623–637 (2019).
16. P. Saha, B. Mathew, K. Garimella, A Mukherjee, "Short is the road that leads from fear to hate: Fear speech in Indian whatsapp groups" in *Proceedings of the Web Conference 2021* (2021), pp. 1110–1121.
17. "screw the optics, i'm going in: Alleged synagogue shooter posts on social media moments before massacre - ABC news." https://abcnews.go.com/US/pittsburgh-synagogue-alleged-mass-shooter-told-swat-officers/story?id=58803485. Accessed 31 March 2022.
18. "Social media site gab is surging, even as critics blame it for capitol violence: Npr." https://www.npr.org/2021/01/17/957512634/social-media-site-gab-is-surging-even-as-critics-blame-it-for-capitol-violence. Accessed 31 March 2022.
19. B. Miroglio, D. Zeber, J. Kaye, R. Weiss, "The effect of ad blocking on user engagement with the web" in *Proceedings of the 2018 World Wide Web Conference, WWW 2018. (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE)* (2018), pp. 813–821.
20. N. Schaffer, D. Boutin, Twitter Followers vs Following: What is the Ideal Ratio? (2009).
21. Y. Mohammad, "How many tweets per day 2022 (number of tweets per day)" (2022). https://www.renolon.com/number-of-tweets-per-day/. Accessed 21 October 2022.
22. B. Mathew *et al*., "Hatexplain: A benchmark dataset for explainable hate speech detection" in *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), vol. **35**, pp. 14867–14875.
23. F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook" in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)* (2017), pp. 86–95.
24. N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, D. Y. Yeung, "Multilingual and multi-aspect hate speech analysis" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, Hong Kong, China, 2019), pp. 4675–4684.
25. "Perspective API." https://www.perspectiveapi.com/. Accessed 12 March 2022.
26. R. Shier, Statistics: 2.3 the Mann-Whitney u test. *Math. Learn. Support Centre*. **15**, 2013 (2004).
27. P. Mishra, M. Del Tredici, H. Yannakoudakis, E. Shutova, "Abusive language detection with graph convolutional networks" in *Proceedings of NAACL-HLT* (2019), pp. 2145–2150.
28. V. Batagelj, M. Zaveršnik, Fast algorithms for determining (generalized) core groups in social networks. *Adv. Data Anal. Classif.* **5**, 129–145 (2011).
29. M. Hoffman, F. Bach, D. Blei, Online learning for latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **23** (2010).
30. S. Phadke, T. Mitra, "Many faced hate: A cross platform study of content framing and information sharing by online hate groups" in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020* (Association for Computing Machinery, New York, NY, USA, 2020), pp. 1–13.
31. M. W. Macy, M. Ma, D. R. Tabin, J. Gao, B. K. Szymanski, Polarization and tipping points. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102144118 (2021).
32. T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language" in *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM 2017* (2017), pp. 512–515.
33. B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, "Spread of hate speech in online social media" in *Proceedings of the 10th ACM Conference on Web Science* (2019), pp. 173–182.
34. A. Founta, et al., Large scale crowdsourcing and characterization of twitter abusive behavior. *Proc. Int. AAAI Conf. on Web Soc. Media* **12** (2018).
35. M. Ribeiro, P. Calais, Y. Santos, V. Almeida, W. Meira Jr., Characterizing and detecting hateful users on twitter. *Proc. Int. AAAI Conf. on Web Soc. Media* **12** (2018).
36. T. Caselli, V. Basile, J. Mitrović, M. Granitzer, "HateBERT: Retraining BERT for abusive language detection in English" in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (Association for Computational Linguistics, Online, 2021), pp. 17–25.
37. A. Koufakou, E. W. Pamungkas, V. Basile, V. Patti, "HurtBERT: Incorporating lexical features with BERT for the detection of abusive language" in *Proceedings of the Fourth Workshop on Online Abuse and Harms* (Association for Computational Linguistics, Online, 2020), pp. 34–43.
38. T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language" in *Proceedings of the International AAAI Conference on Web and Social Media* (2017), vol. **11**, pp. 512–515.
39. K. Wang, D. Lu, C. Han, S. Long, J. Poon, "Detect all abuse! toward universal abusive language detection models" in *Proceedings of the 28th International Conference on Computational Linguistics. (International Committee on Computational Linguistics, Barcelona, Spain (Online)* (2020), pp. 6366–6376.
40. N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, D. Y. Yeung, "Multilingual and multi-aspect hate speech analysis" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 4675–4684.
41. T. Ranasinghe, M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Online, 2020), pp. 5838–5844.
42. S. Ali *et al*., "Understanding the effect of deplatforming on social networks" in *13th ACM Web Science Conference 2021* (2021), pp. 187–195.
43. Y. L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, "CONAN - COunter NArratives through nichesourcing: A multilingual dataset of responses to fight online hate speech" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019), pp. 2819–2829.
44. J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), pp. 4755–4764.
45. M. Fanton, H. Bonaldi, S. S. Tekiroğlu, M. Guerini, "Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2021), pp. 3226–3240.
46. S. Muhammed T, S. K. Mathew *et al*., The disaster of misinformation: A review of research in social media. *Int. J. Data Sci. Anal.* 1–15 (2022).
47. "Raising awareness on hate speech in the republic of Moldova - news." https://www.coe.int/en/web/inclusion-and-antidiscrimination/-/raising-awareness-on-hate-speech-in-the-republic-of-moldova. Accessed 06 July 2022.
48. N. Asimovic, J. Nagler, R. Bonneau, J. A. Tucker, Testing the effects of Facebook usage in an ethnically polarized setting. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022819118 (2021).
49. F. Huszár *et al*., Algorithmic amplification of politics on Twitter. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2025334119 (2022).
50. A. Simchon, W. J. Brady, J. J. Van Bavel, Troll and divide: The language of online polarization. *PNAS Nexus* **1**, pgac019 (2022).
51. S. Zannettou *et al*., "The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources" in *Proceedings of the 2017 Internet Measurement Conference* (2017), pp. 405–417.
52. F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification" in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Association for Computational Linguistics, Online, 2020), pp. 1644–1650.
53. T. Caselli, V. Basile, J. Mitrović, M. Granitzer, "Hatebert: Retraining bert for abusive language detection in English" in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (2021), pp. 17–25.
54. D. Demszky *et al*., "GoEmotions: A dataset of fine-grained emotions" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, 2020), pp. 4040–4054.
55. B. Mathew *et al*., Hate begets hate: A temporal study of hate speech. *Proc. ACM Hum.-Comput. Interact.* **4** (2020).