Identifying Trends in Hate Speech on Gab

Michael Georgariou III

Computer Engineering student, California Polytechnic University, San Luis Obispo

Abstract

As the political landscape of the United States continues to grow more divided and extreme, alternative social media websites such as Gab and Parler have made it easier for users to share controversial and hateful political opinions with many. As these social media platforms continue to gain new members and remain lax on policing the contents of its users, radical ideologies and hate speech are likely to become more accepted and expected on the platforms. Using data scraped from Gab, this paper's goal is to determine if hate speech is becoming more common on Gab after the 2021 US presidential election concluded. By scraping thousands of posts from the site using the hashtag "#maga" from January 2021 and on and using a trained machine learning algorithm to determine hate speech on this data, trends showed that the percentage of "#maga" posts containing hate speech went up, while the total number of posts containing hate speech went down, indicating that discussion within the hashtag on the platform is becoming less popular but more radical.

Introduction

Social media has become a staple in American and worldwide culture. By allowing its users to freely communicate with each other, it has become a space to not only share life updates, but news and opinions. In fact, according to Pew Research from September 2020, 53% of adults get their news from social media sites (Shearer, 2021). Although in some ways this is a positive, there are also negatives that come with people relying on social media for political and social news.

Although social media has been important in politics for over a decade, the Trump presidency has led to social media being at the forefront of political discussion, due in large part to his use of Twitter to divert crucial media from reporting on topics harmful to him and his campaign (Lewandowsky, Jetter and Ecker, 2020). In fact, at least on social media website Twitter, Americans are identifying more closely with their political ideology than they are their religion (Rogers and Jones, 2020).

Following the 2021 storming of the United States Capitol, then-president Donald Trump was banned from Twitter due to "risk of further incitement of violence", according to a blog post detailing the ban (Twitter, Inc, 2021). After this ban, alternative, right-leaning social media sites such as Gab experienced a huge increase in traffic, with Gab gaining 10,000 new users per hour, according to the CEO of Gab (Lonas, 2021). Because of the legitimization of the platform due to these events, and the platform's lack of content policing, it is a great place to understand where far right factions of American politics currently are (Hess, 2016).

In this paper, we will attempt to better understand if right-wing hate speech on Gab increased significantly following the 2021 US presidential election. Using a hate speech corpus compiled by authors at the University of Southern California in their paper titled "The Gab Hate Corpus: A collection of 27k posts annotated for hate speech" (Kennedy et al, 2018), we will train a machine learning algorithm to determine if a social media post contains hate speech or not. Using posts scraped from Gab from January 2021 to April 2021, we will then determine whether the rate of hate speech in posts on Gabs that used Donald Trump-related hashtags changed.

Data

We will briefly describe the methods used to gather data from social media website Gab, as well as the pre-scored hate corpus being used.

Collecting data from Gab

Because we wanted to see if Trump supporters on Gab were becoming more extremist after the 2021 election took place, we chose to scrape all posts from Gab that included the hashtag "#maga" from January 2021 onward. This would allow us to see which posts were being posted with the intention of being seen by other Trump supporters.

We collected data from Gab using an open-source command-line interface for the Gab.com API called Garc. Garc allows searching for posts via hashtag, text, and username. However, as Gab's API is very poorly documented, Garc is very barebones and does not allow for much modification. Garc scrapes from the most recent post and goes back from there, meaning we cannot specify a date range to search on. To get results we can work with, we had to scrape from when we ran the script and let it go back to January on its own. More interesting data could be obtained if in the future we determine a method to specify an exact date range to get data from.

Our data contains 60,600 posts from the site ranging from January 29, 2021 to April 4, 2021. This range and data size allows us to see if hate speech declined or increased post-election on the platform.

All data obtained from Gab had to be stripped of all non-text content, including emojis, HTML, and extra whitespace. This was to ensure the only content we would be analyzing using our machine learning algorithm would be the text of the posts, and not falsely identify certain emojis or HTML tags as hate speech.

The data Garc was in the JSON format, which makes it easily readable by both humans and computers. In the data for each post is contained an ID, date created, URL, reply/reblog count, username, hashtag list, and text, among other categories not useful to us for the purposes of this paper.

Determining what qualifies as hate speech

Thankfully, the question of "what qualifies as hate speech" is not a question this paper attempts to solve. To determine which posts are hate speech, we will be using "The Gab Hate Corpus", compiled by computer scientists, psychologists, and political scientists at the University of Southern California (Kennedy et al, 2018). This corpus contains 27,665 posts from Gab, each annotated by three trained annotators. These annotators labeled whether a post contained hate speech, among other categories. We focused on the hate speech classifications they made. By using this classification, we can determine which Gabs we scraped contain hate speech not by our own judgment, but instead by training a machine learning algorithm on their findings.

Methodology

To determine which of our scraped posts contained hate speech, we needed to create a machine learning algorithm that could take in a Gab post's contents and output how likely it is that the post contains hate speech. By having the pre-annotated posts from the previously mentioned Gab Hate Corpus, this process was made easier for us.

Using Pandas, a popular Python data analysis library, we imported the annotations file provided by the Gab Hate Corpus paper. We then split this data into a training set and a test set, which were 80% and 20% of our data, respectively. The training set was put in a vectorizer, which is the data type expected by machine learning algorithms.

Once we had obtained a vectorizer for all our annotated hate speech, we were able to create a model and train it on the vectorizer. Using TensorFlow, an open-source platform for machine learning, we created a Sequential model with 4 layers. A sequential model was chosen because this is the most appropriate model when each input has a single output. As our input is simply a Gab post, and our output is the likelihood the post contains hate speech, this was a perfect model for this project. The model is created only on the input data (the Gab posts).

Once the model had been created, we then had to fit the model to the outputs as written in the Gab Hate Corpus, meaning we needed to map each input to whether the corpus had identified that post as containing hate speech. We did this using 2 epochs and a batch size of 350, meaning that the model would be fit two times, each time using a sample size of 350 posts. Doing this gave an accuracy of 0.8769 and a loss of 0.3360, where accuracy refers to the percentage of posts that were correctly identified, and loss refers to the amount posts were off by. These numbers are incredibly good and give us confidence on this machine learning model. If we wanted to be more accurate, we could use a greater number of epochs, but we stopped here as we already had a great model to work with. To ensure that this model was predicting hate speech correctly, we tested this newly trained model on the 20% of the data we previously set aside as our test set, which resulted in an accuracy of 0.8739. Again, this gives us confidence that our model can correctly predict when a post contains hate speech most of the time.

The JSON data we scraped from Gab next needed to be prepared so that it can be input into our machine learning algorithm. Again using the Pandas library, we were able to read in the JSON file easily (as a function to read in JSON files already exists in the library). We then dropped all duplicates from our data to ensure we did not count any post more than once. The data then needed to be scraped of all emojis and HTML tags. This was done using functions we wrote ourselves. We then took only the text from each post and put it in a vectorizer similarly to how we handled our training and test sets. Once this was completed, we had our data ready to run through our machine learning algorithm.

Running the data through the algorithm returned a list of percentages, which essentially corresponded to how likely the given post contained hate speech.

We chose two percentages to use as our cut-offs for hate speech – 25% to ensure we captured all instances of hate speech (even if there were some false positives), and 40% to see if trends observed persist with all false positives removed (even if it meant also removing some real hate speech from our analysis).

Results

After retrieving all the results from the machine learning algorithm, we now had the likelihood each of our 60,600 Gab posts we scraped contained hate speech.

The minimum score for our dataset was 0.0000967%, meaning the post was basically guaranteed to not contain hate speech. The maximum score for our dataset was 73.262%, which meant the algorithm was all-but-certain the post contained hate speech. The mean score was 9.028%.

Below are some examples of posts with different scores, included to demonstrate what different scores correspond to in messages. These messages have been selected to show what posts near the top of scoring, near our two chosen thresholds (40% and 25%), and near the bottom of scoring look like. This is done to better understand what it means when the rate of posts above 25% is going up, for example. Any slurs or other profanity have been replaced with asterisks.

CONTENT WARNING: Some of these posts contain very hateful language involving minority groups, including American Jews, African Americans, and undocumented immigrants.

Score of 72.79%:

> “#jews #jewish #parler #gab #nazi #maga #trump #conservative
>
> THE NEXT THING THE JEWS HAVE PLANNED FOR US IS TO STEAL FROM US TO GIVE TO N******. WE HAVE ALREADY LOST OUR SCHOOL ADMISSIONS TO N******, WE HAVE LOST OUR JOBS TO N******, WE HAVE LOST OUR PROMOTIONS TO N******, AND WE HAVE LOST OUR PRIDE AND DIGNITY TO N******, BUT NOW THE JEWS WANT TO RUB IT IN BY STEALING OUR MONEY AND GIVING IT TO N******. FIGHT BACK!!!
>
> <link removed>”

Score of 45.02%:

> “Remember democrats/liberals despise you. Democrats/liberals despise your family your values and your beliefs. Democrats/liberals have found your replacements they’re the illegals. They were invited here by Biden. The illegals are being promised your money your house and your life. WAKE THE F*** UP. #Trump #MAGA #WALL”

Score of 25.60%:

> “Trump pulled back the curtain and revealed the evil Establishment made

> up of Leftists and RINOs who are hell bent on maintaining power.
>
> The swamp is deep and corrupt to the core, but Patriots are willing to fight hard to take back our Republic and stand for truth and honor.
>
> Trump started the fight against the Elitists, and we, the brave and proud Americans, will finish it!
>
> #MAGA #America1st #DrainTheSwamp #USA"

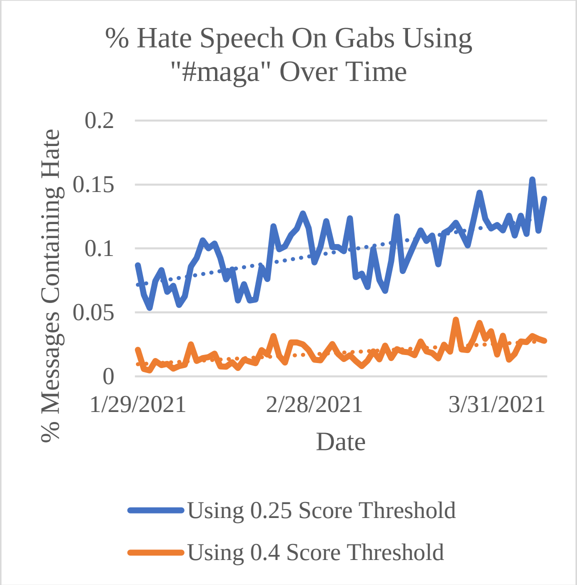
Score of 1.04%:

> "💪Today is a good day to get your keto meal plan, click on the link in my bio. 💥Blessings💥
>
> #change #health #diet #keto #maga #freedom"

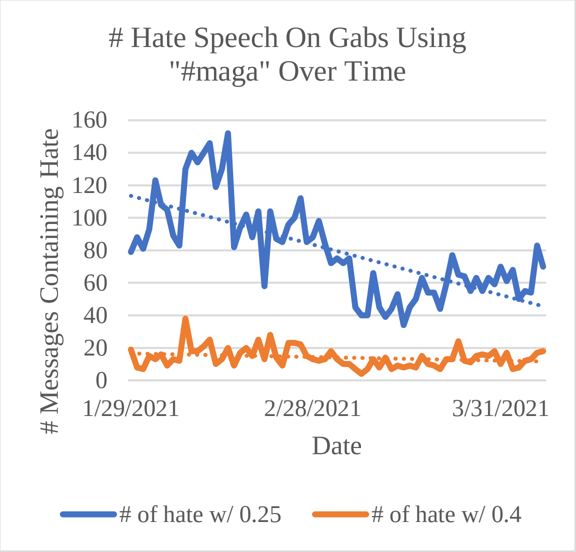
As we can see, the #maga hashtag on Gab has posts ranging from deliberate calls to violence and hate speech towards minority groups in America to advertisements about a diet plan.

What this paper was most interesting in was how the rate of hate speech changed over time. By taking all our data and graphing the date it was posted vs. the percent of messages that day that crossed a given threshold, we can see trends in how common hate speech became within the #maga hashtag. We can also compare this to the raw number of posts and how this trend differs, as well as the total number of posts within the hashtag, whether they contained any hate speech.

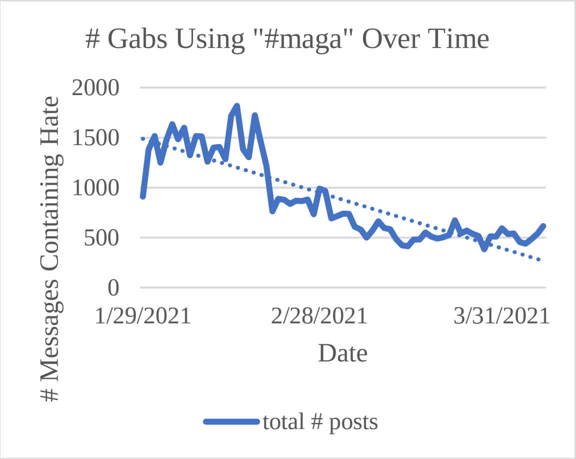
The below graph shows how the rate of hate speech within the hashtag changed over time following the inauguration of Joe Biden in January of 2021. Although it fluctuates, both the 25% and 40% thresholds show a linear trend in the positive direction, indicating that the ratio of hate speech to normal posts in the #maga hashtag has increased over time since January. This graph contains data from January 29, 2021, to April 7, 2021.



Comparing this to the next graph, which charts the number of hate speech posts (not accounting for the total number of posts), it is interesting to note that as the rate of hate speech increased as shown in our previous graph, the number of posts that contain hate speech went down. The linear trendline indicates this. This could signify that while usage of the #maga hashtag has decreased significantly since the election, the percentage of posts within the hashtag containing hate speech has gone up.



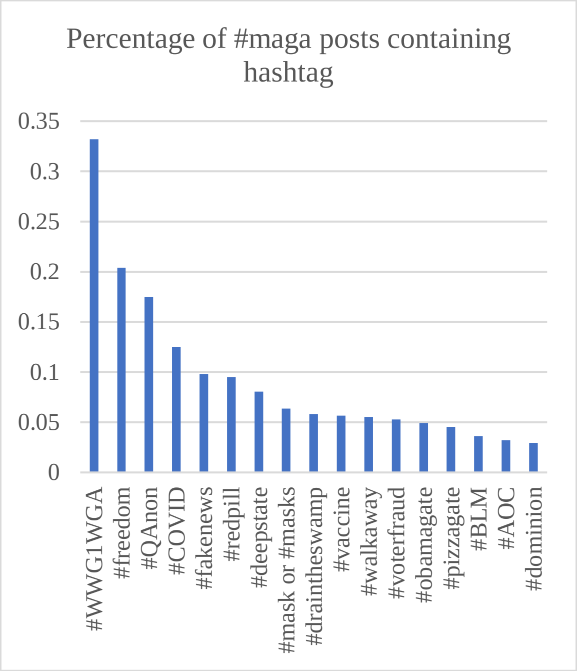
As mentioned before, the number of Gab posts that use the #maga hashtag has significantly decreased since the election occurred. While this does explain the amount of hate speech occurring within the hashtag decreasing in a similar pattern to the hashtag overall, it also indicates that either the accounts still using the hashtag are ones that are more radical and hateful, or that the hate has driven away users who were not hateful from the community.



Worth noting is spikes in this data. February 13-14, 2021, saw the greatest number of posts in the hashtag over this period. This likely corresponds to Donald Trump’s second impeachment trial, which led to the acquittal of the former president on claims of inciting an insurrection at the Capitol (Herb et al, 2021). It is also important to note that while hate speech in the hashtag also peaked on these days, the ratio of hate speech posts to non-hate speech posts was lower these days, meaning a more active use of the hashtag led to a lower rate of hate speech within it.

Another spike to look at is when the number of posts that fell under the 25% category fell while the number of posts that fell under the 40% category rose. This would imply that posts contained more hate on these days. This occurred on February 7, 2021. No major events occurred in politics on this day, but it is still interesting to see this spike occur.

Other data worth at least mentioning is the overlap of other relevant political hashtags in the #maga hashtag over this time frame. Below is a graph indicating popular hashtags that appeared in the same posts as #maga.



It is interesting to note that many of highest percentage of overlapping posts within the #maga hashtag are related to conspiratorial ideas, such as QAnon and “the deep state”. Although this does not directly correlate to the hate speech previously analyzed, it is interesting to consider.

Discussion

While this paper presents many interesting findings about how Donald Trump supporters interact with Gab as a platform, it also raises many more interesting questions that were beyond the scope of this research. Questions such as “how do these trends relate to Gab as a whole?”, “how does Gab compare to other platforms, such as Twitter or Parler?”, and “do we see similar rates and trends in hate speech in left-leaning hashtags?”. Given more time, we feel exploring these questions could lead to more intriguing data and discussion.

We also encountered issues that limited the scope of this research. If solutions to these problems are solved, much more data can be collected and analyzed to see greater trends on Gab as a whole.

Gab API issues

One area holding back our research was the ability to scrape good data from Gab. The API we employed to get posts from the site frequently crashed, with no error logs or message indicating why it stopped scraping data. This made good runs to scrape data very rare. If it were not for this, we would have analyzed more hashtags related to the Donald Trump presidency in similar ways to see if any hashtags were used more than others for spreading hate speech.

The API also limited how we can search through the site. It does not offer an option to specify a date range – only an option to specify the earliest date to search until. This means all calls to the API will return posts starting from the day you run the call and work backwards to the date you gave it. This means if we wanted to get older data, we would need to let the API run for a longer time, and the farther in the future you do this the longer you would need to run the API to receive the same posts. Furthermore, you would have to hope the API does not crash as you perform this scrape, because if it does, you will have to start your search again.

We attempted to modify the API to allow us to specify a date range (as it is written in Python and is available on GitHub as open-source code) but had no luck. With more time, finding a way to scrape Gab data more efficiently and reliably would open many more opportunities for what we can analyze on the platform.

Hate speech trends pre-2021

The biggest question this paper was unable to answer was how rates of hate speech on Gab changed before and after the election. Seeing whether hate speech increased or decreased after the election would be important in determining if far-right circles online further radicalized themselves once Donald Trump had officially lost, if they became more moderate, or if they lost interest in discussing politics online as a whole. This is a question that could shine light on the way the Republican party and its voters are shifting in opinion as well.

Gab compared to other social platforms

Another question that should be researched further is how Gab compares to other social media platforms. It has been shown that other, more mainstream social media platforms have a thriving far-right community as well, including Facebook (Edelson et al, 2021) and Twitter (Bucklin, 2019). Although Gab has gathered a notably more far-right audience than other social media platforms (Kantrowitz, 2016), we do not know from our research whether this also leads to a higher percentage of hate speech on the platform. Gab does not have rules against hate speech (compared to the other major platform previously mentioned, which do), and it would be interesting to see if this different ruleset really does lead to a greater percentage of hate speech on the platform.

Conclusion

As discussed, hate speech on Gab and specifically within pro-Donald Trump posts is ever present. As more extreme right-leaning people move away from traditional mainstream social media websites, such as Twitter and Facebook, alternative social media sites like Gab gain more dedicated users. As the mainstream sites continue to create more rules and regulations regarding hate speech on their platform, this trend will likely continue. The growth of these platforms is aided by prominent figures in right-wing politics being banned from Twitter and Facebook and moving to these alternative platforms. This means that even if these posts and conversations are out of sight for many, we must continue to monitor the trends occurring on these sites and be on the lookout for the potential of hate speech to turn into calls to violence.

Although the total number of posts in the #maga hashtag on Gab has decreased as time goes on, the percentage of these posts that contain hate speech has inversely increased. This indicates that while the MAGA and Trump movement may be losing steam (at least for now), the people who remain active in these communities are more likely to be hateful and extreme.

As Gab continues protecting and giving a platform to these types of messages, more people who may identify as centrist, libertarian, or right-leaning may begin identifying with and agreeing with these more extreme messages on the platform. This could lead to further radicalization and eventually culminate in events like the storming of the Capitol we witnessed in early 2021. The first step in stopping this radicalization is identifying where it occurs.

The #maga hashtag on Gab is slowly becoming more radical, as previously mentioned. As time goes on, the more likely you are to see hate speech if you search for “maga” on Gab. This is harmful not only to the people being spoken of negatively, but the people reading these messages and the people who receive attention for posting them. We hope that by identifying areas online in which hate speech is not only shared, but is increasing every day, we can help determine how it occurs and what can be done about it.

References

Beirich, H. and Agnew, L. (2019). “The Year in Hate: Rage Against Change.” *Southern Poverty Law Center,* https://splcenter.ogr/fighting-hate/intelligence-report/2019/year-hate-rage-against-change

Bucklin, N. (2019). “Exploring Right-Wing Extremism on Twitter.” *Towards Data Science*, https://towardsdatascience.com/exploring-right-wing-extremism-on-twitter-941a9d02825e

Edelson, L. et al (2021). “Far-right news sources on Facebook more engaging.” *Medium: Cybersecurity for Democracy*, https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90

Herb, J. et al (2021). “Trump acquitted for second time following historic Senate impeachment trial.” *CNN Politics*, https://www.cnn.com/2021/02/13/politics/senate-impeachment-trial-day-5-vote/index.html

Hess, A. (2016). “The Far Right Has a New Digital Safe Space.” *The New York Times*, https://www.nytimes.com/2016/11/30/arts/the-far-right-has-a-new-digital-safe-space.html

Kantrowitz, A. (2016). “This New Social Network Promises Almost-Total Free Speech to Its Users.” *BuzzFeed News*, https://www.buzzfeednews.com/article/alexkantrowitz/new-social-network-gab-growing-fast-free-speech

Kennedy, B. et al (2018). “The Gab Hate Corpus: A collection of 27k posts annotated for hate speech.” *PsyArXiv*.

Lewandowsky, S., Jetter, M. and Ecker, U. (2020). “Using the president’s tweets to understand political diversion in the age of social media.” *Nature Communications*.

Lonas, L. (2021). “Social media platform Gab gains traffic, users following Capitol riot fallout.” *The Hill*, https://thehill.com/policy/technology/533502-far-right-social-media-platform-gaining-traction-captiol-riots

Rogers, N. and Jones, J. (2020). “Using Twitter Bios to Measure Changes in Self-Identity: Are Americans Defining

Themselves More Politically Over Time?" *TUP*.

Shearer, E. (2021). "More than eight-in-ten Americans get news from digital devices." *Pew Research Center*, https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

Subel, D. and Madiraju, S. (2020). "The YouTube Radicalization Problem." *California Polytechnic State University*.

Twitter, Inc. (2021). "Permanent suspension of @realDonaldTrump." *Twitter Blog*, https://blog.twitter.com/en_us/topics/company/2020/suspension.html