Gab Hate Speech: Word Counting, Topic Modeling, and Classification

Christina Corrado, Dylan Guzman, Ethan Hsu, Matthew Reardon
Binghamton University

Abstract / Executive Summary

Gab is a social media platform that contains minimal restrictions on what a user can post. This has led to much hate speech being posted on the platform. The Gab Hate Corpus is a dataset containing 27,665 posts that have all been labeled as hate speech or not. To uncover trends in this data, we have implemented word counting, topic modeling, and text classification techniques to help aid our analysis. We found that these posts from Gab contain blatant themes of racism, homophobia, islamophobia, and most of all, anti-Semitism. Thus, leading to the claim that Gab's platform contains obvious far-right, racist, homophobic, and neo-nazi ideologies on its platform.

1. Introduction / Background

The Gab Hate Corpus, published in 2018, is a collection of social media posts from Gab. Gab was founded in 2016 and acted as a platform without speech restrictions, unlike other mainstream social media outlets. Due to the lack of speech restrictions, Gab is " inhabited by deplatformed white nationalists, neo-Nazis, and other hate-mongering ideologue"[1] .Gab continued to remain an outlet for this form of hate speech until the 2018 Pittsburgh synagogue shooting, when it was found that the shooter was an active Gab user. Gab was shut down, and companies like PayPal cut their ties with the site.

This site acted as an example of the online prevalence of hate speech. The growth of the internet has created more outlets for anyone to voice their opinions without repercussions, "increasing the visibility of aggressive, attacking, dehumanizing, and potentially dangerous language"[1]. This increased exposure has redirected researchers to the study of hate speech. In past research, issues arose when trying to create a universal definition for hate speech. The lack of legal consensus among countries on what constitutes hate speech, the unclear motives of those studying it, and the blurred lines between hate speech and freedom of speech. contribute to disparities in its definition. These factors make the need for a sound definition and coding procedures even more important.

Gab no longer holds the mainstream presence it once did. It made headlines for the aforementioned Pittsburgh synagogue shooting in 2018, but then again during the storming of the Capitol in 2021. It turns out that much of the planning for the attack was done out in the open on Gab's platform. The site was criticized by popular news outlets such as the New York Times and ABC News. However, the site saw one of its best weeks ever following the attack, seeing as much as an 800% increase in users[2]. Since the storming of the Capitol, Gab has failed to make it to the mainstream media. Internationally, Gab has "connections to German language speakers, and significant content originating from Russian state-sponsored sources seems to suggest emergent linkages

[1] Kennedy, Brendan, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, et al. 2022. "The Gab Hate Corpus." OSF. April 12. doi:10.17605/OSF.IO/EDUA3.

[2] Allyn, Bobby. "Social Media Site Gab Is Surging, Even as Critics Blame It for Capitol Violence." *NPR.org*, 17 Jan. 2021, www.npr.org/2021/01/17/957512634/social-media-site-gab-is-surging-even-as-critics-blame-it-for-capitol-violence. Accessed 3 Mar. 2025.

[3]Yuchen Zhou, et al. *View of Elites and foreign actors among the alt-right: The Gab social media platform | First Monday*. (n.d.). Firstmonday.org. https://firstmonday.org/ojs/index.php/fm/article/view/10062/8072. Accessed 3 March. 2025.

between online discussions among the alt-right and various foreign entities”[3]. The site is still active, but little information about the site has been reported for the past few years.

2. Data

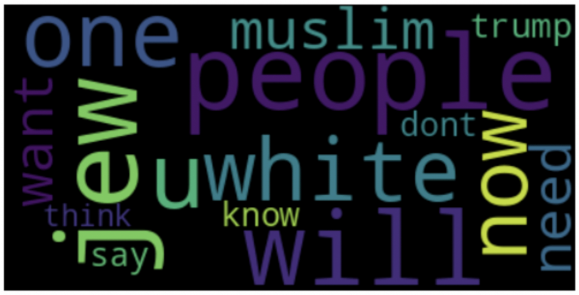
There are 27,665 posts from gab.com annotated by a minimum of three annotators searching for indications of hate-based rhetoric. As previously mentioned, there is no clear-cut definition of what constitutes hate speech, so the annotators answered four questions per post in an attempt to categorize them: Whether it was hateful, if it used vulgar language, if it targeted a group, and whether it was explicit or implicit. Posts were considered hateful if they encouraged a hostile environment or directly attacked human dignity. The other three questions were less subjective. A post was deemed to use vulgar language if a slur or offensive language occurred. It would be analyzed for any stereotypes that targeted a specific group. Last, the post was determined to be explicit or implicit if it was direct in its meaning or only implied its meaning. The annotators were to label their answers to each question in a separate column for each post[3].

The actual dataset uses a variety of acronyms to answer these four questions. To answer if the post was hateful, the post was labeled as hateful or not hateful (NH), and then it was decided if it was a call to violence (CV) or an assault on human dignity (HD). For the next question, if vulgar language is used, it uses the acronym VO to represent derogatory language. It’s important to note a post can be labeled as not hateful (NH) but still contain derogatory language (VO). Concerning the target groups, it gives various acronyms to describe all the possible groups of people being targeted in a hateful post. These are race and ethnicity (RAE), nationality/regionalism (NAT), gender (GEN), religion (REL), sexual orientation (SXO), ideology (IDL), political identification (POL), and mentally/physically disabled (MPH). The question of explicit or implicit would be marked EX(explicit) or IM(implicit).

3. Results and Discussion

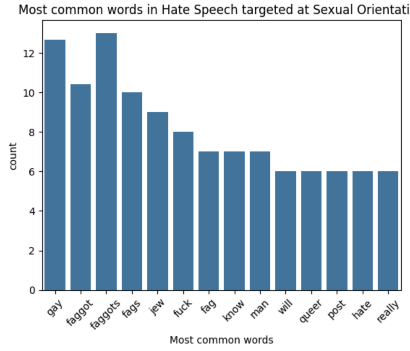
3.1. Word Counting

To apply word counting techniques on the Gab Hate Corpus, we first cleaned the unnecessary columns in the data (GabHateCorpus_annotations.csv) and removed the missing values. Then, we applied the concept of word counting to analyze the overall frequency of words in the dataset. We removed stopwords, such as “and,” “the,” or “is” because these words are not relevant. We utilized the function of value_counts() to measure the occurrences of each word, which identifies the most common terms in the dataset. These methods provided a level of insight into dominant themes and recurring language patterns, as well as a reflection of the type of user base and information talked about. We created a word cloud to highlight the 15 most common words. If a word has a higher frequency, then a larger font size will display it. Unsurprisingly, there were top words about race and politics. The most common racial words are Jew and White. Gab is a far right extremist platform with Neo Nazi users. Thus, from this information about Gab and the word cloud, Jew is a common racial term used in negative connotation compared to other races on this platform. The key word Muslim tells us there is Islamophobia present on this platform as well as anti-semitism. One of the key insights from the word cloud is the identification of the racial group most frequently targeted on Gab.

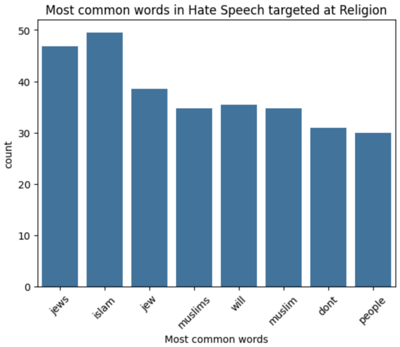


We also wanted to look at the top words used in posts related to all the target groups. This provided very valuable insight into what keywords or phrases would cause these posts to get labeled as targeting certain groups. We decided to visualize Sexual Orientation, Religion, and Race and Ethnicity because they provided the most insight.

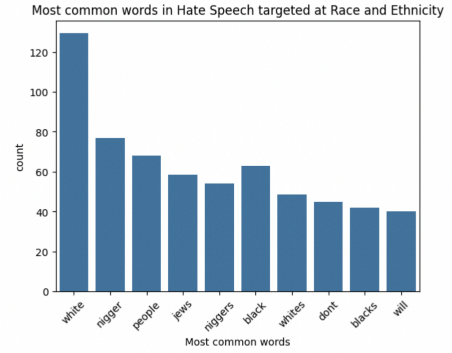
[3] Kennedy, Brendan, et al. “The Gab Hate Corpus.” *OSF*, 18 July 2018, osf.io/edua3/, https://doi.org/10.17605/OSF.IO/EDUA3. Accessed 3 Mar. 2025.



In the visualization above, by assessing the top words, you can see posts that are labeled as hateful that target sexual orientation contain derogatory terms used when referring to someone's sexual orientation. It also shows us some very interesting words when referring to sexual orientation, such as “jew”. In other platforms, Jew would not be commonly referred to in sexual orientation hate speech. This word is specifically symbolic of the anti-semitic presence on Gab. Words like “hate” or “man” also provide us with an interesting insight into what words are being used along with the hateful speech in these posts. Why would man be common but not woman? I looked into the sexual orientation hate speech that contains the word man to find an answer. Some of the appearances of the word were used in a general sense of man and woman. However, most of the hate speech was in fact only referencing gay and straight men. This sheds light on the prominence of the stereotype that all men need to be “manly”. This aligns with Gab’s predominantly far-right extremist and male-dominated user base.



In this graph above, we can see that posts labeled as hate towards Religion seem to be anti-Semitic and Islamophobic. Every one of the top words that directly refers to religion talks about Judaism and Islam. This is also the second time that jew is appearing as the top word in a different category.

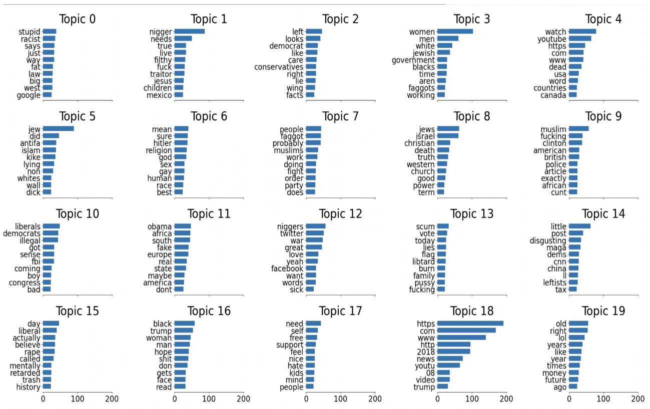


In this bar plot, by looking at the top words, it is pretty evident that this is about race and ethnicity. What is very interesting is that “jews” manages to appear again. This is interesting because the word “jews” seemingly shouldn’t be a top word when referring to certain categories such as sexual orientation or race and ethnicity, yet it is. This reinforced that there seems to be a strong anti-Semitic, far-right, and neo-nazi, male presence on the Gab platform.

Overall, these word count graphs for different targeted groups displayed clear themes that are expressed in every classification of the platform. We were able to gain greater insight than the first general word count we did.

3.2. Topic Modeling

We applied LDA (Latent Dirichlet Allocation) topic modeling to gain an understanding of the different groupings and themes within hate speech terms. LDA allowed us to organize the tokenized words into unnamed topics and analyze the patterns in the dataset. We examined side-by-side grids displaying 20 topics, each highlighting the top ten words. These top ten words would represent which ones are most unique to that unnamed group. Beyond these visualizations, analyzing the distinct weight of each word within a topic helped us infer potential subject matter. For instance, in Topic 10, the words “liberal” and “democrats” had weights of 48 and 43, respectively, suggesting that this topic pertains to hate speech related to political controversies. Similarly, Topic 8 featured extremely high weights for the words “Israel” and “Jew,” alongside terms like “Christian,” “death,” and “church,” indicating a focus on religious hate speech. This aligns with the Gab Hate Corpus classification, which categorizes hate speech by targeted populations, including RAE (race or ethnicity), NAT (nationality/regionalism), GEN (gender), REL (religion), SXO (sexual orientation), IDL (ideology), POL (political identification), and MPH (mental/physical health status).

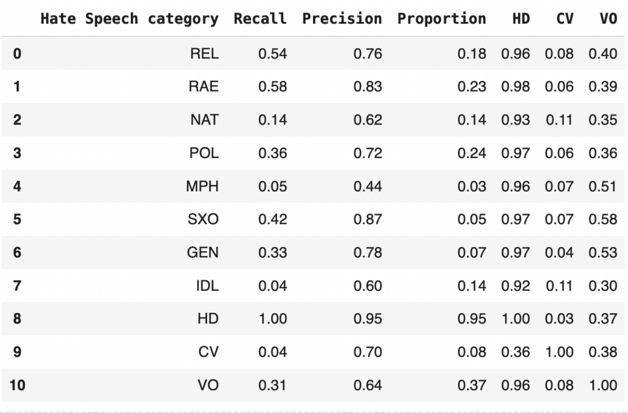


The LDA visualization effectively reveals multiple topics that correspond to these predefined groupings. These groupings only help demonstrate some of our findings in the word counting section. Just upon looking among all the topics, we see the most unique words being slurs and derogatory terms used when in hate speech. Furthermore, we see the word “jew” or some form of it in multiple different groups again. This agrees with our findings in the word count that the term “jew” is being used in all different forms of hate speech.

3.3. Text Classification

In our next step, we performed classification metrics on the dataset. We wanted to determine how accurately the data can be classified into the different hate speech categories. Using an SVM, we trained the classifier model on training sets of our data. We decided to use a SVM because we were trying to predict if a post was hateful towards a group or not, and this binary decision is where a SVM can be particularly useful. We chose the values we wanted to predict as the hate speech category we want the model to predict. The training values were the text data. After comparing the classification metrics of each type of hate speech, we uncovered the most prevalent forms of hate speech and whether these forms of hate speech are easily classified within our model.

.



	Hate Speech category	Recall	Precision	Proportion	HD	CV	VO
0	REL	0.54	0.76	0.18	0.96	0.08	0.40
1	RAE	0.58	0.83	0.23	0.98	0.06	0.39
2	NAT	0.14	0.62	0.14	0.93	0.11	0.35
3	POL	0.36	0.72	0.24	0.97	0.06	0.36
4	MPH	0.05	0.44	0.03	0.96	0.07	0.51
5	SXO	0.42	0.87	0.05	0.97	0.07	0.58
6	GEN	0.33	0.78	0.07	0.97	0.04	0.53
7	IDL	0.04	0.60	0.14	0.92	0.11	0.30
8	HD	1.00	0.95	0.95	1.00	0.03	0.37
9	CV	0.04	0.70	0.08	0.36	1.00	0.38
10	VO	0.31	0.64	0.37	0.96	0.08	1.00

- Most Prevalent Targeted Population in Hate Speech (Proportion):
    - Political Identity (POL): 0.24%
    - Racial/Ethnic Identity (RAE): 0.23%
    - Religious Identity (REL): 0.18%

- Least Prevalent Targeted Population in Hate Speech (Proportion):
    - Mental and Physical Health (MPH): 0.03%
    - Sexual Orientation (SXO): 0.05%
    - Gender Identity (GEN): 0.07%

These proportion numbers mean what proportion of the overall posts did that group make up. Based on the proportion levels, this analysis was not surprising, because Gab was a far-right, dominant platform, where a large amount of the hate speech included speech about political parties and ideals. Racial/Ethnicity Identity and Religious Identity are both incorporated in these political attacks. Furthermore, this is heavily influenced by ideological and cultural divisions, with political ideology and racial/religious tensions. Gab tends to attract individuals who are more likely to express extreme views on politics and identity, which explains why these categories are more dominant in hate speech. In contrast, hate speech targeting Mental and Physical Health, Sexual Orientation, and Gender Identity has lower proportions because the topics users choose to engage with and the broader sociopolitical group within the Gab community can be driven by the ideological grievances and culture war narratives, rather than expressing hostility related to gender or health-based identities.

- Most Prevalent Targeted Population in Hate Speech (Precision):
    - Sexual Orientation (SXO): 0.87%
    - Racial/Ethnic Identity (RAE): 0.83%

- Least Prevalent Targeted Population in Hate Speech (Precision):
    - Mental and Physical Health (MPH): 0.44%

Precision tells us how many of the instances a model predicted as a specific target group were actually that group. SXO and RAE have very high precision because of the consistency in keywords within the text. This makes the model accustomed to the specific target group classification. MPH is the lowest precision because of the low proportion size within the data. There are only 314 cases out of 11249 cases classified as MPH. There is a much smaller training set for the model to detect these cases compared to other groups. Thus, there would be a greater amount of false alarms if certain cases were detected as MPH because of the low 0.03% proportion.

- Most Prevalent Targeted Population in Hate Speech (Recall):
    - Racial/Ethnic Identity (RAE): 0.58%
    - Religious Identity (REL): 0.54%
    - Sexual Orientation (SXO): 0.42%

- Least Prevalent Targeted Population in Hate Speech (Recall):
    - Ideology (IDL): 0.04%
    - Mental/Physical Health (MPH): 0.05%

Recall tells us the percentage of each targeted population that is found. The proportion of cases can be a factor in the recall because when there is a larger number of cases for a specific group, the model might prioritize classifying it. Thus, REL and RAE have the highest proportions and highest recall. However, SXO has a higher recall but a lower proportion. When training the model on SXO cases, the testing data has higher metrics than the training data, meaning the model is overfit. Basically, the model was memorizing the data, and not actually learning it. This is because each hate speech case for SXO had very consistent words such as “gay”, “homosexual”, and “faggot”. The model was memorizing the training data. IDL has a very low recall because of its ambiguity and low proportion; there are scarce consistent keywords in the text, which makes it harder to detect. In addition, MPH has a very low recall because of its low proportion. However, there are more consistent words such as “autistic” and “retard” in the text. So, in this case, I think the drastically low sample size made it difficult for the model to classify.

As far as the types of hate speech, 95% of the dataset was labeled as an attack on human dignity, making it the most common type of hate speech, while 37% of posts were labeled as vulgar or offensive language, and only 8% were labeled as calls to violence. The three types of hate speech had decent to high precision scores, while having a wide range of recall scores:

- Human Degradation (HD): 1.00 Recall, 0.95 Precision
- Calls to Violence (CV): 0.04 Recall, 0.70 Precision
- Vulgar/Offensive Language (VO): 0.31 Recall, 0.64 Precision

We filtered our dataset to only contain instances of hate speech. Attacks on human dignity are defined as instances that degrade a certain group or population, which aligns with most hate speech. Calls to violence are posts that may cause aggression, so it makes sense that there is a small amount of these posts since most hate speech causes aggression. Vulgar or offensive language is commonly used in hate speech, but it is not a requirement. This difference in proportions may explain why there is a large discrepancy in recall scores. 95% of the posts were labeled as HD, which likely led to overfitting. It is possible the model predicted every post as HD, which would explain a recall of 1 and a precision that matches the proportion of posts that are HD. Due to a lack of CV posts in the dataset, the model likely did not have enough data to make accurate predictions. Additionally, CV posts are more subjective than the other types of hate speech, which may make it more difficult for the model to predict. Last, VO posts received a low recall score because posts were labeled as VO if vulgar language was used in a harmful way. Many posts used offensive slurs not at a specific group, which wouldn't be labeled as VO, but would still use language commonly used in VO posts, which would lead to false predictions by the model.

4. Conclusion

Gab is a social media platform that contains a vast amount of hate speech. Through our word counting, topic modeling, and text classification analysis, we were able to uncover some of the trends within the hate speech. Our analysis showed that the platform contained very racist, homophobic, and Islamophobic tendencies. However, what was easily the most apparent was the extreme amount of anti-semitism on the platform. The word "jew", or some form of it, was in multiple topic modeling groups and was in all our word counting analysis. Also, in our text classification models, REL contained one of the highest recalls and precisions. Overall, through our text analysis of the Gab Hate Corpus, we found that Gab harvests tons of far-right, racist, homophobic, neo-nazi ideologies on its platform.