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

By: Christina Corrado, Dylan Guzman, Matthew Reardon, Ethan Hsu

Table of contents

01

Historical Background

Gab Hate Corpus

What is it?

- Largest theoretically justified and annotated hate speech corpus to date, offers valuable resources for training and evaluating hate speech classifiers, as well as for exploring the linguistic and network aspects of hate speech.

Background

- Founded in 2016 by Andrew Torba. Acted as a social media platform that promotes free speech.

- Became a platform for far-right extremists to express controversial and often harmful views with minimal restrictions.

- Gab experienced steady growth in popularity until the 2018 Pittsburgh synagogue shooting. It was revealed that the perpetrator was an active user of the platform. The site was temporarily taken offline for several weeks, and companies like PayPal severed ties with Gab.

Features

- **Limited content moderation**: Allows broader speech freedoms, attracting users that were banned from mainstream platforms.

- **Decentralized integration**: Connects with the Fediverse, enabling cross-platform interaction with services like Mastodon.

- **Independent funding**: Operates without ads or corporate influence, relying on user subscriptions and donations.

Conflict Factors with Hate Speech

- The growing use of the internet and social networking platforms has increased exposure to hate speech, highlighting the need for studies to understand its causes, effects, and prevention methods.

- **Risk factors and previous conflicts:**
  - The absence of a legal consensus among countries makes it challenging to establish a universal definition of hate speech.
  - Companies studying hate speech, such as Twitter and Facebook, have varying motives and objectives.
  - The complex relationship between hate speech and freedom of speech creates inconsistencies in categorizing and identifying hateful content.

Aftermath & Success

Aftermath

- Gab gained popularity within far-right communities.
- Faced significant backlash from mainstream media for fostering a racist and neo-Nazi user base.
- Has not received major media attention since 2021.
- Internationally, lots of content from Russian sponsored sources and german language speakers suggest discourse among alt-right users and foreign elites

Success

- Rebuilt its infrastructure to become self-hosted, ensuring continued operation despite repeated deplatforming.
- Enabled decentralized interactions, reducing censorship risks.
- Retained loyal following, particularly among far-right and libertarian communities advocating free speech.
- The data in GHC can help improve NLP (Natural language processing) models that will identify hate speech

02

Dataset

All 27,665 posts come from gab.ai, all annotated by at least 3 trained annotators

Origin

Labor/People

Researchers from Southern California

July 2018

Published

Data

Methodology

- Contains posts from Gab and whether each post was hateful
- Each post was seen by three annotators who deemed how hateful the post is, which type of group it targeted, and whether it was implicit or explicit

Definitions

- Hate Speech: Any public expression that promotes violence, humiliation, vilification, or incites hatred against specific groups based on religion, race, sexual orientation, or other characteristics.
- Hate-based rhetoric- whether the document is (1) Not-hateful (NH), (2) Incitement to hatred/Call to Violence (CV); and/or (3) Assault on Human Dignity(HD).

Data

Definitions

- **Vulgarity/Offensive Language- whether the document is offensive or abusive(VO)**
- **Targeted Group- Race and Ethnicity(RAE), Nationality/ regionalism(NAT), Gender(GEN), Religion(REL), Sexual Orientation(SXO), Ideology(IDL), Political Identification(POL), Mentally/Physically Disabled (MPH)**

Examples

Text	Hate-Based Rhetoric	Derogatory	Target	Framing
The Black Panther movie is a blatant anti-white ZioJesuit cultural marxist film aimed at further anti-white Christian theft, replacement, and genocide. Send them all back to their native Africa like good ole Abe had the foresight to do and let them find equality and social justice among their own kindred and creeds.	HD, CV	VO	SXO, RAE	EX

Data Cleaning

ID	Annotator	Text	Hate	HD	CV	VO	REL	RAE	SXO	GEN	IDL	NAT	POL	MPH	EX	IM
27232	17	They weren't innocents, they weren't protestors, they were Hamas scum who want to obliterate Jews and Israel. Leftwing MSM needs to realise Hamas would indoctrinate all of them if they lived under its governance. http://t	1	1	0	0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
27233	13	Disgusting to see these fake Palestinians protesting outside Israeli embassies, moaning about Zionism and the Jewish homeland. Newsflash, Palestinians are a fake race, they are fictitious. Arab countries don't want your ter	1	1	0	0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
27233	9	Disgusting to see these fake Palestinians protesting outside Israeli embassies, moaning about Zionism and the Jewish homeland. Newsflash, Palestinians are a fake race, they are fictitious. Arab countries don't want your ter	1	1	0	0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
27233	17	Disgusting to see these fake Palestinians protesting outside Israeli embassies, moaning about Zionism and the Jewish homeland. Newsflash, Palestinians are a fake race, they are fictitious. Arab countries don't want your ter	1	1	0	0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0

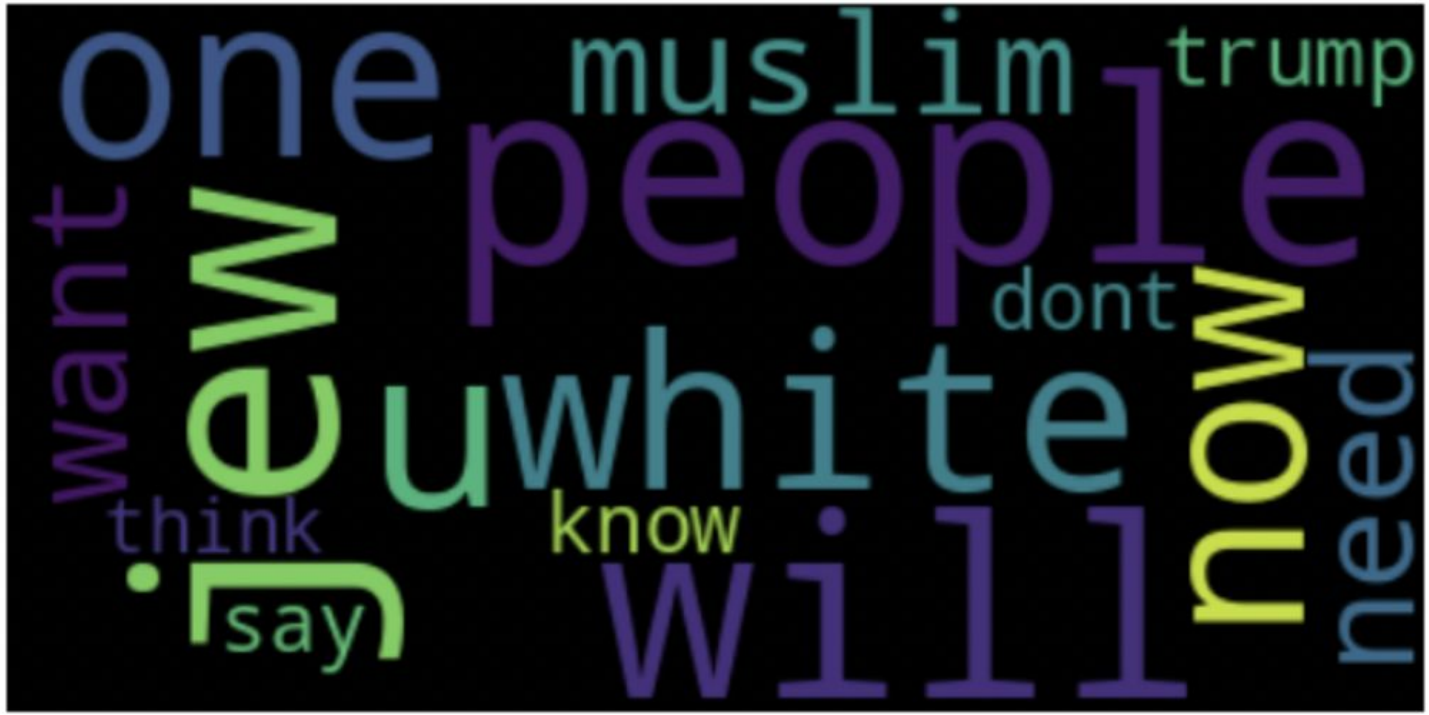
index	Text	Hate	HD	CV	VO	REL	RAE	SXO	GEN	IDL	NAT	POL	MPH	EX	IM
14	#Breaking: Claiming bias against conservatives, @FreedomWatchUSA has filed a class action lawsuit against Google, Facebook, Instagram, Twitter, and Apple.	1	1	0	0	0	0	0	0	0	0	1	0	0	1
16	"These [companies] are run by sociopaths," he said. "These people are complete narcissists. These people ought to be controlled, they ought to be regulated." At one point during the phone call, Bannon said, "These people are evil.' https://www.vanityfair.com/news/2018/08/steve-bann...	1	1	0	0	0	0	0	0	0	1	1	0	1	0
52	Report: Google A.I. Is 'Biased' Against Gay People, Jews http://www.breitbart.com/tech/2017/10/26/report-go...	1	1	0	0	0	1	1	0	0	0	0	0	0	0

- **Removed unnecessary columns**
- **Made each value an integer instead of a float**
- **Dropped all NA values**

03

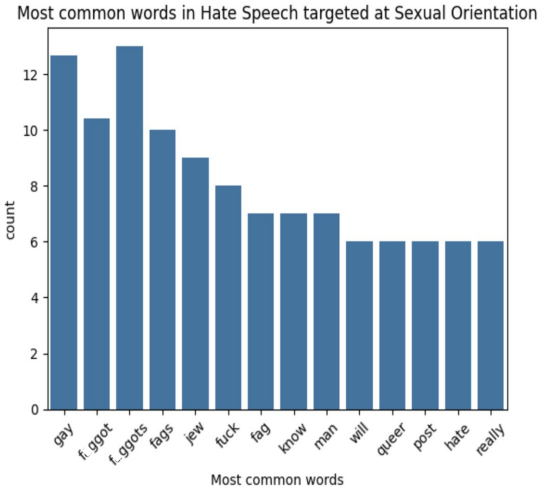
Word Counting

Word Cloud

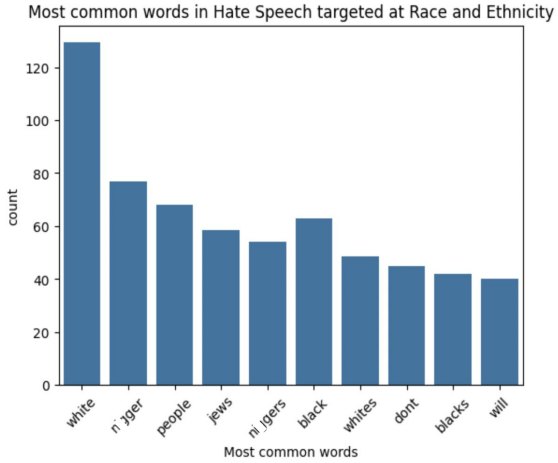

one
muslim
trump
people
want
jew
dont
uwhite
now
need
think
know
will
say

Word Count Visualizations

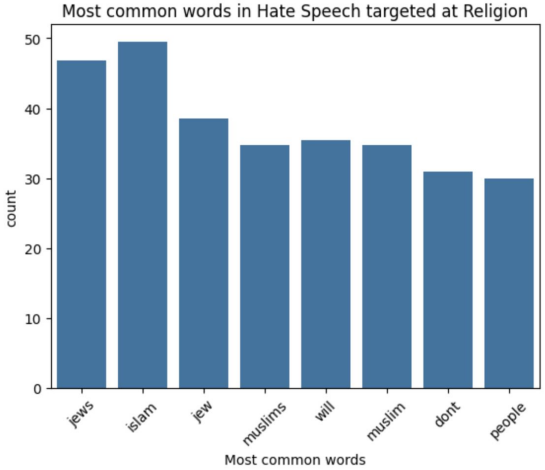
Sexual Orientation (SXO)



Racial/Ethnic Orientation (RAE)



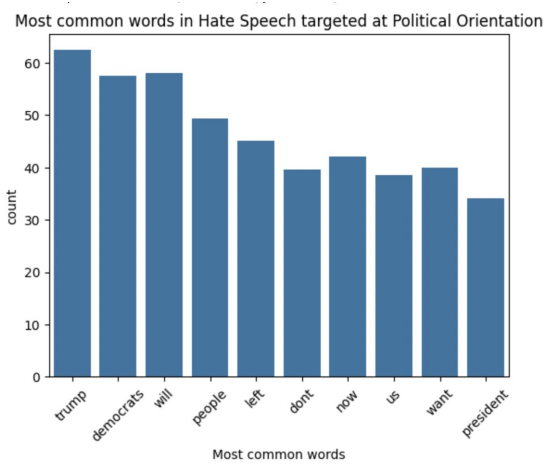
Religious Orientation (REL)



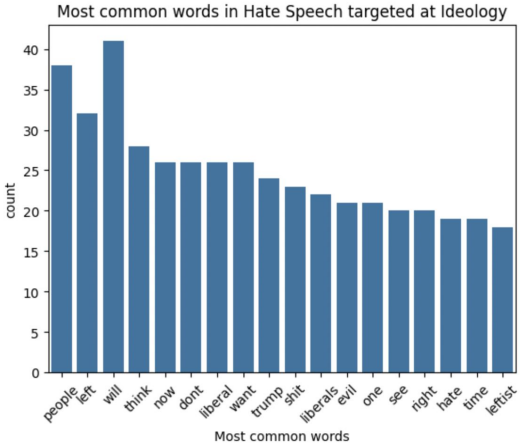
Takeaway: Each of these forms of hate speech include a “Jew” as a common word. There is clear targeted anti-semitic content on Gab.

Word Count Visualizations

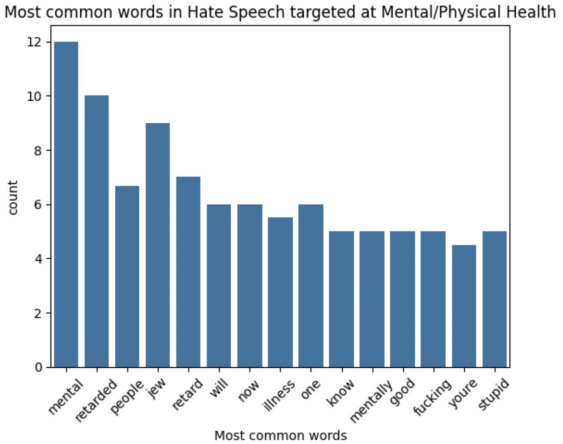
Political Orientation (POL)



Ideology (IDL)



Mental/Physical Health (MPH)



Takeaway: Once again we see Jew as the only ethnicity in the MPH most common words, which sheds light on the anti-semitic standing of the platform. In addition, the POL and IDL most common words are referencing American politics even though Gab is available in other countries, showing the political discourse is heavily influenced by US culture wars

04

Text Classification

Support Vector Machine Modeling

Methodology

- An SVM classifier was trained on text data to categorize hate speech types and targets, using classification metrics.

Recall: The percentage of cases of a Hate Speech category that was classified

Precision: Of the cases detected for a specific category how many of them were correctly classified

Proportion: Percentage of cases per category relative to the whole

HD,CV,VO: Percentage of specific Hate Speech labelled attack on Human Dignity, Call to Violence, and Vulgar/offensive, respectively

Results

	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

Text Classification Summary

Overall:

- Political, racial/ethnic, and religious identities were the most targeted groups, reflecting Gab's far-right, politically driven user base.
- Mental/physical health, sexual orientation, and gender identity were the least targeted, likely due to lower engagement with these topics on the platform.
- Human Degradation (HD) was the most common hate speech type (about 95%) and had the highest recall and precision.
- Calls to violence (CV) and Vulgar/Offensive Language (VO) had the lowest recall due to low data volume and ambiguous labeling.

Precision:

- Sexual orientation and racial/ethnic had the highest precision, based on the consistent keyword frequency in the text.
- Mental/physical health had the lowest precision, due to the limited sample size, leading to more false positives.

Recall:

- Racial/ethnic and religious identities had the highest recall, strictly from higher data proportions
- Ideology and mental health had the lowest recall.

Proportion:

- Human Degradation (HD) was the most common hate speech type (95%) with perfect recall and high precision, likely due to overfitting.
- Vulgar/Offensive Language (VO) appeared in 37% of posts, with moderate precision (0.64) but low recall (0.31) due to ambiguous language use.
- Calls to Violence (CV) were rare (8%) and had low recall (0.04) but decent precision (0.70), likely due to limited data and subjective labeling.

05

Conclusion

Conclusion

- Gab is a social media platform with a large amount of hate speech. Using word counts, topic modeling, and text classification, we identified strong trends of racism, homophobia, and Islamophobia
- The word “jew” appeared frequently across all analyses.
  - Our text classification also showed high accuracy for religion-related hate speech.
- The Gab Hate Corpus reveals a prominent influence from American political culture, with political identity emerging as the most frequent targeted category in hate speech.
  - Keywords like “liberal” and “democrats,” along with anti-semitic and islamophobic language further highlight how the United States political polarization, , especially the far-right, shape the hate speech on the platform.
  - Gab promotes far-right, racist, homophobic, and neo-Nazi content.
- Religious Identity, Political Identity, and Racial/Ethnic Identity emerged as the most prevalent categories of hate speech on the platform. This helps explain why “Jew” was among the most frequently occurring terms across the dataset.