Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

# Shouting into the Void: A Database of the Alternative Social Media Platform Gab

**Gabriel Fair, Ryan Wesslen**
University of North Carolina at Charlotte
gfair,rwesslen@uncc.edu

**Abstract**

As social media platforms have increased their role as content moderators, alternative social media platforms have emerged with fewer rules and policies on moderating content. One such platform is Gab, a social media platform similar to Twitter that champions free speech with minimal standards on content. Early research has linked this platform with alt-right, hate speech, conspiracy theories, and other alternative content that is sometimes marginalized in mainstream social media platforms like Twitter and Facebook. In an effort to provide a means for researchers to study this platform, we introduce a database of 37,012,061 posts (with additional edit histories), 24,551,804 comments, and 819,957 user profiles webscraped from Gab between August 2016 and December 2018. In this paper, we outline our data collection process, describe the data, consider the ethics of our data collection process, and provide suggested avenues of inquiry for researchers interested in analyzing our database.

## Introduction

Social media platforms provide a digital forum for the flow of user-generated content. As the number of users on social media platforms has dramatically grown beyond 3 billion individuals (Social 2018), so too has the rise of hate speech (Nobata et al. 2016), automated processes (e.g., bots) (Ferrara et al. 2016), and misinformation campaigns on such platforms (Allcott and Gentzkow 2017). In response to these issues, many major social media platforms like Twitter, Facebook, and YouTube have started to moderate content to ensure that content adheres to each platform's policies and guidelines. Moreover, alternative platforms have been created that provide looser standards and are subject to higher levels of malicious content (Bib 2019). In this paper, we consider one such platform, Gab, and introduce a database that provides a large-scale collection of its posts,comments, and user profiles for the research community to better understand this platform.

Recently, (Zannettou et al. 2018a) and pushshift.io (Baumgartner 2018) have open sourced similar data of Gab posts. However, a difference between our contribution is that these data sets are limited to compressed flat files (json) of Gab posts, rather than an indexed database. Our contribution includes a larger data set of posts, with comments, and user profile attributes. Our data (e.g., posts and users) incorporates longitudinal observations, in which we include the updates and time stamps when users edited their past content. Our primary contributions of this paper are:

- User account data providing friends and follower information,
- Multiple updated records (e.g., when a user later edits a post) on edited posts and comments,
- Provide code to assist with loading and using the data,
- Data conforms to the FAIR Data Principles[1],
- Data is provided using the dataset sharing service Zenodo.[2]

The structure of this paper is as follows. First, we review past research on Gab and explore unique features about this social media platform. Next, we outline our data collection process to acquire the dataset. Then, we provide a brief outline to consider the ethics and adherence to platform policies in our collection of the data. Last, we outline ideas of possible future research on our dataset.

## Gab and Related work

The first evaluation of Gab was done by (Zannettou et al. 2018a) with a data set of 22 million Gab posts.[3] In their work, they identify the top links and hashtags found in user posts, identify influential users, highlight top terms in profile descriptions, and summarize the creation dates of the user population. They also find that Gab tends to have more hate speech than Twitter, as measured by a Hate-based lexicon; although less when compared to 4chan's Politically Incorrect board (/pol/) (Hine et al. 2016). Subsequent studies have also considered more specific research questions using Gab. For example, (Zannettou et al. 2018c) consider the role of state-sponsored trolls within the Gab platform. Alternatively, (Zannettou et al. 2018b) explores the origin of memes on Gab and related platforms.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] More information about the FAIR data principles can be found here: https://www.force11.org/group/fairgroup/fairprinciples

[2] Code and data can be found at `https://doi.org/10.5281/zenodo.2541323`

[3] `https://zenodo.org/record/1418347`

### Platform affordances: posts, profiles, and users

Gab is a micro-blog platform most similar to Twitter. Users (or gabbers) generate posts that are limited to 300 characters (Gab 2019a). Such posts can include images, links, gifs, polls as well as can be tagged as NSFW, or "Not Safe For Work," to highlight potentially obscene content. A post can then be viewed by other users in their feeds if they are following the user that made the post. Users can respond to other posts with an up-vote that is similar to "likes" or "favorites" on other platforms. By aggregating such user's responses, Gab also provides a favoring system, like facebook, where each post is provided a score (number of points) providing the total number of up-votes and is prominently displayed for other users. Like Twitter, users can also comment directly to posts, repost (similar to retweet), or quote posts as well as include hashtags or mentions within the body of the post.

However, unlike Twitter, Gab offers two different affordances for users. First, users can edit past content (Gab 2019b). Second, users can tag posts to unique topics or categories (Zannettou et al. 2018a). Topics are generated by Gab users and normally reference a unique event or topic that is publicly available to all users. Alternatively, categories are pre-defined themes (e.g., News, Politics, Technology) that users can associate with their content.

Similar to Twitter and Facebook, Gab users also maintain a profile page which includes user-level information (e.g., profile summary, profile and background images) that can be updated at any time (Gab 2019g). Gab also includes a user-level point system that measures the net number of votes on that user user's content. For example, each user earns one point for one up-vote of his content. Alternatively, the user will lose one point for each down-vote of his content as well. The net score of each user is prominently displayed on their profile page next to other summary profile metrics (e.g., number of posts, followers, and followings).[4]

Unlike most mainstream social media platforms whose business model is based on advertising, Gab maintains a user-supported business model in which users can pay for different designations. The first level is Gab Pro, their premium subscription service for users interested in additional features like creating lists, groups, extended character counts, verification, and the ability to be a premium content creator (Gab 2019d). A premium content creator is an individual who can sign up to solicit money for their content (Gab 2019e).

## Data collection

Our data provides three types of site content: posts, comments, and user profiles. Data was collected without creating an account with the site. There is no need to login to view public profiles and posts. We collected our user data in a four step process. First, we scraped[5] the profile information of Andrew Torba's account.[6] From that we got the list of accounts he follows and his followers. Those lists were added to his profile information and saved to a json file. Lastly the next accounts to scrape were taken from this list of people he was following and his followers. The scrape was restarted with this new list. This represented a breath-first-search of the Gab social network. If the account was private then only publicly accessible information was scraped. This includes the month and year the account was created, the number of people the account is following, and their followers, if the account was a premium, investor, pro or donor, and the number of times they have posted to Gab.

The posts and comments were crawled in incremental batches partitioned by the unique post id number. Gab's posts are uniquely identified by an id number that started at 1 with the first post and incremented every time a post is made anywhere on the site. Our data also contains the time in which the post and comment was collected. This makes it easy to reconcile the data with previously published datasets.

## Robots exclusion protocol & platform policies

To ensure the ethical collection of our database, we adhered to Gab's Robots exclusion protocol (i.e., `robots.txt` file), privacy policy, and Terms-of-Service, all of which place no restrictions on web-scraping.

| Restricted | Unrestricted |
|---|---|
| User-agent: * | User-agent: * |
| Disallow: / | Disallow: |

Figure 1: Two examples of `robots.txt` files. The file on the left is a restricted file that prevents all users from crawling any folder. The file on the right is an unrestricted protocol that allows all users to crawl any folders on the server. Gab's `robots.txt` file has continuously maintained an unrestricted protocol.

The `robots.txt` file is a text file used by website owners to give instructions to automated bots (e.g., web scrapers) about using their site.[7] It is divided into two components: the User-agent directives and the Disallow folders. The User-agent directives act to identify to any bot which directive applies. For example, '`User-agent: *`' indicates that the directive applies to all bots. The Disallow information is connected to a User-agent directive and lists what sub-folders the bot cannot visit. The two most comment Disallow statements include either '`Disallow: /`' and '`Disallow:`'. '`Disallow: /`' indicates that the directive restricts bots from the entire server. On the other hand, '`Disallow:`', in which the Disallow is empty, indicates that the directive allows bots to have complete access to the server. Two examples of these files are shown in Figure 1. Using the Wayback machine archive, Gab platform

[4]Users need a score of at least 250 points and verified email address to use downvotes.

[5]`Loosely based on code found at: https://github.com/aaronrudkin/GabScraper`

[6]https://gab.ai/a

[7]`http://www.robotstxt.org/robotstxt.html`

has maintained a robot.txt file to allow all web crawlers access to all content.[8]

In addition, we considered Gab's privacy policy and Terms-of-Service to ensure that our data collection did not violate any conditions (Gab 2019f; 2019c). We found no mention of terms like 'scraping', 'bots', or 'crawling' within either policy and found no conflicts with the policy and our data collection procedure.

## Conclusion & Future Work

Typically, to study social media through large-scale data, most research considers only been provided a limited perspective through samples (e.g., Twitter's streaming API), keyword queries, public pages (e.g., Facebook), or user profiles (e.g., Twitter's REST API). In this paper, we provide a dataset of nearly a full, longitudinal population of a unique social media platform, Gab. Our dataset could have wide use for computational social scientists interested in a platform-level set of word embeddings, bot detection, echo chamber analysis, information framing studies, meme contagion, or cross-platform interactions.

## Acknowledgments

We would like to thank Arunkumar Bagavathi for his assistance in collecting a subset of our data.

## References

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–36.

Baumgartner, J. 2018. *Gab Directory Contents*.

2019. *Gab Is Back, And As Full Of Hate Speech As Ever*.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Gab. 2019a. *Creating, Viewing & Engaging With Content*.

Gab. 2019b. *Editing Content*.

Gab. 2019c. *Gab Privacy Policy*.

Gab. 2019d. *Gab Pro*.

Gab. 2019e. *Gab Pro*.

Gab. 2019f. *Gab Terms of Service*.

Gab. 2019g. *Profile*.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2016. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *arXiv preprint arXiv:1610.03452*.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, 145–153. International World Wide Web Conferences Steering Committee.

[8] `https://web.archive.org/web/*/gab.ai/robots.txt`

Social, W. A. 2018. *Digital in 2018: World's Internet Users Pass the 4 Billion Mark*.

Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; Kwak, H.; and Blackburn, J. 2018a. What is gab? a bastion of free speech or an alt-right echo chamber? *arXiv preprint arXiv:1802.05287*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018b. On the origins of memes by means of fringe web communities. *arXiv preprint arXiv:1805.12512*.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2018c. Who let the trolls out? towards understanding state-sponsored trolls. *arXiv preprint arXiv:1811.03130*.