"Strict Moderation?"

The Impact of Increased Moderation on Parler Content and User Behavior

by

Nihal Kumarswamy

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

University of Texas at Arlington

May 2022

Abstract

Strict Moderation?

The Impact of Increased Moderation on Parler Content and User Behavior

Supervisor: Dr. Shirin Nilizadeh

Committee Members: Dr. Chengkai Li and Dr. David Levine

Social media platforms have brought people from different backgrounds, ethnicity, race, gender, etc together to form a platform to share ideas and opinions and discuss news events among other social events. Unfortunately, these platforms have also been a safe haven for abusive users who harass, bully other users or spread misinformation and disinformation. Social media platforms have a huge incentive to police these abusive users and keep them in check to allow other genuine users to use their platform. Social media platforms employ several different content moderation techniques to perform this task. These techniques vary across platforms, for example, Parler believes in using the least restrictive moderation policies and having open discussion spaces for their users. These policies were used by several members responsible for the 2021 US Capitol Riots.

On January 12, 2021, Parler a social media platform popular among conservative users was removed from the Apple App store, the Google Play Store, and Amazon Web Services. This was blamed on Parler's refusal to remove posts inciting violence following the 2021 US Capitol Riots [12]. To return to the app stores, Parler would have to modify their moderation policies drastically [36]. Shortly before being banned from Amazon Web Services, a Twitter user, donk_enby [20], published frameworks and methodology for scraping Parler using their open API service. Studies like Aliapoulios et al. [6] used this opportunity to collect a dataset of posts from Parler and record user information [5]. After a month of downtime, with a new cloud service provider and a new set of user guidelines, Parler was back online [54]. Our study looks into the moderation changes performed by Parler and studies any noticeable differences in user behavior.

Using Google's Perspective API, we notice a decrease in the toxicity content shared in posts. We also notice similar trends in other labels such as identity attack, insult, severe toxicity, profanity, and threats. We study the most popular topics being talked about on Parler and compare other topics to uncover any changes in the topics of discussion. Finally, the Media Bias Fact Check service also checks the factuality of a sample of news websites being shared. We find an increase in the factuality in the news sites being shared. We also notice a decrease in the number of questionable sources and conspiracy or pseudoscience sources being shared.

Acknowledgements

I would like to thank Dr. Shirin Nilizadeh, my thesis supervisor for guiding me through my thesis. Dr. Nilizadeh has also helped me on other research projects and continues to motivate me to be a better student. I would also like to thank my committee members Dr. David Levine and Dr. Chengkai Li for their time and contribution to this work. Their suggestions have vastly impacted the next steps for this study and have helped me make better decisions for future works. I would also like to thank my fellow research lab members for helping me through the research process.

I would also like to thank my friends, parents, and sister who have given me a reason to work hard and strive to be better.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Abuse, harassment and misinformation are prevalent on social media [22, 31]. These abusers make it harder for genuine users to enjoy the platform by reducing the positive impact of connecting with others on social media platforms. Abusers can misinform users on current news events, and defraud users by stealing their money on the pretense of selling them a product [13]. They also discourage newer users from joining the platform and can contribute to a platform's downfall. Social media platforms, hence use content moderation policies to police these abusive users.

These platforms differ vastly in moderation policies. Some platforms, such as Twitter and YouTube, believe in fact-checking every post to curb false claims while some platforms, such as Facebook, only fact check political advertisements. Similarly, platforms like Reddit, Twitter, and TikTok also moderate content which can be toxic in nature. Other platforms believe in a laissez-faire approach to moderation, where only a specifically targeted attack or severely hateful content is taken action against. Parler was one of these platforms. Ever since its inception in 2018, they have followed a hands-off approach to moderation believing that it fosters better discussion and protects users's free speech [56]. This was until January 6th, 2021 when Parler gained a lot of notoriety for being home to several protesters who stormed Capitol Hill. Parler was then dropped by their cloud service provider Amazon Web Services and removed from the App Store and the Google Play store [23].

This forced Parler users to be disconnected for a considerable amount of time before coming back on February 9, 2021. After a brief hiatus of a month and finding an alternative cloud service provider, Parler also made changes to their user guidelines to be re-instated into the App Store and the Google Play store. After Parler was banned from the Apple App Store and the Google Play store, mobile users were not able to use Parler. Shortly after, their cloud service provider Amazon Web Services also banned Parler due to their inaction on posts inciting violence. Although Parler found an alternative cloud service provider, they had to abide by the App Store and the Play stores rules and guidelines. This meant a change in their moderation policies and new user guidelines.

These changes would introduce more stringent moderation policies which would curb hate speech [36]. The goal of this study is to examine how changes in moderation policies would impact how users would view and continue to use the Parler platform. Several other prior studies and current efforts have focused on topics, such as comparing moderation policies on different online platforms, identifying the impact of de-platforming, identifying small groups of unmoderated communities, or similar topics [49, 27, 9, 62, 66]. Due to the unique nature of Parler's de-platforming and the timely nature, we believe that this area of research can teach us a lot about user behaviors as well as online social media platforms in general. There also exists a lack of work in content moderation post moderation policy changes on Parler. The current study fixes this void by studying the changes noticed on Parler post moderation policy changes. We hope that our research can also shed a light on how policy changes affect the whole platform and its ecosystem and whether these policy changes are effective.

1.1 Research Questions

- Has the prevalence of hateful and toxic content decreased? Our study explores this question to study the effectiveness of the content moderation techniques and compares them from the pre-and post- moderation policy posts.
- Have the topics of discussion changed? If so what are the changes? This question would also help understand the impact of moderation change on user behavior.
- Is there any difference in the credibility of news sources being shared between pre and post moderation policy changes?

To answer the above research questions, we will use data collected about users and posts before as well as after moderation policy changes. Before being taken down, several internet sleuths deciphered Parler's open API system and other studies also collected data including posts, comments, and user information. For example, donk_enby collected about 70 TB of public Parler data from January 6 2021 [20]. This was the first open-sourced Parler public data that was released. Works such as [6, 49, 8, 43] have used either the whole dataset, or used the technique to collect new data to study various themes, such as *vaccines* and *politics*. Our study uses the data from Aliapoulios et al. [6] and also we collect new Parler data using a custom build crawler, to understand how the change in moderation policies affected the hateful rhetoric, the spread of conspiracies, etc. We labeled our dataset as post moderation policy change and the data collected by Aliapoulios et al. [6] as pre-moderation policy change data.

Using the collected data, we conduct several qualitative and quantitative analyses to capture any trends while comparing data from post moderation policy changes and pre-moderation policy changes. These analyses also help us answer our research questions. First, we extracted posts from both of the datasets to analyze the presence of toxic content. We used the Perspective service to score the likelihood of toxic content due to their open source methodology as well as their routinely updated classifier models to match current social media lingo [7]. Using the results obtained from this step, we compared both the pre and post moderation policy change data to answer our first research question. The Perspective service powered by Google uses machine learning to identify hateful or toxic texts by providing a score from 0.0-1.0 based on the attribute. The service offers labels or scores for 6 attributes: i) Toxicity, ii) Severe Toxicity, iii) Identity Attack, iv) Insult, v) Profanity and Threat. Using the scores for each post in each dataset, we were able to compare whether there was any change in the amount of hateful or toxic rhetoric present on the Parler platform.

In the next stage, we explored whether the supposed moderation changes had any impact on the topics of discussion. To investigate this, we used the Latent Dirichlet Allocation (LDA) topic modeling technique to model and extract popular topics of conversations. We used these topics to compare the different datasets and visualize the results using a word cloud. This also helped us answer our second research question about whether the topics of discussion have changed.

To answer our final research question on the credibility of news sources being shared, we extracted links being shared via posts and investigate the websites that users were being redirected to. We used the Media Bias Fact Check service (MBFC) to rate the credibility of the news sources being shared. MBFC is an open source service that uses manual labeling to certify a website's bias, country of origin, and factuality as well as the use of pro-science, conspiracy or questionable sources.

Findings: Using the methods described above, we answered our research questions and uncovered some notable findings shared below:

- Using the Perspective API, we collected and compared scores for the 6 available attributes. We noticed that several posts in the post policy change dataset had a score for toxicity closer to 0. Meanwhile, most of the posts having a higher probability score for toxicity were present in the pre policy change dataset. These scores indicate that toxic posts were more probable to be found in posts before the moderation changes implemented by Parler.
- We also noticed a similar decreasing trend in severe toxicity, profanity, identity attacks, insults, and threats.

- After using LDA to find the most popular topics talked about, we found several topics related to right-wing political groups in both datasets. We also found several Parler-specific words such as Parleys present in the earlier dataset. We theorize that the cause for this might be due to users migrating from other social media platforms like Twitter and Facebook which do not use these terms. These new users could be talking about the Parler-specific terms, introducing them to other users, and learning about them themselves.

- We also found a high usage of terms like *WWG1WGA*, which is a popular term used amongst conservative political groups. It is associated with the political conspiracy theory and personality of QAnon, widely popular on the internet before and after the capital riots. It stands for Where We Go as 1, We Go All which defines the togetherness of the QAnon community [33].

- While studying and comparing the links shared by Parler users, we noticed a steep increase in links directing users to Rumble, which is an online social media platform to share videos. While categorizing the most popular websites, we also noticed a high number of links leading to personal blogs, which were mostly political in nature. Oftentimes, these blogs or websites were making claims without providing sufficient credible evidence.

- Using a service called Media Bias Fact Check (MBFC), we also used the extracted links to add attributes based on the credibility, factuality, country of origin, and general authenticity of the news websites. We notice an increase in the factuality and credibility scores from the pre moderation policy dataset to the post moderation policy. We also notice a decrease in the number of conspiracy-pseudoscience and questionable source links being shared.

- We also compared the number of followers and followings between the datasets as well as the number of badges awarded to users. We noticed an increase in both the number of followers and the followings. This indicates that Parler is still being used by older users who were present before the moderation policies were changed. Similarly, we noticed that the number of users with the verified and gold badge had increased leading us to believe that a large chunk of users has not abandoned or migrated out of Parler. This adds credibility to our study and our findings.

Chapter 2

Related Works

2.1 Parler Dataset

When compared to other popular social media platforms like Twitter, Facebook, and Reddit, Parler is younger. Due to this, we notice that not a lot of studies have focused on collecting or establishing a framework to collect data from Parler. An exception to this is the pre policy change dataset we use extensively in our study to perform comparisons [6]. Other related work on establishing a framework for data collection has focused on older versions of the Parler API, which are now outdated [20, 49].

2.2 Studies About Parler

During our work, we noticed that Parler's users were predominantly conservative. We see this in the topics of discussion as well as their bio's. Works have looked into similar fringe platforms like Voat, 4chan, Gab [46, 27, 9, 62, 66]. There have been studies comparing topics of discussion on Parler and Twitter [49]. Although there exists work in this domain, most of the work is focused on exploring the existence or prevalence of a single topic. Papasavva et al. [46] uses Voat to study the spread of the QAnon movement while Hitkul et al. [49] uses the capitol riots, a pivotal movement in Parler's history, to compare topics of discussion between Parler and Twitter. We believe that our work differs in this aspect as we are studying the changes localized to Parler and how users reacted to a brief hiatus of Parler. Jakubik et al. [29] have used posts on Parler around the time of the US Capitol riots to compare with similar posts from Twitter. The study analyzed emotions surrounding posts on Parler and Twitter and concluded that users on Parler had a comparatively less negative response when compared to users from Twitter.

2.3 Moderation Policy Updates

There have been studies that investigate the moderation policies on other social media platforms like Twitter. Jhaver et al. [30] is one such study that examined how deplatforming users on Twitter could impact their userbase. They studied the effectiveness of this moderation policy by

examining their followers and the shift in talking points. They found that banning significantly reduced the number of conversations about all three individuals on Twitter and the toxicity levels of supporters declined. Ribeiro et al. [53] found that migration of the two subreddits to standalone websites reduced the audience of posts, but also found that users became more active and more toxic. Trujillo & Cresci [61] found that interventions had strong positive effects on reducing the activity of problematic users both inside and outside of r/The_Donald. Some scholarships have examined the effects of deplatforming individuals on the sites that sanctioned influencers move to post-deplatforming [4, 51, 55, 42]. These researchers found a common result, that deplatforming significantly decreased the reach of the deplatformed users, however, the hateful and toxic rhetoric increased.

2.4 Hate Speech Detection and Classification

Research on toxicity has employed machine learning-based detection algorithms to identify and classify offensive language, hate speech, and cyberbully [67, 15]. For example, Koratana and Hu [32] classified comments into seven groups clear, toxic, obscene, insult, identity hate, severe toxic, and threat. ElSherief et al. [21] categorized online hate speech into directed and generalized. Studies like Miok et al. [40] use bayesian attention networks to detect hate speech. Rana et al [50] uses emotion based hate speech detection algorithms. Other studies like Raut et al. [52] and Ahmed et al. [3] use LSTM neural network models for hate speech detection. The machine learning methods use a variety of features, including lexical properties, such as n-gram features [45], character n-gram features [38], character n-gram, demographic and geographic features [64], sentiment scores [17, 57, 24], average word and paragraph embeddings [45, 19], and linguistic, psychological, and effective features inferred using an open vocabulary approach [21]. The state-of-the-art toxicity detection tool is available through Google's Perspective API [7]. Perspective uses a single multilingual token-free Charformer model that is applicable across a range of languages, domains, and tasks. Perspective also claims that through extensive experiments on multilingual toxic comment classification benchmarks derived from real API traffic and evaluation of an array of code-switching, covert toxicity, emoji-based hate, human-readable obfuscation, distribution shift, and bias evaluation settings, they show that their proposed approach outperforms strong baselines [34].

In our study, we have used the Perspective API to label posts with additional attributes like toxicity, insult, profanity, severe toxicity, identity attack, and threat. These attributes contain a probability score ranking how prevalent an attribute can be seen in each post. Studies like Papasavva et al. [47] also use the Perspective API to add more information to their dataset of 4chan posts.

2.5 Media Bias Fact Check

Our current study uses the Media Bias Fact Check (MBFC) service to add additional attributes like credibility, factuality, and the general authenticity of the posts. Gruppi et al. [25] used MBFC service to label websites and the tweets pertaining to COVID-19 and 2020 Presidential elections embedded inside these articles. These websites are part of a seed set, which are used along with their respective RSS feed to collect every article shared by the website. Weld et al. [65] analyzed more than 550 million links spanning 4 years on Reddit. The authors used MBFC to annotate links to news sources with their political bias and factualness. MBFC is widely used for labelling credibility and factuality of news sources for downstream analysis [37, 26, 58, 14, 41] and as ground truth for prediction tasks [18, 60, 48]. Our study examines the websites shared by users, which are in circulation on the Parler social media platform. We believe that this allows us to make conclusions about the user base of Parler as well as evaluate their moderation changes.

.

Chapter 3

Parler

Parler is a social networking and microblogging service initially launched in August 2018 as an online website and later released as a mobile application. Parler markets itself as a free-speech alternative to other mainstream social media platforms such as Twitter and Facebook. In the early years of its existence, several prominent conservative figures joined the platform and brought along several customers with them. A key reason for some of these figures migrating to Parler was banned on other social media sites such as Twitter and YouTube. For example, in June 2019 there were several service disruptions to Parler's services due to the sudden influx of accounts. The new accounts had originated from Saudi Arabia and promoted the use of Parler after accusing Twitter of arbitrarily banning users.

3.1 Parler Terms and Workings

Signing up on Parler: At the time of the study, Parler required new users to sign up by inputting their email, phone number, and name. The provided phone number would be verified by sending a one-time valid password. Currently, Parler has changed their sign-up sheet to only require either an email or a phone number and a secure password. Parler auto-generates and populates the password field with a strong 14-character password consisting of lowercase and uppercase alphabets along with numbers from 0-9 and special characters such as !, -, _, $, % and &. This password can be overwritten with a user-chosen password if desired. After signing up, a user can start view other posts called Parleys on the service. A user can also start posting their own Parleys, comments, echo or upvote on other accessible Parleys. Accessible Parleys constitute Parleys from public users or private users who have accepted a request to be followed. Comments on a Parley are not multi-layered like others, we see on other popular social media platforms like Facebook or Twitter. Users can instead mention another user by using the @ character followed by their username to indirectly reply to their comment. Similarly, upvotes on Parler are comparable to likes on Twitter and Facebook. Echoing a post on Parler can be done in two ways: i) A user can echo a Parley with no additional added text or ii) A user can instead elect to echo a Parley with additional text either adding more information about the original Parley or conveying their opinion on the

original Parley. Upvotes on posts are similar to the like function on platforms like Facebook and Twitter.

User Filters: Parler allows real-time comment moderation for the original poster. The poster can block or hide comments similar to Twitter. The poster can also add a list of words to filter the comments similar to Instagram. Parler users can also block other users to hide their Parleys. Parler also offers a mute option which ensures that the user no longer sees the person's Parleys and comments. There is also a subscribe option that allows users to be notified of new Parleys from the subscribed user. This option is similar to the subscribe and notify option from YouTube. Users can also report inappropriate Parleys and comments which break Parler's rules.

Parler Search Feature: Parler also displays a search box after a user has logged in which can be used to search for Parleys using hashtags and users using their name or username. The search box however cannot be used to search for Parleys using text. There are two kinds of users on Parler: *Private* and *Public* users. Private users can hide their parleys from users who are not approved, followers. Parleys from public users on the other hand can be seen by any Parler user. Parler also has a setting under *Account Privacy Options* to set a public account as an *open* account. Parleys from open accounts can be seen by anyone on the internet and do not require a valid Parler account to be viewed. By default, every new account is an open account.

Notifications & Badges: After posting a Parley, the user will receive notifications for each comment posted. Users can turn off this feature which is turned off by default. Parler also notifies users if their comments were upvoted or if a comment mentions their username using the @ operator. Parler also employs a system of rewarding and identifying other users by presenting them with *Badges*. At the time of this study, Parler awarded 7 badges: i) Verified: Users who have proved to Parler as being real humans and not bots. Users can request this badge by sending Parler an official ID along with a picture to verify their identity. ii) Gold: Parler users cannot request a gold badge, instead, Parler offers this badge to users with a large following. With this badge, Parler intends to help users identify influencers and distinguish them from other accounts impersonating them. iii) Parler Early Adopter: Parler rewards users who were active before December 30th, 2018 with the Parler Early Adopter badge. iv) Parler Employee: This badge can be seen on Parler employees as denoted by the name. v) Parler official: This badge is present on official Parler accounts used to welcome users to the platform, check the status, and make announcements. Like the gold badge, this is used to help users distinguish official Parler channels from other impersonators. vi) RSS: This badge is awarded to accounts that use an RSS feed to post Parleys. vii) Private: This badge is placed on private Parler accounts. A figure of the badges can be seen in fig 3.1

Parler Home & Discover Pages: After logging in, Parler users are welcomed to their homepage

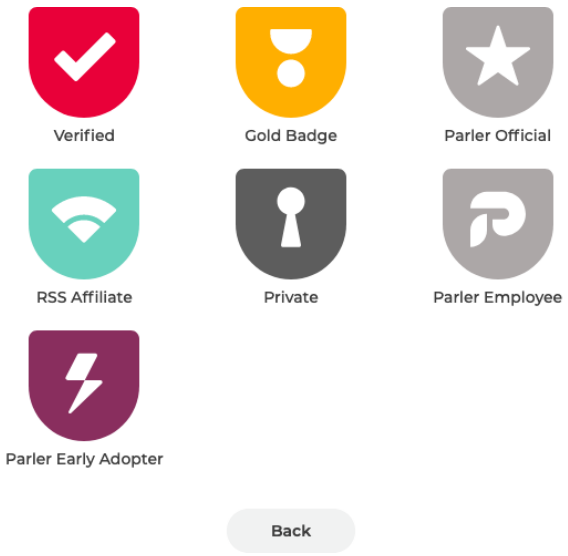


FIGURE 3.1: Figure of available badges

consisting of recent Parleys posted by following users, a menu bar on the left containing quick links to their profile, notifications, a Parler store page, and Parler's Discover page. The homepage also includes two display tabs on the right showing trending users and hashtags. Unlike other major social media platforms, Parler markets and sells merchandise like hoodies, hats, beanies, mugs, shirts, and gift cards. After manually searching for hashtags similar to Parler Store, we could not find any Parleys promoting their merchandise from users. The discover page contains Parleys from other users on the platform. Refreshing the discover page does not populate the page with newer posts which leads us to believe that adding parleys to the discover page is not automated. Scrolling down on the discover page does call Parler's backend services to download more parleys. Parler also doesn't limit the parleys displayed to just verified users. We have noticed a handful of Parleys from users with no badges. Initially, we assumed that the Discover page was user-specific and that the Parleys seen were due to the user's activities on the platform. During our data collection phase, we disprove this theory by noticing the same parleys in the same order from multiple different Parler accounts with different browsing histories.

User Profile: The left pane also includes a link to the user's profile. Users can use this link to view their parleys, look at the number of followers or followings and also edit their profile information. While editing their profile, users can add both: a profile and a cover image. The profile image is displayed on the profile along with their parleys. The cover image is only visible from a user's profile. Users can also add their name, location, their website, and a description consisting of 200 or lesser characters.

Trending: The home page also contains two tabs on the right pane of the screen displaying 5 trending users and 5 trending hashtags. These users are retrieved by using a backend API call to Parler's server. The API call returns 10 users and 10 hashtags unlike only the 5 that get displayed. We did notice a minimum requirement for a user to appear under trending users was to have the verified badge. Initially, we assumed that users appearing under the trending users was based on some automatic mechanism based on the views their profile or their parleys were generating. Similarly, we assumed that trending hashtags were automatically refreshed based on the number of parleys using the hashtags. During our data collection we noticed that this was not the case, however, and we can reasonably establish that the users and hashtags were either manually handpicked or manually updated after a set interval of time.

Chapter 4

Data Collection

4.1 Pre Policy Change Parler API

Initially, prior to being removed from app stores, Parler's websites used an open API format to communicate with the backend Parler servers. These APIs were used to fetch and send data related to user profiles, parleys, comments, etc. These API standards were changed before Parler was reinstated to the app stores and have changed drastically as detailed in Section 4.3. Using these API discoveries from [20], Aliapoulios et al. [6] built a crawler to collect Parler user information, parleys and comments. In the first step, a collection of user unique identifiers (UUID) was established. These identifiers are unique across parler and do not change, unlike the user's username or name. The UUIDs were retrieved using an API request which maps a single monotonically increasing integer to a UUID. These UUIDs were later used to gather posts and user information.

The endpoints used to collect the data were:

- /v1/post: Used with the UUIDs collected from the previous step to collect post information. A similar endpoint was also used to collect comments. Data such as the post text, upvotes, created at, hashtags and URLs if any were present would also be present from the API's response.
- /v1/user: Used with the UUIDs collected from the previous step to collect user metadata. Among the collected information were: user badges, bio, followers, followings, posts, comments, and date joined.

Some fields were collected like the upvotes on a parley, followers or followings of a user, and the number of comments and posts were reformatted from strings to integers to allow for easier numerical analysis. Each response from the API was stored as a JSON object and shared publicly on Zenodo [5].

4.2 Pre Policy Change Data Collected

The data collection tool used in Aliapoulios et al. [6] managed to collect user information for almost all of the users present at the time based on estimates published by Parler. They also managed to collect posts and comments from these users dating back all the way back to 2018 when Parler was created. At the end of the data collection phase, the study collected user information from 4,079,765 users and 98,509,761 posts, and 84,546,856 posts. These 4M users were used as a seed dataset in Section 4.4 stage to collect data post policy changes instituted by Parler.

4.3 Post Policy Change Parler API

Using a custom-built web scraper written in Python, we collected data using the back-end API servicing Parler's website. The API endpoints were discovered while using Parler from a desktop browser. We discovered 10 main endpoints used to share information from Parler's servers.

- /functions/post/fetch/FetchParleyController: Used to display a single post using a unique post_id
- /functions/post/create/PostParleyController: Used to post Parleys or comment on Parleys
- /functions/upload_video and /functions/uploadimage: Used to upload a media file as part of a parley or comment.
- /functions/trending_hashtags: Returns a list of ten trending hashtags. Used to show users a list of trending hashtags on their front page.
- /functions/trending_users: Similar to the trending_hashtags, trending_users return a list of 10 users with the verified badge. These users are displayed to every user on their front page.
- /pages/feed: Used to request parleys posted by a specific user using their username. The endpoint only returns 20 parleys per request. Older parleys need to be requested incrementing the page number counter. This endpoint is also used to retrieve the discover page for the logged-in user.
- /functions/heartbeat: Used to increment the number of views on a Parley. Requires post_id to increase views. Currently, the website version of Parler does not share the number of views on a post although the previous version allowed users to view this.
- /pages/hashtags: Requires a logged-in user to type in a hashtag to search for parleys.

- /pages/profile/view: Requires a valid username of a Parler user. Returns metadata information about users: number of 1) followers, 2) following, 3) parleys, 4) likes, 5) comments, 6) badges, textual information: 1) display name and 2) text in the user bio. The endpoint also returns a link to access the user profile and cover images.
- /pages/search-results: Similar to /pages/hashtags required a text but unlike the former, it can search for users.

We used the Postman API tool extensively to set up request headers, cookies, and request form data and also to test outputs from our scraper for all the endpoints. We used the tool to test several types of request data types and noticed that Parler only supported receiving data from the form data fields.

4.4 Web Scraper

After establishing these endpoints, we built a web scraper using the requests module [28]. Since Parler requires a valid user log in to use these APIs, we started checking the cookies used by Parler's website. We noticed a session token called *parler_auth_token* being set to a 64-bit key. A new session key was initialized and distributed to the user on every log-in but each session key was valid for 60 days from the time of log-in. We used this session key to initialize our cookies along with other required fields auto-generated by the requests module. To collect parleys posted by a user, we used the */pages/feed* endpoint along with the username of a valid user. The username was used as part of the form data encoded in the body of the request. The two required fields in the body: *page* and *user* would be set to the page number and the username, respectively.

Due to changes in the API, we were no longer able to collect data from all users hence we used the dataset of users from Aliapoulios et al. [5] to collect posts from Parler. The study uses a list of 4,079, 765 users to collect posts and we perform a similar step. We divided the 4,079,765 users into batches of 500,000 users each. The scraper used multiple threads to collect multiple user parleys simultaneously but the cumulative outgoing requests were rate limited to avoid overwhelming Parler's servers or causing any disruption in their normal services. The endpoint returns the 20 most recent parleys from the user and a page number counter can be incremented to retrieve other older parleys. The web scraper increments the page number count until there are no parleys left to retrieve. We collected information about the post body, any URLs posted, a URL to the location of any media posted, the date posted, the number of echoes, and other metadata such as the badges of the poster.

The endpoint returned data in the form of a list of JSON objects. After recursively traversing all the pages of a user to collect all the posted parleys, each JSON object was appended to a list and stored as a list of JSON.

For our next phase of data collection, we used the */pages/profile/view* endpoint. This endpoint requires a valid username along with a valid session token and returns the number of followers and following, badges, account creation date, text in bio, whether the account is public or private, account name, and a link redirecting to the profile and cover images stored on Parler's servers. The profile API uses form data stored inside the delivered request to serve user information. The key *user* would be set to a value of a valid username. To collect profile data after the moderation policy change, we use a seed dataset of 13.25M users from Aliapoulias et al. [6] to collect user profile data.

Unlike the previously used post API to collect posts, the profile API returns all the information in a single request and hence does not have to be run recursively. The returned message is encoded as a JSON object with the keys being data headers and the values being corresponding data values.

While testing a small dataset of users to scrape posts and profiles, we noticed that Parler often does not return the values for all headers. We noticed that a lot of posts were missing values for the date of creation. We also noticed some users were missing values for profile and cover images. The scrapper tool built for data collection was modified after noticing this to log and store information about missing fields. After a specific batch of users's data was collected, we used these log messages to re-collect data and fill in the missing fields.

After using the scraper to collect posts and user profile data, we decided to use the *trending_hashtags* and the *trending_users* endpoints to try to collect a sample of users not included in our seed dataset. We verified the trending users and hashtags were not user specific by comparing the results using two different Parler accounts. We noticed the same users and hashtags under trending and therefore concluded that they were not user specific. While looking into adding more users to our dataset, we also explored collecting data from the discover page. Similar to the trending users and hashtags, the discover page was not user specific but only included Parleys by users with a verified badge. To collect data from the discover page, we used the same feed endpoint but instead of sending a username we instead send the text *discover*. The resulting request is similar to collecting Parleys and can be saved as a JSON object as well.

We set up the scraper to collect trending users and hashtags every hour for a 72-hour time period to test the results. After this, we extracted unique users and hashtags and noticed only 100 unique users and hashtags being collected. We determined that the trending users and hashtags were

not being refreshed in a timely manner. The users returned from trending users were also only limited to users with the verified badge. After conducting a similar collection phase for the discover page, similar to the trending users and hashtags, the posts were not updated in a timely manner suggesting some form of manual intervention. Due to these reasons, we decided not to use the trending hashtag and users or the discover page to add users to our dataset.

4.5 Post Policy Change Data Collected

As we can see in table 4.1, after scraping posts from all the users in our dataset, we collected 17,389,610 parleys from 432,654 active users. Several other users part of the initial seed dataset of 4,079,765 were no longer active or had deleted their accounts. We label them as *missing* users, as either, they would have changed their usernames, after Parler was reinstated back, or they were banned by Parler or the users themselves deleted their account. These parleys consisted of users posting around 9M links as well as plain text in the body. A majority of these posts were primary posts that have no parent. If a parley is an original parley and is not an echo of another parley, it is known as a primary post with no parent. For a vast majority of the primary parleys we collected the full text body, a URL if a link was shared, the title of parley, date of creation, flags for trolling, sensitive and self-reported, an upvoted flag, a counter of echoes and likes. We noticed that Parler has a trolling flag which might be set manually by moderators or automatically by the platform.

Table 4.1: Post Policy Change Data Collected

	Count	Users
Posts	17,389,610	4,079,765
User Data	12,497,131	13,248. 086

We also used the profile_view endpoint to collect profile information on 13.25 million users. We found that 12,497,131 of these users still had a valid Parler account and therefore we could collect metadata for these users. We collected and saved all the key-value pairs returned in JSON objects from the API in separate JSON files to allow for easier use. For a vast majority of the accounts, Parler returned the number of followers, the number of following, status(account available or deleted), the number of and types of badges given to the user, and a description of all badges available on Parler at the time of collection, date of parley creation, whether the account is private or public and also whether the account is being followed by or a follower of the user logged in. A minority of profiles have one or more of these fields missing due to changes on the Parler platform from when the user created the account and the time of data collection.

Ethical consideration. Due to the large size and the potentially user-sensitive data collected, we also took several ethical considerations. We only gathered posts from profiles set to public and made no attempt to access private accounts. We also made no attempts to maliciously gather this data, instead of using the same backend APIs that a user browser would request data from. Furthermore, we make no attempts to trace users across other social media platforms or link accounts to real-world persons.

The most recent announcements by Parler have put the number of customers around 20M. We collect user information for 11.7M of these users which ensures that the trends observed in our data are seen on the platform as well. Meanwhile, since we only collect posts from 432,654 users we acknowledge that certain trends and analyses conducted might not be accurately reflected on the platform. However, as of January 2022, months after returning to app stores on both the Android and Apple platforms Parler disclosed that they estimate to have around 700,000 to 1M active users. This ensures that we have collected a significant part of data to study and base our findings on. We also believe that since we are comparing the same users from a previous seed dataset, the trends noticed in our dataset should generally also be present on the platform.

Chapter 5

User Information Metrics

After analyzing data collected about Parleys from both the datasets, we decided to analyze user information collected. Some of the information collected was: badges, names, followers, and followings. Here, we also sort the Parleys based on username to conduct research on topics of discussion. We start by comparing the number of badges present in each dataset as well as comparing the number of common badges between both the datasets that the users have attained.

5.1 Badges

We extracted badges for every user in the dataset of pre moderation policy change as seen in Table 5.1. We notice that badges like Parler employee could help us come to the conclusion that Parler probably employed around 25 employees around the time of data collection. These numbers also suggest a large number of users are private users. We also see quite a large number of users going through Parler's verification process to prove that their account is not a bot. Users submit a picture of a valid ID which would verify them and award them with a *Verified* badge. We also notice a large number of popular influencers who have been awarded a *Gold* badge to show that popular users are using Parler.

In Table 5.2 we can see the badges collected from the post moderation policy change dataset. When comparing the number of users with the verified badge we notice an increase which could prove that users are still active on Parler since receiving the verified badge requires user action and is not automatic. We also notice an increase in the number of *Gold* badges which could mean that existing Parler users have gained popularity to require a *Gold badge*. A sharp increase in the number of verified users can also be seen which might be due to users being fearful of an influx of bots as Parler was growing as a platform and attracting attention from other social media users.

5.2 User Metrics

After looking at the badges, we extracted the metrics *following* and *followers* from both datasets. The resulting set of values did not form a standard distribution and hence we choose to use the

TABLE 5.1: Badges Pre Policy Change

Badge	Description	Pre Policy Change	Disappeared
Verified	This badge means Parler has verified the account belongs to a real person and not a bot. Since verified users can change their screen name, the badge does not guarantee one's identity.	25,734	2,326
Gold	A Gold Badge means Parler has verified the identity of the person or organization. Gold Badges can be influencers, public figures, journalists, media outlets, public officials, government entities, businesses, or organizations (including nonprofits). If the account has a Gold Badge, its parleys and comments come from real people.	589	12
Integration Partner	Used by publishers to import articles and other content from their websites	64	17
RSS feed	These accounts automatically post articles directly from an outlet's website	99	13
Private	If you see this badge, the account owner has chosen to make the account private. This badge may also be applied to accounts that are locked due to community guideline violations	596,824	319,469
Verified Comments	Users with a verified badge who are restricting comments to only other verified users.	4,147	716
Parody	Parler approved parody accounts.	37	9
Parler Employee	This badge is applied to Parler employees' personal accounts, should they wish. Their parleys are their own views and not Parler's.	25	6
Real Name	Users using their real name	2	0
Parler Early	Signifying Parler's earliest members, this badge appears on accounts opened in 2018.	81	177

Mann-Whitney test to reject or accept the null hypothesis that the distribution of followers and followings are different in pre and post moderation policy change datasets. The test resulted in the rejection of the null hypothesis and proved that the distributions were different when compared between by the datasets($p < 0.001$). This along with an increase in the median followers and followings noticed in the post policy change metrics indicate that users are still active.

TABLE 5.3: User Metrics Comparison

	Pre Policy Change				Disappeared				Post Policy Change			
Metric	Min	Max	Mean	Median	Min	Max	Mean	Median	Min	Max	Mean	Median
Followers	0	2,300,000	20.65	1	0	189,000	73.20	1	0	6,048,750	34.8	1
Following	0	126,000	28.28	6	0	112,000	79.79	7	0	479,412	33.4	8

TABLE 5.2: Badges Post Policy Change

Badge	Description	Post Policy Change
Private	If you see this badge, the account owner has chosen to make the account private. This badge may also be applied to accounts that are locked due to community guideline violations.	337,717
Verified	This badge means Parler has verified the account belongs to a real person and not a bot. Since verified users can change their screen name, the badge does not guarantee one's identity.	236,431
Gold	A Gold Badge means Parler has verified the identity of the person or organization. Gold Badges can be influencers, public figures, journalists, media outlets, public officials, government entities, businesses, or organizations (including nonprofits). If the account has a Gold Badge, its parleys and comments come from real people.	668
Parler Early	Signifying Parler's earliest members, this badge appears on accounts opened in 2018.	822
Parler Employee	This badge is applied to Parler employees' personal accounts, should they wish. Their parleys are their own views and not Parler's.	28
RSS Feed	These accounts automatically post articles directly from an outlet's website.	149
Parler Official	These accounts - @Parler, @ParlerDev, and others - issue official statements from the Parler team.	5

5.3 Gender Analysis

From the user data collected in Section 4, we also had a list of names that users had registered with. These names were different from usernames or emails. They would be displayed on user profiles as their real name. To match names derived from users to a gender we follow the U.S. Census Based method used in Nilizadeh et al. [44]. The dataset of names includes a score ranging from 0 to 1 which denotes the likelihood of the name belonging to a person of either male or the female gender. A higher score denotes that the person is likelier to be of the male gender and similarly a lower score denotes that the person is likelier to be of the female gender. We use a threshold of 0.95 or greater to be classified as a male and a threshold of 0.05 or lower to be classified as female [44]. Using this data, we record names that are usually used to address males and females. Historically, most social media services like Facebook and Twitter have reported more female users than male [59]. As we can see in Figure 5.1, Parler does not follow this trend. Instead, we notice

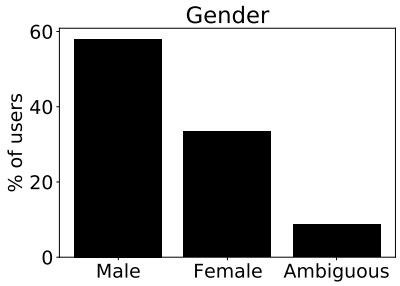


FIGURE 5.1: Histogram showing gender results rounded up

that a large majority of users on Parler are male.

5.4 Summary

We notice an increase in the number of *verified* badges handed out by Parler. This shows us that users are going through Parler's verification process and are willing to show and guarantee other users that they are not bots and are human users. Interestingly, we notice that the number of users with the *Private* badge has decreased post moderation policy. We also notice an increase in the number of other badges like the *Gold* badge as well as an increase in the number of followers and following per user. This might indicate that users are still using Parler. Using the names from user profile data we also analyze their gender. We noticed a large number of male users which suggests that Parler is predominantly a male-dominated social media platform. This is contrary to other social media services like Facebook and Twitter which have reported a higher share of female users when compared to male users [59].

Chapter 6

Parley Analysis Using Persective API

6.1 Introduction to the Perspective API

The Perspectve API is a tool developed by Google, available for users to identify abusive texts or comments. The service uses machine learning models to score phrases or sets of phrases based on the perceived impact the text may have in a conversation [63]. These scores can help users infer whether a text contains harmful aspects or not. The service was mainly marketed towards moderators, who can use the automated service more easily to review comments. Perspective is used by developers at platforms like Disqus and The Wall Street Journal to aid their moderation techniques [63]. It is also used by studies to obtain and sometimes compare different datasets of texts or comments. The Perspective API offers its services in Arabic, Chinese, Czech, Dutch, English, French, German, Hindi, Spanish, Russian, Korean, Portuguese, and several other languages. The newest generation of Perspective algorithms has also been trained to be effective across different usages and styles of text. This is important when monitoring communities, which can develop standard practices and popular words over time that are not seen in other communities.

We notice such occurrences on Parler while manually scanning some users. For example, the phrase *WWG1WGA* was highly active on Parler. The phrase, which stands for Where We Go as 1, We Go as All, is associated with the QAnon conspiracy theory and Parler served as the de-facto communication platform for most of these surrounding conversations [11].

As discussed above, Perspective returns a list of attributes along with scores for each reporting the prevalence of the individual attribute. Perspective allows users to retrieve the scores for the following attributes [16]:

- Toxicity: Toxicity is defined as a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
- Severe toxicity: Severe toxicity on the other hand goes further into the hateful nature and is defined by Perspective as a very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute

is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.

- Identity Attack: Defined by Perspective as negative or hateful comments targeting someone because of their identity.
- Insult: This attribute is defined by Perspective as an insulting, inflammatory, or negative comment towards a person or a group of people.
- Profanity: Perspective defines profanity as swear words, curse words, or another obscene or profane language.
- Threat: Describes an intention to inflict pain, injury, or violence against an individual or group.

6.2 Perspective Scores Data Collection

Perspective provides users with two endpoints: i) AnalyzeComment where a comment or a text is analyzed and returns the likelihood scores for the chosen attributes, and ii) SuggestCommentScore which allows users to suggest scores, often used to change how a comment or piece of text was evaluated by Perspective. For our study, we used the AnalyzeComment API to collect attribute likelihood scores for each post from both the pre moderation policy change and post moderation policy change datasets. Similarly, for each individual post we collected likelihood scores for each of the 6 attributes: i) Toxicity, ii) Severe Toxicity, iii) Identity Attacks, iv) Insults, v) Profanity, and vi) Threats.

For each post, a request formulated with the text from a Parley required attributes, and the language was sent. The response would be a list of attribute likelihood scores which were stored alongside the corresponding posts. While collecting the scores, English was used as the default language for all posts since previous studies showed us that a large majority of Parler's userbase was using English as their language of choice to communicate with other Parler users [6].

Perspective's AnalyzeComment API endpoint returns attribute scores on a successful request as mentioned above. On an unsuccessful request, the response contains an error type, message, and a description of the error failing the request call. While collecting attribute likelihood scores, we noticed some requests resulting in these errors, and the posts were made note of. These posts were separated based on error type and handled in the second stage of sending Perspective requests for the attribute likelihood scores.

6.3 Perspective Score Analysis

After saving these attribute likelihood scores, we performed a simple Mann-Whitney test for each attribute since they were not following a normal distribution. The tests were performed separately for each attribute to compare their scores between pre moderation policy changes and post moderation policy changes. The tests revealed a significant difference in the distribution of all 6 attributes ($p < 0.001$), and we could reject the hypothesis that the distribution of scores is the same in both the pre and post moderation policy changes. The resultant U values are shown in Table 6.1.

Table 6.1: Mann Whitney Test for Perspective Scores

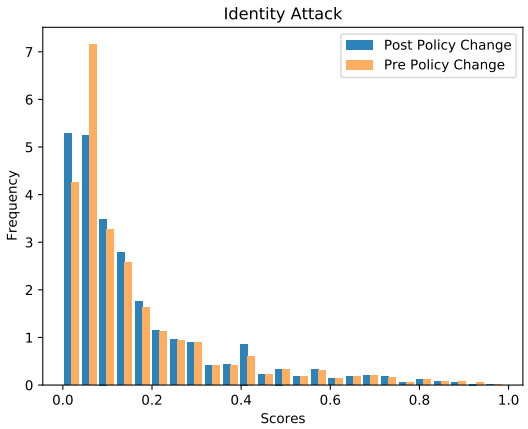
U	Pre Policy Change v/s Post Policy Change
Identity Attack	100,134,726,346,717***
Threat	90,420,711,260,114***
Profanity	96,825,187,118,705***
Toxicity	96,825,187,118,705***
Severe Toxicity	92,914,792,700,764***
Insult	99,313,869,277,933***

Note: *p<0.05, **p<0.01, ***p<0.001

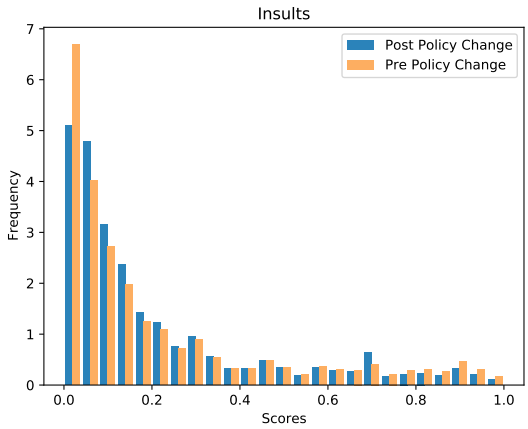
Table 6.2 also shows the min, max, mean and median scores for both the datasets as well as a separate column for disappeared users who are only present in the pre policy change dataset. Through this table, we can also observe that both the mean and median scores for all the attributes are lower post moderation change when compared to the parleys from pre moderation policy change. Interestingly, we also observe that the scores for disappeared users are the lowest on every attribute as well.

Table 6.2: Perspective Score Comparision

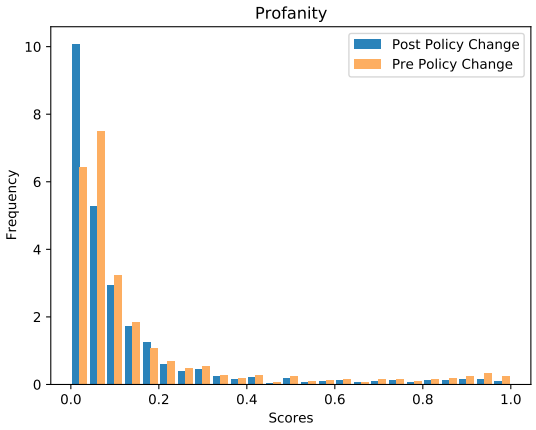
	Pre Policy Change				Disappeared				Post Policy Change			
Metric	Min	Max	Mean	Median	Min	Max	Mean	Median	Min	Max	Mean	Median
Identity Attack	0	0.99	0.18	0.12	0	0.99	0.11	0.06	0	0.99	0.16	0.10
Threat	0	1.0	0.22	0.15	0	1.0	0.16	0.12	0	1.00	0.19	0.12
Profanity	0	1.0	0.18	0.08	0	1.0	0.11	0.04	0	1.00	0.12	0.05
toxicity	0	1.0	0.28	0.16	0	1.0	0.16	0.07	0	1.00	0.22	0.12
Insult	0	1.0	0.25	0.14	0	1.0	0.13	0.04	0	1.00	0.21	0.11
severe_toxicity	0	0.99	0.17	0.07	0	0.99	0.09	0.04	0	0.97	0.12	0.05



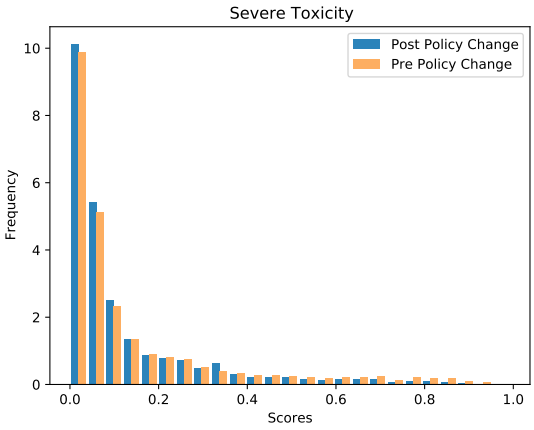
(A) Histogram of Identity Attack Scores



(B) Histogram of Insult Scores



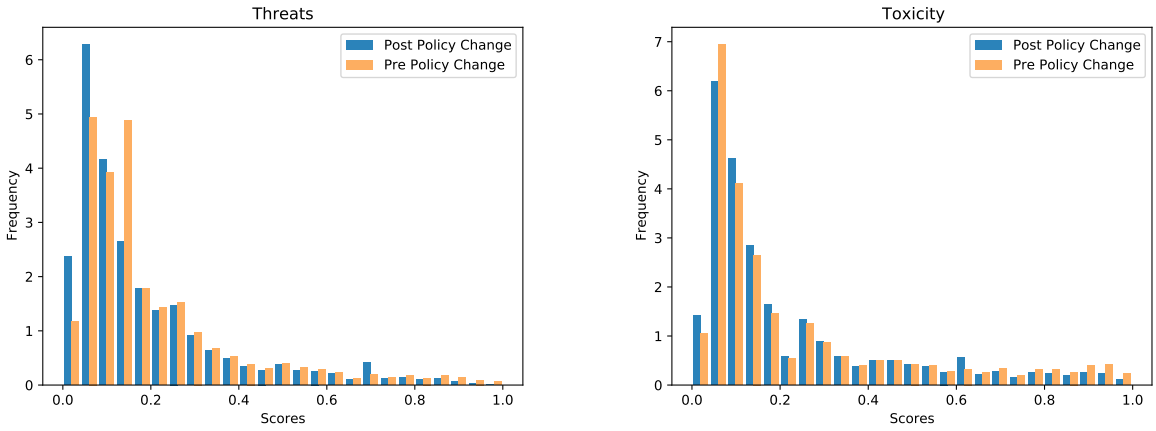
(C) Histogram of Profanity Scores



(D) Histogram of Severe Toxicity Scores

FIGURE 6.1: Histogram of Perspective Attribute Likelihood Scores(I)

We then plot all 6 attribute likelihood scores on a histogram to compare the prevalence of each attribute between pre moderation policy change and post moderation policy change in Figures 6.1 and 6.2. In these histograms, we notice that the labels toxicity 6.2b, severe toxicity 6.1d and profanity 6.1c had a higher number of posts with scores above *0.65* in the pre moderation policy change parleys when compared to parleys from the post moderation policy change. Since Parler's re-entry to the Apple store required action towards posts inciting violence [36], we can conclude with these scores that Parler has worked on reducing toxicity levels shared by their users.



(A) Histogram of Threat Scores

(B) Histogram of Toxicity Scores

FIGURE 6.2: Histogram of Perspective Attribute Likelihood Scores(II)

6.4 Summary

In summary, we notice a severe change in the toxicity as observed in the different Perspective scores toxicity, severe toxicity, identity attack, threat, profanity, and insult. We also noticed that a subset of users who have disappeared show lower toxicity scores than both the pre and post policy change users. After plotting each of the attributes returned by Perspective, we notice additional evidence to further answer our research question. It shows that the prevalence of toxicity did decrease after the moderation changes.

Chapter 7

Topics of Discussion

7.1 Introduction

In this part of the study, we used textual data collected from Parleys to pick the most popular topics. We were initially motivated to study this to observe the most popular topics. Using these topics from each dataset, we compare them to learn about any changes in the discussion. Although we aim to learn about the effectiveness of moderation techniques, we acknowledge that not all of the changes in the topics might be caused due to changes in moderation. We discuss this in more depth in Chapter 9. To extract the most popular topics we used the Latent Dirichlet Allocation (LDA) topic modeling technique. The LDA technique builds a topic including relevant words to form a list of popular topics which can be studied and compared.[10]

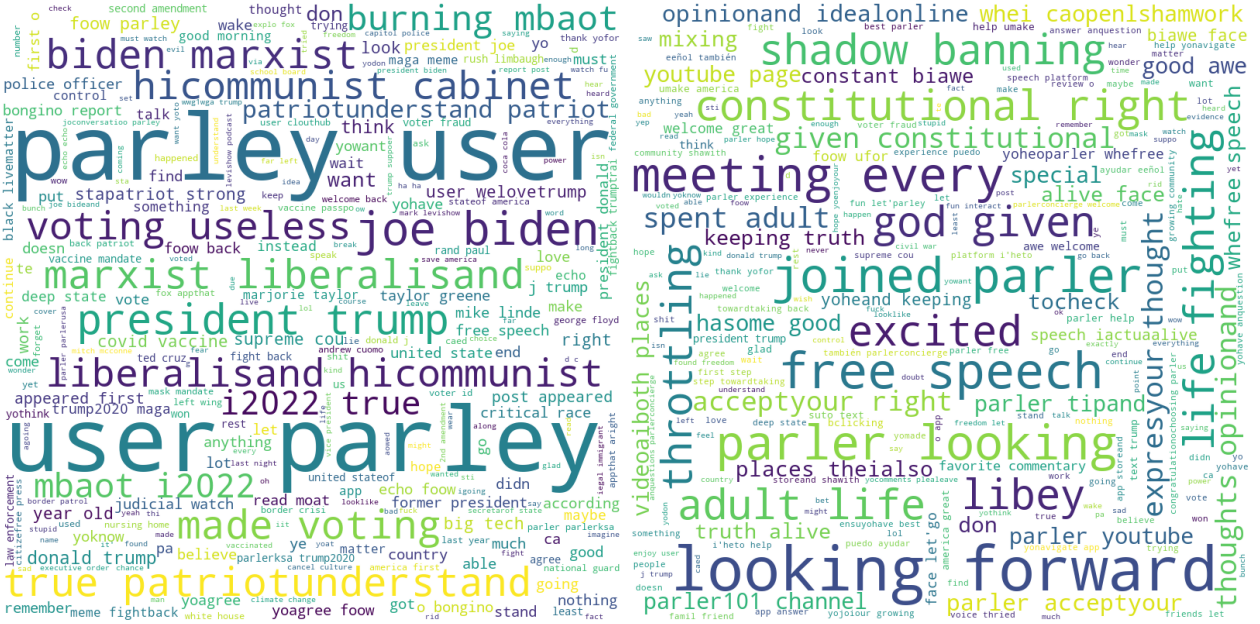
7.2 Methodology

For this phase of our analysis, we used textual data collected with Parleys from our data collection phase as described in chapter 4. First, we removed all the URLs as this would be analyzed later and also removed any unicode characters representing emoticons. After this, we also remove stopwords present in our dataset before extracting popular topics. We use a corpus of stopwords from the Natural Language Toolkit as a list of stopwords to remove from our data. We then use the LDA technique to extract the most popular topics discussed. Using this list of topics we construct a word cloud as seen in Figures 7.1a and 7.1b. .

7.3 Word Cloud

The LDA technique used in the previous step highlights hidden topics that have not been developed yet. This allows us to compare not just the words being used but actual topics of discussion. We noticed a lot of interest in the 2020 US Elections in our pre moderation policy changes. This can be attributed to the fact that the elections took place during the time period of pre moderation policy and most news and discussions about the subject were before post moderation policy parleys were collected.

We also noticed a reduction in the usage of words like WWG1WGA which stands for Where We Go as 1, We Go as All. This term is associated with the QAnon conspiracy movement. We also found several Parler specific words such as Parleys present in the earlier dataset. We theorize that the cause for this might be due to users migrating from other social media platforms like Twitter and Facebook which do not use these terms. These new users could be talking about the Parler specific terms, introducing them to other users, and learning about them themselves.



(A) Word Cloud of Popular topics Pre Moderation Policy Changes

(B) Word Cloud of Popular topics Post Moderation Policy Changes

FIGURE 7.1: Word Cloud comparing popular topics between pre and post moderation policy changes

7.4 Summary

In this section, we uncovered several topics which were widely discussed during several periods of Parler's existence. We noticed some fringe topics being discussed on Parler which have been proven to have originated on other social media platforms. By visualizing the topics in a word cloud we also noticed a lot of users coming to terms with Parler's terms like parleys, echoes, etc.

Chapter 8

Links Shared in Parleys Analysis

8.1 Extracting Links

After the data collection phase, we conducted a brief manual study of the Parleys collected. While viewing the Parleys collected, we observed that several users were sharing links to news websites and other social media websites like YouTube, Twitter, and Rumble. We wanted to further examine this and understand any trends, specifically aligning these trends with the rhetoric around online communities would allow a better understanding of the changes.

To extract links being shared to external websites from Parleys, we used every Parley in both the pre moderation policy change and the post moderation policy change and check for any valid URLs being present. After this, we extracted only the top-level domain names from each URL being shared and add them to our set of links. We also stored the number of times each domain has been shared and use it to measure the popularity of websites in each dataset.

We also use these top-level domains to categorize the top 130 links being shared in Table 8.2. After noticing a large number of news sites being shared we decided to use the Media Bias Fact Check service which labels news sites on how credible they are, country of origin, any political bias or the presence of conspiracy theories, questionable sources, or pro-science sources.

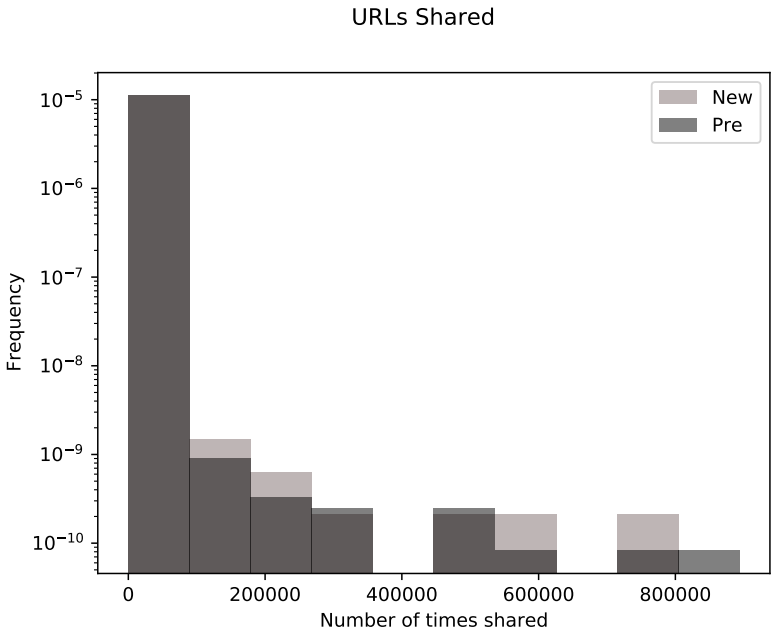


FIGURE 8.1: Histogram of # of Times Websites Any Website is Shared

TABLE 8.1: Most Popular Websites Shared on Parler

Website	Pre Policy	Post Policy	Change(%)
image-cdn.parler.com	7,318,992	1	−99.99
youtube.com	2,499,198	225,562	−83.44
youtu.be	1,812,871	19	−99.99
bit.ly	893,603	5	−99.99
twitter.com	803,514	42,638	−89.92
media.giphy.com	539,389	545	−99.79
i.imgur.com	532,365	5,779	−97.85
facebook.com	520,796	318	−99.87
thegatewaypundit.com	469,855	610,512	+13.01
breitbart.com	328,953	240,547	−15.52
foxnews.com	298,285	136,956	−37.06
instagram.com	168,160	22,932	−75.99
rumble.com	164,949	744,132	+63.71
theepochtimes.com	136,294	33,937	−60.12
hannity.com	13,017	148,026	+83.83
justthenews.com	50,638	147,984	+49.01
www.theblaze.com	2,006	122,111	+96.76
www.westernjournal.com	6,399	119,551	+89.83
bongino.com	17,251	114,334	+73.77
www.bitchute.com	104,462	87,672	−8.73

From Table 8.1 we can see the sharp rise in popularity of Rumble on the Parler social media platform. We can also notice the decline in Twitter links being shared. These observations could be explained by the sharp rise in the rhetoric surrounding censorship on Twitter and other popular social media platforms [1].

8.2 URL Categorization

After extracting links from Parleys, we noticed that we had collected 151,713 links from the pre moderation policy dataset and 61,564 links from the post moderation policy dataset. Out of these, we selected the top 130 links by frequency of shares to categorize into:

- News Websites: Websites that report current news events and opinion pieces from individual journalists. Example: *CNN.com*, *FoxNews.com*, *TheEpochTimes.com*, etc
- Social Media Websites: Websites or online platforms used by users to maintain social connections, share information, and meet new users. Example: *Twitter.com*, *Facebook.com*, etc
- Media Hosting: Websites used to store and share media (images, videos or gifs) primarily. Example: *giphy.com*, *imgur.com*, etc
- Personal Blog: Websites maintained by a single person which shares news reports about live events or other articles. Most of the websites in this category were political commentary by a known influencer. Example: *bonginoreport.com*, *Hannity.com*, etc
- Other: We also noticed websites for online petitions, search engines, etc which were categorized under this.

The results of these categorization steps can be seen in Table **??**

Table 8.2: Website Categories

Category	Pre Policy	Post Policy
News	85	88
Social Media	18	17
Media Hosting	12	4
Personal Blog	8	17
Other	6	4

8.3 Media Bias Fact Check

After categorizing some of the links collected, we noticed several news sites being shared and decided to study these links. We used the Media Bias Fact Check(MBFC) service to measure factuality, presence of any bias, country of origin and presence of conspiracy or pseudoscience, questionable sources, and pro-science sources.

MBFC is an independent organization that uses volunteer and paid contributors to rate and store information about news websites [2]. The organization states its objective as *To determine the bias of media and information sources and the level of overall factual reporting through a combination of objective measures and subjective analysis through the use of our stated methodology.*. We used a list of links shared from both of our datasets to obtain labels for:

- Factuality: Referred to as how factual a website is. Scored between 0-5 where a score of 0 means that a website is not factual and a 5 is very factual. MBFC defines that for a website to be very factual and get a score of 5, it should pass its fact-checking test as well as make sure that critical information is not omitted. Failure of any of these criteria reduces their score [39].
- Bias: MBFC assigns a bias rating of Extreme left, left, left-center, least biased, right-center, right, and extreme right. To assign a bias rating to a website, MBFC contributors check the website's stance on American issues which divides left-biased websites from right-biased websites [35]. Some examples of these issues are:
  - General Philosophy: Left-biased websites often view issues as a collective issue and are generally more collective in their thinking. These websites also value equality, environmental protection, social safety nets, and expanded educational opportunities. On the other hand, right-biased websites view issues as an individual over a community. They prefer a limited government with individual freedom and property rights.
  - Abortion: Left-biased websites believe that abortion is legal in most cases while right-biased websites support the opposite.
  - Economic Policy: Left biased websites believe in income equality, higher taxes on the wealthy, government spending on social programs and infrastructure, and stronger regulations on business and minimum wages. On the other hand, right-biased websites believe in lower taxes, lesser regulations on businesses, reduced government spending, wages being set by the free market, lesser government spending, and charity over social nets.

- Education Policy: Left-biased websites favor lowering the cost of education and government spending while right-biased websites favor the method of homeschooling and are often critical of what is taught in public schools.
- Environmental Policy: Left biased websites believe in strong environmental regulations and a governmental policy to reduce climate change which is man-made. Right biased websites believe that the free market should make decisions about the climate and they will make the best decision compared to the government.
- Gay Rights: Left biased websites support the legalization of gay marriage and protections against discrimination in the workplace. Right-biased websites are generally opposed to gay marriage and believe that some anti-discrimination laws interfere with citizen's freedom of religion.
- Gun Rights: Left biased websites to favor stricter background checks and regulations for gun owners and banning certain high capacity weapons from preventing mass shootings. Right-biased websites believe that owning a gun is a constitutional right from the second amendment and any strong regulation against owning guns is a restriction of their freedom to bear arms.
- Healthcare: Left biased websites to favor universal healthcare and believe that healthcare is a human right. They also support the Affordable Care Act while right-leaning websites oppose it and believe that corporations can provide better healthcare than governmental programs.
- Immigration: Left-leaning websites believe in less restrictive immigration laws, providing a pathway to citizenship and a moratorium on deporting while right-leaning websites believe in stronger borders, deporting any undocumented immigrant, and a more restrictive immigration policy.
- Military: Left-leaning websites critic the government on military spending and believe that it should decrease while right-leaning websites believe in increasing military spending.
- Personal Responsibility: Left-leaning websites believe in stronger laws to protect individuals and to provide structure. Right-leaning websites on the other hand believe in fair competition and the government's role to hold personal responsibility accountable.
- Regulation: Left-leaning websites believe that stronger regulations need to exist to protect consumers and the environment. Right-leaning websites believe in very relaxed

government regulations and that stronger regulations hinder the free market and decrease job growth.

  - Taxes: Left-leaning websites believe in a progressing tax system and are not opposed to raising taxes to increase government spending while right-leaning websites are against increasing taxes for governmental spending. They are also against a progressive taxing system and believe in a flat-rate tax system.

  - Voted ID: Left-leaning websites believe that voter IDs are not required and that it places an undue burden on lower-income voters while voter fraud is virtually non-existent. Right-leaning websites believe that voter IDs are a significant part of a democracy and should be required for all voters.

  - Workers Rights: Left-leaning websites believe in unions, and governmental regulations for minimum wages and support worker's rights while right-leaning websites believe that the free market will force higher-profiting corporations to reward workers with higher wages.

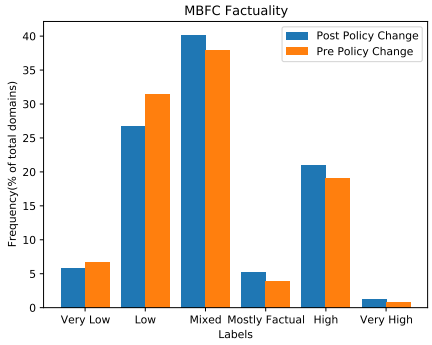
- Presence of conspiracy-pseudoscience: Websites that publish unverified information related to known conspiracies such as the New World Order, Illuminati, False flags, aliens, anti-vaccine, etc. For example, sources that promote human-influenced climate change denial or take anti-vaccination positions will be classified as pseudoscience [39].

- Country of origin: Manually labeled by contributor based on the founding of the website.

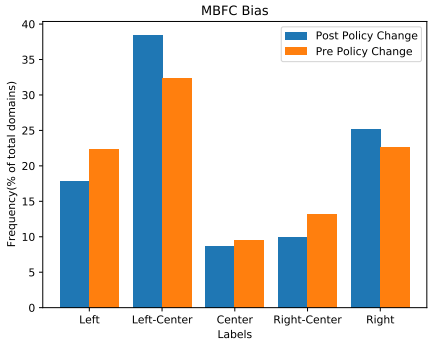
- Usage of questionable sources: MBFC defines this as *A questionable source exhibits any of the following: extreme bias, overt propaganda, poor or no sourcing to credible information, a complete lack of transparency, and/or is fake news. Fake News is the deliberate attempt to publish hoaxes and/or disinformation for profit or influence* [39].

- Usage of pro-science sources: The website's usage of pro-science sources or facts supporting their arguments earns them a pro-science sources label.

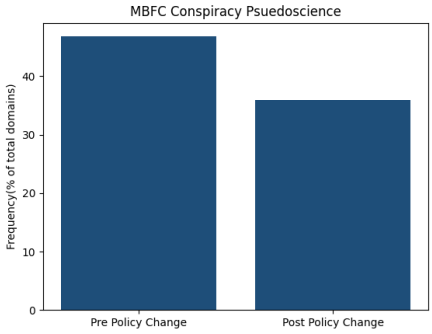
After this step, we collected labels for 7.72% and 1.77% of all links being shared on Parleys from the pre and post moderation policy change dataset respectively. Using these labels, we created histograms to visualize the results. At first, we noticed a decrease in the number of conspiracy-pseudoscience news articles in Figure 8.2c. We also notice the same decrease in the usage of questionable sources and pro-science in Figures 8.2e and 8.2f. We also notice that the country of origin for several websites is unknown in the pre moderation policy change dataset. Often times these websites are obscure sites sharing news articles that are not credible. In Figure 8.2a we also



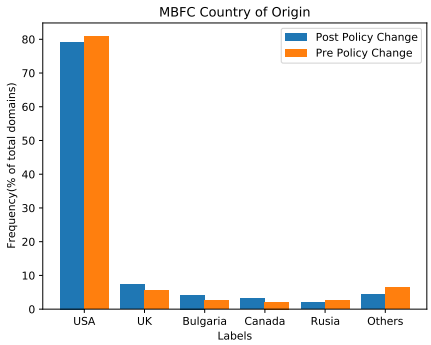
(A) Histogram of Factuality Scores



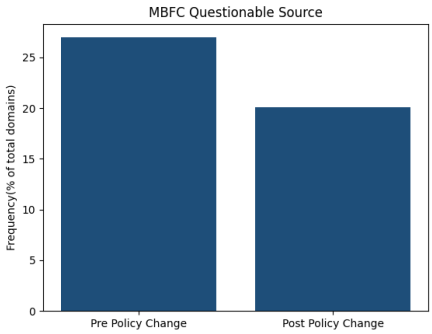
(B) Histogram of Bias



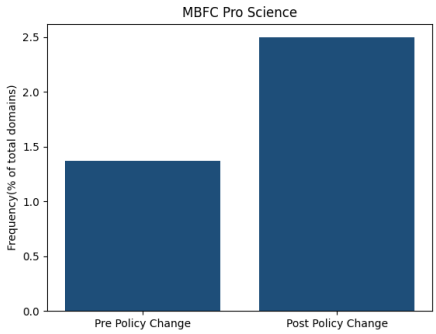
(C) Histogram of Conspiracy-Pseudoscience



(D) Histogram of Origin Countries



(E) Histogram of Questionable Sources



(F) Histogram of Pro Science Sources

FIGURE 8.2: Histogram of MBFC Labels

notice that most links with a score between *1* and *2* were from the pre moderation policy change while post moderation policy links are scattered across higher ranges between *2* and *5*.

8.4 Summary

To study and answer our research question about URLs being shared, we use the MBFC service along with the labels returned for credibility, bias, presence of questionable source, conspiracy-pseudoscience or pro-science, and the country of origin. Using these labels we can come to the conclusion that the credibility of the URLs being shared did increase. We also record other data like the bias of these websites which can be useful for future studies.

Chapter 9

Limitations & Future Work

In our current dataset i.e., the post-policy change dataset, we were not able to collect a random sample of users, hence our analysis might not yield a full-scale impact of the moderation policy change. The other limitation of our work is that users might have changed their usernames when Parler was reinstated back online, as to evade possible detection. We also found that MBFC uses the American definition of the key issues, however, we found about 5% of the links that were originating from the UK, hence MBFC might not have given the full analysis for these links. Another limitation of our study is that users could have changed their profiles to private. We also did not run any analysis for the comments on these posts. Comments can also shed light on the moderation changes that Parler undertook.

The decision to use the Perspective API to answer the research question *RQ# 1: Has the prevalence of hateful and toxic content decreased* was based on the merits of their service such as i) Perspective open sources their methodology and ii) Perspective actively updates their dictionaries to match current lingo being used on social media platforms. The continuous update, while helping us stay informed would not be an accurate assessment of words used in the past. For example, A word used in a Parley from 2018 could have a different connotation or meaning now when compared to its original usage. This brings about only a minor limitation since Parler is a relatively young social media platform(founded in 2018) and hence contains texts with words whose connotations and meanings haven't vastly changed since.

While answering our *RQ #2: Have the topics of discussion changed? If so what are the changes?* we use word clouds to show popular topics over a period of time consisting of the US Presidential Election along with other key events in the US Political field. Some of these events could have led to a change in the popularity of a specific topic. We have also noticed that such events divide users and bring about several toxic traits which are not usually seen. Due to these observations, comparing popular topics between pre and post moderation policy changes can be unreliable.

In the future, we plan to obtain a random sample of the users after Parler was reinstated back online. We assume that this will shed even better light on the effectiveness of the moderation

change. We also plan on comparing the users based on gender, to understand if either of the genders is more active on parler after the moderation changes, compared to the pre-policy change dataset. We also plan to sort the perspective scores into two sets, one below 0.5 and one above 0.5 for all the attributes, in both the pre and post policy change dataset, to understand what topics were getting more toxic and what topics were less toxic. We also plan on doing a more thorough analysis to find the exact countries that we could not find and labeled them as *unknown*. We also plan to study the user comments on posts to understand if the moderation changes are being reflected in the comments also.

Chapter 10

Conclusion

On January 12, 2021, Parler a social media platform popular among conservative users was removed from the Apple App Store and the Google Play Store. Amazon Web Services, Parler's cloud partner also stopped hosting Parler content shortly after. This was blamed on Parler's refusal to remove posts inciting violence following the 2021 US Capitol Riots. Several studies published frameworks and methodology for scraping Parler data using their open API before Parler went offline. Parler was eventually allowed back after they promised that they will strengthen their moderation to remove hateful content. Our study looks into the effect of these policy changes on the user discourse in the pre policy dataset that was collected is published in previous studies, and the post policy dataset, that we collected by first establishing a new framework to collect posts and user information from Parler.

Initially, Parler's open API service was leveraged to collect a large number of parleys and user information by studies like Aliapoulios et al. [5]. The current study, using the newer version of Parler's API, collects a similar dataset to compare the changes in moderation policies. Using observations from this study we aim to help social media platforms make informed decisions on content moderation policies. We also hope to study how effective these content moderation techniques are on curbing harassment, hate speech, and the spread of misinformation and disinformation. Social media platforms can benefit from this because content moderation is widely used to grow their userbase and actively encourage newer users to join their platform.

Our study wanted to answer three main research questions:

- **RQ #1**: Has the prevalence of hateful and toxic content decreased? Our study explores this question to study the effectiveness of the content moderation techniques and compare them from the pre and post moderation policy posts.
- **RQ #2**: Have the topics of discussion changed? If so what are the changes? This question would also help understand the impact of moderation change on user behavior.

- **RQ #3**: Is there any difference in the credibility of news sources being shared between pre and post moderation policy changes? After manually glancing at some posts we noticed a large number of links to news websites being shared. Therefore, we decided to study these links and find any differences in the credibility of the news sources being shared.

Using Google's Perspective API in section 6, we look into answering *RQ #1*. We notice a decrease in trend in the toxicity of Parleys being posted post moderation policy changes. Using this trend we can conclude that yes the prevalence of hateful and toxic content did decrease after moderation techniques were changed on Parler.

From our data collection phase, we used the Parleys collected, and in section 7 use this data along with the LDA topic modeling technique to observe any changes in the topics of discussion. We observe several changes, mostly due to the nature of the events occurring during the time of posting. We observe a decrease in some conspiracy theory rhetoric talked about in section 7.

To answer our *RQ #3* we used the MBFC service which manually labels websites to score their credibility along with certain other labels. We noticed a decrease in the presence of conspiracy pseudoscience as well as questionable sources. We also notice an increase in the factuality as well as pro-science sources which suggest that the credibility of the news sites being shared has increased.

Bibliography

[1] Emily A. Vogels, Andrew Perrin, and Monica Anderson. *Most Americans Think Social Media Sites Censor Political Viewpoints*. Aug. 2020. URL: https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/.

[2] *About MBFC*. Jan. 2022. URL: https://mediabiasfactcheck.com/about/.

[3] Usman Ahmed and Jerry Chun-Wei Lin. "Deep Explainable Hate Speech Active Learning on Social-Media Data". In: *IEEE Transactions on Computational Social Systems* (2022), pp. 1–11. DOI: 10.1109/TCSS.2022.3165136.

[4] Shiza Ali et al. "Understanding the Effect of Deplatforming on Social Networks". In: *13th ACM Web Science Conference 2021*. 2021, pp. 187–195.

[5] Max Aliapoulios et al. "A Large Open Dataset from the Parler Social Network". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021, pp. 943–951.

[6] Max Aliapoulios et al. "An early look at the parler online social network". In: *arXiv preprint arXiv:2101.03820* (2021).

[7] Perspective API. *Using machine learning to reduce toxicity online*. URL: https://perspectiveapi.com/.

[8] Annalise Baines, Muhammad Ittefaq, and Mauryne Abwao. "# Scamdemic,# plandemic, or# scaredemic: What parler social media platform tells us about COVID-19 vaccine". In: *Vaccines* 9.5 (2021), p. 421.

[9] Michael Bernstein et al. "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 5. 1. 2011, pp. 50–57.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[11] Ben Collins. *Increasingly militant 'parler refugees' and Anxious Qanon adherents prep for Doomsday*. Jan. 2021. URL: https://www.nbcnews.com/tech/internet/increasingly-militant-parler-refugees-anxious-qanon-adherents-prep-doomsday-n1254775.

[12] Elizabeth Culliford and Vengattil Munsif. *Explainer: What is Parler and why has it been pulled offline?* https://www.reuters.com/article/us-usa-socialmedia-parler-idUSKBN29H2G2. 2021.

[13] Beth Daley. *The abuse tactics fraudsters use to break the hearts and wallets of those looking online for love*. 2022. URL: https://theconversation.com/the-abuse-tactics-fraudsters-use-to-break-the-hearts-and-wallets-of-those-looking-online-for-love-93663.

[14] Kareem Darwish, Walid Magdy, and Tahar Zanouda. "Trump vs. Hillary: What went viral during the 2016 US presidential election". In: *International conference on social informatics*. Springer. 2017, pp. 143–161.

[15] Thomas Davidson et al. "Automated hate speech detection and the problem of offensive language". In: *Eleventh international aaai conference on web and social media*. 2017.

[16] *Developers*. URL: https://developers.perspectiveapi.com/s/about-the-api-methods.

[17] Karthik Dinakar et al. "Common sense reasoning for detection, prevention, and mitigation of cyberbullying". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2.3 (2012), p. 18.

[18] Yoan Dinkov et al. "Predicting the leading political ideology of YouTube channels using acoustic, textual, and metadata information". In: *arXiv preprint arXiv:1910.08948* (2019).

[19] Nemanja Djuric et al. "Hate speech detection with comment embeddings". In: *Proceedings of the 24th international conference on world wide web*. ACM. 2015, pp. 29–30.

[20] donk_enby. *Parler Tricks*. https://doi.org/10.5281/zenodo.4426283. 2021.

[21] Mai ElSherief et al. "Hate lingo: A target-based linguistic analysis of hate speech in social media". In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[22] *Fake news and Cyber Propaganda: The use and abuse of social media*. URL: https://www.trendmicro.com/vinfo/pl/security/news/cybercrime-and-digital-threats/fake-news-cyber-propaganda-the-abuse-of-social-media.

[23] Brian Fung. *Parler has now been booted by Amazon, Apple and Google | CNN business*. https://www.cnn.com/2021/01/09/tech/parler-suspended-apple-app-store/index.html. 2021.

[24] Njagi Dennis Gitari et al. "A lexicon-based approach for hate speech detection". In: *International Journal of Multimedia and Ubiquitous Engineering* 10.4 (2015), pp. 215–230.

[25] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. *NELA-GT-2020: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles*. 2021. DOI: 10.48550/ARXIV.2102.04567. URL: https://arxiv.org/abs/2102.04567.

[26] Aarash Heydari et al. "YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos". In: *arXiv preprint arXiv:1907.00435* (2019).

[27] Gabriel Emile Hine et al. "Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web". In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[28] *HTTP for humans™¶*. URL: https://docs.python-requests.org/en/latest/.

[29] Johannes Jakubik et al. "Online Emotions During the Storming of the US Capitol: Evidence from the Social Media Network Parler". In: *arXiv preprint arXiv:2204.04245* (2022).

[30] Shagun Jhaver et al. "Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–30.

[31] Sayeed Ahsan Khan, Mohammed Hazim Alkawaz, and Hewa Majeed Zangana. "The Use and Abuse of Social Media for Spreading Fake News". In: *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. 2019, pp. 145–148. DOI: 10.1109/I2CACIS.2019.8825029.

[32] Animesh Koratana and Kevin Hu. "Toxic Speech Detection". In: ().

[33] Anti-Defamation League. *Qanon: A glossary*. https://www.adl.org/blog/qanon-a-glossary. 2021.

[34] Alyssa Lees et al. *A New Generation of Perspective API: Efficient Multilingual Character-level Transformers*. 2022. DOI: 10.48550/ARXIV.2202.11176. URL: https://arxiv.org/abs/2202.11176.

[35] *Left vs. right bias: How we rate the bias of media sources*. May 2021. URL: https://mediabiasfactcheck.com/left-vs-right-bias-how-we-rate-the-bias-of-media-sources/.

[36] Rachel Lerman. *Parler's revamped app will be allowed back on Apple's App Store*. https://www.washingtonpost.com/technology/2021/04/19/parler-apple-app-store-reinstate/. 2021.

[37] Thomas J Main. *The rise of the alt-right*. Brookings Institution Press, 2018.

[38] Yashar Mehdad and Joel Tetreault. "Do characters abuse more than words?" In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2016, pp. 299–303.

[39] *Methodology*. Apr. 2022. URL: https://mediabiasfactcheck.com/methodology/.

[40] Kristian Miok et al. *To BAN or not to BAN: Bayesian Attention Networks for Reliable Hate Speech Detection*. 2020. DOI: 10.48550/ARXIV.2007.05304. URL: https://arxiv.org/abs/2007.05304.

[41] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. "Social media is polarized, social media is polarized: towards a new design agenda for mitigating polarization". In: *Proceedings of the 2018 Designing Interactive Systems Conference*. 2018, pp. 957–970.

[42] Edward Newell et al. "User migration in online social networks: A case study on reddit during a period of community unrest". In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[43] Lynnette Hui Xian Ng, Iain Cruickshank, and Kathleen M Carley. "Coordinating Narratives and the Capitol Riots on Parler". In: *arXiv preprint arXiv:2109.00945* (2021).

[44] Shirin Nilizadeh et al. "Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility". In: *Proceedings of the International AAAI Conference on Web and Social Media* 10.1 (Aug. 2021), pp. 289–298. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14711.

[45] Chikashi Nobata et al. "Abusive language detection in online user content". In: *Proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee. 2016, pp. 145–153.

[46] Antonis Papasavva et al. "" Is it a Qoincidence?": A First Step Towards Understanding and Characterizing the QAnon Movement on Voat. co". In: *arXiv preprint arXiv:2009.04885* (2020).

[47] Antonis Papasavva et al. "Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 885–894.

[48] Victoria Patricia Aires, Fabiola G. Nakamura, and Eduardo F. Nakamura. "A link-based approach to detect media bias in news websites". In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 742–745.

[49] Avinash Prabhu et al. "Capitol (Pat) riots: A comparative study of Twitter and Parler". In: *arXiv preprint arXiv:2101.06914* (2021).

[50] Aneri Rana and Sonali Jha. *Emotion Based Hate Speech Detection using Multimodal Learning*. 2022. DOI: 10.48550/ARXIV.2202.06218. URL: https://arxiv.org/abs/2202.06218.

[51] Adrian Rauchfleisch and Jonas Kaiser. "Deplatforming the far-right: An analysis of YouTube and BitChute". In: *Available at SSRN* (2021).

[52] Raut, Sahil et al. "Hate Classifier for Social Media Platform Using Tree LSTM". In: *ITM Web Conf.* 44 (2022), p. 03034. DOI: 10.1051/itmconf/20224403034. URL: https://doi.org/10.1051/itmconf/20224403034.

[53] Manoel Horta Ribeiro et al. "Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels". In: *arXiv preprint arXiv:2010.10397* (2020).

[54] Adi Robertson. *Parler is back online after a month of downtime*. https://www.theverge.com/2021/2/15/22284036/parler-social-network-relaunch-new-hosting. 2021.

[55] Richard Rogers. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media". In: *European Journal of Communication* 35.3 (2020), pp. 213–229.

[56] Mike Rothschild. *Parler wants to be the 'free speech' alternative to Twitter*. May 2021. URL: https://www.dailydot.com/debug/what-is-parler-free-speech-social-media-app/.

[57] Sara Sood, Judd Antin, and Elizabeth Churchill. "Profanity use in online communities". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2012, pp. 1481–1490.

[58] Kate Starbird. "Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017, pp. 230–239.

[59] Published by Statista Research Department and Jan 28. *U.S. social reach by gender 2019*. Jan. 2022. URL: https://www.statista.com/statistics/471345/us-adults-who-use-social-networks-gender/.

[60] Peter Stefanov et al. "Predicting the topical stance and political leaning of media using tweets". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 527–537.

[61] Amaury Trujillo and Stefano Cresci. "Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald". In: *arXiv preprint arXiv:2201.06455* (2022).

[62] Marc Tuters and Sal Hagen. "(((They))) rule: Memetic antagonism and nebulous othering on 4chan". In: *New media & society* 22.12 (2020), pp. 2218–2237.

[63] *Using machine learning to reduce toxicity online*. URL: https://perspectiveapi.com/how-it-works/.

[64] Zeerak Waseem and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[65] Galen Weld, Maria Glenski, and Tim Althoff. "Political Bias and Factualness in News Sharing across more than 100,000 Online Communities". In: *ICWSM* (2021).

[66] Savvas Zannettou et al. "What is gab: A bastion of free speech or an alt-right echo chamber". In: *Companion Proceedings of the The Web Conference 2018*. 2018, pp. 1007–1014.

[67] Justine Zhang et al. "Conversational flow in Oxford-style debates". In: *arXiv preprint arXiv:1604.03114* (2016).