Parlez-vous le hate?: Examining topics and hate speech in the alternative social network Parler

by

Ethan Ward

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Over the past several years, many "alternative" social networks have sprung up, with an emphasis on minimal moderation and protection of free speech. Although they claim to be politically neutral, they have been a haven for conservatives who feel mistreated by popular social networks, and often those who have had false posts deleted from those sites.

Parler is the latest in this trend, a Twitter alternative that grew in popularity when Donald Trump and many other conservatives had posts deleted on Twitter for spreading false information. We are among the first to analyze comments and posts made on Parler, attempting to characterize the kinds of posts users made on the site, as well as the amount of hateful words used. We also compare these to randomly sampled Twitter posts from the same time period to determine how Parler differs from conventional social networks.

Acknowledgements

I would like to thank my supervisor, Dr Jesse Hoey, for his continued support during all stages of my thesis, especially during the pandemic.

I would also like to thank the CHIL lab group for feedback on my thesis presentation and regular entertaining presentations. I'd also like to thank the University of Waterloo for the generous scholarships and financial aid.

Finally, I'd like to thank my partner, Elliott, and friends and family for their support and feedback on my thesis.

Table of Contents

List of Figures

Chapter 1

Introduction

There have long been right-wing or conservative communities on the internet, as well as those that promote conspiracy theories or false narratives. These have largely been smaller, relatively unpopular subcultures, such as boards on 4chan [27] or communities on Reddit[36],[22]. However, over the past several years, these communities have spread their hateful ideas to larger, more popular social networks such as Facebook and Twitter[54] [53]. In response to this and to concerns about the use of social media to spread false narratives, popular social networks have increased moderation of the content allowed on their sites. This has caused a crackdown on blatantly false news stories (particularly with respect to elections [13] [17]) and hateful subcommunities [37] [47]. In response to these new policies, "alternative" social networks have sprung up, with an emphasis on minimal moderation and protection of free speech. Although they claim to be politically neutral, they have been a haven for conservatives who feel mistreated by popular social networks, and often those who have had false posts deleted from those sites. Among these have been Gab [52], which advertised itself as a Twitter alternative, and Voat[38], which was meant to be a Reddit alternative. Both of these sites ended up full of hateful and inflammatory content, with Gab eventually removed from app stores and deplatformed after connections to a mass shooting at a synagogue in 2018. [26].

Parler is the latest in this worrying trend, claiming to be a Twitter alternative with an emphasis on free speech. It gained substantial attention initially in 2019, and eventually became one of the most downloaded apps on the Apple App Store after being endorsed by then-President Donald Trump. It was eventually connected with the insurrection attempt at the US Capitol building on Jan 6, 2021 [10], and on Jan 11, 2021, had its security and hosting site pull support, temporarily making the site unusable. The dataset that we use in this analysis only goes up until this date. After unsuccessful lawsuits, Parler has changed

its policies and is now available for download again on the Apple App Store[24], although not on the Google Play Store. It has not regained its popularity, however, especially with much of the election-related momentum that caused its popularity surge gone.

In our analysis of Parler data, we seek to answer four major research questions:

RQ1 - How do Parler users use the platform on average? More specifically, what kind of phrases do Parler users use the most, and what are the common topics of discussion? How does this change between parleys (main posts made by a user) and comments (replies to others' posts)?

RQ2 - How hateful are Parler users, i.e., how many hate words do they use overall?

RQ3 - How do topics of discussion and hatefulness of parleys and comments vary with respect to time and their popularity? Are more hateful posts more likely to be popular?

RQ4 - How do the above findings compare to sampled Twitter posts from the same period? Are Parler posts more hateful than tweets, and are the topics discussed more political?

1.1 Related Work

There is a recent surge in interest in the literature of analyses of hateful and conservative social media. We split these analyses into four rough categories, as well as separating Parler on its own:

- Alternative Social Media: Analyses of social media websites that are created explicitly as free-speech promoting alternatives to more popular sites.
- Subcommunities: Analyses of social media sites that are defined by being split into many subcommunities, thus lending themselves to a more granular analysis of these subcommunities rather than a discussion of the site as a whole.
- Mainstream Social Media: Analyses of extremely popular social media sites.
- Cross-Site Analysis: Analyses that examine trends across different websites, and the impact of some social media websites on others.

We collect information about the social media sites referenced here in Appendix A for those who wish for more context. We additionally discuss the existing literature on hate speech detection.

1.1.1 Parler

This analysis builds on the work done by Aliapoulios et al [6], who created the dataset of Parler posts that we used. In their paper, they explain the structure of the dataset that they created. They then analyze the biographies of users on Parler, finding that the userbase overwhelmingly describes itself as conservative. They also analyze the growth of the userbase and posts on Parler, finding three major increases - June 1, 2019, when a substantial Saudi Arabian population left Twitter for Parler, June 16, 2020, when a prominent conservative politician bought shares in Parler, and November 4, 2020, the US presidential election. They then examine the most popular hashtags and the most popular sites that users shared news articles from.

1.1.2 Alternative Social Media

Gab

Zannettou et al. [52] analyze 22M posts from Gab, a similar free-speech promoting alternative to Twitter. Gab was created in 2016 before the US presidential election that year, and had its mobile app barred from both the Google and Apple app stores for hate speech. Similar to Parler and Twitter, users can create short posts that can then be reposted by their followers. It, like Parler, also has a voting component on posts similar to upvotes and downvotes on Reddit. They find that the most followed users on the site are all prominent conservative or alt-right personalities, with no left-leaning popular users. Similarly, many of the most used words and hashtags relate to Donald Trump and his slogans. Using Hatebase, they find that roughly 5.5% of posts on the site contain a hate word. They compare this to findings from Hine et al.[27] (discussed below), concluding that posts on Gab are less hateful than posts from the 4chan /pol/ board, but more hateful than tweets from the same time period.

Voat

Similarly, Papasavva et al [38] analyze a subcommunity of Voat, billed as a free-speech Reddit alternative. The subcommunity is /v/GreatAwakening, which refers to the popular conspiracy theory known as QAnon. This theory, spread by a 4chan user known as "Q", hints at the existence of a "deep state" that is secretly controlling the American government, and contends that many prominent politicians are part of a secret pedophile ring. The researchers collect 150K posts from /v/GreatAwakening, as well as 350K other

posts from popular "subverses" on the site for comparison. These posts range from May to October 2020. They find that the QAnon-related subverse gets substantially more submissions than other popular subverses. They run LDA (Latent Dirichlet Analysis) for topic discovery on the dataset, finding that both types of subverse discuss US political content, but that GreatAwakening has more of a focus on Donald Trump. They also run a toxicity analysis on the site using the Google Perspective API, finding that GreatAwakening contains less toxicity and profanity than their baselines. They attribute this to the focus on discussion of specific parts of the QAnon conspiracy theory, while other parts of the site are more generally hateful.

1.1.3 Subcommunities

4chan

Another popular source of these analyses is 4chan, an online forum known primarily for its toxicity and anonymity, but also its power to spread ideas across the Internet. One of the first such analyses is that of Bernstein et al. [12], who analyzed 5M posts on the /b/ (Random) board to discern how forums such as 4chan succeed when most posts are anonymous and are deleted from the site rapidly after inactivity. They find that /b/ has roughly 35,000 new threads and 400,000 posts in those threads every day, with each thread lasting a median of 3.9 minutes and a mean of 9.1 minutes. They also find that over 90% of posts on the site are by users who choose to remain anonymous. This inspires a number of other more specific analyses over the coming years, and starts an interest in analyzing fringe subcommunities.

Hine et al. [27] analyze 8M posts over 2 and a half months from /pol/ (Politically Incorrect), a subforum of 4chan that discusses politics. They attempt to characterize the posts, finding that many have links to other websites, with the mots prevalent being YouTube, Wikipedia, and Pastebin. Using a modified version of the Hatebase dictionary, they find that 12% of posts from the /pol/ board of 4chan contained hate words, and that only 2.2% of a subset of tweets from the same time period used hate words. They also analyze some more heavily moderated 4chan boards - /sp/ (Sports) and /int/ (International), of which 6.3% and 7.3% contain hate words respectively. They finish by talking about "raids" on other sites that /pol/ users often perform, analyzing a specific raid on YouTube comments and finding a significant correlation between threads about a raid and hateful comments around the same time.

Reddit

Flores-Saviaga et al. [22] examine 16 million comments from r/The_Donald, a subreddit for supporters of Donald Trump, to understand both how community members act and how they mobilize supporters for group action. They found extensive use of conservative slang, a large portion of which came from widely-used bots on the subreddit, which often helped drive engagement. Next, they used lists of action verbs to gather a large subset of posts, and then used human validation to find a total of roughly 3.3K posts that they classified as a call to action. They then used a clustering algorithm to find similarities between the posts, and found three distinct styles. The first they characterized as "troll slang", using common alt-right slang terms like "kek" (a variant of "lol") and "centipedes" (from a popular Trump-promoting YouTube series). The second was a "Viral News" style, revolving around sharing a link and asking for users to amplify it. The third was a "Historian" style, which had long sections of background detail about the issue they wanted action on.

Mittos et al. [36] analyze the toxicity of conversations about genetic testing from 1.3M comments collected from both 4chan and Reddit. They collect this data by searching for keywords related to genetic testing, focusing on subcommunities that discuss genetic testing regularly. They categorize posts by groups of subreddits, including a general "Hate" category for subreddits decided to be largely hate-related. They use the Google Perspective API to determine the levels of toxic and hateful content, and find that the most toxic communities are those that were deemed as part of the "Hate" category. Some of these communities are also those that have the highest levels of genetic testing-related content. LDA analysis of each category shows larger amounts of discussion of specific race-related topics as well as anti-Semitic keywords in the "Hate" category. The authors perform similar analysis on posts from the /pol/ board of 4chan, finding that many of the posts associated with genetic testing contain racist language or images.

Rajadesingan et al. [41] examine political content and toxicity on subreddits that are not explicitly about politics. They collect 2.8B comments from all subreddits, and train a human-based classifier to determine whether a given comment is political, as well as to determine which subreddits are political. They find that roughly half of all political comments on Reddit are in non-political subreddits, identifying further areas for research data. They then classify users as either left-leaning or right-leaning depending on their posting profile in well-known political subreddits, and analyze the toxicity of these users comments using Google Perspective. They find that interactions between users of different ideologies are less toxic on non-political subreddits, but that political interactions overall are still more toxic than non-political ones.

1.1.4 Mainstream Social Networks

Twitter

Arviv et al. [8] look at the usage of one specific anti-Semitic meme, the "echo", which involves triple parentheses being used around one word, i.e. (((*echo*))). They find users who use this syntax often, and collect 18M tweets by these users to examine both how they use the echo and how they behave otherwise. They find that users of the echo tend to mention each other often, indicating the existence of close-knit hateful communities. Additionally, the most prevalent users of the echo are largely prominent white supremacists, such as David Duke, a former leader of the KKK. They also find that users who use the echo often use various slurs or other hateful slang.

1.1.5 Cross-Site Analysis

Ribiero et al. [42] examine activity, migration, and toxicity of the loosely spread misogynist community known as the Manosphere. They collect 7M posts from 6 different forums and 22M posts from 51 subreddits. They find clear trends in the migration from older communities that were often related to giving each other tips on how to pick up women, to communities related to men wanting to not interact with women, to a final group of communities related to "incels", or men who consider themselves unable to have sex due to societal forces. They then use the Google Perspective API and a misogyny-based hate dictionary to determine the toxicity of these communities, showing a worrying trend of gradually increasing toxicity over time.

Zannettou et al. [54] examine the flow of news articles between Twitter, the /pol/ board of 4chan, and selected subreddits. They find that although all three sources have large amounts of alternative news articles, Twitter has the largest proportion relative to mainstream news articles. They also model the spread of articles between different aggregators as a Hawkes process, where each event in one location can cause events in others. In doing so, they find that articles are often shared first on Twitter and Reddit, and that Twitter is the largest source of alternative news articles being spread. However, they do find that a significant amount of articles posted on /pol/ and /r/The_Donald result in those articles being spread on Twitter, indicating that these hateful sources have an impact on news on Twitter.

In a similar analysis, Zannettou et al. [53] look at the spread of image-based memes from fringe communities (/pol/, Gab, and /r/The_Donald) to Twitter and more mainstream

subreddits. They start by clustering images found in the above fringe communities and annotating the clusters based on matches to metadata from Know Your Meme, a thorough database of memes. They find that /pol/ contains especially anti-Semitic memes, such as the "Happy Merchant", a Jewish caricature, and memes about Adolf Hitler. They also find that Donald Trump and Pepe the Frog related memes are extremely popular on all three fringe communities. They then use Hawkes processes to estimate the impact of each fringe community, finding that /pol/ creates the most new memes that get spread, while /r/The_Donald creates less but is more efficient at spreading the ones that it does create.

1.1.6 Hate Speech Detection

The field of hate speech detection is quite new, with most literature in the field only dating back to 2016. However, there are already several surveys of existing literature, as well as a number of prominent papers that show common methods of analysis.

Learning Classifiers for Hate Speech

The most common (and best performing) method for automatic detection of hate speech is using modifications of existing machine learning methods to learn classifiers to distinguish between posts that contain hate speech and those that do not. One of the first papers to use this method was by Badjatiya et al. [9] They used an existing dataset of 16K tweets that were annotated as either sexist, racist, or neither. [49] They created multiple feature space representations of these posts, ranging from character n-grams to a simple bag of words approach. They then trained multiple state-of-the-art types of neural networks (LSTM, CNN, etc) to create classifiers to distinguish between racist post, sexist posts, or neither. With these techniques, they were able to create a model with an overall F1 score of 93%. This substantially outperformed earlier methods which used techniques such as SVMs.

Agrawal et al. [5] performed a similar analysis across a broader spectrum of datasets. This included the above mentioned Twitter dataset [49], a dataset of 100K Wikipedia discussion comments that were labeled on whether they contained a personal attack, and a dataset of 12K question and answer pairs from Formspring that were labeled according to whether they contained cyberbullying. They used a variety of both classic machine learning techniques such as SVMs and deep learning models. Similar to [9], they found that deep learning models outperformed other techniques.

Davidson et al. [19] propose a more nuanced form of hate classification, using three categories - a post that contains hate speech, a non-hateful post that contains offensive

language, and a post that contains neither. They argue that other researchers may be incorrectly classifying non-hateful posts as hate speech and thus biasing their classifiers. They create a new dataset using 25K tweets found by searching for terms in Hatebase [25] that they then ask CrowdFlower workers to annotate according to their new specification. They follow this by using logistic regression to train a classifier to distinguish between these three labels, and find that although it correctly identifies the offensive and neither categories with a 91% and 95% recall respectively, it has difficulties distinguishing between offensive and hateful content, identifying only 61% of hateful content correctly and labeling 31% of the remainder as simply offensive. According to the authors, this highlights the challenge in identifying content with hate speech that does not contain offensive words.

A more recent analysis by Mathew et al. [34] both releases a new dataset of annotated social media posts and proposes new methods for improving performance of existing classifiers. They use a lexicon-based approach based on [19] and others to automatically search for tweets and posts on Gab that contain hate speech. They combine these into a single dataset that contains 20K posts, and freshly annotate these as containing either hate speech, offensive language, or neither using Mechanical Turk workers, removing those in which all respondents choose a different class. They also asked the workers to annotate their answers with what specific words helped them reach the label that they came to. They then used various neural network architectures (CNN, BiRNN, and BERT) to train two types of classifiers - one type that only uses the labels of each post, and one that uses an attention-based representation of the words selected by the workers as additional input. They find that although BERT models using their new dataset have high accuracy and F1 scores, they have low explainability. They use this to discuss the different tradeoffs between models, and recommend that other researchers determine whether they want to use either more explainable or more accurate models.

Issues and Surveys

As the field of hate speech detection has matured, there has been increased criticism of existing models, as well as attention directed towards identifying some of the common problems inherent in detecting hate speech.

Arango et al [7] examine the results of previous models [5] [9] and discover issues both in the techniques used and the generalizability of the created models. They start by critically evaluating the datasets most commonly used in the literature. They find that one of the most commonly used datasets, one created by Waseem et al. [49], has one user that generated over 90% of the tweets labeled as racist, and another than generated over 40% of the tweets labeled as sexist. Thus, classifiers trained on this dataset may reduce

the general problem of hate speech detection to simply detecting the speech patterns of a few particular users. They then describe an issue in [9] in which the authors used feature extraction on the entire dataset, both testing and training, to create their features, thus increasing overfitting. When this is fixed, the overall F1 score of the model proposed in the paper drops from 93.1 to 73.1. They find a similar issue in [5], in which the authors replicated posts labeled as hate speech in order to increase the small sample size before splitting into testing and training. As before, when this issue is resolved, the overall F1 score of their model drops from 94.5 to 79.6. Additionally, when the models proposed in these papers are tested on a different dataset than the one trained on, they both drop to an F1 score of roughly 47. Overall, this indicates substantial overfitting issues in the literature, and the authors recommend increased attention towards the methods used.

MacAveney et al. [31] survey the existing hate speech literature and list a number of existing challenges with hate speech detection overall. The first is in creating a single definition of what exactly constitutes hate speech. Various authors come up with different definitions, which then impacts the results of their analysis. A specific example is that of Davidson et al. [19], which as discussed above, believe that speech with offensive language can be distinct from hate speech. MacAveney et al. expand this by mentioning issues with effectively defining and identifying subtle hate speech, as well as types of speech that may rely on outside knowledge to classify, such as praising certain groups that are known to be hateful. The authors also discuss the large variety in existing datasets, which range in the labels used to classify posts, the source of the data, the size of the data, and the relative amount of hate in each dataset. They then propose their own classifier, which trains an SVM on each feature of a post, and then combines them all into one SVM. They find that this approach works well on some datasets, but not on others. They conclude with additional examples of edge cases in hate speech classification, such as quotes of old hate speech, slight variations in language that result in different classification, and the continual effort by bad actors to change their methods to evade existing hate speech detection.

Poletto et al. [40] systematically review over 60 works related to hate speech detection. As with MacAveney et al, they find that the literature varies substantially on the specific focus and labels used, ranging from overall hate speech to cyberbullying to only focusing on specific forms of hate such as racism. However, they argue that many authors do not properly define what they mean by hate speech, which makes it unclear exactly what they are trying to detect. They also discuss the over-representation of data from Twitter, and the possible consequences of relying too much on only one source of comments. Next, they address the wide variety of both annotators used and labels used. Many papers use crowd-sourced annotators to label their posts as hateful or not, with an unclear level of expertise. Others claim that they rely on expert judges, without necessarily clarifying what

that means. Still more are unclear on how they annotate their dataset. Further, many that do discuss their annotation methods are unclear on how they deal with agreement of multiple annotators. Overall, the authors see a clear tradeoff between large volumes of simple annotations and smaller datasets with more expert annotation. They conclude with a desire for more information on how hate speech is defined and annotated in future papers, as well as worries that existing techniques will create biased models or datasets.

Most recently, Yin and Zubiaga [51] review existing hate speech literature in the context of creating hate speech detection models that will generalize to multiple data sources. They find that in the field overall, the performance of various models has been over-estimated, and that most or even all show a substantial performance drop when applied to unseen datasets. This even extends to BERT, the most recent advance in NLP neural network architecture, although some studies have shown that BERT-based models do generalize better than previous ones. However, they do mention that there is a limited amount of studies that do a systematic cross-analysis of a number of popular models and datasets, and so these results may be overstated. The authors then list various issues that they believe contribute to issues with generalization. The first is the use of non-standard grammar and spelling, which depends both on the overall lack of consistent grammar on social media and the use of different terms by bad actors to evade detection. They discuss a few solutions to this problem - creating dictionaries of code words used to evade detection, using character-level features to minimize the impact of simple misspellings, and models of language that include additional context. The second is issues with variance in the datasets, labels used, and the types of annotators used, as discussed in the previous surveys. They also mention existing biases in datasets, such as the under-representation of African-American slang in datasets causing such slang to be falsely identified as hate speech. Finally, the authors discuss implicit hate speech, or hate speech that does not contain distinctive keywords, and as such is very hard to either detect or classify.

1.2 Disclaimers on Hate Speech

We choose in this paper to focus on identifying instances of hateful words and slurs, as determined by Hatebase (discussed in more depth later). This approach was decided on instead of more sophisticated methods for a number of reasons. First, as discussed above, recent research has shown that many existing methods have issues with generalization to unknown datasets. Since many of them were trained on tweets, this could end up giving more correct results for only part of our analysis, biasing our results. Second, we are concerned more with providing a high-level overview of this new dataset, rather

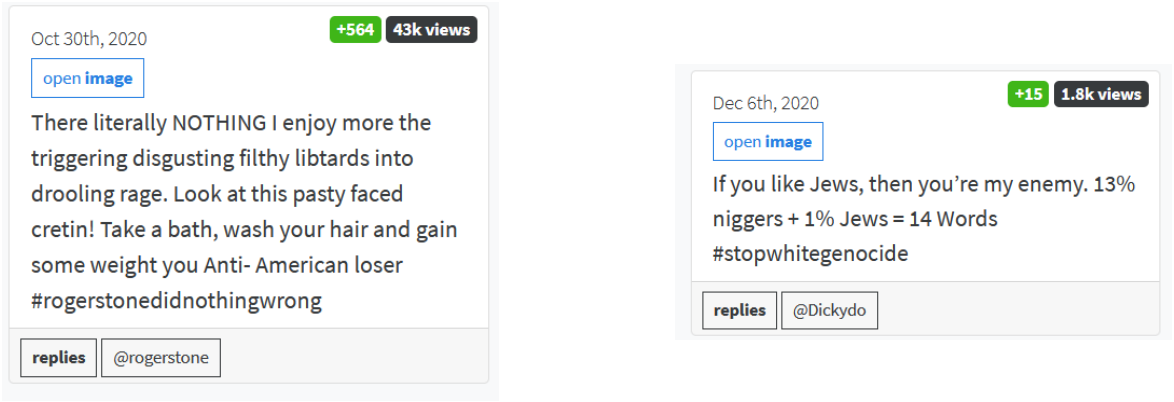

Figure 1.1: Examples of hateful parleys that are correctly included in our results, as they contain hate words. Thanks to the online Parler Archive [46] for visualization of posts.

than a more in-depth look into hate speech specifically. Third, many existing methods are relatively uninterpretable, while a simpler method like incidence of hate speech gives clearer results, and does not have issues with possibly over-representing the validity of the results. However, we note that that this chosen identification process is very subjective and context-dependent. Hateful content can take a number of different forms, from calls for violence to racial slurs to more individualized hate. Furthermore, some benign users may choose to use similar language as hateful users in either a joking manner or in order to reduce the power of some types of language, as discussed in [19]. Our approach correctly identifies a number of hateful parleys, such as those in Figure 1.1. However, because of the aforementioned issues, this will necessarily not account for all of the content many would describe as hateful. For example, violent posts such as those in Figure 1.2 do not contain hate words. It will also include some content intended in a non-hateful way that includes the types of words that we are searching for. For example, some words that are often used hatefully can also be used a non-hateful way, as in Figure 1.3. We provide results that account for some of this last issue by providing separate analysis for words that have no other possible use than as a slur. However, this still does not account for uses of hateful words in a sarcastic way intended to defang or reclaim the word (as in Figure 1.4). Given the difficulty of identifying the context in which these words are used, we have no way of knowing the extent to which we find false positives in our results. However, we are also excluding some varieties of hateful posts. Given this, we advise readers to examine our results only in the specific context of our analysis, i.e. amount of posts that contain potentially hateful words. We do note that since we use the same methodology when comparing both content from Parler and content from Twitter, the results should be

Figure 1.2: Examples of hateful parleys that do not contain instances of hate speech, and thus are not included in our results. Thanks to the online Parler Archive [46] for visualization of posts.

comparable to each other. However, readers should be cautioned away from completely generalizing these specific results to the general character of each platform, and further analysis should be performed in order to get a more full picture of the types of posts that are common on each site.

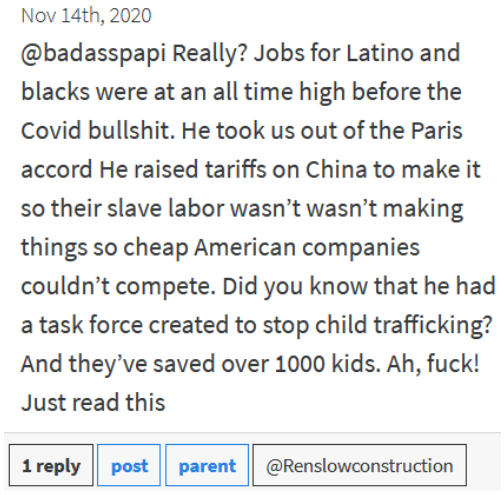

Figure 1.3: An example of a parley that uses a hateful word (slave) in its more ambiguous sense.

Figure 1.4: An example of a parley that uses a hateful word (libtard) in a sarcastic sense.

Chapter 2

Data and Methods

Parler is an alternative social network started in August 2018. It advertises itself as "the world's premier free speech platform" [39], with an emphasis on light moderation and lack of ideological bias. It rose to prominence in 2020, as more popular social networks such as Twitter and Facebook started removing posts that promoted or spread "fake news". This largely impacted ideologically conservative users, as there was a substantial population of them on those platforms. Many popular figures, such as US Senator Ted Cruz and then-President Donald Trump, advocated leaving Twitter and turning to Parler instead, creating a substantial influx of new users. [30]

Parler is very similar to Twitter, with users posting 1000-character limit messages that can then be seen by their followers on an updating newsfeed. Users can upvote posts that they agree with, which affects the order that posts appear on the feed. They can also "echo" (unrelated to the anti-Semitic meme mentioned above) them, which puts them onto their own feed like retweets in Twitter. They can also comment on posts, and can carry on new conversations in these comments. These comments can be both upvoted and downvoted, displaying a total score. However, they cannot be echoed onto one's feed.

Although Parler emphasizes free speech, it does have some moderation policies, and base guidelines for what is considered an acceptable post. It largely emphasizes individual moderation, with a large variety of options for filtering replies to their content that each user can set, such as only allowing verified users to comment on their posts. Users can also individually moderate comments on their posts, and block specific users from commenting on their posts. For site-wide moderation, it largely relies on manual review of reported posts by a team of volunteer moderators, rather than any sort of automated filtering. At points, this led to a large influx of spam and adult content on the site [18], which caused

worries that illegal content such as child pornography could be spread on the site without any sort of automatic detection.

As noted by Aliapoulios *et al.* [6], although the user base of Parler is primarily American, there is a substantial population of both Saudi Arabian and Brazilian users. I focus exclusively on the English words used in Parler, and filter out these non-English posts when necessary.

2.1 Parler Data

The data we use is sourced from Aliapoulios *et al.* [6], who released an open-source dataset of Parler parleys, comments, and user data. They sampled roughly 4M users at random, and then downloaded their posts, consisting of approximately 98.5M parleys and 84.5M comments. Here, a parley is a publicly available message that a user posts to their feed, while a comment is a reply to a specific post. The data is split into 167 separate files, each file with a JSON representation of a parley or comment on each line. In this paper we separate our analysis of parleys and comments. This allows us to compare the results in like kind when analyzing based on popularity, as comments are unable to be echoed and parleys do not have a score.

We will now explain the structure of the data. Figure 2.1 shows an example of the JSON representation of a post, with any specific URLs removed. The fields shown are limited to those that we consider relevant, with the removed fields largely pertaining to internal state, automatically generated links to the post, the depth of the post (if it is a comment), and other irrelevant information. We list the relevant fields and what they represent below.

- comments: The number of comments on the post.
- bodywithurls: The full text of the post, along with associated URLs that the user included in their post.
- body: The full text of the post with URLs excluded. This field is what we use for our analysis in order to limit the analysis to just the words in the post, rather than generated URL strings or names of websites.
- createdAt: The full date and time of the post, with any punctuation or spaces omitted.

```
{
    "comments":-1,
    "bodywithurls":
        "[VIDEO] PA Election Whistleblower Says
        Anti-Trump AG Sent Two Special Agents to Her House
        Unannounced After Her Testimony - *name omitted*.com
        https://www.*name omitted*.com/2020/11/
        election-whistle-blower-pa-ag/",
    "body":
        "[VIDEO] PA Election Whistleblower Says Anti-Trump AG
        Sent Two Special Agents to Her House Unannounced After Her
        Testimony - *name omitted*.com",
    "createdAt":"20201128181326",
    "createdAtformatted":"2020-11-28 18:13:26 UTC",
    "creator":"c34fca9c55ba437ba36d3eebf74d01ee",
    "datatype":"posts",
    ...
    "hashtags":[],
    ...
    "impressions":3300,
    ...
    "reposts":10,
    "upvotes":20,
    "downvotes":7,
    "score":13
}
```


Figure 2.1: An example of a JSON formatted Parler post. URLS, identifying information, and irrelevant fields, such as internal state, are omitted.

- createdAtformatted: As above, but formatted into a human-readable string.
- datatype: The type of the post, i.e. whether it is a post or a comment.
- hashtags: A list of hashtags extracted from the post.
- impressions: The total number of users who have seen the post. This value is rounded in the following way: above 100, it is rounded to the nearest 10s place, above 1000, it is rounded to the nearest 100s place, and so on. As an example, if the true value of impressions was 1368, it would be rounded to 1400. If it was 14489, it would be rounded to 14000.
- reposts: The total number of echoes the post has, i.e., the number of times the post has been reposted onto another user's feed. This field displays the same rounding as impressions. Note that only posts can be reposts, not comments.
- up(down)votes: The total number of times the post has been up(down)voted. Upvotes are present on both posts and comments, but downvotes are only present on comments. Again, this field has the same rounding as above.
- score: The displayed total score of the post, i.e. the sum of the downvotes and upvotes. This field is only present on comments, not posts. This field has the same rounding as discussed above.

2.2 Hatebase Data

Hatebase[25] , a company based in Toronto, maintains a collaborative dictionary of hate speech for use in detecting and preventing hate speech usage online. They also maintain a database of sightings of each term, and where each was used. As of the time of writing, their dictionary contains about 3800 terms in 98 languages. The dictionary also contains metadata about each instance of hate speech for better use in classification. We limit ourselves only to terms in English, and prune the dataset to remove terms that are too ambiguous, as discussed in section 2.4.

We will now explain the metadata associated with each piece of hate speech. Figure 2.2 shows an example of the JSON representation of an entry in the Hatebase dictionary. We omit metadata associated with the reported sightings of each term, which we do not concern ourselves with.

```
{
    "term": "fruit",
    "hateful_meaning": "A homosexual.",
    "nonhateful_meaning": "A plant",
    "is_unambiguous": false,
    "average_offensiveness": 50,
    "language": "eng",
    "is_about_nationality": false,
    "is_about_ethnicity": false,
    "is_about_religion": false,
    "is_about_gender": false,
    "is_about_sexual_orientation": true,
    "is_about_disability": false,
    "is_about_class": false
},
```


Figure 2.2: An example of a Hatebase entry.

- term: The hate term that the dictionary entry refers to.
- hateful_meaning: The definition of the hate term, as well as often its origin.
- nonhateful_meaning: The possible definitions of the term if it is not used as a hate term.
- is_unambiguous: A boolean variable that represents whether the term is unambiguously hateful, i.e., whether the only possible meaning it can have is hateful. In our analysis, we use the prevalence of unambiguous hate words as an overall floor on how hateful posts are.
- average_offensiveness: A score from 1-100 that represents how offensive the term is on average. This is crowdsourced from contributors to Hatebase, where each contributor can vote on how hateful they think the term is.
- language: The language that the term is in. We restrict ourselves to only terms that are in English.
- is_about_: These seven fields represent the particular way in which a term is hateful, divided into seven classes. These classes are hateful against nationality, ethnicity,

religion, gender, sexual orientation, disability, and class. Each term can belong to multiple classes.

2.3 Twitter Data

The data we use is sourced from the Internet Archive [1], who release a collection of tweets obtained via the Twitter Sample Stream API [48]. According to Twitter, this gives a real-time stream of roughly 1 percent of tweets. This archive is split by month, and the Internet Archive acknowledges that there may be small portions of it missing, such as when the Twitter API or the script they were using was unavailable. For the purposes of this study, we use data from October to December 2020, both for the purposes of space (even 3 months of data is roughly 200 gigabytes of raw JSON) and to attempt to characterize the common topics discussed during a politically charged time.

Figure 2.3 shows an example of some fields in the JSON representation of a tweet. We remove any identifying information, and also exclude the actual text of the tweet, as per the terms of the Twitter API. As Twitter maintains large amount of metadata on each tweet, such as location, user metadata, and more, we limit ourselves to showing only a subset of the fields. We list the relevant fields and what they represent below.

- created_at: The timestamp of when the tweet was created.
- text: The full text of the tweet, which includes any URLs posted, as well as a generated link to any media attached to the tweet.
- display_text_range: Indices that demarcate the edges of the actual body of the tweet. When used to index into the text, this removes any users that the creator of the tweet has mentioned, as well as removes any URLS.
- in_reply_to_status_id: This gives the ID for the tweet that this tweet is replying to, if any. We use this to determine if a given tweet is an original post or a reply to another tweet.
- reply_count: The total number of replies to the tweet.
- retweet_count: The total number of times the tweet has been retweeted.
- favorite_count: The total number of times the tweet has been liked (or favorited, the previous term Twitter used for this function)

```
{
    "created_at": "Sat Oct 10 16:10:00 +0000 2020",
    "text": "RT @MikeHudema: Wow. This is what #climate leadership
    looks like. #Denmark's new gov't unveils one of the world's
    most ambitious green plan...",
    "display_text_range": [0, 140],
    "in_reply_to_status_id": null,
    "user": {},
    "reply_count": 0,
    "retweet_count": 0,
    "favorite_count": 0,
    "entities": {},
    "retweet_status": {},
    "lang": "en"
}
```


Figure 2.3: An example of a JSON formatted tweet. URLS, identifying information, and irrelevant fields, such as internal state, are omitted.

- entities: A dictionary that contains any hashtags, user mentions, and URLS, as well as where in the text of the tweet they occur.
- retweet_status: Only appears if the tweet is a retweet. A dictionary that contains the JSON representation of the original tweet.
- lang: The BCP-47 code representing the language that Twitter has classified the tweet to be in. This can also take the value 'und', representing tweets where Twitter is unable to classify the language. This is often the case in tweets with no non-URL text or very short text.

2.4 Methods

We start by processing the Parler data into a more efficient form. We parse the original JSON files into python dictionaries, extract the body of the post and other relevant metadata, and write it to disk as a binary representation. When doing this, we also remove posts that have an empty body or just have a URL, as well as perform some preprocessing

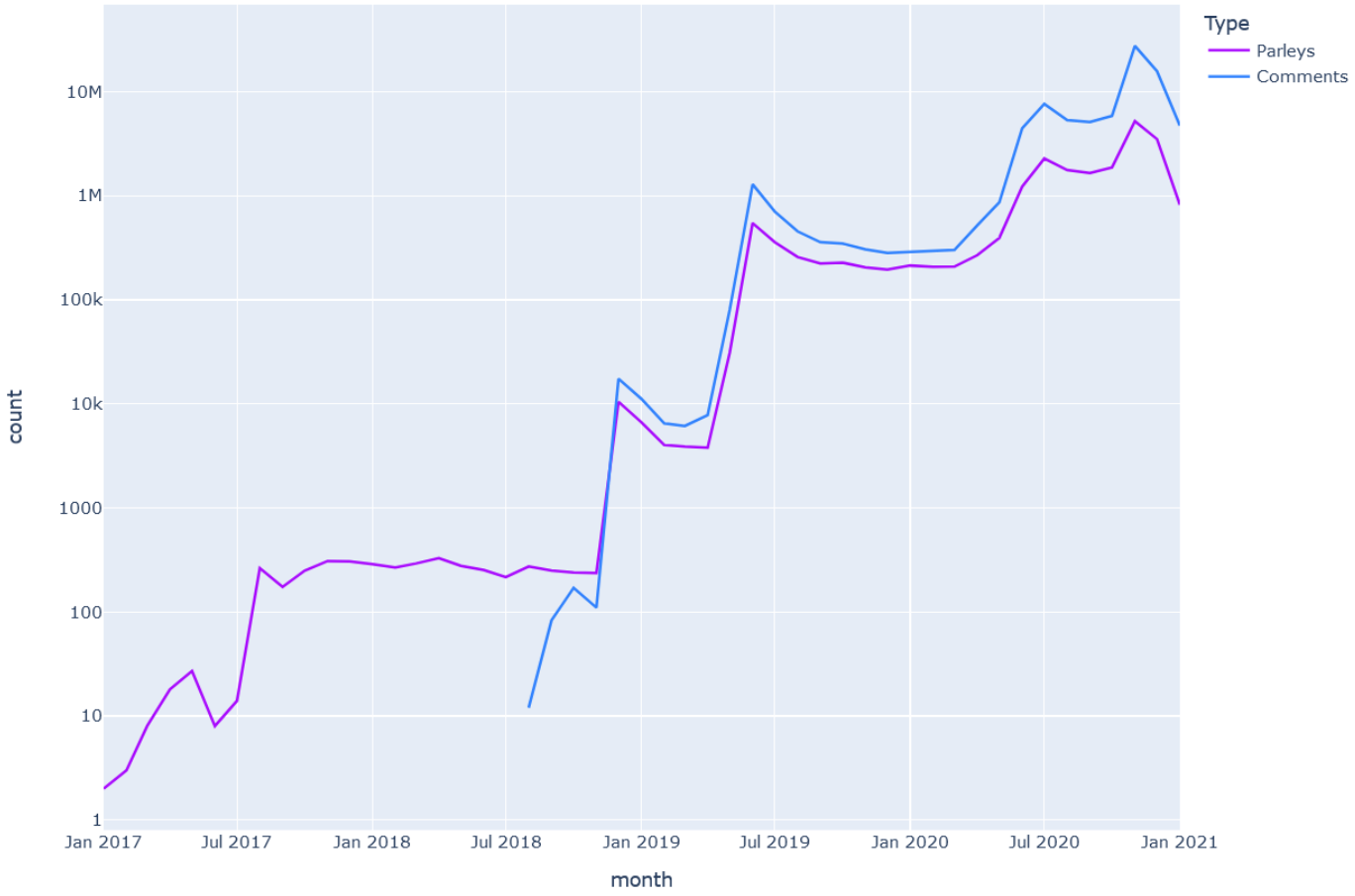

Figure 2.4: The number of posts and comments per month on Parler over time. Note the log scale on the y-axis.

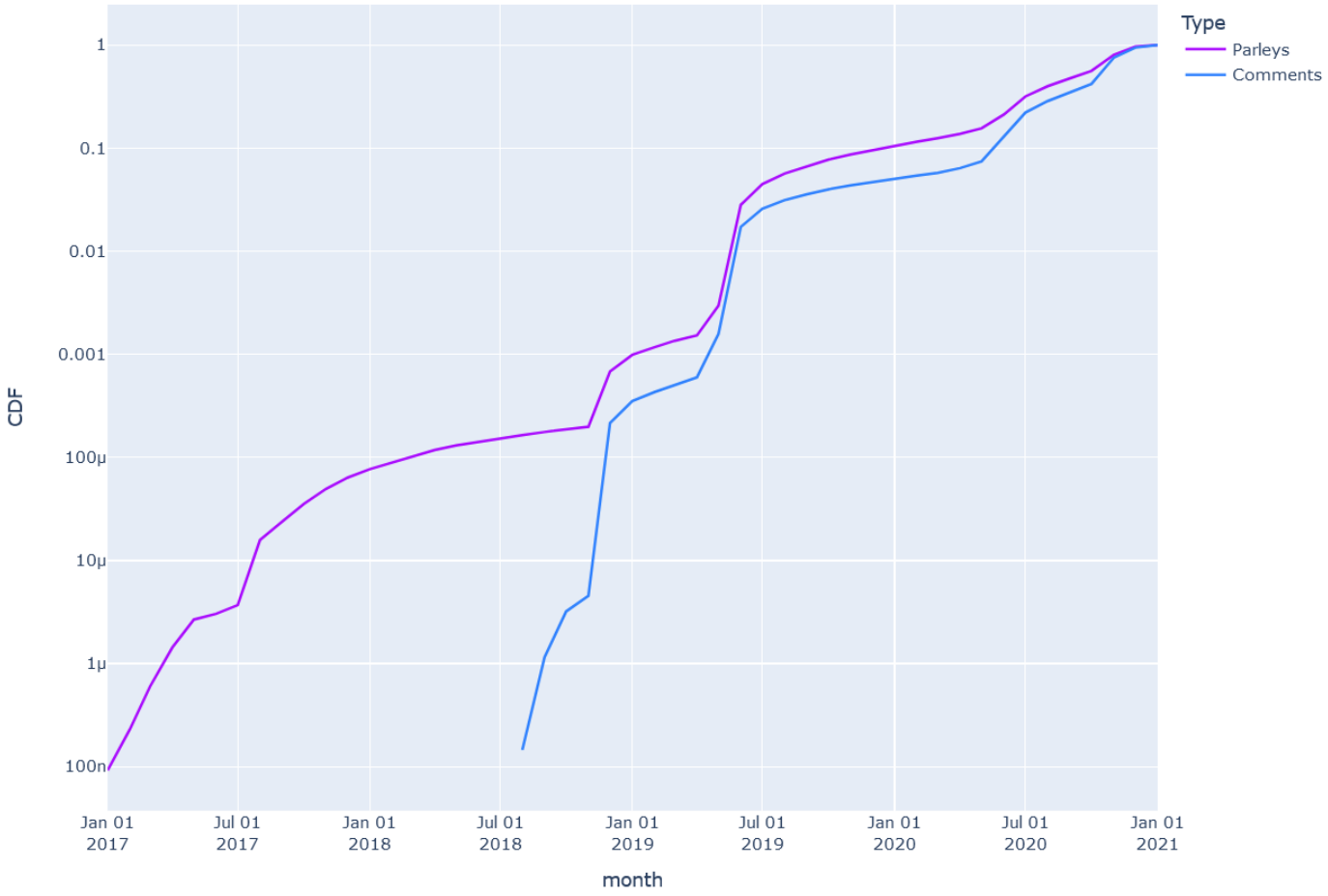

Figure 2.5: The CDF of posts and comments on Parler over time. Note the log scale on the y-axis.

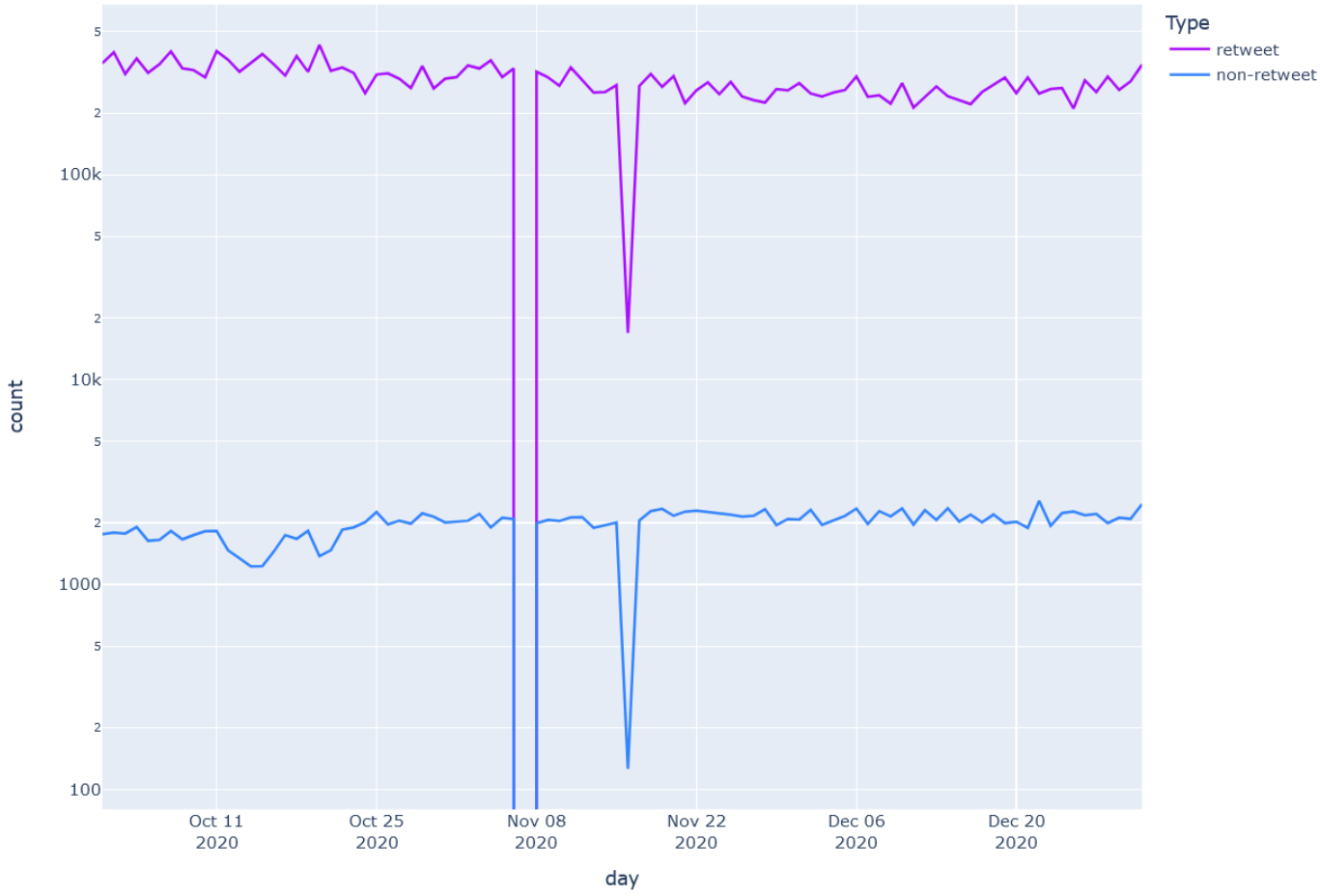

Figure 2.6: The number of original tweets and retweets over time. Note the log scale on the y-axis.

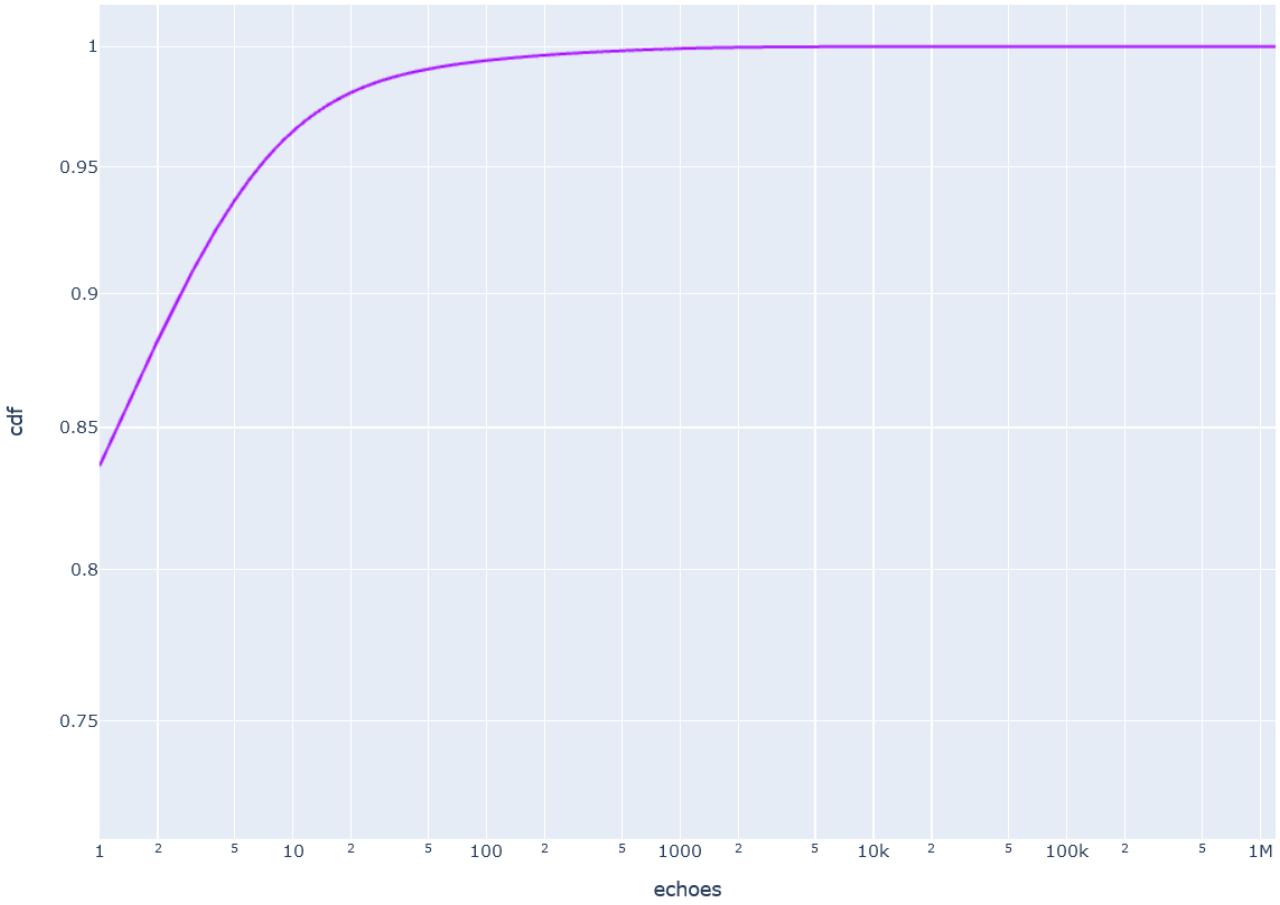

Figure 2.7: The CDF of the distribution of reposts on Parler posts in our dataset. Note the log scale on the x-axis.

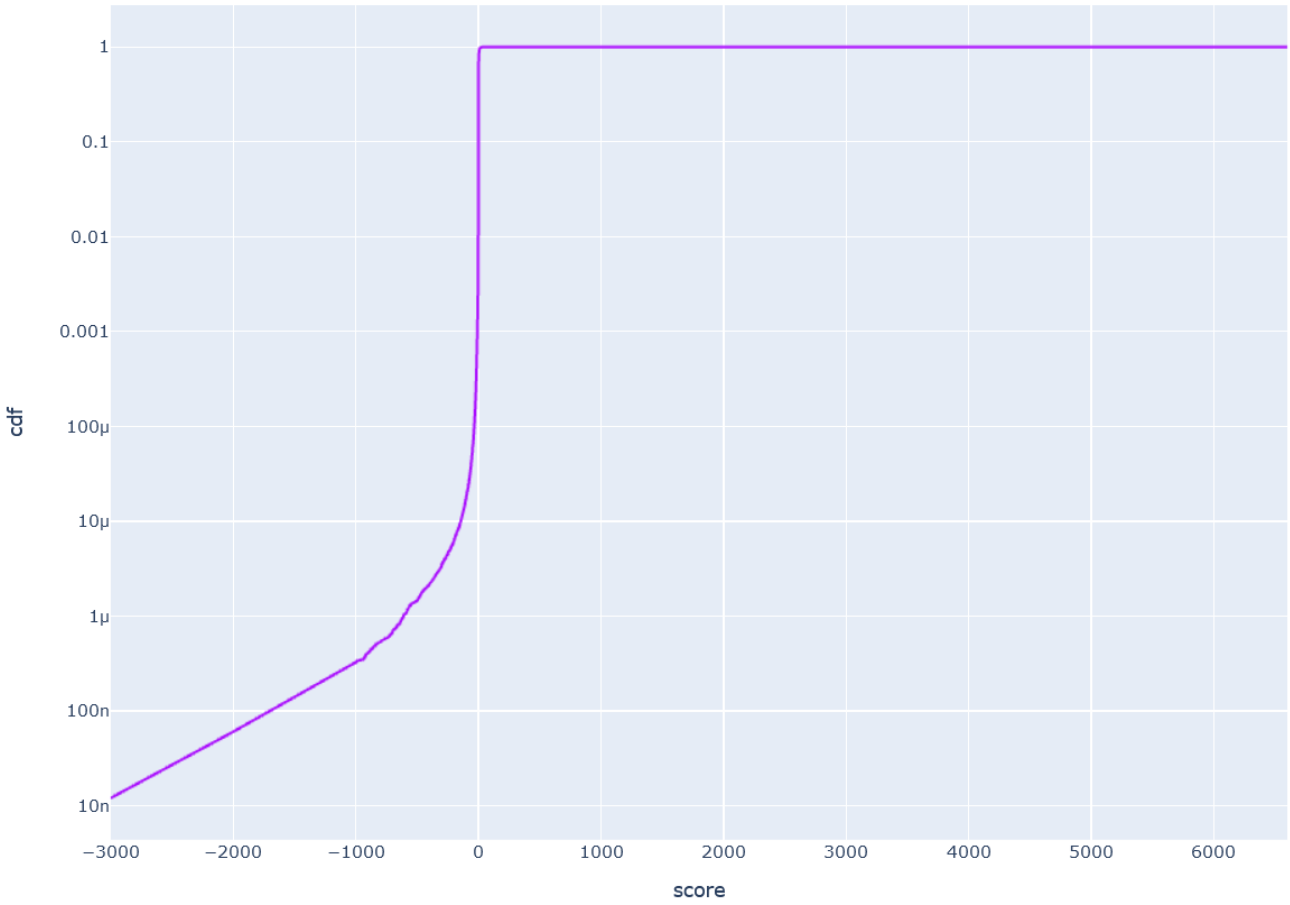

Figure 2.8: The CDF of the distribution of the score (sum of upvotes and downvotes) on Parler replies in our dataset. Note the log scale on the x-axis and y-axis.

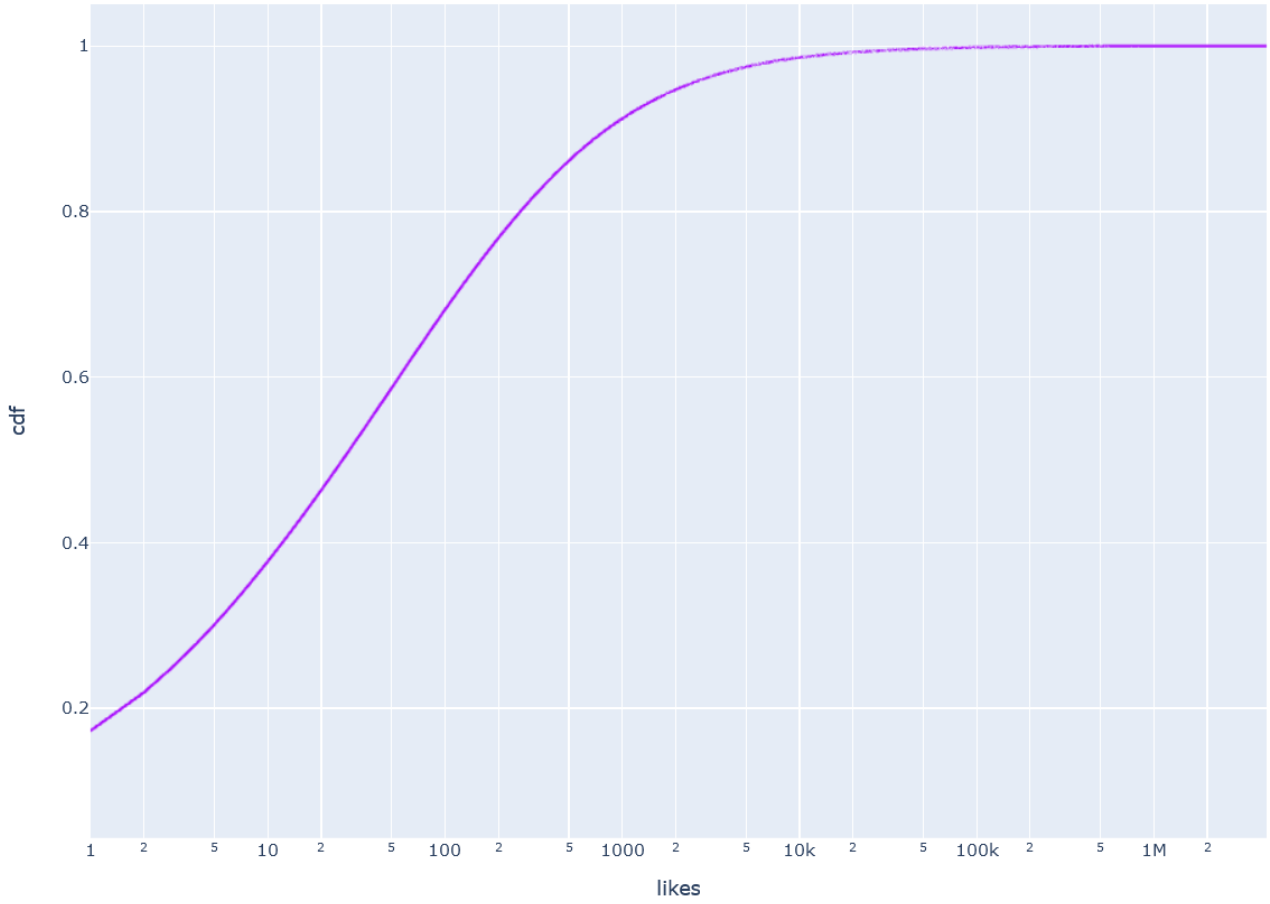

Figure 2.9: The CDF of the distribution of likes on tweets in our dataset. Note the log scale on the x-axis.

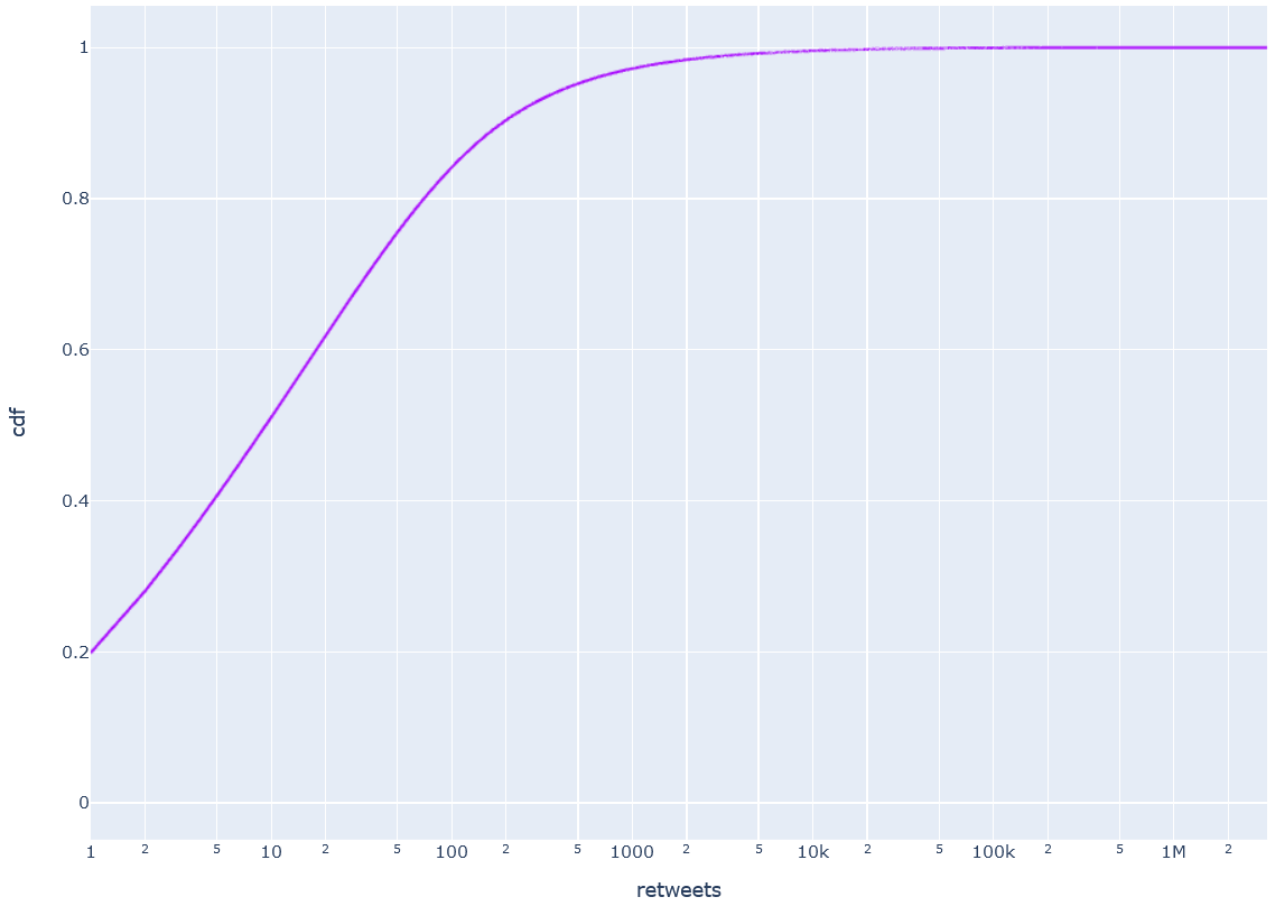

Figure 2.10: The CDF of the distribution of retweets on tweets in our dataset. Note the log scale on the x-axis.

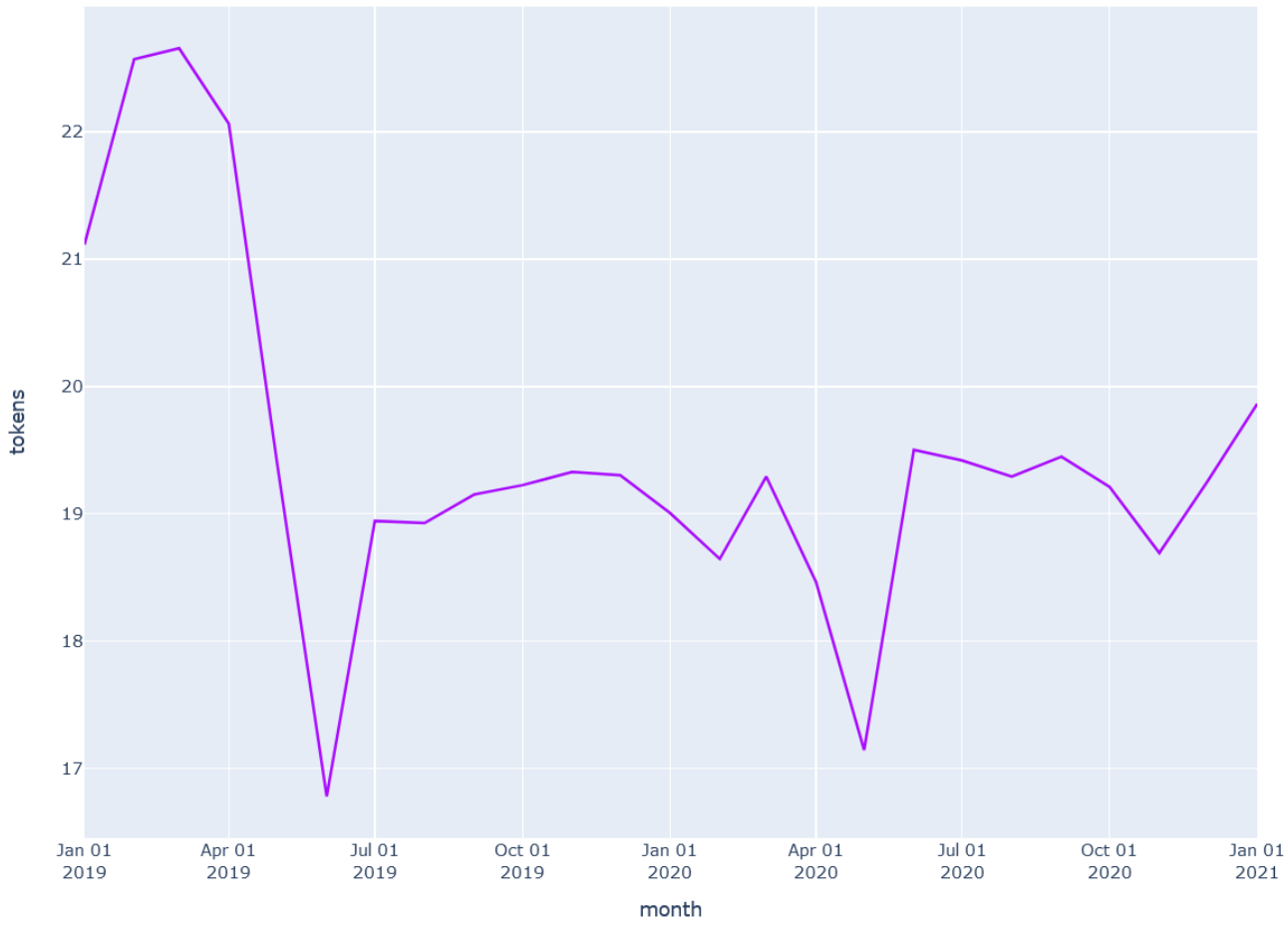

Figure 2.11: The average number of tokens per post over time.

on the text. We remove the phrases "I just joined Parler! Looking forward to meeting everyone here." and "Explore the Fox News apps that are right for you at...", which can be automatically posted when a user joins Parler for the first time or posts a Fox News link respectively. These phrases were present in posts enough that they substantially affect analysis, especially biasing the bigram analysis. We note that even after this processing, there is still some similar phrasing (i.e. "Explore these Fox News apps") that ends up having a statistical impact, but that is harder to automatically remove. We also find a large number of automatically posted comments on posts which are identical in text, which we also remove from the dataset. We discuss these further in B. We then use NLTK [14] to tokenize the dataset into words, and then remove English stopwords and tokens of length 1 and 2. After tokenization and removal of stopwords, on average each post contains roughly 8.67 tokens, which stays constant over time, as seen in Fig 2.11. Next, we use WordNet [20] to lemmatize the words into their base form to avoid issues with pluralization. We also split the data into parleys (original messages written by a user) and comments (replies to a post). After this preprocessing, we end up with roughly 22M parleys and 78M comments. This preprocessing reduces the total space taken up by parleys from 147GB to 16GB. We note that according to Aliapoulios et al. [6], the original dataset contains roughly 98.5M parleys and 84.5M comments. This means that the majority (close to 80%) of parleys contain either only URLS or automatic phrases that we have removed, while only roughly 8% of replies get removed this way.

For analysis purposes, we also further subdivide the data by month and by the popularity of the post as determined by the number of echoes. We choose months as the split for time-based analysis in order to see noticeable changes in responses to events while also having enough data in each division for significant analysis. We also note that we discard the data from 2017-2018 for multiple reasons. First, the small number of posts (see 2.4 and 2.5) makes them unsuitable to draw conclusions from - specifically, posts from before January 2019 only account for 0.1% of the data. Second, the comment data only starts in mid-2018, with similar data amount issues. For popularity, we note that of the roughly 22M total posts, around 17M of them have 0 echoes. The distribution of the rest of the posts is shown in 2.7, which is log plotted for visibility. The vast majority of posts have a small number of echoes, with roughly 100K having more than 100 echoes. We split these posts into three categories - posts with zero echoes, posts with 1-249 echoes, and posts with 250 or more echoes. This split is chosen to give enough data in the popular posts for significant analysis, while still focusing on a high enough popularity to be a significant difference. As discussed above, posts can be echoed, while comments cannot. In order to determine the popularity of comments, we instead look at the total score of the post. This lends itself to a natural split into three categories - those with a negative score, those

with a score of zero, and those with a positive score. We note that the majority of comments have a score greater than or equal to zero, as seen in Fig 2.8. These subdivisions allow us to determine both the change in posts over time and whether there are noticeable characteristics of posts that become popular.

We perform a similar preprocessing step on the Twitter data that we have obtained. We iterate through the JSON files to extract the relevant metadata and process the data into a more efficient form. While doing this, we also limit ourselves to only tweets in English, which we determine based on the automatic 'lang' label that Twitter provides from its API. We manually strip URLs from the body of each tweet. We also observe that a large majority of the tweets in the dataset are retweets of other users' tweets, as shown in 2.6. We can see that there are consistently roughly 300K tweets and 2K non-retweets on most days of the dataset. We note that on November 7th we have zero English tweets, and that there is a roughly ten-fold drop in the number of English tweets on November 16th. Looking at the full dataset, the drops in number of English tweets on these days are proportional to the number of tweets in all languages, meaning these were likely due to errors in collecting data either from Twitter or from Archive Team. In order to both capture the existence of popular tweets while not heavily duplicating them, we store the IDs of tweets and ensure that we only have one copy of each tweet in the dataset. We also store the original retweet and like count of the tweet for analysis, rather than the values for the retweet. We then tokenize and lemmatize each body using the same processes described for the Parler data.

Initially, we planned to split up tweets into original tweets and replies to other tweets to compare them to Parler posts and comments. However, we found that replies are not represented enough in the dataset for this analysis to make sense. For example, for the month of October 2020, there are roughly 10M tweets in our dataset, but only 18K replies to tweets. Thus, we simply combine tweets and replies in our analysis. When analyzing the statistics of the dataset, we note that the distribution of retweets and likes of tweets has a substantially different shape than that of Parler posts and comments, as seen in 2.9 and 2.10. Only roughly 20% of tweets have zero retweets, compared to 80% having zero echoes in Parler posts. This is likely due to the method by which this dataset was gathered. Since it only contains 1% of tweets, and treats retweets as tweets, tweets with larger amounts of retweets are more likely to be represented, since they are effectively copied a number of times equal to their retweet count. This is supported by the counts shown in 2.6, as the majority of data in our dataset is from retweets. For analysis purposes, we divide our data into three categories of popularity - those that have less than 10 retweets, those that have between 10 and 2000 retweets, and those that have greater than 2000 retweets. We base these categories on the observed CDF for the distribution of retweets in the dataset (see 2.10), with approximately 50% of the dataset having less than 10 retweets, and with

2000 retweets being close to the 99th percentile mark. This gives similar cutoff points for those chosen for the Parler dataset. We also choose days for the subdivision of time-based analysis, as we see larger individual trends by day in the Twitter dataset.

Next, we query the Hatebase API for all hate words that they have in English. This gives us 1556 words and phrases as a starting point, 362 of which are considered unambiguous. Most of these are related to ethnicity or nationality - 1084 are related to ethnicity, and 354 are related to nationality (only 149 of those are not also considered to be related to ethnicity). When examining the phrases provided by Hatebase, we noted that some of them seemed to be outdated or unlikely to be used in a hateful way. In order to cut down on false positives on hate classification, we manually inspected each Hatebase entry and selected 116 that were likely to not be used in a hateful way in our context. In order to verify that these words were indeed not being used hatefully, we then selected 20 random posts from the dataset that contained each word or phrase, and inspected them to determine if the words were being used in a hateful context. After this inspection, we reduced the number of false positive Hatebase entries down to 109. Examples of these words are "ninja", "ABCs", "apples", and "Charlies". We then remove these Hatebase entries from the set that we use for analysis, bringing us down to 1448 hateful phrases. In our initial look through the Parler dataset, we also noticed some variants of hateful words that are not in the Hatebase dictionary, mostly simple combinations of a normal word and a hateful word. This is likely because although clearly hateful, these variants are not used often enough across platforms to be picked up and officially categorized. In order to capture these in our analysis, we looked also for any word that contains "tard", "fag", "nigger", and "faggot" - the four base words that were often used to create these variants.

After processing the Parler data and Twitter data, we perform frequency-based analysis to determine the most common words in each dataset. We also find the most common bigrams using likelihood ratios.[32] This ensures that we find 2 words that occur together more than they appear with other words, indicating that they are part of an actual phrase, rather than just common sentence structure. Next, we look at the total number of posts that contain a hate word from our modified Hatebase dictionary, as well as the total number of times each word is used. We use Hatebase's classification system to determine the types of hate words that are being used most often. We also perform these types of analysis on our score-divided datasets and month-divided datasets to observe trends over time and differences in what kinds of posts become popular.

Finally, we perform Latent Dirichlet Analysis (LDA) [15] on both datasets to discover common topics of discussion among posts. The central assumption in LDA is that we can treat documents in a corpus as a random mixture of some number of latent topics. A topic here is defined as a probability distribution over some words, with the goal being

to interpret the meaning of each topic by its common words. This approach does rely on a bag of words model, meaning that each document is treated as a collection of words, without the context of nearby words or structure. In order to add some context back into this model, we add to each document common bigrams joined by a separator, i.e. "Donald Trump" would become "donald_trump". This lets the discovered topics capture more meaning, as they have additional context. The final challenge for LDA involves tuning the hyperparameters of the model in order to find the most interpretable output. The most important of these is the number of latent topics assumed to exist in the corpus. For each of the three datasets, we tested using topic numbers of 8-20, and selected the number of topics with the highest average coherence score for each dataset. A topic coherence score is the output of different automatic methods that attempt to determine how understandable a given topic is to humans. Each topic is represented by its top words, i.e. some number of the most probable words to appear in that topic. Roder et al. [43]) devised a framework to represent the space of different coherence measures as compositions of various parts. They find the best performing measure is C_v, which looks at all words in the corpus to determine whether words in a topic are correlated not just with each other but with similar groups of words. It additionally calculates these correlations using a sliding window over each document to create a number of virtual documents, which captures context within the document. We tested a number of these different topic coherence measures and ended up settling on C_v, as it gave the most interpretable topics.

We find that the length of Parler posts (limited to 1000 or less characters) lets us treat each post as a document for LDA, but that tweets are on average too short. To solve this, we adopt hashtag pooling, as discussed by Mehrotra et al. in [35]. We find a list of hashtags used in the tweets in our dataset, and then treat each tweet with the same hashtag as being part of the same "document". We note that tweets with multiple hashtags will be treated as part of the document for each of its hashtags.

Chapter 3

Results and Analysis

3.1 Unigrams and Bigrams

We start the analysis by finding the most popular words and bigrams from among all of the data. We find both the total count of words and bigrams as well as the percent of all posts that each is used in. Popular words are found simply by using those that appear most often, excluding words that are overall commonly used in English sentences. Popular bigrams are calculated using maximum likelihood, thus finding bigrams that are likely to appear with each other more than with other words.

The left side of Figure 3.1 shows that most of the most common topics discussed on Parler were political, with the most commonly used word being "trump", with "biden" and "president" close behind it. This corresponds with Parler as a haven for conservatives (especially Donald Trump supporters) that were expelled from Twitter. Other popular words such as "vote", "country", "election", "democrat" and "america" are similarly political, and the large amount of election-related words makes sense with the influx of users to Parler around the time of the 2020 US Presidential election. Most of the other words do not seem to point to anything in particular.

Figure 3.1 also shows common bigrams that are even more political. The phrases "donald trump", "president trump", and "trump supporter" make the top list, showing the high level of support for the former president shared on the app. "joe biden", "hunter biden", and "deep state" show discussion of perceived enemies to Trump, as well as an interest in perceived conspiracies. "god bless" indicates a community interested in expressing Christan beliefs, possibly as a backlash to the perceived secularization of the US. The fact that

Word	Count
trump	2314015
people	1354675
biden	1118345
president	1097465
get	1003254
one	917860
like	905868
election	891189
democrat	874196
time	869495
need	860633
news	794267
know	754705
god	738250
state	737905
america	734219
right	731501
vote	730905
american	696515
say	660695

Phrase	Count
president trump	435871
joe biden	242283
fox news	214095
god bless	164818
donald trump	153887
united state	113952
supreme court	104894
voter fraud	103790
look like	92942
news apps	89676
apps right	89588
explore fox	89581
trump supporter	88430
new york	87589
white house	84727
life matter	83437
deep state	83335
american people	83094
hunter biden	78894
echo echo	78315

Figure 3.1: The top 20 most used words and phrases in Parler posts.

Word	Count
trump	5332822
people	5151511
like	4383366
get	4175672
one	3555215
need	3534168
know	3263365
would	3038899
president	2862147
time	2861148
right	2534125
think	2436770
biden	2426720
good	2340032
see	2265687
want	2262996
election	2175518
god	2115278
vote	2111054
democrat	2066290

Phrase	Count
president trump	910820
god bless	584104
look like	442256
joe biden	311501
trump supporter	273210
american people	270125
deep state	262654
united state	251075
donald trump	250700
supreme court	228803
social medium	222744
voter fraud	222476
fake news	219377
sound like	205314
trump 2020	193050
fox news	182039
year ago	171285
free speech	167244
welcome parler	164388
election fraud	162195

Figure 3.2: The top 20 most used words and phrases in Parler comments. Since we stem words, we get some odd-sounding phrases, i.e. "social medium" instead of "social media".

Word	Count
like	1314328
amp	1099387
one	1021048
people	912366
day	882959
get	874604
time	782748
love	799463
trump	633376
new	760295
know	649335
want	637756
today	633631
year	778622
please	583751
let	587427
good	596778
make	574372
need	576192
see	557373

Phrase	Count
2020mama voted	168582
mamavote 2020	145477
2020 mama	141869
mama 2020	139302
2020 sun	124121
sun mnetmama	112541
good morning	111807
happy birthday	110282
joe biden	83283
look like	77637
year old	59984
feel like	58450
president trump	58062
year ago	56212
donald trump	51641
artist year	51369
make sure	49863
new year	49165
year amas	49113
fan choice	45606

Figure 3.3: The top 20 most used words and phrases in all tweets.

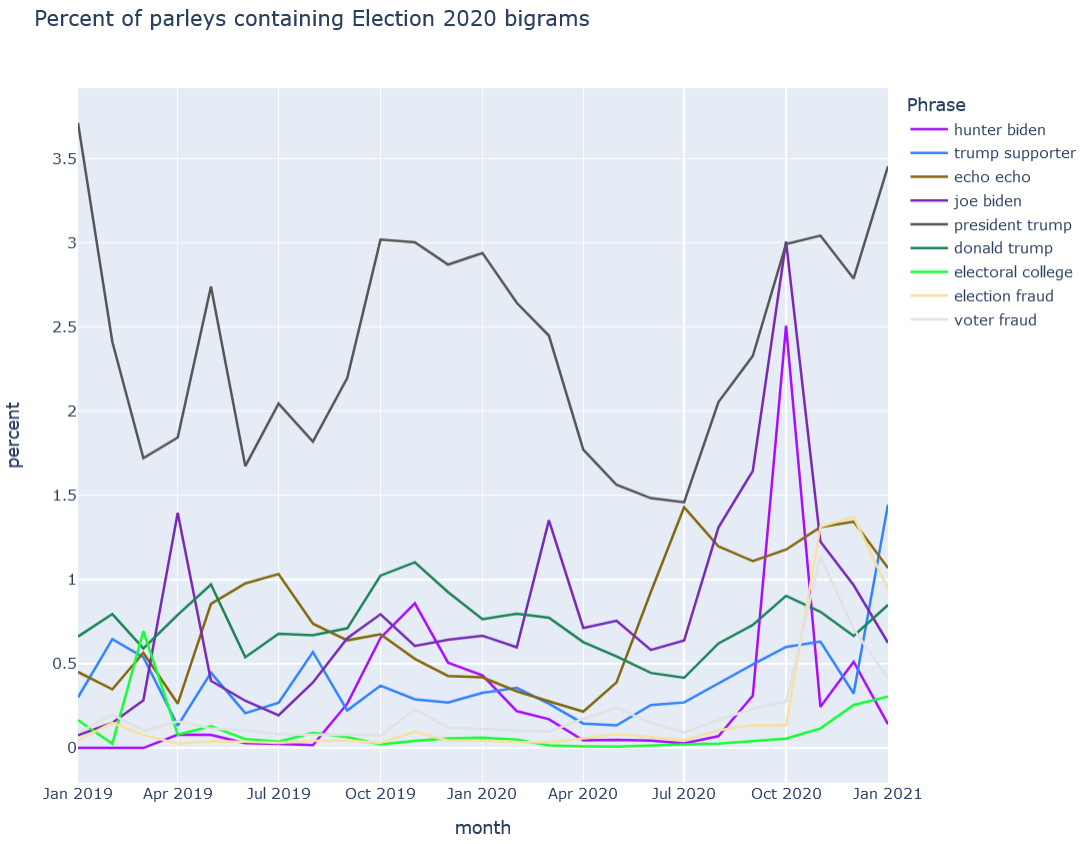

Figure 3.4: Change in the percent of parleys that use election phrases over time.

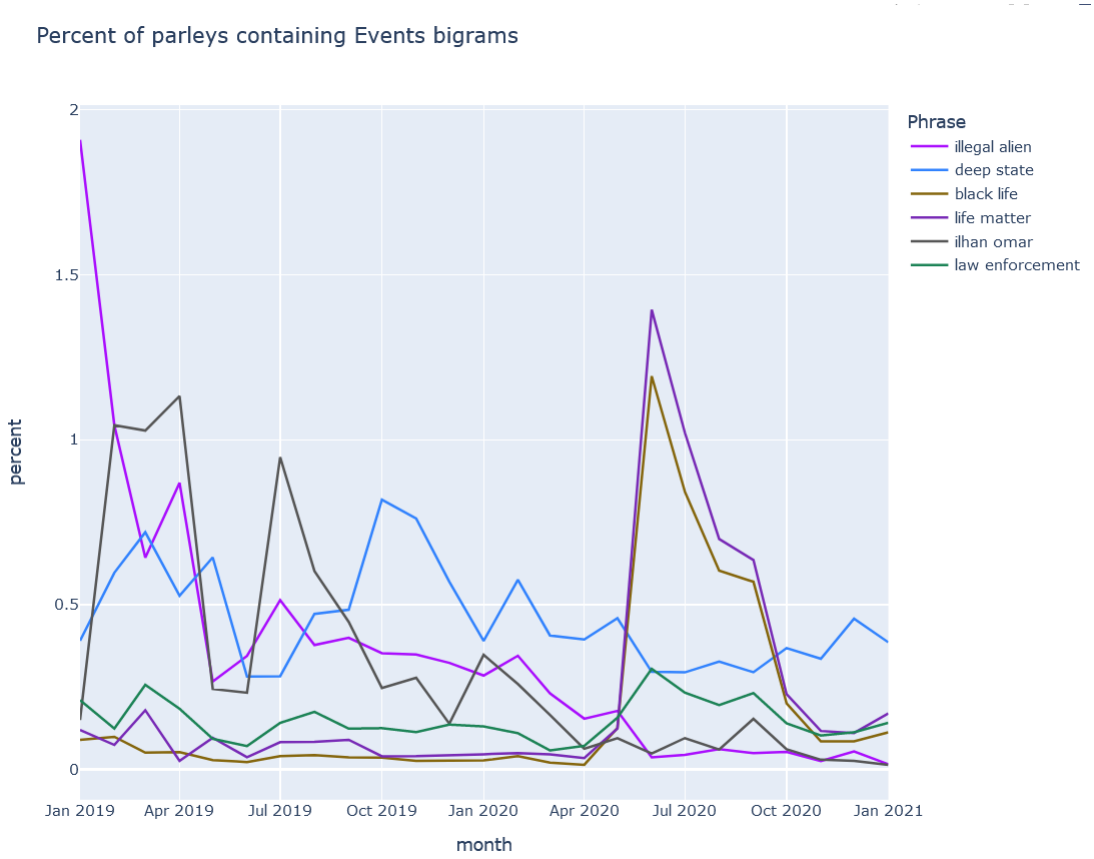

Figure 3.5: Change in the percent of parleys that use event-related phrases over time.

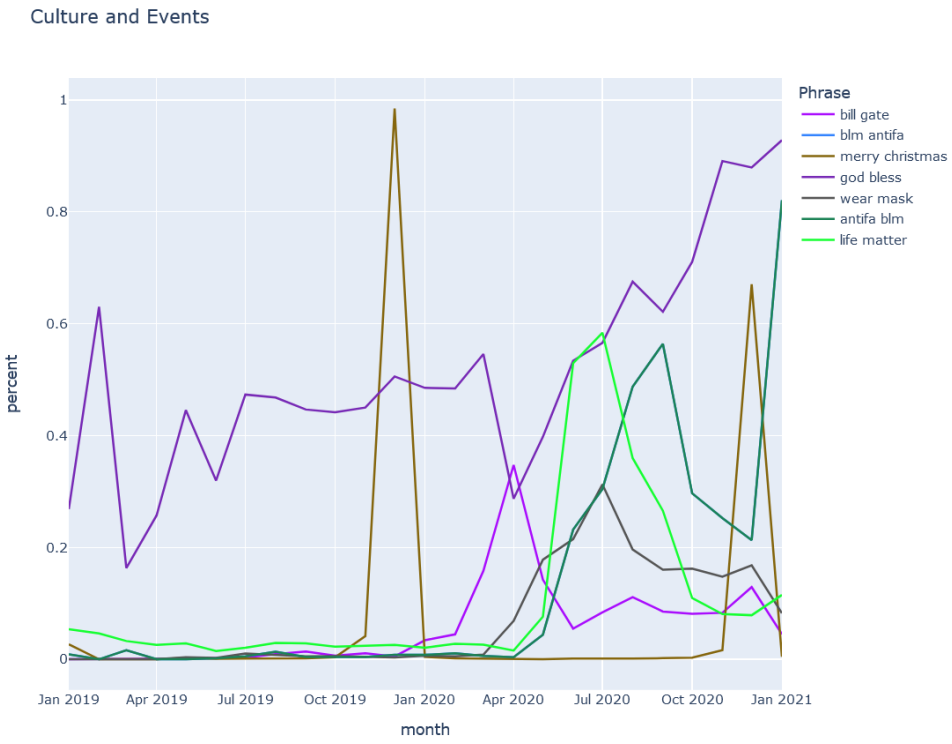

Figure 3.6: Change in the percent of Parler comments that use culture-related bigrams over time.

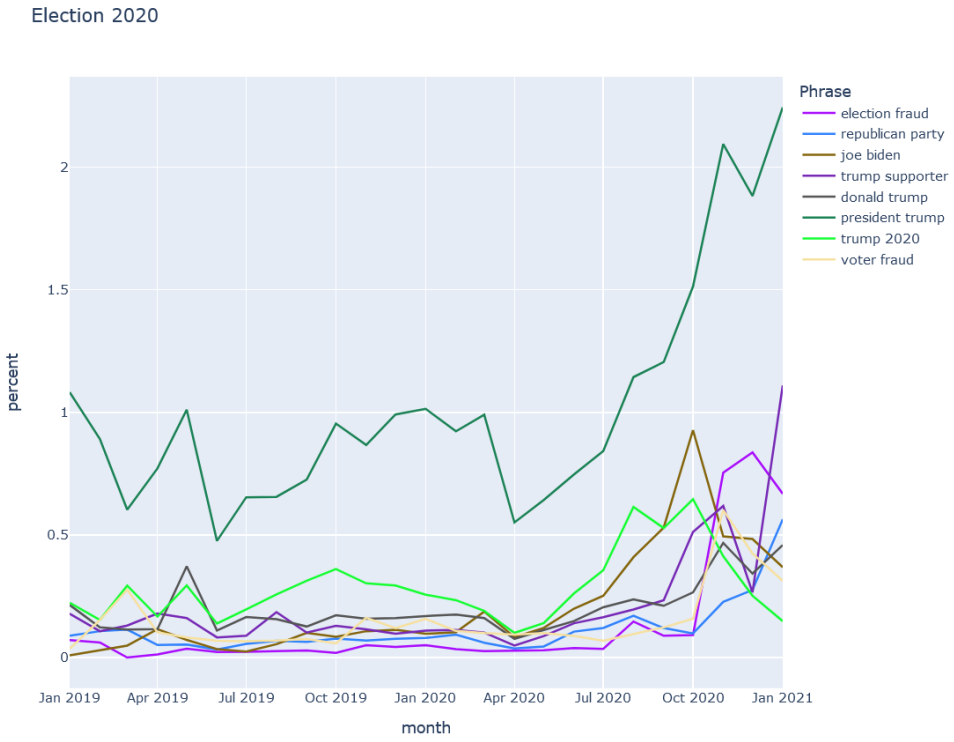

Figure 3.7: Change in the percent of Parler comments that use election-related bigrams over time.

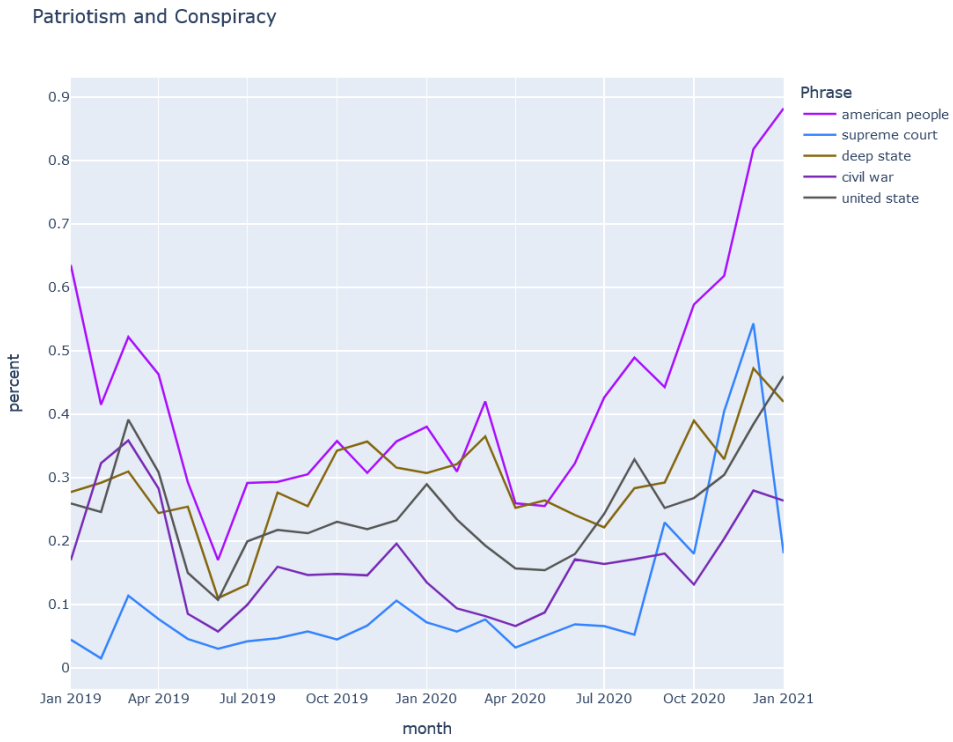

Figure 3.8: Change in the percent of Parler comments that use patriotism-related bigrams over time.

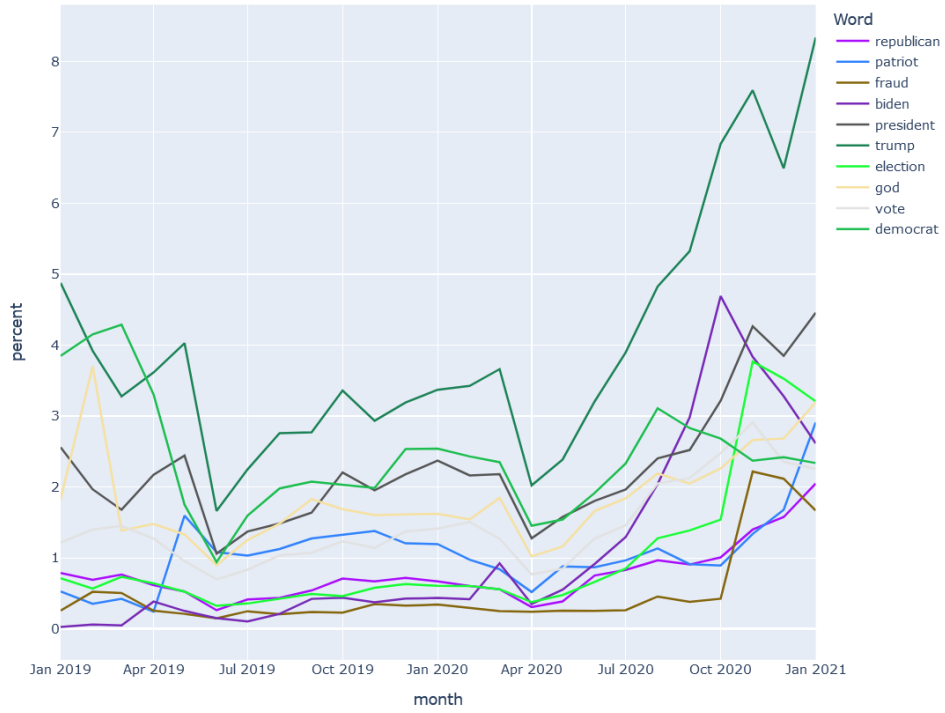

Figure 3.9: Change in the percent of Parler comments that use selected words over time.

"voter fraud" makes the list emphasizes the high level of increased activity around the 2020 election, as this was when the idea of voter fraud for that election started to take off. The bigrams "news apps", "apps right", and "explore fox" seem hard to interpret initially, but inspection of the dataset shows that these are all part of the same commonly used phrase "explore the fox news apps that are right for you", a phrase automatically generated by sharing a Fox News story to the app. Finally, the phrases "supreme court" and "life matter" show the interest in recent political events, namely the nomination of Amy Coney Barrett to the US Supreme Court in late 2020 and the Black Lives Matter protests in the summer of 2020.

Figure 3.2 shows the most common unigrams and bigrams used in Parler comments. Somewhat unsurprisingly, the overall top words used are roughly the same, with some minor changes in the order. The bigrams show some more interesting changes, with "social medium", "fake news", and "free speech" all newly showing up in the list. This indicates that there was more of a discussion of social media in the comments, with an emphasis on the prevalence of fake news and the importance of free speech. "echo echo" also no longer shows up in the top bigrams, which makes sense, as comments cannot be echoed and so there is no need to call for users to spread the post.

Figure 3.3 shows the most common unigrams and bigrams used in tweets. The top words used are substantially different from those in Parler, and are largely more generic. "trump" still shows up as a top word, but no other political words. Similarly, although the

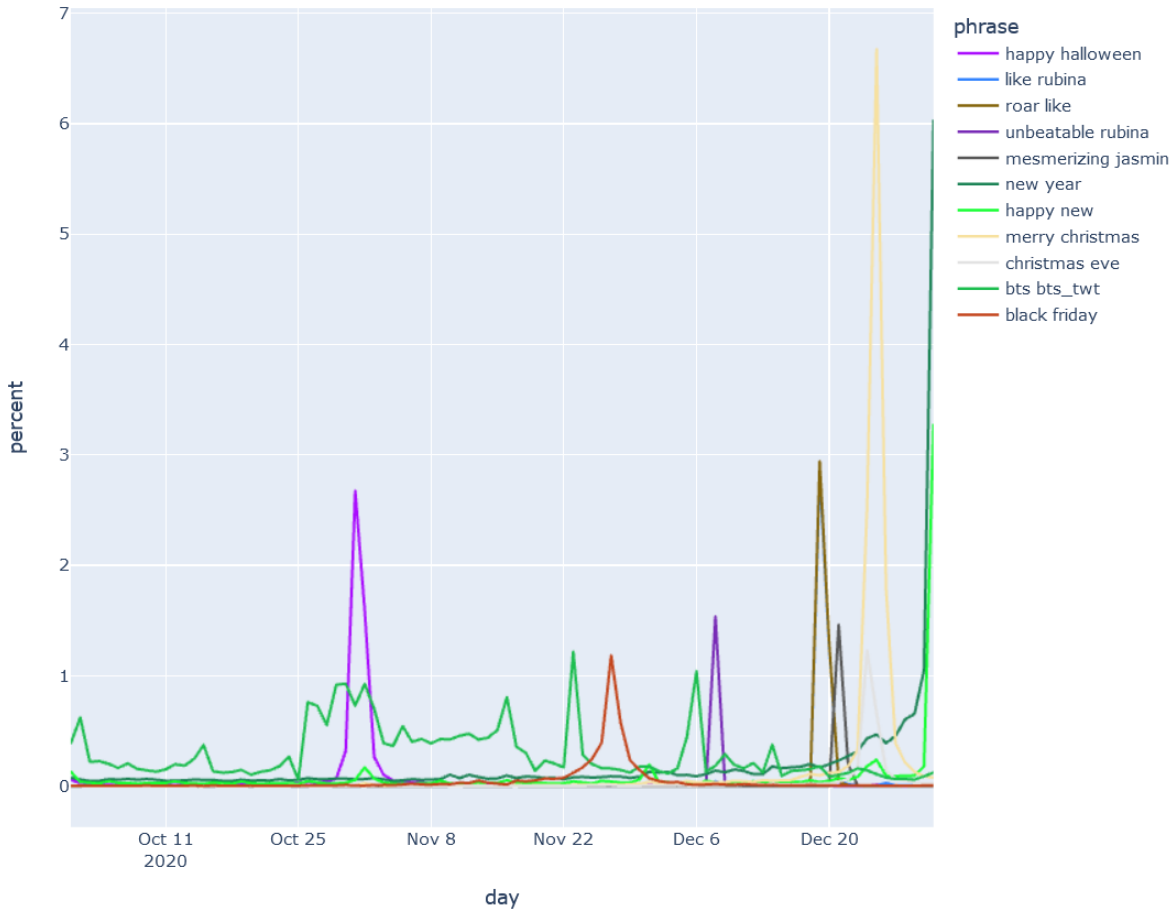

Figure 3.10: Change in the percent of tweets that use event-related bigrams over time.

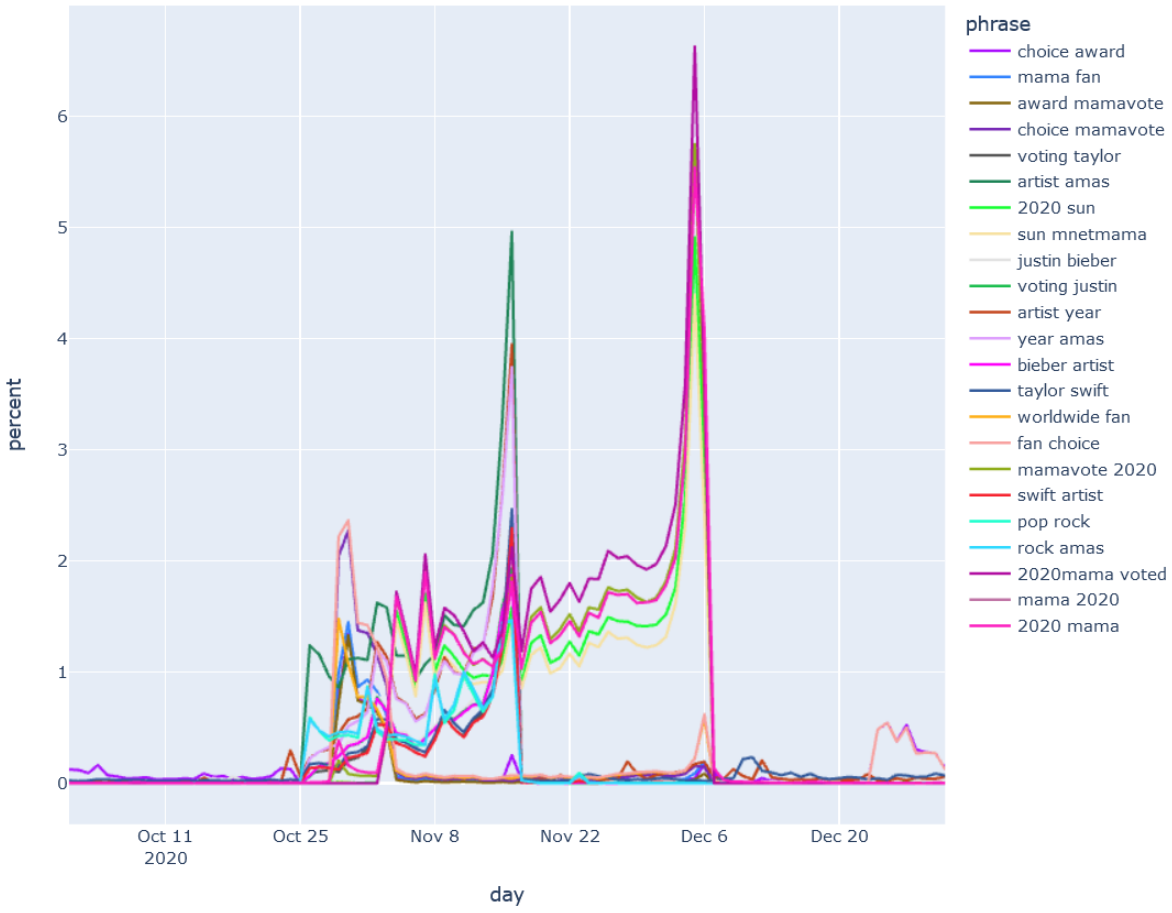

Figure 3.11: Change in the percent of tweets that use music awards-related bigrams over time.

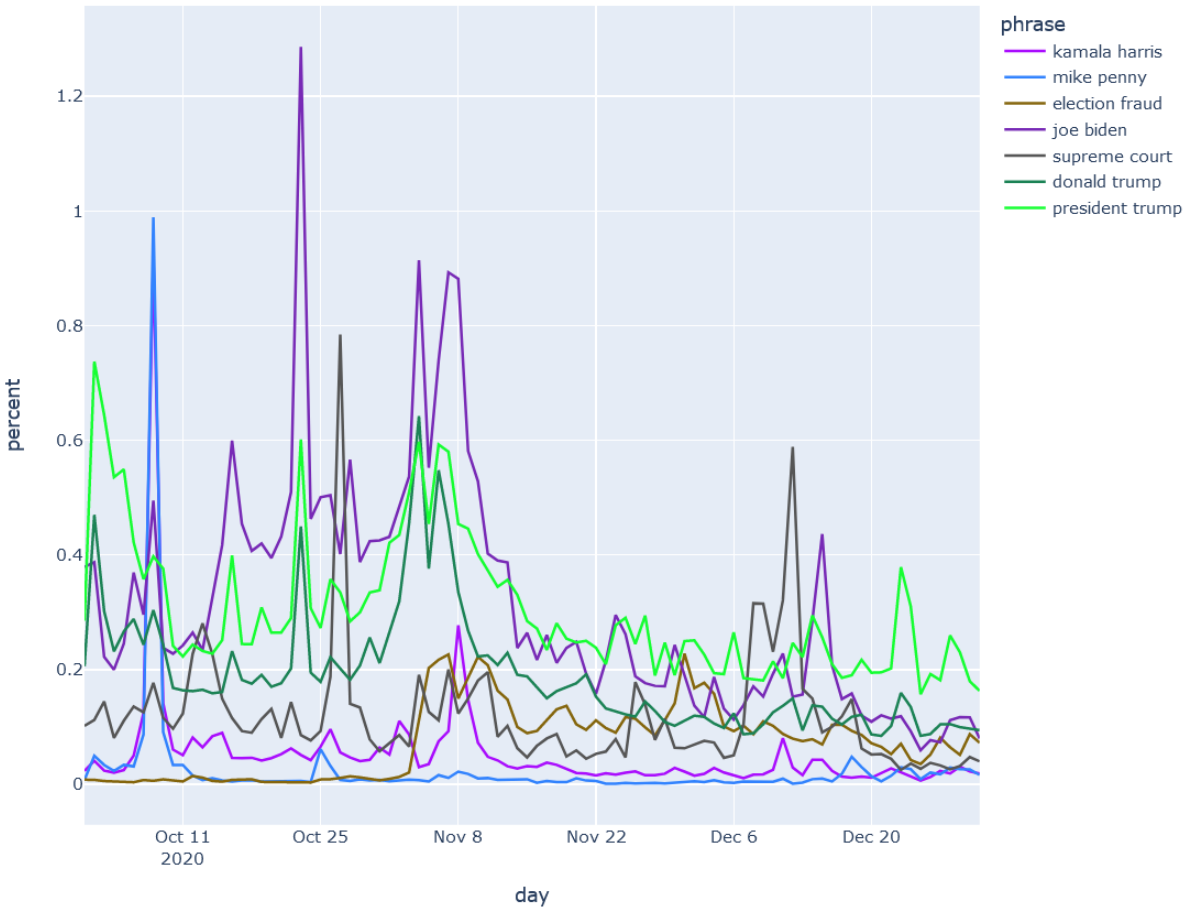

Figure 3.12: Change in the percent of tweets that use politics-related bigrams over time.

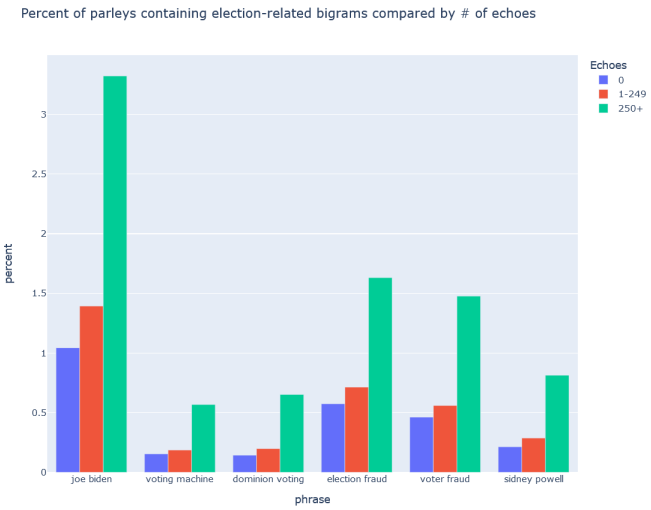

Figure 3.13: The change in election-related bigrams by popularity of posts.

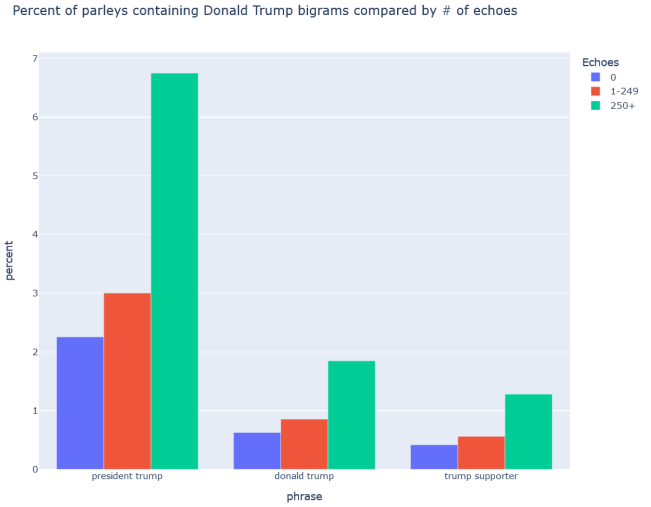

Figure 3.14: The change in trump-related bigrams by popularity of posts.

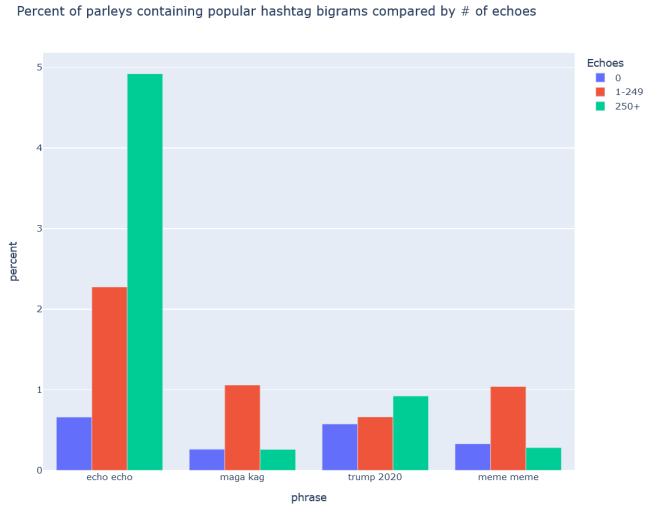

Figure 3.15: The change in popular hashtag bigrams by popularity of posts.

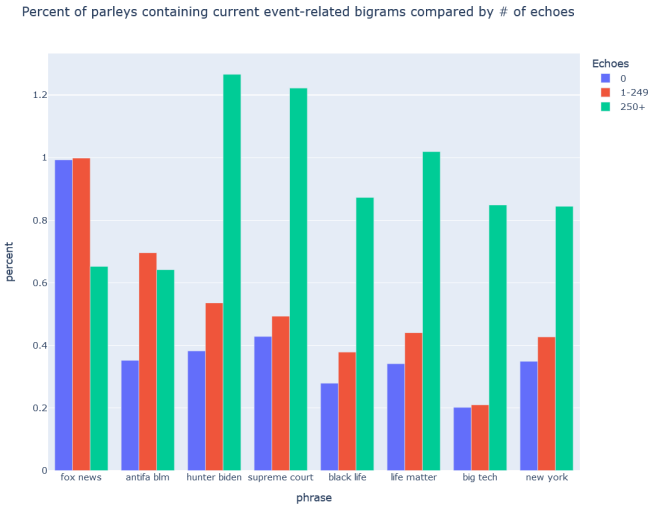

Figure 3.16: The change in current events bigrams by popularity of posts.

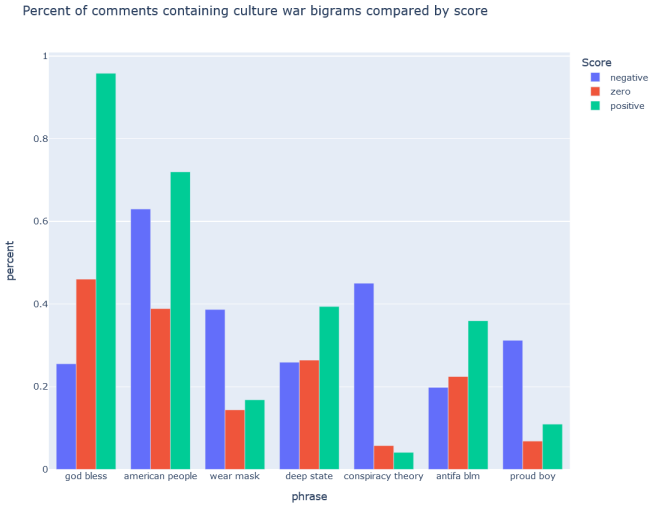

Figure 3.17: The change in culture and event bigrams by score of comments.

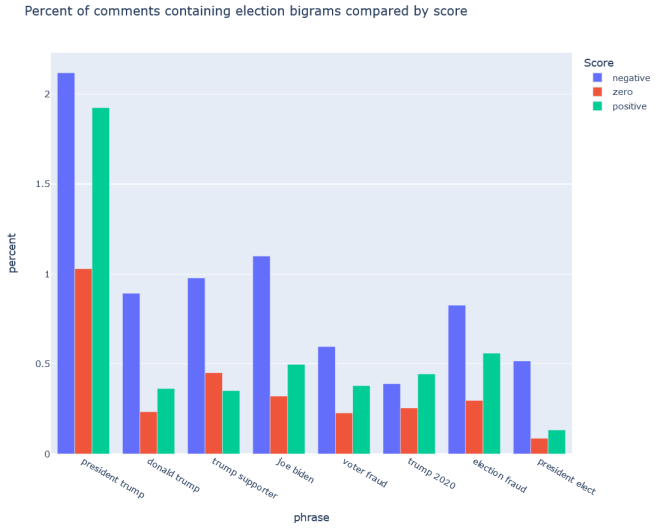

Figure 3.18: The change in election bigrams by score of comments.

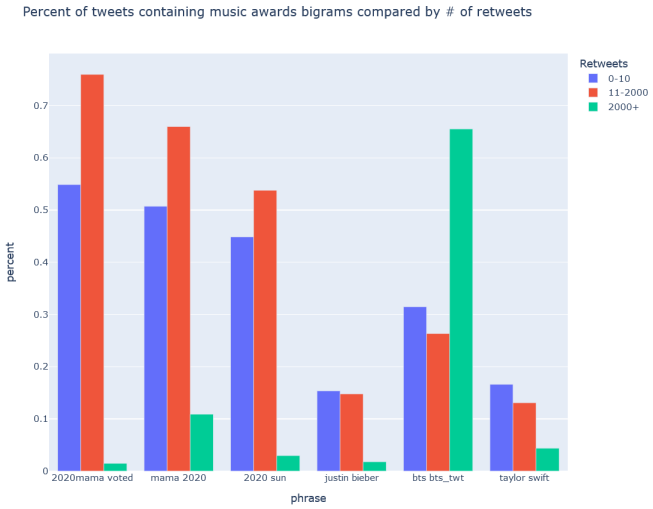

Figure 3.19: The change in music bigrams by number of retweets.

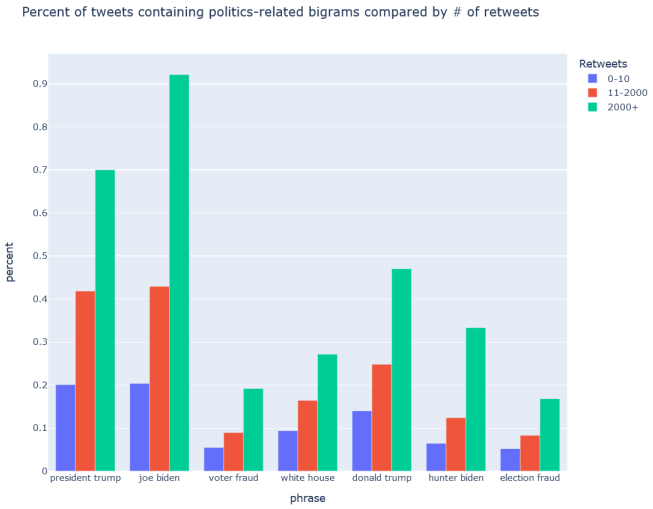

Figure 3.20: The change in politics bigrams by number of retweets.

political phrases of "donald trump", "joe biden", and "president trump" are in the top 20 of common Twitter phrases, the most common are related to voting in a 2020 Korean music awards ceremony known as MAMA. The American Music Awards, or AMAs, also show up. Finally, there are a couple of widely used common phrases such as "happy birthday" or "good morning". This all indicates a community with a wider focus than that of Parler, with people talking about common topics other than just politics.

Next, we analyze the change in the popularity of words and bigrams over time. When analyzing this change, we use only the percent of posts that contain certain words, not the total count. This is because the substantial increase in users over time in the Parler dataset means the total counts increase over time, making the graphs hard to interpret.

Figs 3.4 (Events) and 3.5 (Election) graph phrases that show substantial change over time. "hunter biden" and "joe biden" spike substantially in October 2020, as the Trump campaign promoted the story of Hunter Biden's leaked emails, and attempted to link his supposed fraud to his father. We see an increase in the phrases "voter fraud" and "election fraud" after the election, as the narrative that Donald Trump had the election stolen from him started to gain force among his supporters. These changes map onto real-world politically charged events very closely. For example, in May 2020, there is a spike in the phrases "law enforcement", "black life", and "life matter", all around the time of increased Black Lives Matter protests. An important note is that the bigram "life matter" shows much more of a jump than the bigram "black life", as the phrases "blue lives matter" and "all lives matter" increased as a backlash to the protests.

We split our time-based analysis of phrases in comments into three rough categories for ease of presentation: Culture/Events (3.6), the 2020 US election (3.7), and Patriotism (3.8). Culture and event-related bigrams largely map onto specific events - there is a large spike in "merry christmas" around each December, "bill gate" and "wear mask" both spike in spring of 2020, as they are both discussed with respect to the COVID-19 pandemic, and "life matter" and "blm antifa" both spike in summer of 2020 when BLM protests were popular in the US. Election-related phrases are all fairly low during 2019, and then spike in summer and fall 2020, as discussion of the election heats up. Of note is the fact that use of the phrase "president trump" nearly doubles in the last part of 2020, especially after the election, likely used to support the idea that Trump actually won the election. Similarly, after November 2020, "election fraud" and "voter fraud" also see large increases, and the phrase "trump supporter" goes from appearing in 0.3% of posts to 1.1% of posts as politicization increased. Finally, patriotic discussion overall seems to increase after the election, with the phrase "american people" nearly doubling in usage in the last few months of 2020 and the phrase "united state" also increasing substantially. We can also see increased interest in the "supreme court" at the end of of 2020, when Amy

Coney Barrett was nominated to the court. Analysis of unigrams over time in Fig 3.9 show similar effects, where usage of the word "trump" increases greatly in the last half of 2020, and election-related words such as "biden" and "vote" peak in October and November of 2020, followed by an increase in "fraud". Patriotic trends also appear here, with "patriot" increasing in December 2020 and January 2021.

As before, we split Twitter phrase analysis into relevant categories - Events (3.10), Music Awards (3.11), and Politics (3.12). The event-related phrases show even more holiday-related trends than in the Parler data, with significant spikes for Halloween, Christmas, Black Friday, and New Year's Day. It also shows an interest in other pop culture topics - here "roar like rubina", "unbeatable rubina", and "mesmerizing jasmin" are all trending Twitter phrases related to the 14th season of Bigg Boss, the Bollywood version of Big Brother [11]. We can see similar interest in pop culture in phrases related to music awards, where we can see three distinct peaks - the first in late October for the fan choice awards for the Korean MAMA awards, the second in mid November for the American AMAs, and the last in early December for the overall Korean MAMA awards. These are a widely discussed topic on Twitter, especially compared to political discussion. As we can see in the political section, the most popular topic of discussion is Joe Biden, especially in the weeks just before the elction. Discussion of Donald Trump is right behind him in popularity, and discussion of the two vice presidential candidates also peak in early October (note that Mike Pence is shown as "mike penny" due to stemming). However, this discussion largely dies down after the election, rarely rising above 0.25% of tweets (compared to the 6% of tweets that mention music awards around the same time). Discussion of election fraud also stays steady, and Supreme Court discussion does peak around the nomination of Barrett.

Overall, we can see that popular Parler phrases tend to be much more political than on Twitter, with Twitter being used to discuss a wider range of topics, especially pop culture and music. We also see political discussion on the two platforms display opposite trends - the level of discussion peaks around the election on Twitter, while it only increases over time on Parler, as users become increasingly convinced by claims of election fraud.

Finally, we partition each dataset by the popularity of posts to find differences in the words used by popular and unpopular posts. Ae use the total number of echoes, or reposts to one's own timeline, as the measure of popularity for parleys. We split the parleys into three categories - parleys with zero echoes, parleys with 1-249 echoes, and parleys with 250 or more echoes. As Parler comments cannot be echoed, we instead use score (combination of upvotes and downvotes) as our measure for popularity, and use the straightforward categories of posts with a positive overall score, those with a score of zero, and those with a negative overall score. Last, for consistency with parleys, we use number of retweets as our measure for Twitter popularity, splitting the dataset into tweets with 0-10 retweets, those

with 11-2000 retweets, and those with greater than 2000 retweets. In our comparisons, we choose phrases to examine that show interesting trends for sake of graph readability, omitting many others.

As with time-based analysis, we similarly break up our popularity analysis by groups for ease of presentation. For parleys, we choose the categories of Trump (3.14) the 2020 election (3.13), current events (3.16), and common hashtags (3.15). We split Trump-related hashtags into their own category because of their comparative popularity, observing a substantial increase in the use of all three phrases in posts with large numbers of echoes. This trend continues when observing election-related bigrams, with discussion of election fraud in various forms especially prominent in popular parleys. These include simply the phrase "election fraud" as well as mentions of voting machines, a specific company that manufactures them, and one of the main lawyers that promoted claims of election fraud. Comparison of some popular hashtags shows that the phrase "echo echo", which promotes spread of a parley, understandably mostly shows up in popular parleys. However, the hashtag "maga kag" mostly show up in posts with 1-249 echoes, indicating a semipopular template of parleys that was not spread especially far. Finally, looking at phrases related to current events, we can see that again, much of the discussion of Hunter Biden, Black Lives Matter, and the Supreme Court takes place in especially popular parleys. The two main exceptions are discussion of Antifa/BLM, which takes place in posts with 1 or more echoes, and discussion of Fox News, which largely occurs in posts with 0-249 echoes.

Our popularity-based comment analysis is broken into two categories - American culture (3.17) and the 2020 election(3.18). Culture bigrams exhibit some interesting dichotomies - "god bless", "deep state", and "antifa blm" are all more likely to appear in upvoted comments, while comments with the phrases "wear mask", "conspiracy theory", and "proud boy" are all more likely to appear in downvoted comments. This illustrates what seems to be a divide in the userbase between users that think the deep state is a real concern and those that seem to label some popular discussions as conspiracy theories. This also makes sense in terms of the specific phrasing, as the term conspiracy theory is usually used in a derogatory sense, and usually to label topics of the sort characterized by discussions of deep states. Similarly, it seems that discussion of wearing masks is normally downvoted, which also makes sense, as it was politicized by conservatives during the COVID-19 pandemic. Finally, we can see that the phrase "american people" was used by both negative and positive comments, likely making opposite points. Interestingly, most of the election-related bigrams are more likely to appear in negative comments versus positive ones. The main exceptions are "trump 2020" and "president trump", which appear at roughly the same frequency in positive and negative comments. Again, this points to a similar split in the userbase, as users commented about election fraud both positively and negatively.

Finally, when analyzing the Twitter data when split by popularity, we examine bigrams in two categories, music (3.19) and politics (3.20). We note that most of the music award-related bigrams appear in tweets with lower amounts of retweets, indicating that they are likely either automatically generated or not widely spread for other reasons. The main outlier here is discussion of the specific Korean boy band BTS, which appears in more popular tweets than unpopular ones. Contrary to the non-BTS music awards tweets, we see that politics-related phrases are more likely to appear the more popular the tweet is. This indicates that although more tweets overall mentioned popular music awards than political phrases, the tweets that were political were more widely spread.

Overall, we see that Parler posts and comments tend to focus on a variety of political topics, while Twitter has a wider range of popular topics. We can also see that while political discussion in tweets decreases substantially after the election, discussion on Parler instead increases substantially until January. Finally, more popular tweets and parleys are more likely to have political content and election discussion, while Parler comments tend to have more election discussion in unpopular comments.

3.2 LDA Topics

We now present the results of our Latent Dirichlet Allocation(LDA) analysis. We start by looking at Fig 3.21, the topics discovered in parleys. They are largely easy to interpret, with the exception of topic 2, which seems to capture more generic words and posts. Topic 1 centers around the Trump 2020 campaign as well as the QAnon conspiracy theory, with a lot of use of specific hashtags around both. It likely captures parleys around these topics that are designed to be easy to find and to share. Topic 3 is similar, but with more a focus on discussion of the election, with no hashtags. Topic 4 clearly captures discussion of popular Democratic politicians, as well as popular news sources. Topic 5 involves discussion of COVID-19, while topic 6 uses keywords around patriotism and religion. Topic 7 is the only one that contains Spanish words, and is clearly indicative of the Brazilian subcommunity of Parler, with discussion of Jair Bolsonaro. Finally, Topic 8 looks at discussion of popular social media sites. Overall, the topics are very political, with 3 of the top four focusing on the US election.

Next, we look at the topics discovered from Parler comments in Fig 3.22. Here, we found 8 topics to be the best performing number. Most of the topics are straightforwardly interpretable. Topic 1 clearly corresponds to the US 2020 presidential election, although mixed with some COVID-19 related words such as 'case' and 'vaccine'. Topic 2 is the least interpretable, and is likely more generic posts without a clear mention of the other topics.

Number	Coherence	Top Words
1	0.904	trump2020, maga, echo, wwg1wga, stopthesteal, meme, parler, qanon, freedom, trump, kag, usa, patriot, news, thegreatawakening, parlerksa, electionfraud, trumptrain, voterfraud, maga2020
2	0.828	the, and, this, that, you, for, are, they, not, have, all, will, with, what, people, our, can, their, who, but
3	0.769	trump, election, vote, state, president, fraud, the, for, 2020, voter, ballot, court, voting, republican, biden, democrat, georgia, donald, donald trump, win
4	0.642	biden, news, trump, joe, fox, obama, joe biden, fox news, president, his, pelosi, fbi, for, hunter, wa, cnn, clinton, harris, with, house
5	0.634	the, covid, for, police, new, mask, china, from, coronavirus, vaccine, virus, state, after, with, bill, death, city, business, school, home
6	0.599	you, god, love, your, for, thank, our, please, and, jesus, bless, president, good, god bless, fuck, patriot, lord, as, pray, happy
7	0.558	que, yes, com, amen, não, para, newuser, por, george, uma, mais, brasil, bolsonaro, soros, como, do, está, presidente, george soros, tem
8	0.538	twitter, watch, parler, post, video, facebook, this, wow, check, share, please, medium, youtube, like, here, speech, new, social, account, look like

Figure 3.21: The topics discovered from parleys.

Number	Coherence	Top Words
1	0.792	trump, president, election, vote, biden, fraud, republican, state, win, court, voter, house, ballot, evidence, voting, 2020, office, case, voted, vaccine
2	0.776	the, you, that, and, they, this, for, are, not, have, with, all, what, can, but, wa, just, your, like, get
3	0.767	the, and, our, will, for, are, their, democrat, country, american, from, america, all, left, medium, people, state, law, must, god
4	0.699	them, money, black, white, child, their, mask, covid, police, antifa, blm, home, pay, death, business, kid, city, criminal, crime, virus
5	0.583	his, shit, joe, fuck, as, obama, biden, fucking, traitor, bitch, jail, boy, swamp, treason, bill, piece, soros, po, lying, disgusting
6	0.582	you, love, thank, great, god, parler, please, follow, your, twitter, welcome, thanks, post, bless, here, friend, god bless, comment, glad, facebook
7	0.443	she, her, pelosi, yep, nancy, lmao, kamala, bye, york, aoc, luck, new york, wood, good luck, rope, lin, mitch
8	0.379	yes, news, amen, fox, fake, awesome, cnn, fake news, tucker, agreed, fox news, newsmax, network, wallace, hahaha, hannity

Figure 3.22: The topics discovered from comments.

Topic 3 seems to correspond to American patriotism and religion. Topic 4 combines general cultural flashpoints, such as Antifa, police, COVID, Black Lives Matter, and crime. Topics 5 and 7 both involve specific prominent Democratic political figures, although Topic 5 is more centered around calling male political figures negative words, while Topic 7 seems to be more focused on female political figures. Topic 6 is the most positive topic, seemingly centered around positive interactions between Parler members as well as discussion of social media. Finally, Topic 8 discusses fake news and various news sources, with a specific focus on Fox News. Overall, only topics 2 and 6 are arguably non-political, with the rest corresponding to major US conservative issues.

For our Twitter dataset, we use hashtag pooling in order to improve the quality of the topics found, as suggested by Mehrotra et al. [35]. We note that this limits the dataset to roughly 23% of its original size, as we only use tweets that contain at least one hashtag. Figure 3.23 shows the topics found via this method. In substantial contrast to the Parler topics, these topics are almost all non-political, and are instead primarily focused on popular culture. There are several that involve popular bands - topic 1 focuses on the 2020 MAMA awards, topic 4 focuses on the Korean boy band BTS, while topic 10 focuses on the Korean boy band NCTsmtown. Topic 2 is clearly focused on pornographic content. Topic 3 focuses on the Indian reality show Bigg Boss, which also spawned noticeable bigram trends. Although both contain some seemingly unconnected words, topics 5 and 9 seem the most political, with 5 involving COVID and its effect on businesses, while 9 focuses on Donald Trump and requests for specific action. Topic 6 focuses on Indian and Pakistani politics. Finally, Topic 7 is similar to Topic 6 in Fig 3.22, which showed positive words, with this one focused around Christmas. Finally, Topic 8 focuses on the celebration of New Year's Day, and also seems to lump in other time-related words. Overall, although the Twitter topics are harder to interpret than the Parler topics, they still suggest that the most popular topics of discussion are non-political, and instead focus more on popular culture. Of course, this may also result from the hashtag pooling method, as cultural trends such as those shown in the topics tend to create specific hashtags, while political events may not.

3.3 Hate Words

The next part of the analysis relies on analyzing the usage of hate words in each of our three datasets. Figure 3.24 shows the comparisons of the total percentage of posts that contain either a unambiguous hate word or an ambiguous one. Parler comments are the highest at 2.5%, followed by parleys at 2.05% and tweets at 1.44%. This means that Parler

Number	Coherence	Top Words
1	0.828	2020, fan, award, choice, treasure, got7, global, vote, ateezofficial, pledis 17, seventeen, got7official, mamavote, ateez, 2020mama, mama, drop, voted, twice, redvelvet
2	0.668	video, follow, like, full, hot, amp, gay, retweet, nsfw, sexy, onlyfans, link, want, fuck, foot, cum, http, get, sex, part
3	0.642	tweet, tag, trending, trend, master, fan, guy, retweet, let, get, movie, speed, colorstv, like, amp, sidharth shukla, biggboss, king, alygoni, abijeet
4	0.592	vote, bts twt, reply, bts, exo, stray kids, voting, blackpink, kai, txt members, weareoneexo, taehyung, year, let, artist, best, award, global, retweet, pop
5	0.555	amp, new, help, check, via, today, project, covid, work, covid19, need, great, business, online, health, free, community, learn, join, http
6	0.539	farmer, justice, pakistan, god, support, india, people, amp, one, please, want, sushant, life, government, right, today, must, leader
7	0.523	love, day, like, one, christmas, good, let, thank, time, see, look, best, happy, always, know, make, today, much, want, beautiful
8	0.489	new, year, 2021, amp, 2020, live, today, back, day, first, win, december, welcome, week, time, coming, join, watch, game, tonight
9	0.476	please, need, let, get, people, help, deserve, retweet, tacha, erica, trump, better, realdonaldtrump, know, amp, say, make, queen, trending, titan
10	0.306	happy, birthday, happy birthday, day, taeyong, nct, nctsmtown, well, love, soon, album, get, 2020, get well, song, update, enhypen members, sb19official, thank

Figure 3.23: The topics discovered from tweets.

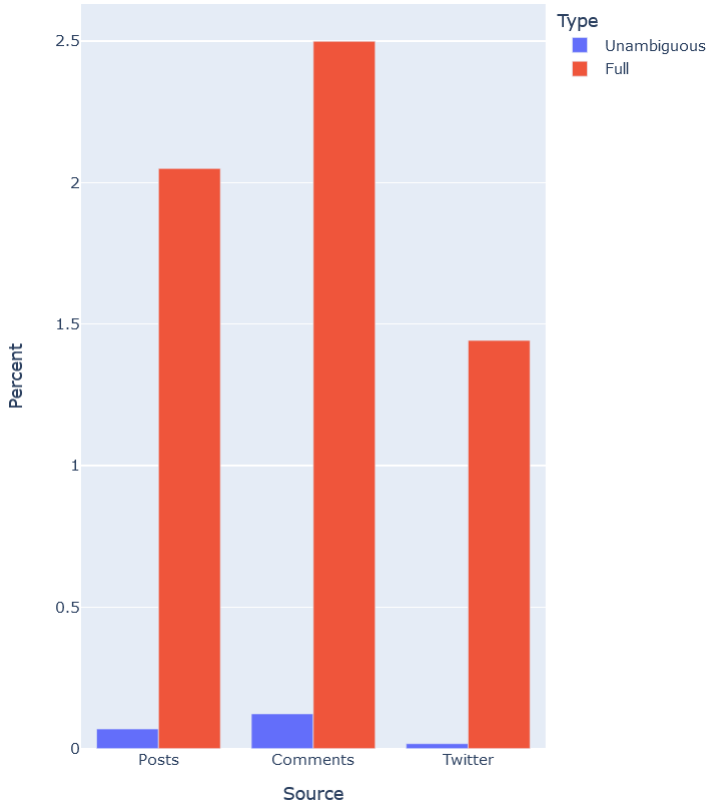

Figure 3.24: A comparison of the amount of posts that contain hate words from all three sources.

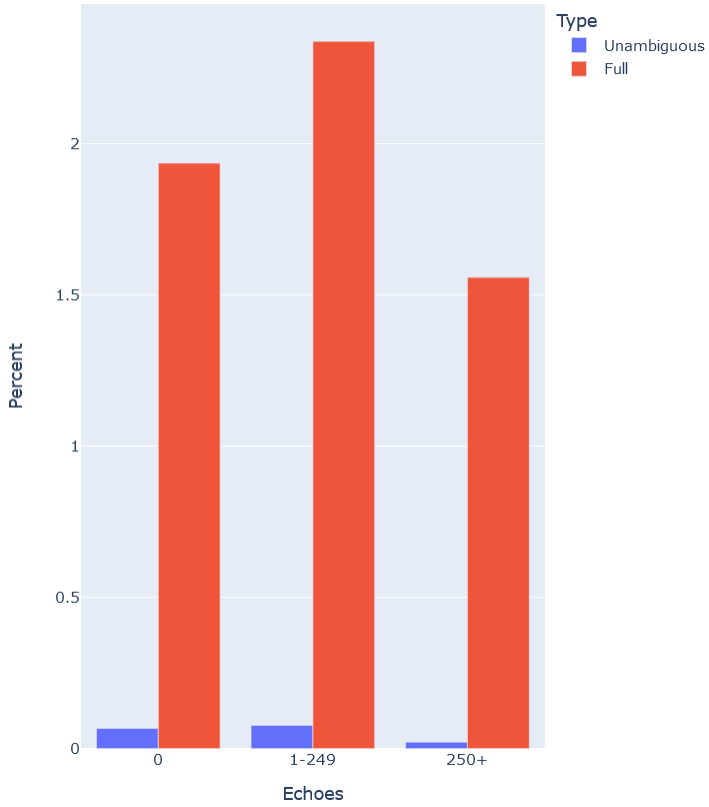

Figure 3.25: The percentage of parleys that contain a hate word when partitioned by popularity.

comments have 70% more posts that contain hate words than tweets, and 20% more than parleys. This number is considerably higher when taking into account unambiguous words - they have 76% more than parleys, and 600% more than tweets.

We do note that in all three datasets, there are considerably fewer posts that contain unambiguous hate words, which makes sense, as these are likely seen as excessive by a majority of the population of both communities.

Next, we look at the differences in hate percentages when split by popularity. Figure 3.25 shows the percent of parleys that contain a hate word when split by the number of echoes each parley got. Interestingly, more posts of intermediate popularity contain hate words of both types than those with zero echoes, indicating that a substantial amount of the hate in parleys is at least somewhat amplified by the community. Both types then

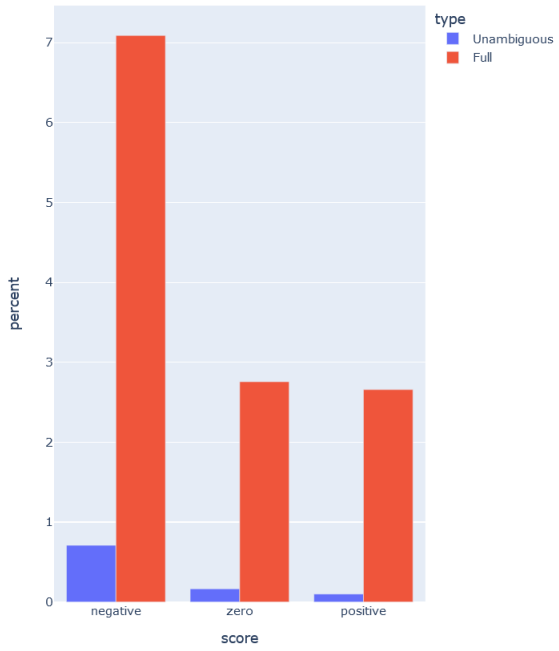

Figure 3.26: The percentage of comments that contain a hate word when partitioned by popularity.

drop significantly for the most popular parleys, although more so for unambiguous words - parleys with 1-249 echoes had 50% more hate words overall, but 250% more unambiguous words. This indicates that popular parleys use less hate words, and are less explicit about it when they do. This negative correlation between popularity and hate words is also shown when looking at the average number of echoes parleys have. On average, a parley has 5.8 echoes, which decreases to 4.2 if it contains a hate word, and further to 1.8 with an ambiguous hate word.

Parler comments show a similar trend of more popular posts having less hate. Here, there are roughly 160% more posts with a hate word that have a negative overall score than others - 7.09% compared with 2.75% and 2.65%. This is exaggerated when it comes to unambiguous hate words - there are 326% more comments with a negative score than those with a score of zero, and 600% more compared to this with a positive score. Overall, this indicates that Parler users are critical of the use of hate words, especially explicit slurs. Looking at it from a different perspective, we find that the average score of a comment is 2.36, but that decreases to 2.13 if it contains a hate word and further to 1.13 if it contains an unambiguous hate word.

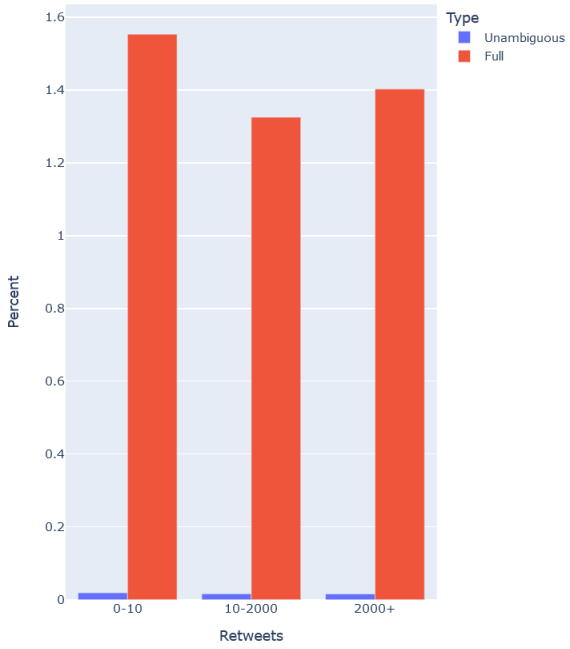

Figure 3.27: The amount of tweets that contain a hate word when partitioned by popularity.

Finally, tweets have similar levels of hate between levels of popularity. 1.55% of tweets with 0-10 retweets contain a hate word, compared with 1.32% for 10-2000 and 1.40% for 2000+. This is similar for unambiguous hate words, with the biggest change being a 16% increase in 0-10 compared to 2000+. Interestingly, there seems to be little to no difference between the three levels of popularity we analyze here. However, we can see a more clear trend looking overall - on average in our dataset, a tweet has 281.3 retweets, but a tweet with a hate word has only 246.5, and a tweet with an unambiguous hate word has 187 retweets. This indicates a negative correlation between popularity and hate, similar to the Parler dataset.

The total percentage of comments with hate words peaks at 4% at the start of 2019, then stays roughly level around 3% except for a few months where it dips to around 2% - June 2019, April-May 2020, and November 2020. June 2019 is when there was a large influx of Saudi Arabian users to Parler, and November 2020 was when the US Presidential election happened, so those two months likely had more discussion than normal around those two topics that did not involve hate words. The percentage of parleys exhibits similar trends, although with a fairly consistently lower percentage. The levels of unambiguous hate also show drops at around the same time, although there is a peak in January 2020 for parleys and in January 2019 for comments.

The percentage of tweets with a hate word fluctuate substantially from day to day, although it stays within a range of 1.3 to 1.8%. There are two noticeable outliers - a sharp increase to 1.8% on November 20th, and a sharp decrease 1.25% on December 6th. Referring back to 3.11, we can see that December 6th was a day with a spike in music awards-related tweets, likely accounting for this drop with an increase in tweets around a specific topic. The November 20th spike is less explainable, and looking at the actual words used, the only noticeable change is an increase in the word "nigga". The unambiguous hate similarly fluctuates substantially from day to day, although with less standout highs or lows. It is likely that much of this is attributable to variance, both from the actual day-to-day posting on Twitter and the variance in sampling from the Twitter stream.

3.4 Hate Classification

In order to more fully understand the types of words being used in the dataset, we use Hatebase's classification system, which classifies each word into being discriminatory based on nationality, ethnicity, religion, gender, sexual orientation, disability, and class. For reference, we include a graph of the total number of words in each of these classes as

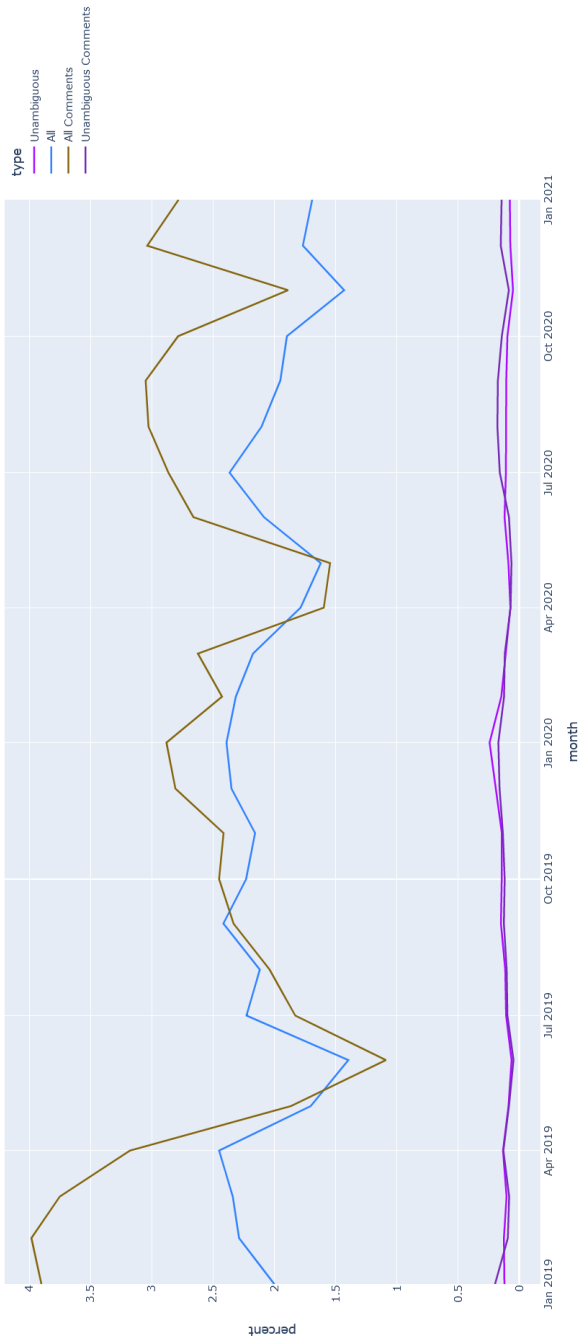

Figure 3.28: The change in the percent of parleys and comments that contain hate words over time.

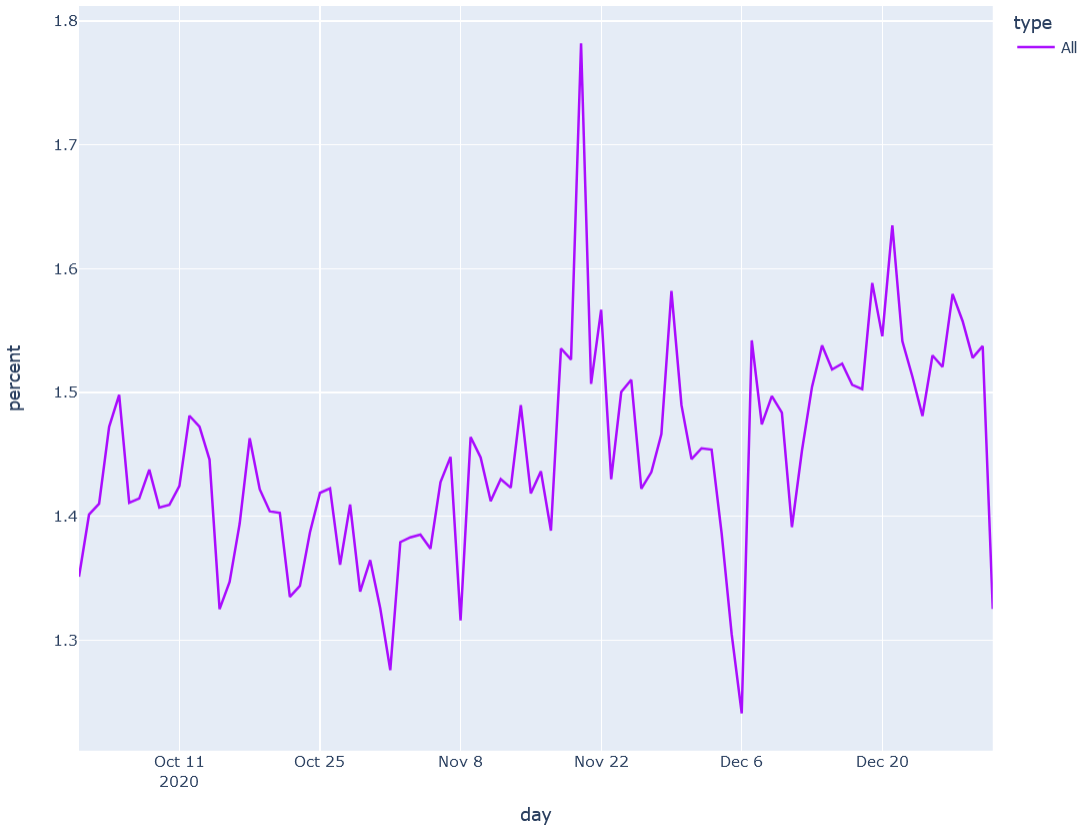

Figure 3.29: The amount of tweets that contain a hate word over time.

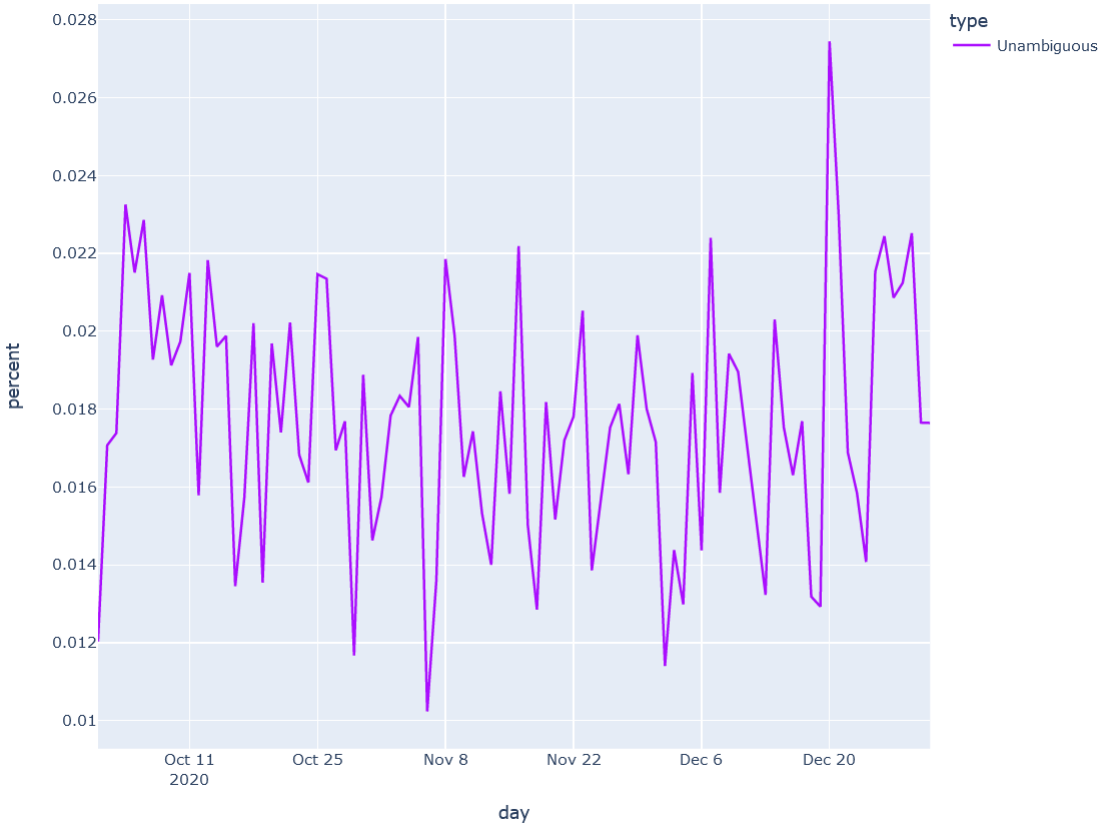

Figure 3.30: The amount of tweets that contain an unambiguous hate word over time.

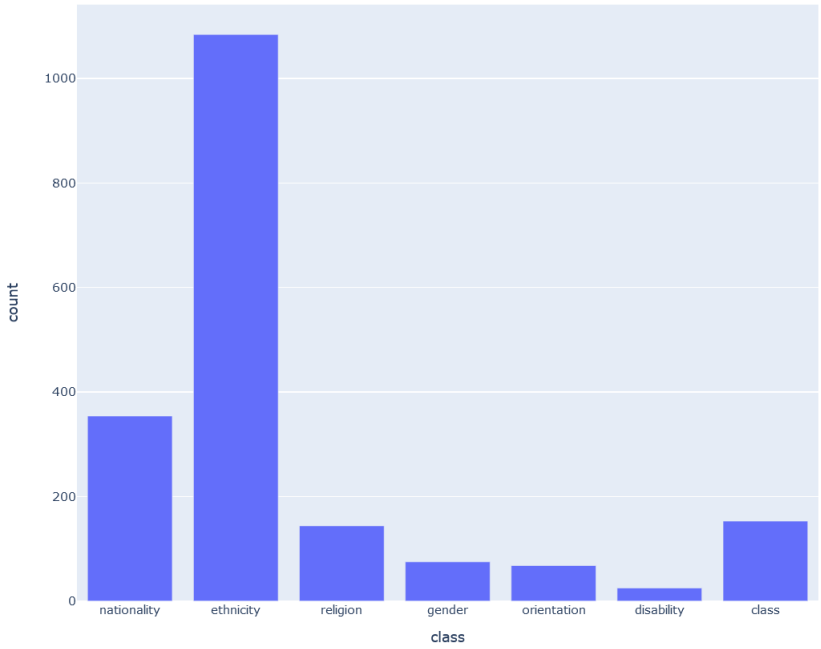

Figure 3.31: The total number of terms in each type of class as categorized by Hatebase.

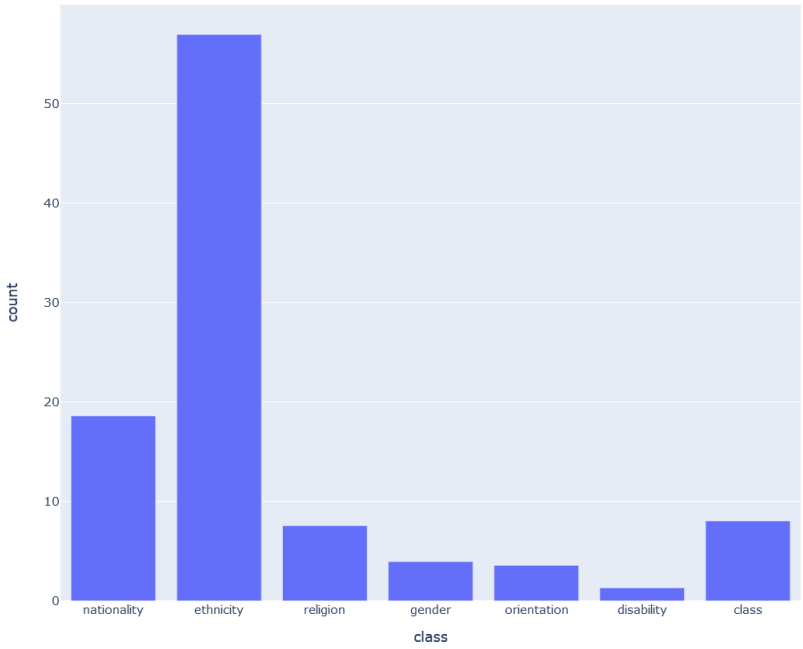

Figure 3.32: The total number of unambiguous terms in each type of class as categorized by Hatebase.

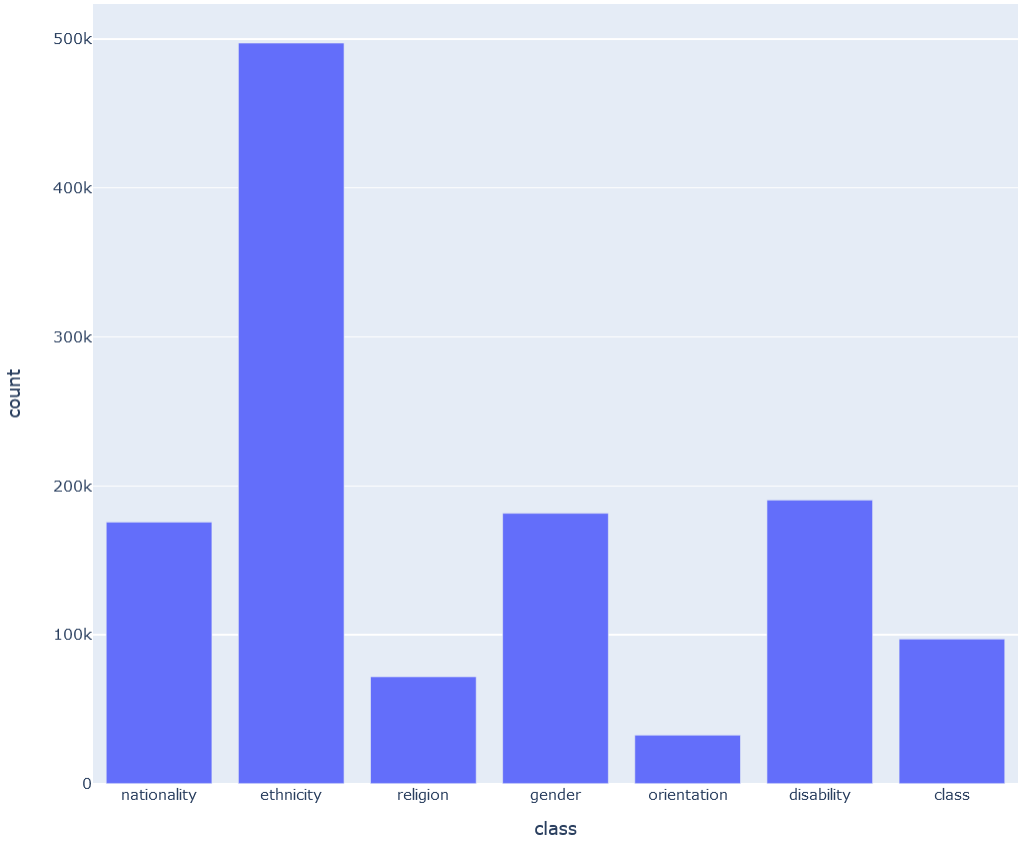

Figure 3.33: The classification of all hate words in all parleys.

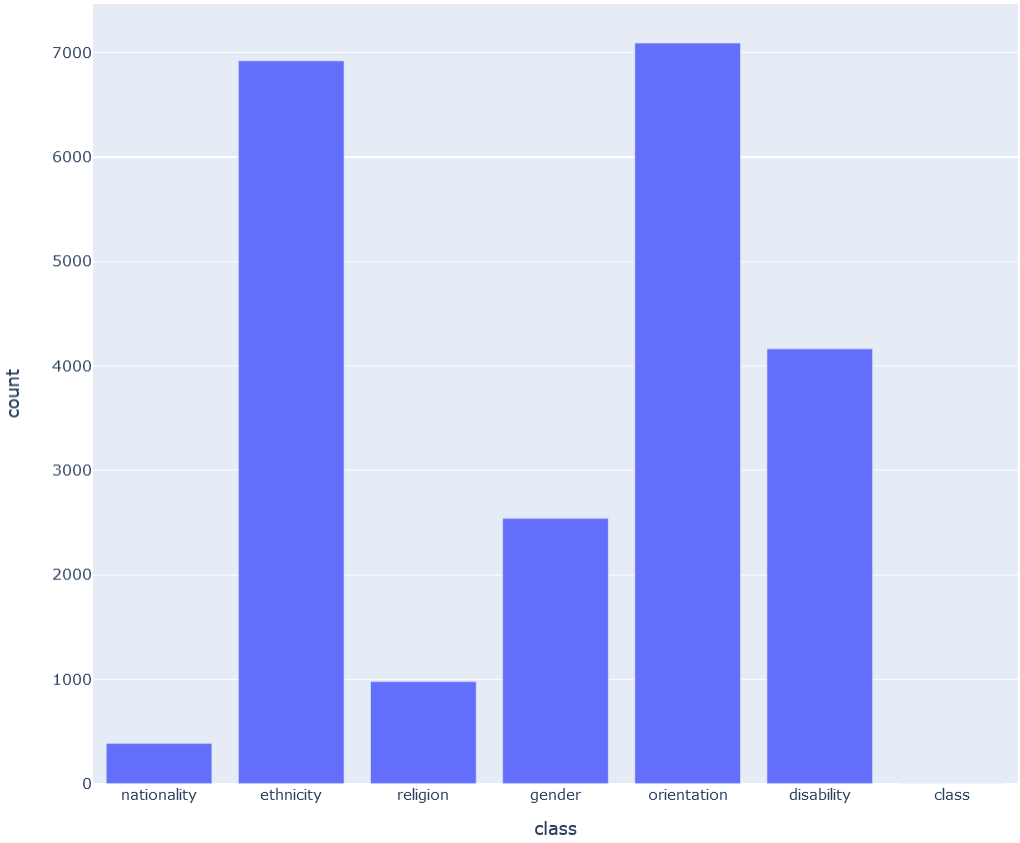

Figure 3.34: The classification of unambiguous hate words in all parleys.

provided by Hatebase in Fig 3.31 and that of the total number of unambiguous words in Fig 3.32. We examine the total amount of each type present in each dataset, as well as the actual words used.

Figure 3.34 shows the classification of unambiguous hate words in parleys. The majority of slurs used are those discriminatory to orientation and ethnicity, with gender and disability next. The two most used types, ethnicity and orientation, are almost wholly represented by the words "nigger", "faggot", and "tranny". The disability category is only the word "retard", while the gender category is also largely the word "tranny".

Figure 3.33 shows the categorization of all hate words in parleys. Here, the largest category is ethnicity-related words, such as "slave", "trash", and "spic". Nationality-related and religon-related words are largely taken up by "globalist", often used as a coded word for Jewish people, while gender-related words are those such as "bitch", "whore" and "cunt". Disability-related words are largely variations of the word "retard", especially the term "libtard", used as a pejorative against liberals.

Figure 3.35 shows the categorization of all hate words in comments. The distribution here is substantially different than in parleys, with the highest category being gender, followed by disability and then ethnicity. The actual top words used are fairly similar, with "bitch", "pussy", and "cunt" topping the gender list, variants of "retarded" topping the disability list, and "trash" and "slave" topping the ethnicity list. We do note that the third and fourth most popular ethnicity-related words are "snowflake" and "lefty", which are defined as related to Jewish people and people of Arabic descent respectively. However, it is likely that these words instead are being used to disparage people on the political left, with "special snowflake" often being used by conservatives to insult people they perceive as wanting to be catered to. Nationality and religion are both largely taken up by "globalist", orientation is largely "faggot" and "tranny", and class is largely "trash", "redneck" and "cracker".

Figure 3.36 shows the distribution of unambiguous hate words in comments, which is more similar to that of parleys. We see relatively more disability words than in parleys, and less ethnicity and religious words. Again though, we see the words "retarded", "tranny", "faggot" and "nigger" account for the majority of these words.

Figure 3.37 shows the categorization of all hate words in tweets. As with comments, we see ethnicity and gender-related words comprising a majority. Unlike Parler, however, the most common ethnicity-related word is "nigga", often used as a reclaimed and more informal version of the word "nigger". This is followed by "trash" and "slave". Gender-related words are primarily "bitch", "pussy", and "slut", again similar to those used in Parler. Disability words are almost exclusively just "tard", while orientation words are

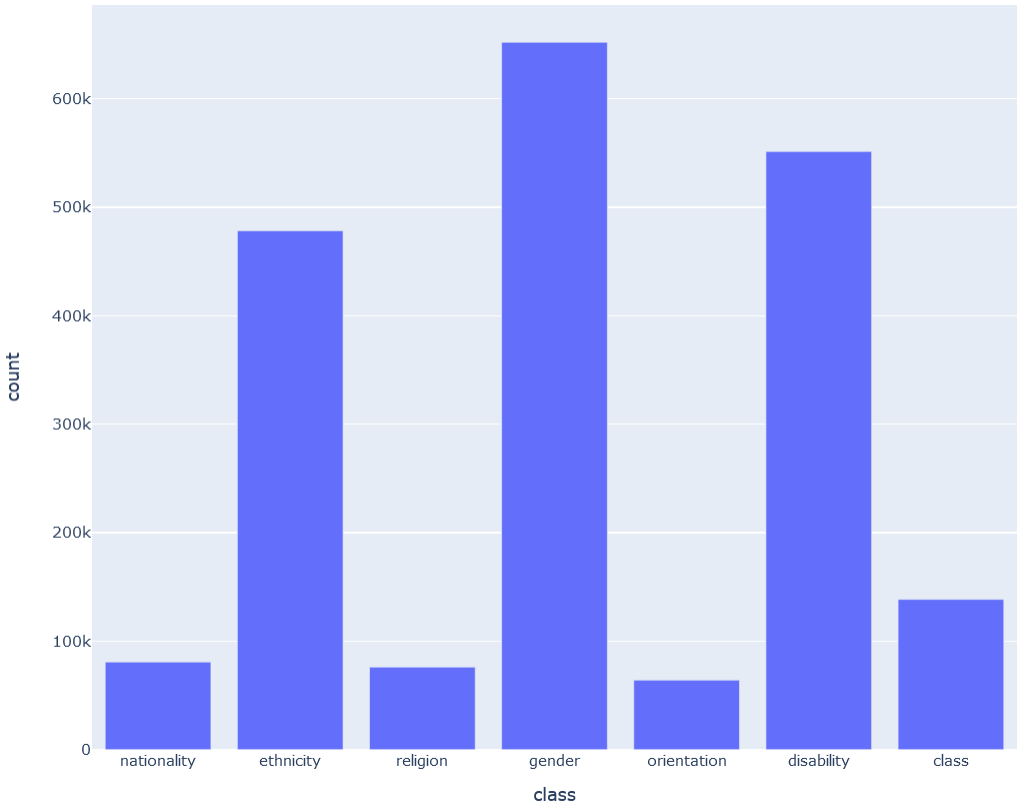

Figure 3.35: The classification of all hate words in Parler comments.

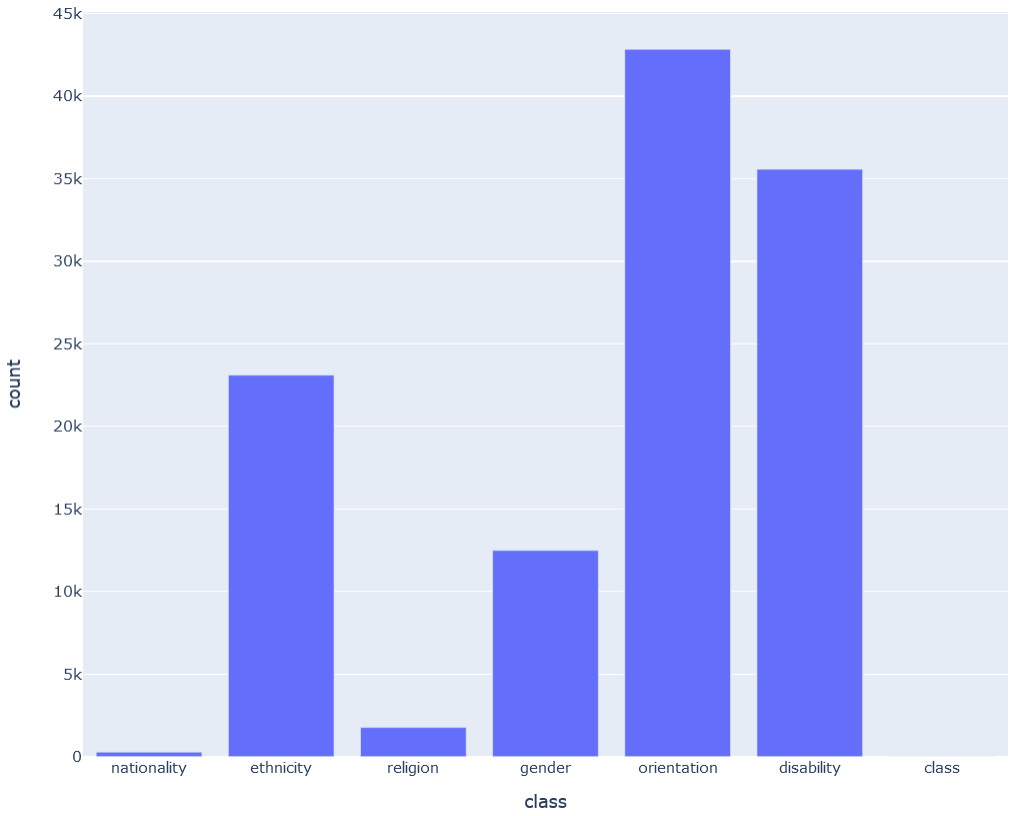

Figure 3.36: The classification of unambiguous hate words in Parler comments.

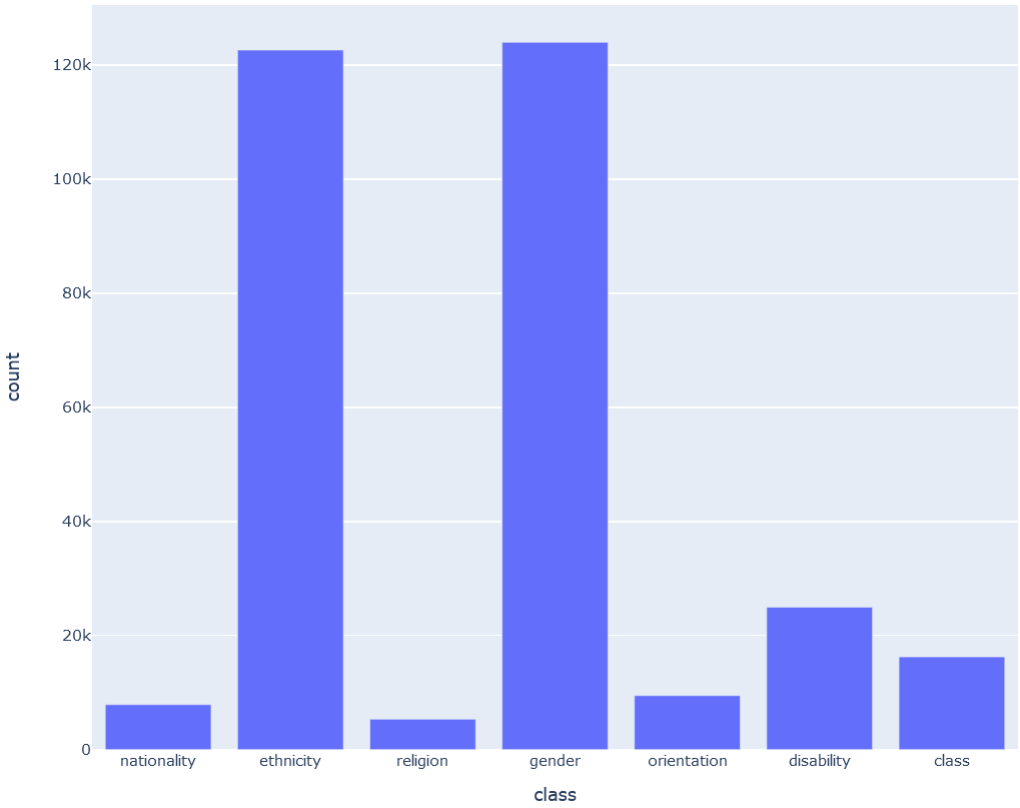

Figure 3.37: The classification of all hate words in tweets.

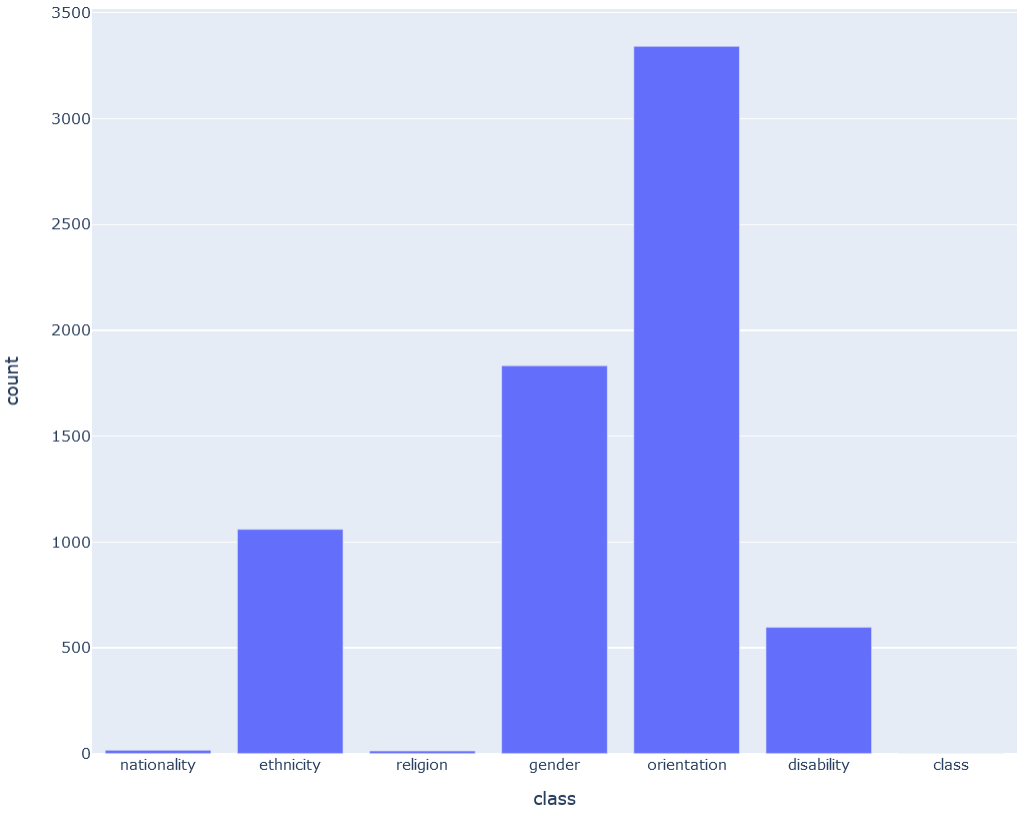

Figure 3.38: The classification of unambiguous hate words in tweets.

"queer" and "faggot". Transgender-related hate words do not make the list, other than "shemale". Finally, "globalist" still tops the nationality and religion-related words.

Figure 3.38 shows the distribution of unambiguous hate words in tweets. Gender and orientation share "shemale" and "tranny", while orientation gets a boost from "faggot". Finally, "nigger" and variants are the only slurs in ethnicity, and "retarded" is the only word that shows up in class.

Chapter 4

Conclusion

4.1 Research Findings

We summarize our results with respect to each of our research questions.

RQ1 - How do Parler users use the platform on average? More specifically, what kind of phrases do Parler users use the most, and what are the common topics of discussion? How does this change between parleys (main posts made by a user) and comments (replies to others' posts)?

Overall, we find that both parleys and comments from Parler use primarily political words and phrases, with a substantial emphasis on Donald Trump and the 2020 US election. Similarly, most of the most common topics found by LDA in both portions of the Parler dataset are political, especially related to the election.

RQ2 - How hateful are Parler users, i.e., how many hate words do they use overall?

We find that 2.5% of comments contain a hate word, while 2.05% of parleys contain a hate words. This means that comments have 20% more hate words than parleys, which increases to 76% more when looking at unambiguous hate words. When looking more specifically at the types of words used, we find that both parleys and comments tend to have the most hate words related to gender, disability, and ethnicity, although parleys have more hate words directed at nationality and religion. Overall, the top words themselves are largely similar among both parts of the dataset.

RQ3 - How do topics of discussion and hatefulness of parleys and comments vary with respect to time and their popularity? Are more hateful posts more likely to be popular?

Parleys show sharp increases in discussions of certain terms around specific political events such as the Black Lives Matter protests during the summer of 2020 and the release of emails from Hunter Biden in October 2020. They also show an increase in some election-related terms after the election was over, showing the spread of the narrative that the election was stolen. This trend is exaggerated in Parler comments, which see huge increases in many of these terms after the election. We see that in parleys, election-related phrases are more likely to show up in the more popular posts. However, many similar phrases instead show up more often in comments with an overall negative score, possibly indicating a backlash to the popular narrative. Other phrases such as "conspiracy theory" primarily showing up in downvoted comments also indicate a non-conforming population on Parler.

In respect to hate speech, we see that hate words appear more often in unpopular parleys and comments, and decrease in prevalence as the popularity increases. This is especially true in respect to Parler comments. This indicates that although there is a substantial community that uses hate words on Parler, it does not necessarily reach to a mainstream status.

RQ4 - How do the above findings compare to sampled Twitter posts from the same period? Are Parler posts more hateful than tweets, and are the topics discussed more political?

Many of the most common words and phrases in tweets are more related to trends and events in popular culture. Similarly, most of the topics discovered by LDA are related to popular culture. It is possible that some of this result comes from the design of Twitter, which makes popular trends and hashtags easy to find and join in on. We also find that many of these popular culture phrases on Twitter spike in usage over a few days, indicating a short-lived trend rather than an extended topic of discussion. In contrast, many of the election-related phrases occur in a consistent amount of posts until the election happens, at which point they decrease substantially. Finally, we see that many of the popular culture phrases primarily show up in tweets with few retweets, while many of the election-related phrases show up in popular tweets.

Only 1.44% percent of tweets contain hate words, substantially lower than either part of the Parler dataset. There is a more marked difference when it comes to unambiguous hate words, as Parler comments have 600% more than tweets. These results do not show much of an interpretable trend with respect to either time or popularity.

In conclusion, we see that Parler is both more hateful and more political than Twitter. This is perhaps unsurprising, given its founding as a less-moderated alternative to Twitter and subsequent endorsements by a variety of US conservative personalities. We do note that the use of hate words seems to be more confined to unpopular posts and users.

Additionally, there seems to be a significant backlash to some of the more outlandish claims about election fraud and larger conspiracies, indicating a subcommunity that does not share the same political opinions as the majority of Parler.

4.2 Further Work

Our paper provides an initial high-level characterization of this new dataset, but there is a lot of room for more fine-grained analysis. For example, we see a large amount of discussion of Donald Trump on Parler, but in light of the indications of a dissenting subcommunity, it could be interesting to perform sentiment analysis on these posts to get a look at how people view him. This could then be contrasted with a similar analysis of tweets that discuss Donald Trump to attempt to get an overall picture of how each platform feels about him. This deeper analysis could also be done on other election-related topics, such as voter fraud, Joe Biden, and more.

Another area of further study would be around the analysis of hate itself. As mentioned in section 1.2 of the introduction, our use of Hatebase is fairly limited in scope, and merely looks at the usage of words. There could be further work around both reducing the amount of false positives found by this analysis and including types of hatred that this analysis does not find. A simple addition could be using a dictionary of violent words and phrases to augment our search for hate words. This could capture more posts that could be seen as unacceptable, but would inevitably run into the same issues as Hatebase - some phrases would either be used in a sarcastic manner or simply to mean something different than the violent meaning. In order to more robustly find and classify hateful posts, we think that some sort of human element would need to be added. This could be as simple as a larger-scale version of the analysis we performed for trimming down hate words, where we give some subset of the posts in a dataset that use each word/phrase to humans and ask them to identify whether they are hateful. This could then be used to determine a level of severity for each term, and perhaps create a cutoff for what words should actually be considered hateful, i.e. only if 40% or more of participants consider it to be hateful. Alternatively, we could randomly sample some subset of all posts and determine our hateful vocabulary from the posts most deemed as hateful, although this could run into issues of sample size. Although these methods would likely improve the accuracy of our analysis, they are of course limited by the makeup of people who rank these posts, and what exactly they consider hateful. Although likely infeasible, one of the best methods would be to gather all posts from a given social media site and have each post rated by a diverse group of people on some straightforward scale as hateful, violent, etc. A robust machine learning

classifier could then be trained on this data in order to more accurately predict whether a post is hateful and thus determine the level of hate on a given site. Even in this ideal case, however, it would be unlikely to be perfect. The content and structure of posts often changes substantially both between sites and over time, requiring changes and updates.

One aspect of the Parler dataset that we do not use is the data collected on the users themselves. Although we perform analysis of posts based on the popularity of individual posts, it could be interesting to look at posts through the lens of popular and unpopular users instead. Additionally, it may be interesting to look at changes in the makeup of the userbase over time, both in terms of how often they use the platform and in the types of posts that they make.

There could also be additional refinements to the datasets themselves. As mentioned in section 2.4, we removed a large number of automated posts from the Parler dataset. However, it is possible that there are additional automated posts on a smaller scale still remaining. A similar cleanup could be done on the Twitter dataset, although we did not find similar evidence of automated posts on the same scale as in Parler. The Twitter dataset could also be expanded, as it only includes 1% of posts from October to December 2020. This fails to capture some important events that were captured in the Parler dataset, such as the Black Lives Matter protests. It also could be expanded to use a larger version of the stream, thus hopefully reducing the variance of the sampling methods used to create our dataset.

4.3 Thoughts

Our methods overall are fairly simplistic, and although they do capture the most popular topics of discussion, they will necessarily miss more diffuse or less discussed topics. They also do not say anything about the context of the topics, and in what sense people are discussing them (critically, positively, angrily, etc). Similarly, although we are able to see the amount of hate words used on each site, as well as what kinds of words are used, we do not have the full context for each usage. These methods are thus more suited to providing a high-level picture of the general trends and attitude towards specific hate words of different social media sites than characterizing a site or userbase as a whole. We do think this is still worthwhile, and can be used to examine general patterns of use in new social media sites. In particular, examining how these trends change both over time and with respect to measures of popularity help give nuance to this type of analysis, and can be used to see how the userbase of a site changes over time.

Another topic of discussion comes from an obvious question - what can we do with analysis of this kind? It seems obvious that hate speech is a problem on social media overall, and various social media sites have started to focus more and more on stopping it on their sites. There has thus been an increasing interest in identifying and removing hate speech. However, two issues arise from this. First, we can see from many of the sites discussed here that even if mainstream sites successfully crack down on language of this kind, more sites will spring up that will cater to the people banned or censored by these crackdowns. As with this analysis, analysis of these kinds of sites are unlikely to convince their creators that anything needs to change, but they can be used to examine the effects of different moderation policies and advocate for external change. Second, as we have discussed substantially, actually identifying hateful posts in any kind of automated manner is extremely difficult, and will likely lead to false positive identifications as well as miss posts intended to harm others. This challenge is exacerbated by the tendency of language to change over time. People who intend to spread hate invent new terms over time, and will adapt their behavior to get around systems of identification. On the other end, attitudes towards some words and sentiments will change over time, as some hateful words become archaic and rarely used while others are reclaimed by those intended to be their targets.

Given these difficulties, it is worthwhile to take a step back and look at what exactly the goals of hate speech policies are. Given the Western focus on the importance of freedom of speech, what is the actual harm in allowing hate speech to exist on platforms? The negative consequences of hate speech seem to largely fall into two categories. The first is hate speech as a means of harassment or way of signalling to certain demographics that they do not belong in a community. In this case, most of the consequences are still in a certain online community. The goal here would thus be ensuring that all members of an online community feel safe and respected, and are not afraid to speak about their viewpoints. The second consequence would be the use of hate speech to normalize violent viewpoints. The argument here is that allowing communities that speak in hateful ways to exist will eventually result in members of that community enacting those viewpoints in the outside world, and so hurting or even killing people. There is some evidence for this, such as a user on Gab posting violent anti-Semitic content and then becoming a mass shooter [26] and the planning of portions of the US Capitol insurrection attempt on Parler [10]. However, it is extremely difficult to actually approximate someone's mental state is based on the text that they post online, and it is of course nearly impossible to say what exactly influences people to behave in exactly the ways they do. It is also worth noting that most of these incidents come from sites that have been created precisely for users to post whatever sentiments they choose. In this case, analysis of the kind in this paper is

unlikely to matter to how they manage their site, as it is not seen as an actual problem.

To us, one of the main solutions to this kind of problem lies in combining the sort of overall analyses we use in this thesis with more fine-grained analyses in order to get a large-scale look at how users use a site and what the major issues are. For sites that are interested in change, this would have to be combined with a healthy amount of internal human intervention to help provide the context that can often be hard to introduce fully into automated systems. It is also important to explicitly set the goals for what exactly a particular social media site should be, and what kinds of behavior are thought of as acceptable. These analyses can also be used to provide an objective look at the conditions created by different varieties of moderation policies in order to understand what the actual effects of these policies are.

References

[1] Archive Team: The Twitter Stream Grab. https://archive.org/details/twitterstream, accessed 10-15-2021.

[2] GAB, Jul 2017. https://web.archive.org/web/20170712215207/https://www.startengine.com/startup/gab, accessed 11-08-2021.

[3] How twitter handles abusive behavior | twitter help, 2021. https://help.twitter.com/en/rules-and-policies/abusive-behavior, accessed 10-25-2021.

[4] Twitter's policy on hateful conduct | twitter help, 2021. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, accessed 10-25-2021.

[5] Sweta Agrawal and Amit Awekar. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer, 2018.

[6] Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. An early look at the parler online social network.

[7] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019.

[8] Eyal Arviv, Simo Hanouna, and Oren Tsur. It's a thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 2020.

[9] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[10] Aleszu Bajak, Jessica Guynn, and Mitchell Thorson. When Trump started his speech before the Capitol riot, talk on Parler turned to civil war, Feb 2021. https://www.usatoday.com/in-depth/news/2021/02/01/civil-war-during-trumps-pre-riot-speech-parler-talk-grew-darker/4297165001/, accessed 08-02-2021.

[11] Urmimala Banerjee. Bigg Boss 14: Rubina Dilaik fans trend 'roar like Rubina' to applaud the actress' fearless stance on the show - read tweets, Dec 2020. https://www.bollywoodlife.com/bigg-boss/bigg-boss-14-rubina-dilaik-fans-trend-roar-like-rubina-to-applaud-the-actress-fearless-stance-on-the-show-read-tweets-1748370/, accessed 09-18-2021.

[12] Michael Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 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*, volume 5, 2011.

[13] Kayvon Beykpour and Vijaya Gadde. Additional steps we're taking ahead of the 2020 US election, Oct 2020. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes, accessed 10-21-2021.

[14] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[16] Kate Cox. Reddit clone Voat, home to hate speech and Qanon, has shut down, Dec 2020. https://arstechnica.com/tech-policy/2020/12/reddit-clone-voat-home-to-hate-speech-and-qanon-has-shut-down/, accessed 11-08-2021.

[17] Bridget Coyne and Sam Toizer. Helping you find accurate US election news and information, Sep 2020. https://blog.twitter.com/en_us/topics/company/2020/2020-election-news, accessed 10-21-2021.

[18] Drew Harwell Craig Timberg. Parler's got a porn problem: Adult businesses target pro-Trump social network, Dec 2020. https://www.washingtonpost.com/technology/2020/12/02/parler-pornography-problem/, accessed 08-01-2021.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[20] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[21] Caitlin Fitzsimmons. YouTube besieged by Porn Videos, May 2009. https://www.theguardian.com/media/2009/may/22/youtube-porn-day, accessed 11-08-2021.

[22] Claudia Flores-Saviaga, Brian Keegan, and Saiph Savage. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[23] The Smoking Gun. Another 4chan user gets busted by FBI, Feb 2011. https://www.thesmokinggun.com/documents/internet/another-4chan-user-gets-busted-fbi, accessed 11-08-2021.

[24] Robert Hart. Parler's Popularity Plummets As Data Reveals Little Appetite For Returning 'Free Speech' App Favored By Conservatives, Jun 2021. https://www.forbes.com/sites/roberthart/2021/06/02/parlers-popularity-plummets-as-data-reveals-little-appetite-for-returning-free-speech-app-favored-by-conservatives/?sh=4bbbc3295e13, accessed 08-10-2021.

[25] Hatebase. Hatebase API. https://hatebase.com/, accessed 08-21-2021.

[26] Laura Hensley. Right-wing platform Gab taken down after pittsburgh shooting, says it's been 'smeared' by media - national, Oct 2018. https://globalnews.ca/news/4606576/gab-com-officially-taken-offline/, accessed 10-24-2021.

[27] Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[28] Steve Huffman. Update to Our Content Policy, Jun 2020. https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/, accessed 10-20-21.

[29] Mike Isaac and Kate Conger. Reddit bans forum dedicated to supporting Trump, and Twitter permanently suspends his allies who spread conspiracy theories., Jan 2021. https://www.nytimes.com/2021/01/08/us/politics/reddit-bans-forum-dedicated-to-supporting-trump-and-twitter-permanently-suspends-his-allies-who-spread-conspiracy-theories.html, accessed 09-12-2021.

[30] Rachel Lerman. The conservative alternative to Twitter wants to be a place for free speech for all. it turns out, rules still apply., Nov 2020. https://www.washingtonpost.com/technology/2020/07/15/parler-conservative-twitter-alternative/, accessed 08-01-2021.

[31] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

[32] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[33] Adrienne Massanari. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3):329–346, 2017.

[34] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

[35] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892, 2013.

[36] Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. "and we will fight for our race!" a measurement study of genetic testing conversations on reddit and 4chan. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 452–463, 2020.

[37] Abby Ohlheiser. Fearing yet another witch hunt, Reddit bans 'Pizzagate', Nov 2016. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/23/fearing-yet-another-witch-hunt-reddit-bans-pizzagate/, accessed 10-24-2021.

[38] Antonis Papasavva, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. "is it a qoincidence?": An exploratory study of qanon on voat. In *Proceedings of the Web Conference 2021*, pages 460–471, 2021.

[39] Parler. About Parler. https://parler.com/main.php, accessed 07-20-2021.

[40] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47, 2020.

[41] Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 2021.

[42] Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 2020.

[43] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[44] Everett Rosenfeld. Mountain Dew's 'dub the dew' online poll goes horribly wrong, Aug 2012. https://newsfeed.time.com/2012/08/14/mountain-dews-dub-the-dew-online-poll-goes-horribly-wrong/, accessed 11-08-2021.

[45] Twitter Safety. Updating our rules against hateful conduct, Dec 2020. https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate, accessed 10-25-2021.

[46] A Data Scientist. A Parler Archive. https://parler.adatascienti.st/, accessed 10-20-2021.

[47] Bijan Stephen. Reddit's qanon ban points to how it's tracking toxic communities, Sep 2018.

[48] Twitter. Sampled stream introduction | Docs | Twitter Developer Platform. https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction, accessed 08-28-2021.

[49] 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*, pages 88–93, 2016.

[50] Emma Woollacott. Users flock to Voat as Reddit shuts harassing groups, Jul 2015. https://www.forbes.com/sites/emmawoollacott/2015/06/11/users-flock-to-voat-as-reddit-shuts-harassing-groups/?sh=4ffd83ff1199, accessed 11-08-2021.

[51] Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.

[52] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014, 2018.

[53] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202, 2018.

[54] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 internet measurement conference*, pages 405–417, 2017.

[55] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 786–797, 2020.

APPENDICES

Appendix A

Glossary of relevant social media websites

- **Reddit**: Founded in 2005, Reddit is one of the most popular sites in the world. It is split into a number of "subreddits", each of which is focused on discussion of specific topics. Users can subscribe to any number of these, customizing the kinds of posts they see. Moderation on the site is largely by volunteer moderators for each subreddit, with Reddit itself mostly only stepping in in extreme cases. Reddit has gotten increased media attention over the last few years for the number of hateful subreddits on the site. Some researchers have blamed the structure of Reddit for these groups [33], especially related to extended harassment campaigns such as GamerGate. This has resulted in recent rules changes [28] and bans of various subreddits [37] [47] to try to stop this behavior. It can be found at `reddit.com`.

- **4chan**: Created in October 2003, 4chan is most noted for its anonymity and ephemerality. Users do not need an account to post on the site, and so by default are completely anonymous. The threads of posts that users post are also deleted after enough inactivity, making it a place where users can post any kind of content without it being traced back to them or any record of their post existing. It is split into a number of imageboards, each dedicated to discussion of a specific topic, ranging from general (such as the /b/ or "Random" board) to specific (such as the /sp/ or "Sports" board). It has a few base rules (such as not posting illegal content), but is largely free of moderation. 4chan has been linked to a number of coordinated attacks [21] or pranks [44] on other websites, and has been the subject of media scrutiny around posting of offensive or illegal content [23]. It can be found at `4chan.org`.

- **Gab**: Launched in August 2016, Gab bills itself as a social media alternative for "conservative, libertarian, nationalist, and populist internet users from around the world who are seeking an alternative to the current social networking ecosystems." [2]. Its design is very similar to Twitter, with users able to create posts that are displayed on an updating feed to other users who follow them. As discussed above, Gab was eventually deplatformed and left by many after the discovery of its use by an anti-Semitic mass shooter. [26] As of the publishing of this work, Gab is still active, and can be found at gab.com.

- **Voat**: Created in April 2014 as Whoaverse, Voat was structurally very similar to Reddit, and was split into "subverses", each of which was based around discussion of a particular topic. It was designed to have looser moderation than Reddit, and experienced large user growth after Reddit banned various hateful subreddits [50]. Voat was shut down on December 25, 2020 due to lack of funding [16].

Appendix B

Discussion of automatically generated Parler comments

Here we look at the Parler comments we found that were clearly automated in some way, as they contain exactly the same text repeated by the same user hundreds of thousands of times. We noticed them after looking at the most popular phrases, where some phrases were clearly part of a sentence that seemed unlikely to organically be repeated. Fig B.1 shows the text of the comments that we found, as well as a count of how many times each was used. We note that there were likely more comments repeated on a smaller scale than the ones listed here, but that they are difficult to find without either time-prohibitive analysis of every single comment or extremely detailed examination of trends in phrases. We can see that most of them are welcoming comments, likely created in response to new users making their first post. This also includes a team of volunteers called the Parler Concierge, who posted helpful and welcoming content to new users. Many of the users making the posts are Parler staff, including the CEO. However, some are not, including the official Team Trump Parler account.

Figures B.2 and B.3 show the total number of identified automatic comments over time and the percentage of all comments that are an automatic comment over time. We can see two clear increases, one around June 2020 and one around November 2020. This matches with the content of the posts, since those months also correspond to the two greatest increases in users over time. Since these posts only start around that time, it is possible that they were implemented by Parler staff in response to complaints about the usability of the site by a large influx of new users. It is still unclear why only a few non-staff users have made them, however, and we were unable to find any official justification or acknowledgement of these posts.

Post	Count
"I've spent my adult life fighting for Liberty, and I am excited about this new space where I can openly share my work with special people like you. Parler accepts your right to express your thoughts, opinions and ideals online. No "throttling." No shadow banning. Just Free Speech—our God-given, Constitutional Right."	1624667
Welcome! For people looking for Parler tips and how tos Check out @parlersupport or the Parler Youtube page. The videos are in both places. There is also a Parler101 channel which has some good data as well.	1580986
Glad to see you here on Parler where free speech is actually alive and well. Looking forward to mixing it up with you here and keeping the truth alive in the face of the constant bias we face. Let's Go!	1490300
Welcome. Great to have you. Follow us for your favorite commentary.	1459760
Welcome to Parler! Help us MAKE AMERICA GREAT AGAIN by clicking the link below. Be sure to text TRUMP to 88022!	870983
I'm here to help you navigate the app, answer any questions, and ensure you have the best Parler experience. Puedo ayudar en Español también! #parlerconcierge	537365
Welcome to Parler. The first step towards taking back your voice. Thrilled to have you join our growing community! Share with your family & friends. Let's have some fun. Let's Parley!	466988
Welcome to Parler. I will see you in the comments. Please leave a review on the app stores and share with your friends!	250666
Welcome to Freedom!Let me know if you have any questions! #parlerconcierge	201834
Welcome to Parler. Hope you enjoy your new found freedom. Have fun, interact and enjoy. #newuser #truefreespeech #parlerconcierge	147674
Welcome to #Parler! We are glad you are here...please enjoy your freedom to speak! If you have any questions feel free to ask....someone will try to answer it for you #parlerconcierge #parlerusa #freedomfromcensorship	37

Figure B.1: Automatic Parler comments and the amount of times each appear. Some emojis are removed.

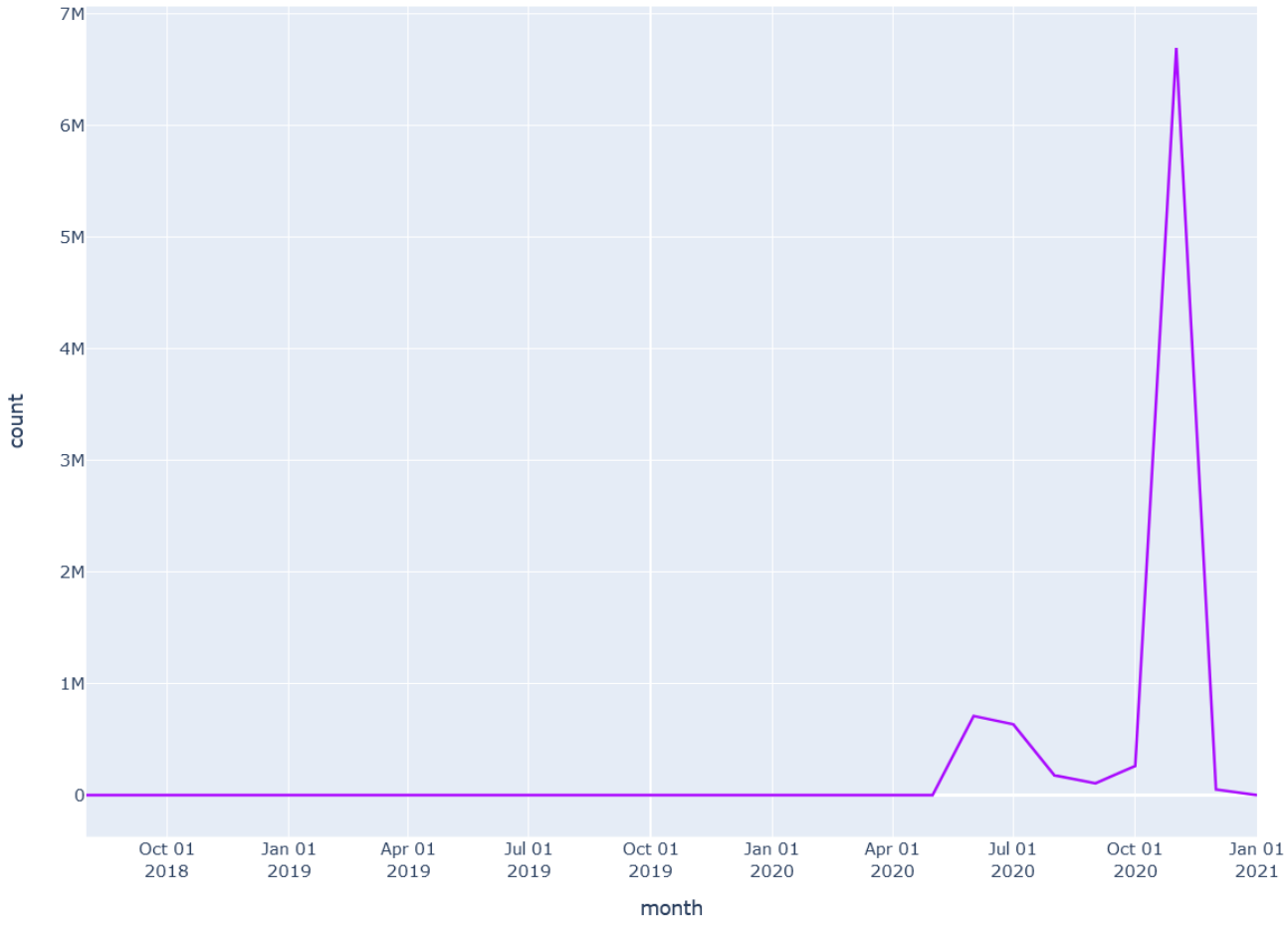

Figure B.2: The counts of identified automatic comments over time.

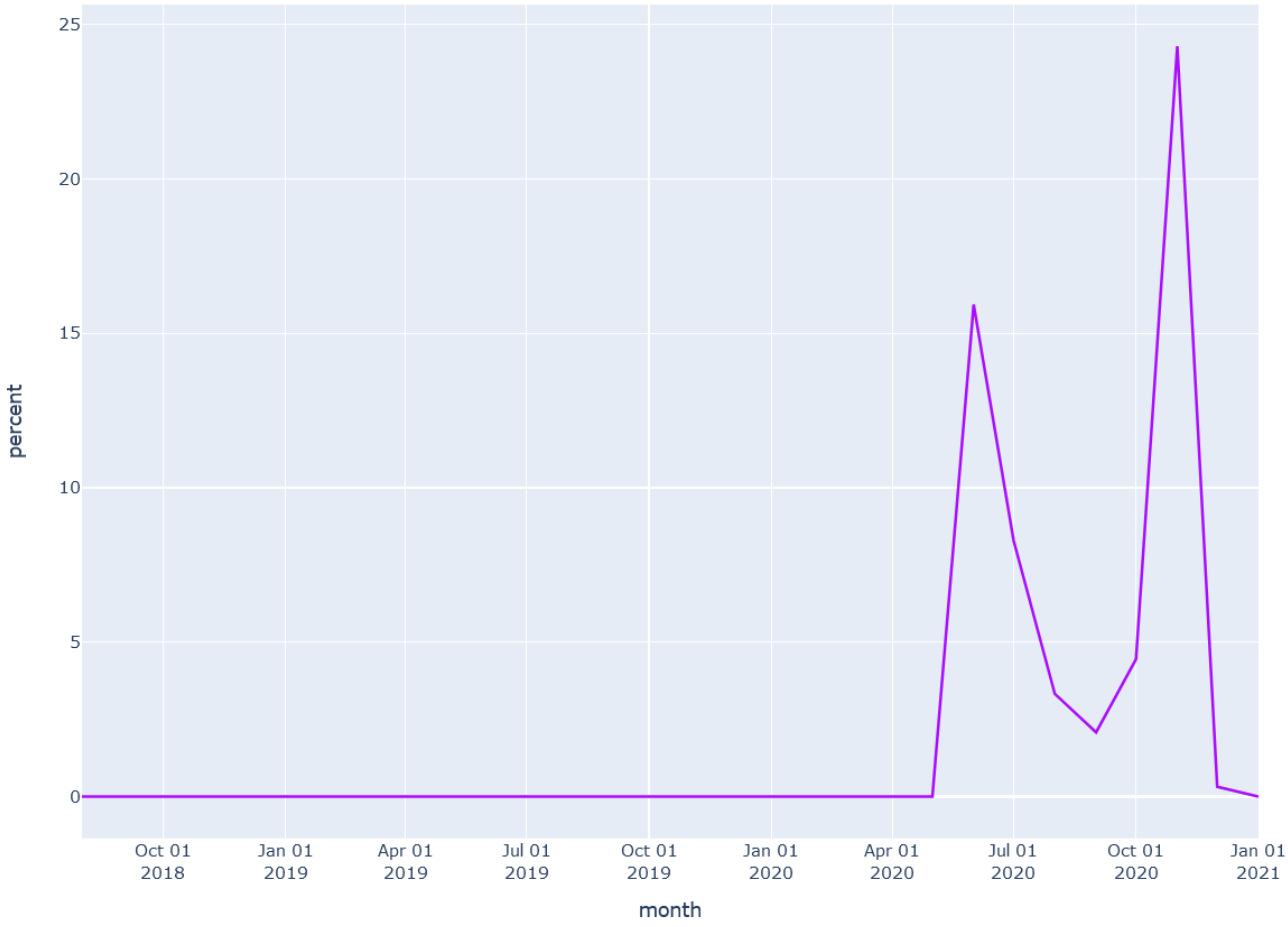

Figure B.3: The percent of posts in each month that are identified automatic comments.