Free speech or Free Hate Speech?
Analyzing the Proliferation of Hate Speech in Parler

Abraham Israeli isabrah@post.bgu.ac.il
Oren Tsur orentsur@bgu.ac.il
Department of Software and Information System Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel

Abstract

Social platforms such as Gab and Parler, branded as 'free-speech' networks, have seen a significant growth of their user base in recent years. This popularity is mainly attributed to the stricter moderation enforced by mainstream platforms such as Twitter, Facebook, and Reddit. In this work we provide the first large scale analysis of hate-speech on Parler. We experiment with an array of algorithms for hate-speech detection, demonstrating the limitations of transfer learning in that domain, given the illusive and ever changing nature of the ways hate-speech is delivered. In order to improve classification accuracy we annotated 10K Parler posts, which we use to fine-tune a BERT classifier. Classification of individual posts is then leveraged for the classification of millions of users via label propagation over the social network. Classifying users by their propensity to disseminate hate, we find that hate mongers make about 16% of Parler active users, and that they have distinct characteristics comparing to other user groups. We find that hate mongers are more active, more central, express distinct levels of sentiment, and convey a distinct array of emotions like anger and sadness. We further complement our analysis by comparing the trends observed in Parler to those found in Gab. To the best of our knowledge, this is among the first works to analyze hate speech in Parler in a quantitative manner and on the user level.

1 Introduction

Social platforms like Twitter, Facebook, and Reddit have become a central communication channel for billions of users.[1] The immense popularity of social platforms resulted in a significant rise in the toxicity of the discourse, ranging from cyberbullying to explicit hate speech and calls for violence against individuals and groups (Waseem and Hovy, 2016; Mondal et al., 2017; Laub, 2019; Ziems et al., 2020). Women, people of color, the LGBT community, Muslims, immigrants, and Jews are among the most targeted groups. Recent studies report on a surge in Islamophobia (Akbarzadeh, 2016; Sunar, 2017; Osman, 2017; Chandra et al., 2021), antisemitism (ADL, 2020; Zannettou et al., 2020), xenophobia (Iwama, 2018; Entorf and Lange, 2019), hate of Asians (An et al., 2021; Vidgen et al., 2020a) and hate crimes (Dodd and Marsh, 2017; Levin and Reitzel, 2018; Edwards and Rushin, 2018; Perry et al., 2020).

Facing an increased public and legislature scrutiny, mainstream social platforms (e.g., Facebook, Twitter, Reddit) committed to a stricter enforcement of community standards, curbing levels of hate on the platform.[2]

The stricter moderation of content drove many users into joining alternative social platforms such as Parler and Gab. Touting their commitment to 'free speech' and 'no moderation' policy, these platforms attract users that were suspended from mainstream platforms, conspiracy theorists, extremists, unhinged users, free-speech advocates, political activists as well as others.

User migration to Parler and Gab was not only grass-root. The platforms were promoted by prominent news anchors and political figures. For example, U.S. Senator Ted Cruz (R-TX) tweeted *"I'm proud to join @parler_app – a platform gets what free speech is all about – and I'm excited to be a part of it. Let's speak. Let's speak freely. And let's end the Silicon Valley censorship"* (6/25/2020). Sean Hannity, a popular host and commentator on Fox news, informed the viewers of his daily show that *"I saw that the president had joined it. At least there is a place, it's like Twitter, it's called Parler, I have an account there... good for you because the president joined, because they are censoring him and Dan Scavino and everybody else"* (1/8/2021).

[1] E.g., Facebook 2021 Q2 report (Meta, 2021a).

[2] e.g., Facebook 2021 report on hate-speech (Meta, 2021b), and the Time magazine cover of hate-speech in Twitter: (Time, 2021).

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 109 - 121
July 14, 2022 ©2022 Association for Computational Linguistics

Hate, brewing online, often spills to the streets (Hankes and Amend, 2019; Munn, 2019; Malevich and Robertso, 2019; Thomas, 2019). Thus, defending 'hate speech' under the right for 'free speech' may result in very concrete actions in real life. The perpetrator of the Pittsburgh synagogue shooting[3] was active on Gab, referring to "kike infestation" and "the children of satan". His final post, minutes before opening fire in the synagogue, was *"I can't sit by and watch my people get slaughtered. Screw your optics, I'm going in."*. Similarly, the storming of the U.S. Capitol on January 6, 2021 was found by the U.S. Senate Investigation Committee to be encouraged and coordinated on Parler (Peters et al., 2021).

In this work we focus on Parler, investigating the proliferation of hate speech on the platform, both on the post level and on the user level. We identify three distinct groups of users, denoted as hate mongers, standard users and hate flirts. We show significant differences between the groups in terms of language, emotion, activity level and role in the network. We further compare our results to the hateful dynamics reported for Gab.

2 Related Work

A growing body of work studies the magnitude and the different manifestations of hate speech in social media (Chandrasekharan et al., 2017; Zannettou et al., 2018; Zampieri et al., 2020; Ranasinghe and Zampieri, 2020), among others. Here, we present an overview of the current literature in three different perspectives: (i) The detection of hate speech on the *post* level, (ii) The detection of hate-promoting *users*, and (iii) The characterization of hate speech on the *platform* level.

Post-level classification Most previous works address the detection of hate in textual form. Keywords and sentence structure in Twitter and Whisper were used in (Mondal et al., 2017; Saleem et al., 2017), demonstrating the limitations of a lexical approach. The use of code words, ambiguity and dog-whistling, and the challenges they introduce to text-based models were studied by (Davidson et al., 2017; Ribeiro et al., 2017; Arviv et al., 2021). The detection of implicit forms of hate speech is addressed by Magu et al. (2017) which detects the use of hate code words (e.g., google, skype, bing and skittle to refer to Black people, Jews, Chinese, and Muslims, respectively) using an SVM classifier based on bag-of-words. ElSherief et al. (2021) introduced a benchmark corpus of 22.5K tweets to study implicit hate speech. The authors presented baseline results over this dataset using Jigsaw Perspective,[4] SVM, and different variants of BERT (Devlin et al., 2018).

The use of demographic features such as gender and location in the detection of hate speech is explored by Waseem and Hovy (2016). User meta features, e.g., account age, posts per day, number of followers/friends, are used by Ribeiro et al. (2017).

Computational methods for the detection of hate speech and abusive language range from SVM and logistic regression (Davidson et al., 2017; Waseem and Hovy, 2016; Nobata et al., 2016; Magu et al., 2017), to neural architectures. Recently, Transformer-based architectures (Mozafari et al., 2019; Aluru et al., 2020; Samghabadi et al., 2020; Salminen et al., 2020; Qian et al., 2021; Kennedy et al., 2020; Arviv et al., 2021) achieved significant improvements over RNN and CNN models (Zhang et al., 2016; Gambäck and Sikdar, 2017; Del Vigna12 et al., 2017; Park and Fung, 2017). In an effort to mitigate the need for extensive annotation some works use transformers to generate more samples, e.g., (Vidgen et al., 2020b; Wullach et al., 2020, 2021). Zhou et al. (2021) integrate features from external resources to support the model performance.

In order to account for the often elusive and coded language and for the unfortunate variety of targeted groups (Schmidt and Wiegand, 2017; Ross et al., 2017), a set of functional test was suggested by Röttger et al. (2020), allowing an quick evaluation of hate-detection models.

Classification of hate *users* Characterizing *accounts* that are instrumental in the propagation of hate is gaining interest from the research community and industry alike, whether in order to better understand the social phenomena or in order to suspend major perpetrators instead of removing sporadic content. Detection and characterization of hateful Twitter and Gab users was tackled by Ribeiro et al. (2018); Mathew et al. (2018, 2019) and Arviv et al. (2021), among others. An annotated dataset of a few hundreds Twitter users was released as part of a shared task in CLEF 2021, see (Bevendorff et al., 2021) for an overview of the data and the submissions. Das et al. (2021) intro-

[3] ADL report on the attack: https://tinyurl.com/yz87jn69 (accessed: 4/17/22)

[4] https://www.perspectiveapi.com

duced a user-level annotated dataset of 798 Gab users which we use for evaluation and comparison.

Hate speech on Parler and Gab While most prior work focus on the manifestations of hate in mainstream platforms, a number of works do address alternative platforms such as Gab and Parler. Two annotated Gab datasets were introduced by Kennedy et al. (2018) and by Qian et al. (2019). We use these datasets in this work as we compare Parler to Gab.

Focusing on users, rather than posts, Das et al. (2021) experiment with an array of models for hate users classification. Lima et al. (2018) aims to understand what users join Gab and what kind of content they share, while Jasser et al. (2021) conduct a qualitative analysis studying Gab's platform norms, given the lack of moderation. Gallacher and Bright (2021) explore whether users seek out Gab in order to express hate, or that the toxic attitude is adopted after joining the platform. The diffusion dynamics of the content posted by hateful and non-hateful Gab users is modeled by Mathew et al. (2019) and by Mathew et al. (2020).

Parler, launched in August 2018 and experiencing its impressive expansion of user base from late 2020, is only beginning to draw the attention of the research community. Early works analysed the language in Parler in several aspects such as QAnon content (Sipka et al., 2021), COVID-19 vaccines (Baines et al., 2021), and the 2021 Capitol riots (Esser, 2021). The first dataset of Parler messages was introduced by Aliapoulios et al. (2021), along with a basic statistical analysis of the data, e.g., the number of posts and the number of registered users per month, along with the most popular tokens, bigrams, and hashtags. We use this dataset in the current work to analyze hate speech on Parler. Ward (2021) used a list of predefined keywords (hate terms), assessing the level of hate-speech on the platform.

Our work differs from these works in a number of fundamental aspects. First, we combine textual and social (network) signals in order to detect both hateful posts and hate-promoting accounts. Second, we suggest models that rely on state-of-the-art neural architectures and computational methods, while previous work detects hate speech by matching a fixed set of keywords from a predefined list of hate terms. Furthermore, we provide a thorough analysis of the applicability of different algorithms, trained and fine-tuned on various datasets and tasks. Third, we provide a broader context to our analysis of the proliferation of hate in Parler, as we compare and contrast it to trends observed on Gab.

3 Data

In this section we describe the datasets used for this work – starting with a general overview of the platforms, then providing a detailed description of the datasets and the annotation procedure.

3.1 Parler and Gab Social Platforms

Parler Alluding to the french verb 'to speak', Parler was launched on August 2018. The platform brands itself as "The World's Town Square" a place in which users can *"Speak freely and express yourself openly, without fear of being "deplatformed" for your views"*.[5]

Parler users post texts (called *parlays*) of up to 1000 characters. Users can reply to parlays and to previous replies. Parler supports a reposting mechanism similar to Twitters retweets (called 'echos'). Throughout this paper we refer to echo posts as *reposts*, not to confuse with the ((())) (echo) hate symbol (Arviv et al., 2021).

Parler's official guidelines[6] explicitly allow "trolling" and "not-safe-for-work" (NSFW) content, include only two "principles" prohibiting "unlawful acts", citing "Obvious examples include: child sexual abuse material, content posted by or on behalf of terrorist organizations, intellectual property theft".

By January 2021, 13.25M users have joined Parler and its application was the most downloaded app in Apple's App Store. This growth is attributed to celebrities and political figures promoting the platform (see Section 1) and the stricter moderation enforced by Facebook and Twitter, culminating with the suspension of Donald Trump (`@realDonaldTrump`), the 45th President of the United States, from Twitter and Facebook.

Gab Gab, launched on August 2016, was created as an alternative to Twitter, positioning itself as putting "people and free speech first", welcoming users suspended from other social networks (Zannettou et al., 2018). Gab posts (called *gabs*) are limited to 300-characters, and users can repost, quote or reply to previously created gabs. Gab permits

[5] Parler branding on its landing page (accessed: 3/10/2022)

[6] https://parler.com/documents/guidelines.pdf (accessed: 4/17/2022)

	Parler	Gab
Users	4.08M	144.3K
Posts	20.59M	7.95M
Replies	84.55M	5.92M
Reposts	77.93M	8.24M
Time-Span	08/2018 – 01/2021	08/2016 – 01/2018

Table 1: Datasets Statistics. Replies are responses to main posts. Reposts are equivalent to Twitter retweets.

pornographic and obscene content, as long as it is labeled *NSFW*. Previous work finds the majority of Gab users to be Caucasians-conservatives-males (Lima et al., 2018). For more details about Gab usage, users and manifestations of hate see references at Section 2.

3.2 Parler and Gab Corpora

We use the Parler and Gab datasets published by Aliapoulios et al. (2021) and Zannettou et al. (2018), respectively. The Parler dataset is unlabeled, therefore annotation is required. We describe the annotation procedure and label statistics in Section 3.3.

Both datasets include posts and users' meta data, though the Parler dataset is richer, containing more attributes such as registration time. Each of the datasets is composed of millions of posts and replies, see Table 1. The Parler dataset is bigger, containing more posts and more users, however, on average, Gab users post more content per user. We note that there is no temporal overlap between the two datasets. In Section 7 we discuss this point and its possible impacts on our analysis.

We use three Gab *annotated* datasets which are all sampled from the unlabeled Gab corpus we use: (i) The Gab Hate Corpus – 27.5K Gab posts published by Kennedy et al. (2018), (ii) 9.5K Gab posts published by Qian et al. (2019), and (iii) 5K posts published by Arviv et al. (2021).[7]. In total, we collect a corpus of 42.1K annotated Gab posts. 7.7K (18.4%) of the posts are tagged as hateful.

3.3 Annotation of Parler Data

Hate speech takes different forms in different social platforms (Wiegand et al., 2019) and across time (Florio et al., 2020). It is often implicit (ElSherief et al., 2021), targeting a variety of groups. Consequently, transfer learning remains a challenge for hate-speech detection, and an annotated Parler dataset is needed in order to achieve accurate classification. These challenges and the significant improvements in performance achieved by proper fine-tuning are demonstrated through extensive experimentation in Section 4.1. In the remainder of this section we describe the annotation procedure.

The annotation task was designed as follows: 10K posts were sampled from the Parler corpus. All posts are: (i) in English; (ii) at least 10 characters long; (iii) neither a repost nor a comment; and (iv) do not contain a URL.

The 10K annotated posts *were not* randomly selected from the Parler corpus. A random selection of posts would have led to an extremely imbalanced dataset as most of the posts are not expected to express hate. Hence, we opt to stratified sampling. This sampling process relies on an approximation of the likelihood of each post to include hateful content. We used a pretrained hate speech prediction model to approximate this likelihood.

Annotation was done by 112 student (more than half of them are graduate students), who were provided detailed guidelines and training involving the various types of hate speech, the elusiveness of hate expressions using coded language, how to detect it, and a number of examples of different types. Each of the annotators was prompted with a list of 300 posts and had to assign each with a Lickert score ranging from 1 (not hate) to 5 (extreme or explicit hate). We provided annotators only with the textual content of the post. Each of the 10K posts was annotated by three annotators. Annotators presented a satisfying agreement level of 72% and a Cohen's Kappa of 0.44. Labels of posts with a low agreement level[8] were ignored (~7% of the annotated posts). We define a post as hateful (non-hateful) if its average score is higher (lower) than three. We omit posts with an average score of exactly three. Accordingly, 3224 of the 10K posts (32.8%) were labeled as hateful and 6053 (59.8%) as non-hateful.

We make this annotated corpus available under our public GitHub repository[9] – the first public annotated corpus of Parler.

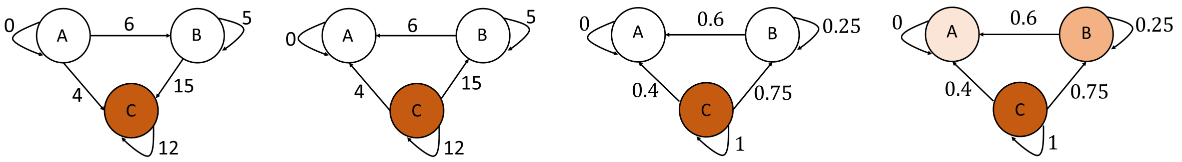
4 Methods

In this work we are interested in the detection of hate, both on the post level and the account

[7] This work models Twitter data but also published an annotated dataset of Gab

[8] Low agreement is defined as either an annotation with at least three different Likert values, or a difference greater than 2 between the Likert values.

[9] https://github.com/NasLabBgu/parler-hate-speech



(a) Repost Network (b) Belief Network (c) Belief Network (Normalized) (d) Belief Network (after diffusion)

Figure 1: An illustration of the diffusion model over three nodes. Self loops represent the total number of posts per node. In step (a) The repost network is built and nodes are assigned an initial belief – seed hate mongers with a value of 1 (orange) and others with a value of 0 (white). In steps (b) and (c) The network is converted to a belief network – reversing edges direction and normalizing weights. In step (d) The diffusion process is simulated. Belief updates are indicated by darker shades.

level. Our interest in the post level classification is twofold. Given an accurate classifier, we can: (a) Approximate the hate degree in different aggregation levels – e.g., over the full network or and per user, and (b) Use the post-level predictions to support training a user level classifier. A review of the various post level classifiers is provided in Section 4.1 and our modifications to a diffusion-based model for user classification are presented in Section 4.2. Ethical considerations related to user classification are discussed at the end of Section 7.

4.1 Post Level Classification Models

We fine-tune the DistilBERT (Sanh et al., 2019) transformer on each of the datasets, obtaining two fine-tuned models (referred to as Our-FT BERT). We compare the models performance on the respective datasets against four competitive models:

1. **Jigsaw Perspective**: A widely used commercial model geared toward detection of hateful and toxic content, developed by Google. Jigsaw was found to perform well in an array of tasks related to hate-speech detection (Röttger et al., 2020). Jigsaw implementation is not public and the service is provided as a black-box through an online API.[10]

2. **deHateBERT** (Aluru et al., 2020): An adaptation of the BERT Transformer for hate-speech detection – the pretrained transformer was fine-tuned on a corpus of 96.3K text snippets from Twitter and from the white supremacist forum Stormfront.org. The authors indicate that 15.01K (15.6%) training samples were labeled as hate-speech.

3. **Twitter-roBERTa** (Barbieri et al., 2020): This model uses the RoBERTa (Liu et al., 2019) architecture, specifically fine-tuned on the task of hate-speech detection of micro-messages. The authors used a corpus of 13K tweets, 5.2K (40%) of them are labeled as hate speech.

4. **HateBase** (Tuckwood, 2017): HateBase is a multilanguage vocabulary of hate terms that is maintained in order to assist in content moderation and research. We use 68 explicit hate terms that were used in prior works (Mathew et al., 2018, 2019). These terms were manually selected from HateBase's English lexicon. All the terms in the list are *explicit*, e.g., 'kike' (slur targeting Jews), 'paki' (slur against Muslims, especially with Pakistani roots), and 'cunt'. Text is labeled as hate if it contains at least a single hate term.

4.2 User Level Classification

In order to leverage the network structure, we view each platform as a social network with users as nodes and *reposts* as directed edges. Edges are weighted to reflect levels of engagement, as illustrated in Figure 1(a): a directed edge (A, B) with a weight of 6 indicates that user A reposted 6 posts originally posted by user B.

We modify the diffusion-based approach for the detection of hate mongers proposed by Ribeiro et al. (2018) in order to achieve a more accurate classification. The basic diffusion-based classification is performed in two stages: (a) Identifying a *seed* group of hate mongers; and (b) Applying a diffusion model over the social network. We use the DeGroot's hate diffusion model (Golub and Jackson, 2010) which outputs an estimated belief value (i.e., "hate") per user, over the [0,1] range. A toy example of the diffusion process is illustrated in Figure 1. In our experiments we set the number of diffusion iterations to three. One clear advantage of this approach over fully supervised methods is that it does not require a large dataset annotated on the user level.

[10] https://www.perspectiveapi.com

Modified Diffusion Model We introduced two modifications to the diffusion model used by Ribeiro et al. (2018) and Mathew et al. (2019): (i) *Seed definition*: Instead of taking a lexical approach in order to identify users posting more than *k* hateful posts, we use our fine-tuned Transformers. We argue that fine-tuning the classifiers for each social network significantly improves the classification on the post level (as demonstrated in Section 5.1), and ultimately, improves the performance of the diffusion model; and (ii) *Hateful users definition*: In the original diffusion process, hate (as well as "not-hate") labels are diffused through the network. This way, seed hate mongers may end with a low hate score, which in turn propagates to their neighbours. However, seed users were chosen due to the fact that they post a significant number of undoubtedly hateful posts. Fixing the hate score of seed users results in a more accurate labeling of the accounts in the network.

5 Classification Results

5.1 Post Level Results

We use the annotated corpora (see Sections 3.2 and 3.3) to fine-tune the pretrained Transformer on each social platform, splitting the labeled data to train (60%), validation (20%), and test (20%) sets.

The precision-recall curves of the Parler and Gab models are presented in Figure 2. Our fine-tuned models significantly outperforms the other models in both datasets. We wish to point out that while the popular keywords base approach (HateBase) achieves a high precision and a moderate recall on the Gab data, outperforming all Transformer models except the platform fine-tuned ones, it collapses in both measures on the Parler dataset. These results revalidate the limitations of lexical approaches, and of neural methods that are not fine-tuned for the specific dataset.

5.2 User Level Results

As described in Section 4.2, in order to classify accounts we use a diffusion model. The diffusion process is seeded with a set of hateful accounts. The choice of seed accounts involves the following steps: (i) After establishing the accuracy of the fine-tuned models (Section 5.1) we use these models to label *all* the posts in the respective datasets; (ii) Opting for a conservative assignment of seed users, we consider only posts with hate score (likelihood) over 0.95 (0.9) in the Parler (Gab) dataset to be hateful. This threshold setting yields a precision of 0.801 (0.902) and a recall of 0.811 (0.903) over the Parler (Gab) dataset.[11]; Finally, (iii) Users posting 10 or more hateful posts are labeled as seed accounts. We take the conservative approach in steps (ii) and (iii) in order to control the often noisy diffusion process.

Simulating the modified diffusion process described in Section 4.2 we obtain a hate score per *user*. For analysis purposes we divide users to three distinct groups – hate mongers (denoted HM), composed of the users making the top quartile of hate scores; Standard users (denoted S) making the bottom quartile; the rest of the users (denoted $\widetilde{HM}$) suspected as "flirting" with hate mongers and hate dissemination. Users with a low level of activity (less than five posts or users who joined the network less than 60 days prior to data collection) were not considered.[12] The distribution of *active* users by type in Parler is 16.1%/42.4%/41.5% per $HM/\widetilde{HM}/S$ populations, and 10%/41.7%/48.3% in Gab.

Evaluation of the diffusion model A user-level annotated dataset of 798 Gab users was shared by Das et al. (2021). We use this dataset to validate the performance of the diffusion models – both the standard model and our modified model (see Section 4.2). We find our modified model to outperform the standard model, achieving precision/recall/F1-scores of 0.9/0.54/0.678, comparing to 0.95/0.34/0.5. Therefore, results and analysis in the remainder of the paper are based on the modified diffusion model.

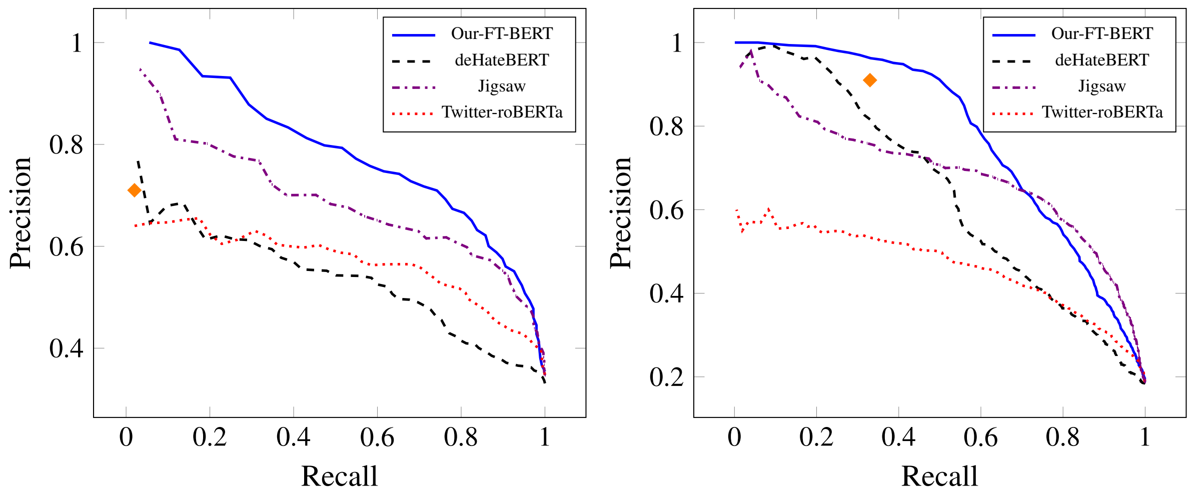
6 Analysis: The Propensity of Hate

6.1 Hate on the Post Level

Taking our conservative approach, we find that the frequency of hate posting is higher in Parler (3.29%), compared in Gab (2.13%). However, we find that 13.95% of Parler users share at least one hateful post – a significantly lower number compared to Gab (18.58%). We find that 65.5% of the hate content in Parler is posted as a reply to other parlays. This reflects a significant over-representation of replies compared with full corpus distribution (46.2% of posts are replies, see Table 1). Similarly, 38.9% of the hate content on Gab are replies.

[11] These measures are the weighted average precision/recall over both hate/non-hate classes.

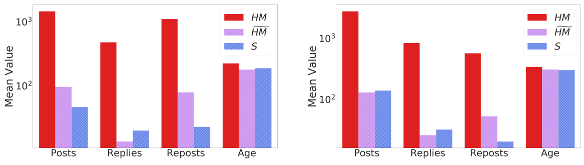
[12] 87.1% (63.4%) of the users in Parler (Gab)



Parler Classifiers

Gab Classifiers

Figure 2: Post level Precision-Recall (PR) curves. FT-BERT: Fine Tuned BERT; Orange diamond (◆) marks the PR performance of the lexical-based approach (HateBase). Unlike the other four methods, this approach cannot be controlled by a threshold parameter, hence only a single PR value is available.



(a) Parler

(b) Gab

Figure 3: Activity measures per user group. Numbers are averaged per measure and group. We use a log-scale over the y-axis.

6.2 Hate on the User Level

We provide an analysis of the characteristics of the HM, $\widetilde{HM}$ and S accounts on an array of attributes, ranging from activity levels to centrality, sentiment and the emotions they convey.

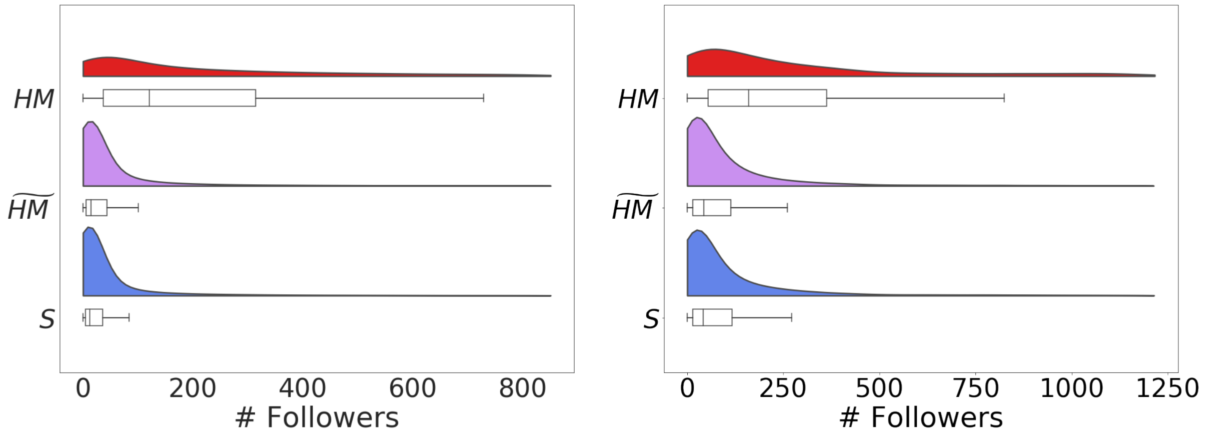
Activity Level Activity levels are compared via four features – number of posts, replies, reposts, and users' age (measured in days).

HM are the most active user group in both platforms across all activity types (see Figure 3). We find that the $\widetilde{HM}$ users have similar characteristics in both platforms – overall, they post less content than the HM users, repost more content than the S group, and their tendency to reply is lower compared to the S users.

Interestingly, although the HM make only 16.1% (10%) of the active users in Parler (Gab) – they generate a disproportional number of posts: 30.6% (59.45%) of the posts in Parler (Gab). The same holds for replies – the HM users post 36.68% (75.57%) of the replies in Parler (Gab). When aggregating all activity types (post/reply/repost) – the HM users generate 41.23% (71.38%) of the content in Parler (Gab).

User *Age* (days from account creation to the most recent post in the data), is an exceptional feature. We observe only insignificant differences between the three user groups. This observation holds for both platforms. However, collapsing the groups – we do find a significant difference between the two platforms. Gab users are "older" with an average age of 323.9 compared to 189.6 of the Parler users. We hypothesize that the difference is a result of the way both platform evolve over time, given the unfolding of events driving users to these platforms (see Sections 1 and 3.1).

Popularity and Engagement We quantify the popularity level of users based on the number of *followers* they have. Figure 4 presents numbers for both platforms. On both platforms hate mongers (HM) are significantly more popular compared to users in other user groups. In Parler, the median number of followers is 121 compared to 15 and 12 of $\widetilde{HM}$ and S, respectively. The same holds for Gab – a median value of 160 for HM users compared to 43 and 41 of the other two user groups. Interestingly, although Parler is a much larger social platform (mainly in terms of registered users, see Section 3 and Table 1) we do not see a significantly higher number of followers in Parler. Moreover, when calculating the number of followers over the



(a) Parler Followers

(b) Gab Followers

Figure 4: Followers distributions. The extreme percentiles (2.5%) of the data are omitted for visualization purposes. Rectangles indicate the ± standard division around the average; The median is indicated by a vertical line.

whole population, the median in Gab is three times higher – 48 vs. 16.

Engagement level is measured by the number of *followees* each account has (the number of accounts a user follows). We find that HM are highly engaged in both platforms, compared to other user groups. In Parler, the median number of followees of HM users is 106 – significantly higher than 46 and 36 median values of the $\widetilde{HM}$ and S users, respectively.

Account's Self Description Analogous to the account's description in Twitter, Parler users can provide a short descriptive/biographical text to appear next to the user's avatar. For example, the biography that is associated with a specific Parler user is: *"Conservative banned by mainstream social media outlets for calling the leftists out for what they really are! Been awake for YEARS! #trump2020"*.

We use this content to further assess users' commitment to the Parler platform,[13] assuming more engaged users are, the more likely they add the description to their profile. We find that while only 35.8% of the S users use the biography field, 59.6% of the HM users provide the description in their profile. We also find that the average (median) biographical text length of HM users is 128.6 (134). This is considerably longer, compared to $\widetilde{HM}$ and S users who included the description in their profile, with an average (median) of 99.4 (90) and 94.6 (84) text length, respectively.

Social Structure Analysing the degree distribution of users provides an interesting difference between the platforms. As observed in Figure 5, HM users have the most distinctive distribution in both Parler and Gab. However, while the $\widetilde{HM}$ and the S group distributions are inseparable in Gab, the Parler user groups have distinct distributions. These distributions highlight the distinctiveness of the position of HM users in the network, as well the role of the $\widetilde{HM}$ compared to S users.

Emotional Features We compare the sentiment expressed and the emotions conveyed by different user groups. We use pretrained BERT models for both the sentiment[14] and emotion[15] predictions. Results are presented in Table 2. Looking at the Parler users, we find a small though significant (p-value $< 10^{-3}$) tendency of HM to express a more negative sentiment. The same holds for Gab, although the sentiment expressed by $\widetilde{HM}$ is closer to the sentiment of the HM users, rather to that of the S users. Aggregating the emotion predictions, we find that HM users tend to convey more *Anger* and *Sadness* than the other groups. This observation holds for both Parler and Gab, although *Anger* is more prominent.

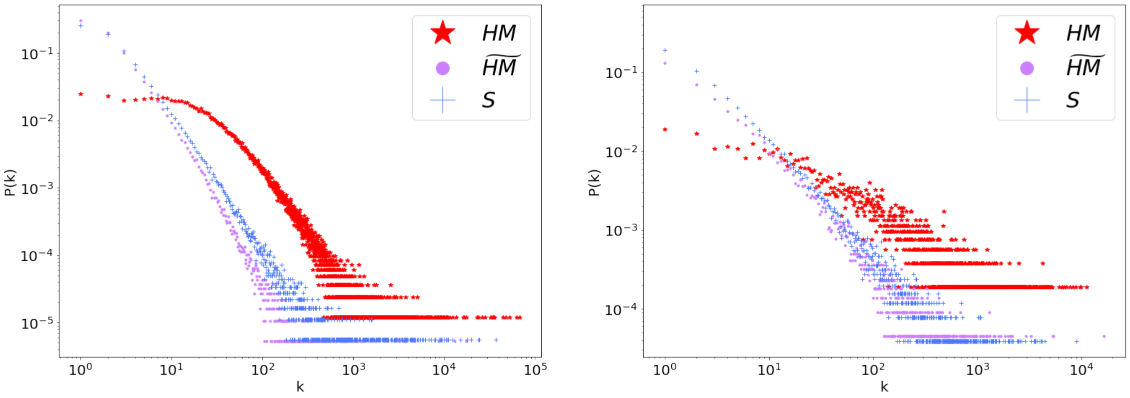
7 Discussion

Time span Given that we provide a comparison between trends in Parler and Gab, it is im-

[13] In this part, we do not compare Parler to Gab since account's self description is not available for the Gab corpus.

[14] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[15] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion



(a) Parler

(b) Gab

Figure 5: Social networks degree distribution. We present the in-degree distributions. Network is based on reposts. p(k) (y-axis) is the probability value per a each node's degree (x-axis). We use a log-scale over both axis.

		Anger	Joy	Sad	Fear	Sentiment
Parler	HM	48%	37.9%	7.4%	5.1%	2.63
	$\widetilde{HM}$	41.9%	44.3%	6.7%	5.3%	2.84
	S	33.6%	55.7%	5%	4.3%	2.84
Gab	HM	40.0%	44.5%	7.2%	6.3%	2.55
	$\widetilde{HM}$	35.9%	49.7%	5.9%	7.1%	2.56
	S	35.5%	51.1%	6.0%	5.7%	2.67

Table 2: Emotions and sentiment analysis. The four leftmost columns are the distribution of emotions per user group while the rightmost column is the median sentiment score. The sentiment spans over [1,5] (i.e., 5 is the highest score).

portant to note the datasets span different and non-overlapping time-frames (see Table 1). Therefore, the comparison we provide should be read cautiously. We do note, however, that each of the datasets was crawled from the early days of the social platform and spans over a similar time range (17 months). Moreover, the temporal disparity between the dataset could be considered as an advantage – allowing us to examine the generalization performance of hate speech models, as we report in Section 5.1.

Ethical Considerations Analysing and modeling hate speech in a new social platform such as Parler is of great importance. However, classifying *users* as hate mongers, based on the output of an algorithm, may result in marking users falsely (which may result in suspension or other measures taken against them). While we always opted for a conservative approach, as well as focusing on aggregated measures characterizing the trends of a *platform*, we note that user labeling should be carefully used, ideally involving a 'man-in-the-loop'.

Considering the annotation task – the annotation process did not include any information about the identity of the users. In addition, we warned our human annotators about the possible inappropriate and triggering content of the posts. We also make sure to remove users' information from the annotated dataset that we publish.

8 Conclusion and Future Work

To the best of our knowledge, we present the first large-scale computational analysis of hate speech on Parler, and provide a comparison to trends observed in the Gab platform.

We tag and share the first annotated Parler dataset, containing 10K posts labeled by the level of hate they convey. We used this dataset to fine-tune a transformer model to be used to mark a seed set of users in a diffusion model, resulting in user-level classification. We find significant differences between hate mongers (HM) and other user groups: HM represent only 16.1% and 10% of the active users in Parler and Gab respectively. However, they generate *41.23%* of the content in Parler and *71.38%* of the content in Gab. We find that HM show higher engagement levels and they have significantly more followers and followees. Other differences are manifested through the sentiment level expressed and the emotions conveyed.

Future work takes two trajectories: (i) Comparison of the current results with a more traditional social platform (e.g., Twitter); and (ii) An early detection of hate mongers – building a classifier to detect hate mongers based on their very first steps in the social platform.

References

ADL. 2020. Antisemitic incidents hit all-time high in 2019.

Shahram Akbarzadeh. 2016. The muslim question in australia: Islamophobia and muslim alienation. *Journal of Muslim Minority Affairs*, 36(3):323–333.

Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. An early look at the parler online social network. *arXiv preprint arXiv:2101.03820*.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting anti-asian hateful users on twitter during covid-19. *arXiv preprint arXiv:2109.07296*.

Eyal Arviv, Simo Hanouna, and Oren Tsur. 2021. It'sa 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, pages 61–70.

Annalise Baines, Muhammad Ittefaq, and Mauryne Abwao. 2021. # scamdemic,# plandemic, or# scaredemic: What parler social media platform tells us about covid-19 vaccine. *Vaccines*, 9(5):421.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Janek Bevendorff, Berta Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Rangel Francisco, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Wiegmann Matti, Magdalena Wolska, and Eva Zangerle. 2021. Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection. In *12th International Conference of the CLEF Association (CLEF 2021)*. Springer.

Mohit Chandra, Manvith Reddy, Shradha Sehgal, Saurabh Gupta, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2021. " a virus has no religion": Analyzing islamophobia on twitter during the covid-19 outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 67–77.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021. You too brutus! trapping hateful users in social media: Challenges, solutions & insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 79–89.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vikram Dodd and Sarah Marsh. 2017. Anti-muslim hate crimes increase fivefold since london bridge attacks. *The Guardian*, 7.

Griffin Sims Edwards and Stephen Rushin. 2018. The effect of president trump's election on hate crimes. *Available at SSRN 3102652*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Horst Entorf and Martin Lange. 2019. Refugees welcome? understanding the regional heterogeneity of anti-foreigner hate crimes in germany. *Understanding the Regional Heterogeneity of Anti-Foreigner Hate Crimes in Germany (January 30, 2019). ZEW-Centre for European Economic Research Discussion Paper*, (19-005).

Arne C Esser. 2021. How does the language of corpora from radicalized communities discovered on parler compare to online conversations on twitter regarding the 2021 capitol riots and election fraud? Master's thesis.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

John Gallacher and Jonathan Bright. 2021. Hate contagion: Measuring the spread and trajectory of hate on social media.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Benjamin Golub and Matthew O Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49.

Keegan Hankes and Alex Amend. 2019. Aspi explains: 8chan.

Janice A Iwama. 2018. Understanding hate crimes against immigrants: C onsiderations for future research. *Sociology compass*, 12(3):e12565.

Greta Jasser, Jordan McSwiney, Ed Pertwee, and Savvas Zannettou. 2021. 'welcome to# gabfam': Far-right virtual community on gab. *New Media & Society*, page 14614448211024546.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*.

Zachary Laub. 2019. Hate speech on social media: Global comparisons.

Brian Levin and John David Reitzel. 2018. Report to the nation: hate crimes rise in us cities and counties in time of division and foreign interference.

Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 515–522. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the hate code on social media. In *Eleventh International AAAI Conference on Web and Social Media*.

Simon Malevich and Tom Robertso. 2019. Violence begetting violence: An examination of extremist content on deep web social networks.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Meta. 2021a. Facebook reports second quarter 2021 results. `https://tinyurl.com/2p8r4wd6`. Accessed: 2022-04-17.

Meta. 2021b. Update on our progress on ai and hate speech detection. `https://tinyurl.com/muvn4hma`. Accessed: 2022-04-17.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th acm conference on hypertext and social media*, pages 85–94.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Mohamed Nawab Bin Mohamed Osman. 2017. Retraction: Understanding islamophobia in asia: The cases of myanmar and malaysia. *Islamophobia Studies Journal*, 4(1):17–36.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Barbara Perry, Davut Akca, Fatih Karakus, Mehmet Fatih Bastug, et al. 2020. Planting hate speech to harvest hatred: How does political hate speech fuel hate crimes in turkey? *International Journal for Crime, Justice and Social Democracy*, 9(2).

Gary Peters, Rob Portman, Amy Klobuchar, and Roy Blunt. 2021. Examining the u.s. capitol attack: a review of the security planning and response failures.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. *arXiv preprint arXiv:2106.02821*.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2017. "like sheep among wolves": Characterizing hateful users on twitter. *arXiv preprint arXiv:1801.00317*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1.

Niloofar Safi Samghabadi, Parth Patwa, PYKL Srinivas, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Andrea Sipka, Aniko Hannak, and Aleksandra Urman. 2021. Comparing the language of qanon-related content on parler, gab, and twitter. *arXiv preprint arXiv:2111.11118*.

Lütfi Sunar. 2017. The long history of islam as a collective “other” of the west and the rise of islamophobia in the us after trump. *Insight Turkey*, 19(3):35–52.

Elise Thomas. 2019. Aspi explains: 8chan.

Time. 2021. Twitter penalizes record number of accounts for posting hate speech. `https://time.com/6080324/twitter-hate-speech-penalties/`. Accessed: 2022-04-17.

Christopher Tuckwood. 2017. Hatebase: Online database of hate speech. *The Sentinal Project. Available at: https://www. hatebase. org*.

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020a. Detecting east asian prejudice on social media. *arXiv preprint arXiv:2005.03909*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020b. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Ethan Ward. 2021. Parlez-vous le hate?: Examining topics and hate speech in the alternative social network parler. Master’s thesis, University of Waterloo.

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Tomer Wullach, Amir Adler, and Einat Minkov. 2020. Towards hate speech detection at large via deep generative modeling. *IEEE Internet Computing*, 25(2):48–57.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. *arXiv preprint arXiv:2109.00591*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. 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. International World Wide Web Conferences Steering Committee.

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

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2016. Hate speech detection using a convolution-lstm based deep neural network.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7158–7166.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counter-hate in social media during the covid-19 crisis. *arXiv preprint:2005.12423*.