Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

The Pushshift Reddit Dataset

Jason Baumgartner,[1,*] **Savvas Zannettou,**[2,☺] **Brian Keegan,**[3] **Megan Squire,**[4] **Jeremy Blackburn**[5,☺]
[1]Pushshift.io, [2]Max-Planck-Institut für Informatik, [3]University of Colorado Boulder, [4]Elon University, [5]Binghamton University
[*]Network Contagion Research Institute, [☺]iDRAMA Lab
jason@pushshift.io, szannett@mpi-inf.mpg.de, brian.keegan@colorado.edu, msquire@elon.edu,
blackburn@cs.binghamton.edu

Abstract

Social media data has become crucial to the advancement of scientific understanding. However, even though it has become ubiquitous, just collecting large-scale social media data involves a high degree of engineering skill set and computational resources. In fact, research is often times gated by data engineering problems that must be overcome before analysis can proceed. This has resulted recognition of datasets as meaningful research contributions in and of themselves.

Reddit, the so called "front page of the Internet," in particular has been the subject of numerous scientific studies. Although Reddit is relatively open to data acquisition compared to social media platforms like Facebook and Twitter, the technical barriers to acquisition still remain. Thus, Reddit's millions of subreddits, hundreds of millions of users, and billions of comments are at the same time relatively accessible, but time consuming to collect and analyze systematically.

In this paper, we present the Pushshift Reddit dataset. Pushshift is a social media data collection, analysis, and archiving platform that since 2015 has collected Reddit data and made it available to researchers. Pushshift's Reddit dataset is updated in real-time, and includes historical data back to Reddit's inception. In addition to monthly dumps, Pushshift provides computational tools to aid in searching, aggregating, and performing exploratory analysis on the entirety of the dataset. The Pushshift Reddit dataset makes it possible for social media researchers to reduce time spent in the data collection, cleaning, and storage phases of their projects.

1 Introduction

Understanding complex socio-technical phenomena requires data-driven research based on large-scale, reliable, relevant data sets. Web data, particularly data from application programming interfaces (APIs), has been an enormous boon for researchers using online social platforms' databases of user-generated activity and content (Freelon 2014; Golder and Macy 2014; Hampton 2017; Lazer and Radford 2017). The ability to "crawl" and "scrape" large-scale and high-resolution samples of publicly-accessible user data stimulated emerging fields like social computing (Wang et al. 2007) and computational social science (Lazer et al. 2009), and developed new fields like crisis informatics (Palen and Anderson 2016). But following major scandals around data privacy and ethics, social media platforms like Facebook and Twitter changed previously permissive data access provisions of their public APIs (Walker, Mercea, and Bastos 2019). As a consequence, the ability for researchers to collect timely data, share tools, instruct students, and reproduce findings has been curtailed.

This "post-API age" is characterized by the deprecation of data resources used for research and teaching (Freelon 2018; Puschmann 2019), increased stratification of data access based on social, technical, and financial capital (boyd and Crawford 2012; Manovich 2011), and greater fear of prosecution around violating terms of service in the course of research (Halavais 2019; Patel 2018). These changes have had a profoundly chilling effect on researchers' use of API-derived data to investigate behavior like discrimination, harassment, radicalization, hate speech, and disinformation. Furthermore, researchers have struggled in systematically studying the role that platforms' changing features, design affordances, and governance strategies play in sustaining these forms of "turpitude-as-a-service" (Bruns 2019; Keegan 2018). Faced with conflicting incentives between protecting their users' data from abuse and maintaining their commitments to values of openness, online social platforms are exploring alternative data sharing models like "trusted third party" models that still carry significant technical and reputational risks (Bruns 2019; Gibney 2019; Ingram 2019; Mervis 2019; Puschmann 2019).

Even if the "golden age" of API-driven computational social science and social computing research had not closed in the shadow of privacy scandals, it was nevertheless characterized by enormous inefficiencies in data collection and inequalities in access (Manovich 2011; Puschmann 2019), ethically-suspect methods and implications (boyd 2016; Tufekci 2014; Olteanu et al. 2019), a lack of concern for data sharing or reproducibility (Borgman 2012; Weller and Kinder-Kurlanda 2016), and failures to validate constructs or generalize to off-platform behavior (Ekbia et al. 2015; Howison, Wiggins, and Crowston 2011; Japec et al. 2015). Facebook's and Twitter's changes in data access were significant, however the enclosure of previously open big social data sources is not ubiquitous among platform providers (Boyle 2017; Hess and Ostrom 2003; Hunter 2003). Social platforms and online communities

like Wikipedia (Foundation 2019), Stack Exchange (Archive 2019), GitHub (Gousios 2013), and Reddit (Reddit 2019) continue to offer open APIs and data dumps that are valuable for researchers.

In this paper, we assist to the goal of providing open APIs and data dumps to researchers by releasing the Pushshift Reddit dataset. In addition to monthly dumps of 651M submissions and 5.6B comments posted on Reddit between 2005 and 2019[1], the Pushshift Reddit dataset also includes an API for researcher access and a Slackbot that allows researchers to easily interact with the collected data. The Pushshift Reddit API enables researchers to easily execute queries on the whole dataset without the need for downloading the monthly dumps. This reduces the requirement for substantial storage capacity, thus making the data more available to a wider range of users. Finally, we provide access to a Slackbot that allows researchers to easily produce visualizations of data from the Pushshift Reddit dataset in real-time and discuss them with colleagues on Slack. These resources allow research teams to quickly begin interacting with data with very little time spent on the tedious aspects of data collection, cleaning, and storage.

2 Pushshift

Pushshift is not a new or isolated data platform, but a five year-old platform with a track record in peer-reviewed publications and an active community of several hundred users. Pushshift not only collects Reddit data, but exposes it to researchers via an API. Why do people use Pushshift's API instead of the official Reddit API? In short, Pushshift makes it much easier for researchers to query and retrieve historical Reddit data, provides extended functionality by providing full-text search against comments and submissions, and has larger single query limits. Specifically, because, at the time of this writing, Pushshift has a size limit five times greater than Reddit's 100 object limit, Pushshift enables the end user to quickly ingest large amounts of data. Additionally, the Pushshift API offers aggregation endpoints to provide summary analysis of Reddit activity, a feature that the Reddit API lacks entirely.

The Pushshift Reddit dataset provides not just a *technical* infrastructure of software and hardware for collecting "big social data" but also a *social* infrastructure of organizational processes for responsibly collecting, governing, and discussing these research data.

Data collection process

Pushshift uses multiple backend software components to collect, store, catalog, index, and disseminate data to end-users. As seen in Fig. 1, these subsystems are:

- The **ingest engine**, which is responsible for collecting and storing raw data.
- A PostgreSQL **database**, which allows for advanced querying of data and meta-data storage.
- An Elastic Search **document store** cluster, which performs indexing and aggregation of ingested data.

[1]Available at https://files.pushshift.io/reddit/

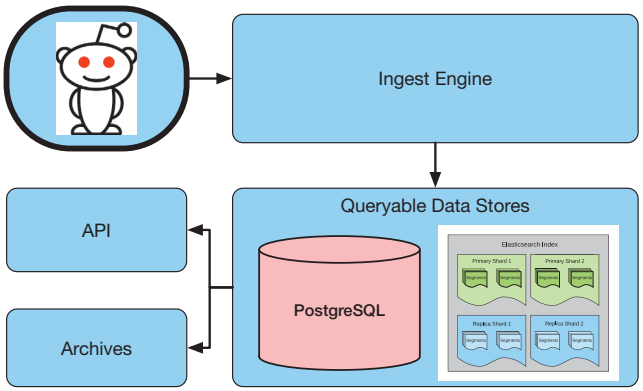

Figure 1: Pushshift's Reddit data collection platform.

- An **API** to allow researchers dynamic access to collected data and aggregation functionality.

Ingest Engine. The first stage in the Pushshift pipeline is the ingest engine, which is responsible for actually collecting data. The ingest engine can be thought of as a *framework* for large scale collection of heterogeneous social media data sources. The ingest engine orchestrates the execution of a multiple data collection programs, each designed to handle a particular data source. Specifically, the ingest engine provides and manages a job scheduling queue, and provides a set of common APIs to handle the data storage. Currently, Pushshift's ingest engine works as follows:

First, the program runner starts each ingest program (*i.e.*, the programs that actually collect the data). The ingest engine is agnostic to the particulars of the individual ingest programs: no particular programming language is required, and there is no particular expectation of how an ingest program works, modulo its interactions with the remainder of the ingest engine. Typically, an ingest program will directly interact with Web APIs, scrape content from HTML pages, use data streams where available, *etc*. Next, the ingest program inserts the raw data retrieved into a database as well as into a document store. Behind the scenes, each piece of collected data is added to an intermediate queue (currently implemented via Redis), which serves as a staging area until the data is processed by any custom processing scripts the ingest program's creator might require. Finally, the raw data is periodically flushed to disk. The data storage format can be specified by the ingest program creator via the custom processing scripts previously mentioned, or a standard, Pushshift-implemented format can be used (*e.g.*, `ndjson`).

PostgreSQL & ElasticSearch. Pushshift currently uses Elasticsearch (ES) as a scalable document store for each data source that is part of the ingest pipeline. ES offers a number of important features for storing and analyzing large amounts of data. For example, ES achieves *ease-of-scaling* by utilizing a cluster approach for horizontal expansion. It ensures *redundancy* by creating multiple replicas for each index so that a node outage does not af-

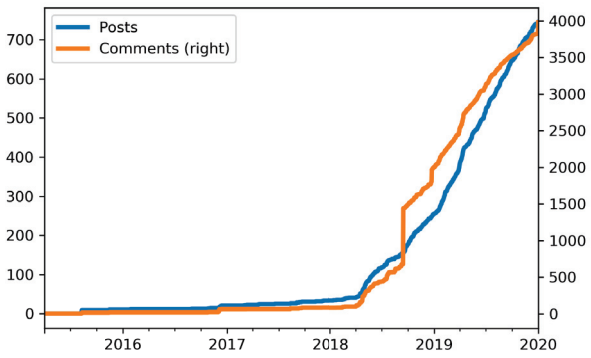

Figure 2: Activity on the r/pushshift subreddit.

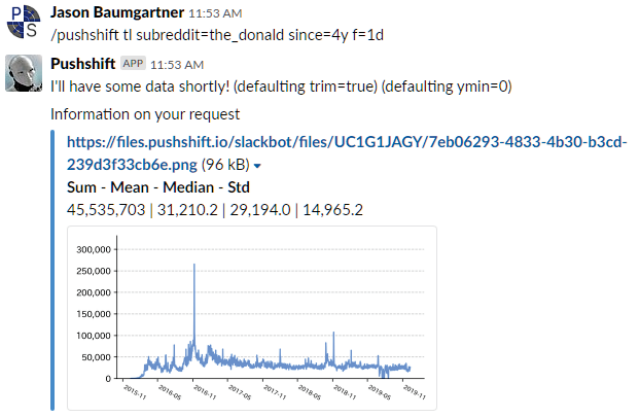

Figure 3: The Pushshift chatbot in Slack.

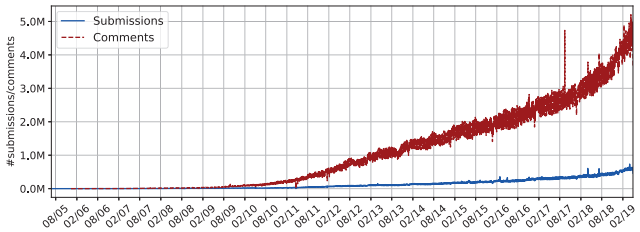

Figure 4: The number of submissions and comments for each day of our dataset.

fect the overall health of the cluster. The ES robust dynamic mapping tools allow *easy modification and expansion* of indexes to accommodate changes in data structure from the source. This is useful because Reddit's API does not implement any type of versioning, yet there are constant additions and modifications made to the API when new features and data types are added to the response objects. By using dynamic mapping types, Pushshift can easily add new fields to existing indices. This enables us to quickly modify the corresponding mappings to allow search and aggregation on those new fields. Pushshift also makes use of the ICU Analysis plug-in for ES (Committee 2019; Elasticsearch 2019), which provides support for international locales, full Unicode support up through Unicode 12, and complete emoji search support.

API Pushshift currently allows users to search Reddit data via an API. Right now, this API exports much of the search and aggregation functionality provided by Elastic Search. This functionality supports dozens of community applications and numerous research projects. The API is the major workload of handled by Pushshift's computational resources, serving 500M requests per month. Although in this paper we focus on a description of the data (Section 3) due to space limitations, we provide online API documentation at https://pushshift.io/api-parameters/.

Community In addition to Pushshift's website, which features an interactive dashboard of current activity trends, Pushshift also has two active user communities on Reddit and Slack. The /r/pushshift subreddit was created in April 2015 and is used for sharing announcements, answering questions, reporting bugs, and soliciting feedback for new features. There are more than 2,100 subscribers to this subreddit, an active team of 10 moderators, and more than 700 posts (with more than 4,000 comments) from over 350 unique users (see Fig. 2).

The Pushshift Slack team has nearly 300 registered users and more than 260,000 messages across 53 channels discussing data science and visualization. Custom tools have also been developed to integrate the Pushshift archive into these Slack communities. For example, users can interact with a Slack chatbot in realtime. The bot can analyze and visualize Pushshift data based on queries made in the Slack channel, and return those visualizations to the channel for discussion and observation. In Fig. 3, a user queried the total number of daily comments to the /r/the_donald subreddit by day over the past four years and received a time series plot and summary statistics from the chatbot within a few seconds. The chatbot can also be shared to other non-Pushshift workspaces, allowing researchers in other Slack workspaces to use the data. This extends the reach of Pushshift data even further.

3 Description of the Pushshift Reddit Dataset

Pushshift makes available all the submissions and comments posted on Reddit between June 2005 and April 2019. The dataset consists of 651,778,198 submissions and 5,601,331,385 comments posted on 2,888,885 subreddits. Fig. 4 shows the number of submissions and comments per day. We observe that the number of submissions and comments increase over the course of our dataset. After August 2013, we have consistently over 1M comments per day, while by the end of our dataset (April 2019) we have 5M comments per day. Also, while submissions are substantially fewer than comments, submissions have reached reached a consistent level of over 500K per day in this dataset.

The Pushshift Reddit dataset is made up of two sets of files: one set of files for the submissions and one for the

Field	Description
id	The submission's identifier, e.g., "5lcgjh" (String).
url	The URL that the submission is posting. This is the same with the permalink in cases where the submission is a self post. E.g., "https://www.reddit.com/r/AskReddit/
permalink	Relative URL of the permanent link that points to this specific submission, e.g., "/r/AskReddit/comments/5lcgj9/what_did_you_think_of_the_ending_of_rogue_one/" (String).
author	The account name of the poster, e.g., "example_username" (String).
created_utc	UNIX timestamp referring to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the submission is posted. Note that it excludes the prefix /r/. E.g., 'AskReddit' (String).
subreddit_id	The identifier of the subreddit, e.g., "t5_2qh1i" (String).
selftext	The text that is associated with the submission (String).
title	The title that is associated with the submission, e.g., "What did you think of the ending of Rogue One?" (String).
num_comments	The number of comments associated with this submission, e.g., 7 (Integer).
score	The score that the submission has accumulated. The score is the number of upvotes minus the number of downvotes. E.g., 5 (Integer). **NB:** Reddit fuzzes the real score to prevent spam bots.
is_self	Flag that indicates whether the submission is a self post, e.g., true (Boolean).
over_18	Flag that indicates whether the submission is Not-Safe-For-Work, e.g., false (Boolean).
distinguished	Flag to determine whether the submission is posted by moderators or admins. "null" means not distinguished (String).
edited	Indicates whether the submission has been edited. Either a number indicating the UNIX timestamp that the submission was edited at, "false" otherwise.
domain	The domain of the submission, e.g., self.AskReddit (String).
stickied	Flag indicating whether the submission is set as sticky in the subreddit, e.g., false (Boolean).
locked	Flag indicating whether the submission is currently closed to new comments, e.g., false (Boolean).
quarantine	Flag indicating whether the community is quarantine, e.g., false (Boolean).
hidden_score	Flag indicating if the submission's score is hidden, e.g., false (Boolean).
retrieved_on	UNIX timestamp referring to the time we crawled the submission, e.g., 1483228803 (Integer).
author_flair_css_class	The CSS class of the author's flair. This field is specific to subreddit (String).
author_flair_text	The text of the author's flair. This field is specific to subreddit (String).

Table 1: Submissions data description.

comments. Below, we describe the structure of each of the files in these two sets.

Submissions. The submissions dataset consists of a set of newline delimited JSON[2] files: we maintain a separate file for each month of our data collection. Each line in these files correspond to a submission and it is a JSON object. Table 1 describes the most important key/values included in each submission's JSON object.

Comments. Similarly to the submissions, the comments' dataset is a collection of ndjson files with each file corresponding to a month-worth of data. Each line in these files correspond to a comment and it is a JSON object. Table 2 describes the most important keys/values in each comment's JSON object.

FAIR principles. The Pushshift Reddit dataset aligns with the FAIR principles.[3] Our dataset is *Findable* as the monthly dumps are publicly available via Pushshift's website[4]. We also upload a small sample of the dataset to the Zenodo service, so that we obtain a persistent digital object identifier (DOI): 10.5281/zenodo.3608135.[5] Note that we were unable to upload the entire dataset to Zenodo, since the service has a limit of 100GB and our dataset is in the order of several terabytes. The Pushshift Reddit dataset is *Accessible* as it can be accessed by anyone visiting the Pushshift's website. Furthermore, we offer an API and a Slackbot that allow researchers to easily execute queries and obtain data from our infrastructure without the need to download the large monthly dumps. Also, our dataset is *Interoperable* because it is JSON format, which is a widely known and used format for data. Because the provenance for the collected data is very clear, and users are simply asked to cite Pushshift in order to use the data, our dataset is also *Reusable*.

4 Dataset Use Cases

The Pushshift Reddit dataset has attracted a substantial research community.As of late 2019, Google Scholar indexes over 100 peer-reviewed publications that used Pushshift data

[2]http://ndjson.org/

[3]https://www.go-fair.org/fair-principles/

[4]https://files.pushshift.io/reddit/

[5]https://zenodo.org/record/3608135

Field	Description
id	The comment's identifier, e.g., "dbumnq8" (String).
author	The account name of the poster, e.g., "example_username" (String).
link_id	Identifier of the submission that this comment is in, e.g., "t3_5l954r" (String).
parent_id	Identifier of the parent of this comment, might be the identifier of the submission if it is top-level comment or the identifier of another comment, e.g., "t1_dbu5bpp" (String).
created_utc	UNIX timestamp that refers to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the comment is posted. Note that it excludes the prefix /r/. E.g., 'AskReddit' (String).
subreddit_id	The identifier of the subreddit where the comment is posted, e.g., "t5_2qh1i" (String).
body	The comment's text, e.g., "This is an example comment" (String).
score	The score of the comment. The score is the number of upvotes minus the number of downvotes. Note that Reddit fuzzes the real score to prevent spam bots. E.g., 5 (Integer).
distinguished	Flag to determine whether the comment is made by the moderators or admins. "null" means not distinguished (String).
edited	Flag indicating if the comment has been edited. Either the UNIX timestamp that the comment was edited at, or "false".
stickied	Flag indicating whether the submission is set as sticky in the subreddit, e.g., false (Boolean).
retrieved_on	UNIX timestamp that refers to the time that we crawled the comment, e.g., 1483228803 (Integer).
gilded	The number of times this comment received Reddit gold, e.g., 0 (Integer).
controversiality	Number that indicates whether the comment is controversial, e.g., 0 (Integer).
author_flair_css_class	The CSS class of the author's flair. This field is specific to subreddit (String).
author_flair_text	The text of the author's flair. This field is specific to subreddit (String).

Table 2: `Comments` data description.

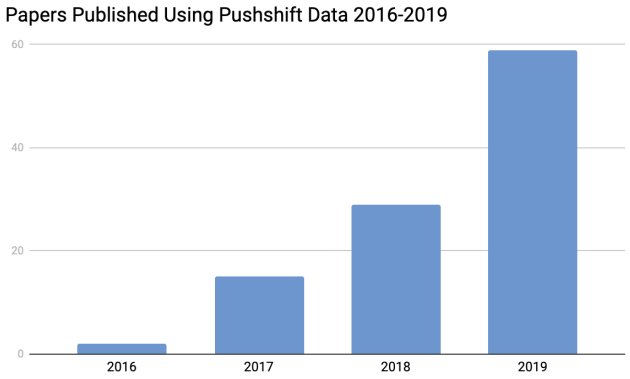

Figure 5: Over 100 peer-reviewed papers have been published using Pushshift data.

(see Fig. 5). This research covers a diverse cross-section of research topics including measuring toxicity, personality, virality, and governance. Pushshift's influence as a primary source of Reddit data among researchers has attracted empirical scrutiny (Gaffney and Matias 2018), which in turn has led to improved data validation efforts (Baumgartner 2018). We note that there is some difficulty in ascertaining our dataset's full contribution to the scientific community due to a previous lack of deliberate efforts to conform to FAIR principles, which we address in this paper.

Online community governance. Reddit's ecosystem of sub-reddits are primarily governed by volunteer moderators with substantial discretion over creating and enforcing rules about user behavior and content moderation (Fiesler et al. 2018; Squirrell 2019). This distributed and volunteer-led model stands in contrast to the centralized strategies of other prominent social platforms like Facebook, Twitter, and YouTube (Seering et al. 2019). These differences between centralized versus delegated moderation make ideal case studies for comparing the effectiveness of responses to difficult issues like social movements, fringe identities, hate speech, and harassment campaigns (Massanari 2017; Matias 2016; 2019). Pushshift data has already been instrumental for researchers exploring the spillover effects of banning offensive sub-communities (Chandrasekharan et al. 2017a), identifying common features of abusive behavior across communities (Chandrasekharan et al. 2017c), similarity in norms and rules across communities (Chandrasekharan et al. 2018; Fiesler et al. 2018), perceptions of fairness in moderation decisions (Jhaver et al. 2019; Jhaver, Bruckman, and Gilbert 2019), and improving automated moderation tools (Chandrasekharan et al. 2019).

Online extremism. The political extremism research community currently faces significant challenges in understanding how mainstream and fringe online spaces are used by bad actors. Despite widespread agreement that recent increases in online radicalization are due to "a globalised, toxic, anonymous online culture" operating largely outside mainstream social media platforms (Oboler, Allington, and Scolyer-Gray 2019), much of the research on extremist use of social media still focuses on mainstream sites like Facebook or Twitter (Burris, Smith, and Strahm 2000). Access to these rapidly-changing online spaces is difficult, and many research teams end up using out-of-date data, or relying on the data they have, rather than the data they need. Many social media platforms face pressure to monetize their data (Botta, Digiacomo, and Mole 2017) or remove access to it entirely (Bastos and Walker 2018), making research access to these spaces expensive and difficult. Yet, extremism researchers agree that data access is a key limitation to under-

standing online radicalization as a phenomenon. Online extremism researchers top recommendation is to "invest more in data-driven analysis of far-right violent extremist and terrorist use of the Internet." (Terrorism 2019) Pushshift data has already been used to understand the phenomenon of hate speech and political extremism (Johnston and Marku 2020; Chandrasekharan et al. 2017b; Fair and Wesslen 2019; Farrell et al. 2019; Grover and Mark 2019) and trolling and negative behaviors in fringe online spaces (Almerekhi et al. 2019; Zannettou et al. 2018a; 2018b).

Online disinformation. The online disinformation research community has focused its attention on how social media facilitates the spread of deliberately inaccurate information (Narayanan et al. 2018; Starbird 2017). The use of social media platforms to spread this "fake news" and biased political propaganda was particularly concerning given the events surrounding Russian interference in the 2016 US presidential election. Researchers studying disinformation acknowledge that mainstream platforms, particularly Facebook, are still the main place where disinformation campaigns take place (Bradshaw and Howard 2019) and that a lack of data access is significantly limiting their efforts (Alba 2019). While mainstream sites are the largest amplifiers of disinformation content, the content itself is often created on fringe sites that serve as proving grounds (Funke 2018; Gonimah 2018; Marwick and Lewis 2017). As with extremism and terrorism research, data access and data sharing in the disinformation research community is an ongoing struggle. Pushshift data has already been used in a number of papers on disinformation and social media trustworthiness (Crothers, Japkowicz, and Viktor 2019; Horne and Adali 2017; Zannettou et al. 2019; Zhou et al. 2019).

Web science. Datasets like Pushshift are critically important for researchers who answer questions at the intersection of Internet and society. How does technology spread? What is the impact of each interface or design choice on the efficacy of social media platforms? How should we measure the success or failure of an online community? Pushshift data has already been used in studies of user engagement on social media (Aldous, An, and Jansen 2019), social media moderation schemes (Shen and Rose 2019; Srinivasan et al. 2019), measuring success and growth of online communities (Cunha et al. 2019; Tan 2018), conflict in online groups (Datta and Adar 2019; Datta, Phelan, and Adar 2017; Kumar et al. 2018), the spread of technological innovations (Glenski, Saldanha, and Volkova 2019), modeling collaboration (Kasper et al. 2017; Medvedev, Delvenne, and Lambiotte 2018), and measuring engagement and collective attention (An et al. 2019; Lorenz-Spreen et al. 2019).

Big data science. As one of a few easily-accessible, very large collections of social media data, Pushshift enables data-intensive research in foundational areas like network science (Fire and Guestrin 2019; Sarantopoulos et al. 2018; Tsugawa and Niida 2019), and new algorithms for cloud computing (Kunft et al. 2018) and very large databases (Kunft et al. 2017; 2019; Ozcan 2017).

Health informatics. Because of the relative anonymity allowed by certain social media platforms, large social media datasets are useful for researchers studying topics in health informatics including sensitive medical issues, atypical behaviors, and interpersonal conflict. Pushshift data has been used by researchers studying eating disorders and weight loss (Enes et al. 2018), addiction and substance abuse (Balsamo, Bajardi, and Panisson 2019; Barker and Rohde 2019; Bowen, O'Donnell, and Sumner 2019; Brett et al. 2019; Lu et al. 2019; Zhan et al. 2019), sexually transmitted infections (Lama et al. 2019), difficult child-rearing problems (Ammari, Schoenebeck, and Romero 2019), and various mental health challenges (Chakravorti et al. 2018; Delahunty, Wood, and Arcan 2018; Fraga, da Silva, and Murai 2018; Grant et al. 2018; 2017; Pirina and Çöltekin 2018; Rezaii, Walker, and Wolff 2019).

Robust intelligence. Intelligent systems that can augment and enhance human understanding often require large amounts of human-generated text data generated in a social context. Social media data collected by Pushshift has been used already by researchers in computational linguistics and natural language processing (Fulda 2019; Gamallo et al. 2019; Hidey and McKeown 2019; Jiang et al. 2018; Wang et al. 2019; Zheng and Zhou 2019; Zhuang et al. 2018), recommender systems (Buhagiar, Zahir, and Abhari 2018; Eberhard et al. 2019; Halder, Kan, and Sugiyama 2019; Hessel, Lee, and Mimno 2017), intelligent conversational systems (Ahmadvand et al. 2019; Golovanov et al. 2020; Jonell et al. 2019), automatic summarization (Völske et al. 2017), entity recognition (Derczynski et al. 2017), and other fields associated with the development of systems that can sense, reason, learn, and predict.

5 Related Work

Existing Data Collection Services

Promising alternatives to the aforementioned model of "storage buckets of open data hosted by cloud providers" exist that are better-tailored towards the needs of researchers.

Pushshift is not the first large-scale real-time social media data collection service aimed towards researchers. Table 3 summarizes the social and organizational features of other similar services. While not an exhaustive list, the following have heavily influenced the research community as well as motivated Pushshift's own goals and design.

Media Cloud is an "open source platform for studying media ecosystems" that tracks hundreds of millions of news stories and makes aggregated count and topical data available via a free and semi-public API (Chuang et al. 2014). The Media Cloud platform has been used to study digital health communication, agenda-setting, and online social movements. Researchers can use the API to get counts of stories, topics, words, and tags in response to queries by keyword, media source, and time window using a Solr search platform (Cloud 2019).

Feature	Media Cloud	GDELT	StatsExchange	Wikimedia	Pushshift (now)
Public API	◗	●	●	●	●
Data archive/dump	❍	●	●	●	●
Regularly updated	●	●	●	●	●
Interactive computing	❍	◗	◗	●	❍
Tutorials & demos	❍	◗	●	◗	◗
Online community	❍	❍	◗	◗	●
Outreach	❍	◗	❍	◗	❍

Table 3: Features of big social data analysis cyber-infrastructures. ● fully, ◗ partially, and ❍ not supported.

GDELT is a free open platform monitoring global news media tracking events, related topics, and imagery. The platform offers a database and knowledge graph accessible both through dumps and an "analysis service" for filtering and visualizing subsets of the complete dataset (Leetaru and Schrodt 2013).

Stats Exchange is a platform of social question answering communities, including Stack Overflow. While data dumps of the platform are hosted by the Internet Archive (Archive 2019), Stack Exchange offers both an API of activity as well as a "Data Explorer" allowing users to write SQL queries via a web interface against a regularly-updated database (Exchange 2019).

Wikimedia is the parent organization of projects like Wikipedia. It hosts data dumps of revision histories, content, and pageviews; makes data available through robust APIs; and offers a variety of interactive services. Wikimedia's deployment of Jupyter Notebooks can access replication databases of revisions and content. This enables researchers focus on analyzing data rather than system and database administration.

Other dataset papers. Considering the challenges in the post-API age, the collection, curation, and dissemination of datasets is crucial for the advancement of science. To that end, it is worth exploring other works whose primary contribution has been the dataset they provide. For example, (Fair and Wesslen 2019) released a dataset that includes 37M posts and 24M comments covering August 2016 through December 2018 from Gab, a Twitter-like social media platform that after being de-platformed by major service providers ported their codebase to use the federated social network protocol from the Mastodon project. As it turns out, (Zignani et al. 2019) released a dataset focused around Mastodon itself. Their dataset contains 5M posts, along with a crowdsourced (by Mastodon users) label that indicates whether or not the post contains inappropriate content. Research into other types of computer-mediated communication platforms have also been enabled by dataset contributions. (Garimella and Tyson 2018) released a dataset from 178 WhatsApp groups that includes 454K messages from 45K different users.

6 Discussion & Conclusion

In this paper, we presented the Pushshift Reddit Dataset, which includes hundreds of millions of submissions and billions of comments from 2005 until the present. In addition to offering Pushshift's data as monthly dumps, we also make this dataset available via a searchable API, as well as additional tools and community resources. This paper also serves as a more formal and archival description of what Pushshift's Reddit dataset provides. Having already been used in over 100 papers from numerous disciplines over the past four years, the Pushshift Reddit dataset will continue to be a valuable resource for the research community in the future.

References

Ahmadvand, A.; Sahijwani, H.; Choi, J. I.; and Agichtein, E. 2019. Concet: Entity-aware topic classification for open-domain conversational agents. In *CIKM*.

Alba, D. 2019. Ahead of 2020, facebook falls short on plan to share data on disinformation. https://nyti.ms/39ner99.

Aldous, K. K.; An, J.; and Jansen, B. J. 2019. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In *ICWSM*.

Almerekhi, H.; Kwak, H.; Jansen, B. J.; and Salminen, J. 2019. Detecting toxicity triggers in online discussions. In *Hypertext*.

Ammari, T.; Schoenebeck, S.; and Romero, D. 2019. Self-declared throwaway accounts on reddit: How platform affordances and shared norms enable parenting disclosure and support. *CSCW*.

An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *ICWSM*.

Archive, I. 2019. Stack exchange data dump. https://archive.org/details/stackexchange.

Balsamo, D.; Bajardi, P.; and Panisson, A. 2019. Firsthand opiates abuse on social media: Monitoring geospatial patterns of interest through a digital cohort. In *WWW*.

Barker, J. O., and Rohde, J. A. 2019. Topic clustering of e-cigarette submissions among reddit communities: A network perspective. *Health Education & Behavior* 46(2_suppl):59–68.

Bastos, M., and Walker, S. 2018. Facebook's data lockdown is a disaster for academic researchers. https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533.

Baumgartner, J. 2018. My response to the paper highlighting issues with data incompleteness concerning my reddit corpus. https://www.reddit.com/r/datasets/comments/884vkh/.

Borgman, C. L. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6):1059–1078.

Botta, A.; Digiacomo, N.; and Mole, K. 2017. Monetizing data: A new source of value in payments. https://mck.co/2WHYgAx.

Bowen, D. A.; O'Donnell, J.; and Sumner, S. A. 2019. Increases in online posts about synthetic opioids preceding increases in synthetic opioid death rates: a retrospective observational study. *Journal of general internal medicine* 34(12):2702–2704.

boyd, d., and Crawford, K. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15(5):662–679.

boyd, d. 2016. Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics* 12(1):4–13.

Boyle, J. 2017. The second enclosure movement and the construction of the public domain. In *Copyright Law*. Routledge. 63–104.

Bradshaw, S., and Howard, P. N. 2019. The global disinformation order: 2019 global inventory of organised social media manipulation. https://bit.ly/39i03yZ.

Brett, E. I.; Stevens, E. M.; Wagener, T. L.; Leavens, E. L.; Morgan, T. L.; Cotton, W. D.; and Hébert, E. T. 2019. A content analysis of juul discussions on social media: Using reddit to understand patterns and perceptions of juul use. *Drug and alcohol dependence* 194:358–362.

Bruns, A. 2019. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society* 22(11):1544–1566.

Buhagiar, N.; Zahir, B.; and Abhari, A. 2018. Using deep learning to recommend discussion threads to users in an online forum. In *IJCNN*.

Burris, V.; Smith, E.; and Strahm, A. 2000. White supremacist networks on the internet. *Sociological Focus* 33(2):215–235.

Chakravorti, D.; Law, K.; Gemmell, J.; and Raicu, D. 2018. Detecting and characterizing trends in online mental health discussions. In *ICDMW*.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017a. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.* 1.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017b. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *CSCW*.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017c. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 3175–3187. ACM.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *CHI*.

Chandrasekharan, E.; Gandhi, C.; Mustelier, M. W.; and Gilbert, E. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *CHI*.

Chuang, J.; Fish, S.; Larochelle, D.; Li, W. P.; and Weiss, R. 2014. Large-scale topical analysis of multiple online news sources with media cloud. In *NewsKDD: Data Science for News Publishing*.

Cloud, M. 2019. API specifications. https://github.com/berkmancenter/mediacloud/.

Committee, I. P. M. 2019. International components for unicode. http://site.icu-project.org/home.

Crothers, E.; Japkowicz, N.; and Viktor, H. L. 2019. Towards ethical content-based detection of online influence campaigns. In *MLSP*.

Cunha, T.; Jurgens, D.; Tan, C.; and Romero, D. 2019. Are all successful communities alike? characterizing and predicting the success of online communities. In *WWW*.

Datta, S., and Adar, E. 2019. Extracting inter-community conflicts in reddit. In *ICWSM*.

Datta, S.; Phelan, C.; and Adar, E. 2017. Identifying misaligned inter-group links and communities. *CSCW*.

Delahunty, F.; Wood, I. D.; and Arcan, M. 2018. First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. In *AICS*, 327–338.

Derczynski, L.; Nichols, E.; van Erp, M.; and Limsopatham, N. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 140–147.

Eberhard, L.; Walk, S.; Posch, L.; and Helic, D. 2019. Evaluating narrative-driven movie recommendations on reddit. In *IUI*, 1–11.

Ekbia, H.; Mattioli, M.; Kouper, I.; Arave, G.; Ghazinejad, A.; Bowman, T.; Suri, V. R.; Tsou, A.; Weingart, S.; and Sugimoto, C. R. 2015. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology* 66(8):1523–1545.

Elasticsearch. 2019. Icu analysis plugin. https://www.elastic.co/guide/en/elasticsearch/plugins/current/analysis-icu.html.

Enes, K. B.; Brum, P. P. V.; Cunha, T. O.; Murai, F.; da Silva, A. P. C.; and Pappa, G. L. 2018. Reddit weight loss communities: do they have what it takes for effective health interventions? In *WI*.

Exchange, S. 2019. Data explorer help. https://data.stackexchange.com/help.

Fair, G., and Wesslen, R. 2019. Shouting into the void: A database of the alternative social media platform gab. In *ICWSM*.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring misogyny across the manosphere in reddit. In *WebSci*.

Fiesler, C.; McCann, J.; Frye, K.; Brubaker, J. R.; et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *ICWSM*.

Fire, M., and Guestrin, C. 2019. The rise and fall of network stars: Analyzing 2.5 million graphs to reveal how high-degree vertices emerge over time. *Information Processing & Management*.

Foundation, W. 2019. Wikimedia downloads. https://dumps.wikimedia.org/.

Fraga, B. S.; da Silva, A. P. C.; and Murai, F. 2018. Online social networks in health care: A study of mental disorders on reddit. In *WI*.

Freelon, D. 2014. On the Interpretation of Digital Trace Data in Communication and Social Computing Research. *Journal of Broadcasting & Electronic Media* 58(1):59–75.

Freelon, D. 2018. Computational Research in the Post-API Age. *Political Communication* 35(4):665–668.

Fulda, N. E. 2019. Semantically aligned sentence-level embeddings for agent autonomy and natural language understanding.

Funke, D. 2018. Misinformers are moving to smaller platforms. so how should fact-checkers monitor them? https://bit.ly/33NTFhx.

Gaffney, D., and Matias, J. N. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE* 13(7):e0200162.

Gamallo, P.; Sotelo, S.; Pichel, J. R.; and Artetxe, M. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics* 1–27.

Garimella, K., and Tyson, G. 2018. Whatapp Doc? A First Look at Whatsapp Public Group Data. In *ICWSM*.

Gibney, E. 2019. Privacy hurdles thwart Facebook democracy research. *Nature* 574(7777):158–159.

Glenski, M.; Saldanha, E.; and Volkova, S. 2019. Characterizing speed and scale of cryptocurrency discussion spread on reddit. In *WWW*.

Golder, S. A., and Macy, M. W. 2014. Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology* 40(1):129–152.

Golovanov, S.; Tselousov, A.; Kurbanov, R.; and Nikolenko, S. I. 2020. Lost in conversation: A conversational agent based on the transformer and transfer learning. In *The NeurIPS'18 Competition*. Springer. 295–315.

Gonimah, D. 2018. Storyful's guide to the social media landscape: Beyond the iceberg metaphor. https://bit.ly/3akqFAw.

Gousios, G. 2013. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR '13. IEEE Press.

Grant, R.; Kucher, D.; León, A. M.; Gemmell, J.; and Raicu, D. 2017. Discovery of informal topics from post traumatic stress disorder forums. In *ICDMW*.

Grant, R. N.; Kucher, D.; León, A. M.; Gemmell, J. F.; Raicu, D. S.; and Fodeh, S. J. 2018. Automatic extraction of informal topics from online suicidal ideation. *BMC bioinformatics* 19(8):211.

Grover, T., and Mark, G. 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *ICWSM*.

Halavais, A. 2019. Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society* 22(11):1567–1581.

Halder, K.; Kan, M.-Y.; and Sugiyama, K. 2019. Predicting helpful posts in open-ended discussion forums: A neural architecture. In *NAACL*.

Hampton, K. N. 2017. Studying the Digital: Directions and Challenges for Digital Methods. *Annual Review of Sociology* 43(1):167–188.

Hess, C., and Ostrom, E. 2003. Ideas, artifacts, and facilities: information as a common-pool resource. *Law and contemporary problems* 66(1/2):111–145.

Hessel, J.; Lee, L.; and Mimno, D. 2017. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *WWW*.

Hidey, C., and McKeown, K. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *NAACL*.

Horne, B. D., and Adali, S. 2017. The impact of crowds on news engagement: A reddit case study. In *ICWSM*.

Howison, J.; Wiggins, A.; and Crowston, K. 2011. Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems; Atlanta* 12(12):767–797.

Hunter, D. 2003. Cyberspace as place and the tragedy of the digital anticommons. *California Law Review* 91:439.

Ingram, M. 2019. Silicon Valley's Stonewalling. *Columbia Journalism Review*.

Japec, L.; Kreuter, F.; Berg, M.; Biemer, P.; Decker, P.; Lampe, C.; Lane, J.; O'Neil, C.; and Usher, A. 2015. Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly* 79(4):839–880.

Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019. "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.* 3.

Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *CHI*.

Jiang, J.-Y.; Chen, F.; Chen, Y.-Y.; and Wang, W. 2018. Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In *NAACL*, 1812–1822.

Johnston, A., and Marku, A. 2020. Identifying extremism in text using deep learning. *Development and Analysis of Deep Learning Architectures* 267–289.

Jonell, P.; Fallgren, P.; Doğan, F. I.; Lopes, J.; Wennberg, U.; and Skantze, G. 2019. Crowdsourcing a self-evolving dialog graph. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, 14. ACM.

Kasper, P.; Koncar, P.; Walk, S.; Santos, T.; Wölbitsch, M.; Strohmaier, M.; and Helic, D. 2017. Modeling user dynamics in collaboration websites. In *Dynamics on and of Complex Networks*, 113–133. Springer.

Keegan, B. 2018. Discovering the Social. http://www.brianckeegan.com/2018/03/discovering-the-social/.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *WWW*.

Kunft, A.; Katsifodimos, A.; Schelter, S.; Rabl, T.; and Markl, V. 2017. Blockjoin: efficient matrix partitioning through joins. *VLDB Endowment*.

Kunft, A.; Stadler, L.; Bonetta, D.; Basca, C.; Meiners, J.; Breß, S.; Rabl, T.; Fumero, J.; and Markl, V. 2018. Scootr: Scaling r dataframes on dataflow systems. In *SoCC*.

Kunft, A.; Katsifodimos, A.; Schelter, S.; Breß, S.; Rabl, T.; and Markl, V. 2019. An intermediate representation for optimizing machine learning pipelines. *VLDB Endowment*.

Lama, Y.; Hu, D.; Jamison, A.; Quinn, S. C.; and Broniatowski, D. A. 2019. Characterizing trends in human papillomavirus vaccine discourse on reddit (2007-2015): An observational study. *JMIR public health and surveillance* 5(1):e12480.

Lazer, D., and Radford, J. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* 43(1):19–39.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Alstyne, M. V. 2009. Computational Social Science. *Science* 323(5915):721–723.

Leetaru, K., and Schrodt, P. A. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA*.

Lorenz-Spreen, P.; Mønsted, B. M.; Hövel, P.; and Lehmann, S. 2019. Accelerating dynamics of collective attention. *Nature communications* 10(1):1759.

Lu, J.; Sridhar, S.; Pandey, R.; Hasan, M. A.; and Mohler, G. 2019. Investigate transitions into drug addiction through text mining of reddit data. In *KDD*.

Manovich, L. 2011. Trending: The promises and the challenges of big social data. *Debates in the digital humanities* 2:460–475.

Marwick, A., and Lewis, R. 2017. Media manipulation and disinformation online: Case studies. https://bit.ly/2JgCkEP.

Massanari, A. 2017. #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19(3):329–346.

Matias, J. N. 2016. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *CHI*.

Matias, J. N. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*.

Medvedev, A. N.; Delvenne, J.-C.; and Lambiotte, R. 2018. Modelling structure and predicting dynamics of discussion threads in online boards. *Journal of Complex Networks* 7(1):67–82.

Mervis, J. 2019. Privacy concerns could derail Facebook data-sharing plan. *Science* 365(6460):1360–1361.

Narayanan, V.; Barash, V.; Kelly, J.; Kollanyi, B.; Neudert, L.-M.; and Howard, P. N. 2018. Polarization, partisanship and junk news consumption over the us. https://bit.ly/2WKrfnl.

Oboler, A.; Allington, W.; and Scolyer-Gray, P. 2019. Hate and violent extremism from an online subculture: The yom kippur terrorist attack in halle, germany. https://bit.ly/3amqBQV.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2.

Ozcan, F. 2017. *Bayesian Nonparametric Models on Big Data*. Ph.D. Dissertation, UC Irvine.

Palen, L., and Anderson, K. M. 2016. Crisis informatics—New data for extraordinary times. *Science* 353(6296):224–225.

Patel, K. S. 2018. Testing the Limits of the First Amendment: How Online Civil Rights Testing is Protected Speech Activity. *Columbia Law Review* 118(5):1473–1516.

Pirina, I., and Çöltekin, Ç. 2018. Identifying depression on reddit: The effect of training data. In *EMNLP Workshop SMM4H*.

Puschmann, C. 2019. An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society* 22(11):1582–1589.

Reddit. 2019. API documentation. https://www.reddit.com/dev/api/.

Rezaii, N.; Walker, E.; and Wolff, P. 2019. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ schizophrenia* 5.

Sarantopoulos, I.; Papatheodorou, D.; Vogiatzis, D.; Tzortzis, G.; and Paliouras, G. 2018. Timerank: A random walk approach for community discovery in dynamic networks. In *Complex Networks*.

Seering, J.; Wang, T.; Yoon, J.; and Kaufman, G. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21(7):1417–1443.

Shen, Q., and Rose, C. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, 58–69.

Squirrell, T. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society*.

Srinivasan, K. B.; Danescu-Niculescu-Mizil, C.; Lee, L.; and Tan, C. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *CSCW*.

Starbird, K. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *ICWSM*.

Tan, C. 2018. Tracing community genealogy: how new communities emerge from the old. In *ICWSM*.

Terrorism, T. A. 2019. Insights from the centre for analysis of the radical right's inaugural conference in london. https://bit.ly/39gKFCQ.

Tsugawa, S., and Niida, S. 2019. The impact of social network structure on the growth and survival of online communities. *ASONAM*.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM*.

Völske, M.; Potthast, M.; Syed, S.; and Stein, B. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, 59–63.

Walker, S.; Mercea, D.; and Bastos, M. 2019. The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society* 22(11):1531–1543.

Wang, F.-Y.; Carley, K. M.; Zeng, D.; and Mao, W. 2007. Social Computing: From Social Informatics to Social Intelligence. *IEEE Intelligent Systems* 22(2):79–83.

Wang, A.; Hula, J.; Xia, P.; Pappagari, R.; McCoy, R. T.; Patel, R.; Kim, N.; Tenney, I.; Huang, Y.; Yu, K.; et al. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *ACL*.

Weller, K., and Kinder-Kurlanda, K. E. 2016. A manifesto for data sharing in social media research. In *WebSci*.

Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; and Stringhini, G. 2018a. Understanding web archiving services and their (mis) use on social media. In *ICWSM*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018b. On the Origins of Memes by Means of Fringe Web Communities. In *IMC*.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019. Who let the trolls out?: Towards understanding state-sponsored trolls. In *WebSci*.

Zhan, Y.; Zhang, Z.; Okamoto, J. M.; Zeng, D. D.; and Leischow, S. J. 2019. Underage juul use patterns: Content analysis of reddit messages. *Journal of medical Internet research* 21(9):e13038.

Zheng, W., and Zhou, K. 2019. Enhancing conversational dialogue models with grounded knowledge. In *CIKM*.

Zhou, Y.; Dredze, M.; Broniatowski, D. A.; and Adler, W. D. 2019. Elites and foreign actors among the alt-right: The gab social media platform. *First Monday* 24(9).

Zhuang, Y.; Xie, J.; Zheng, Y.; and Zhu, X. 2018. Quantifying context overlap for training word embeddings. In *EMNLP*.

Zignani, M.; Quadri, C.; Galdeman, A.; Gaito, S.; and Rossi, G. P. 2019. Mastodon Content Warnings: Inappropriate Contents in a Microblogging Platform. In *ICWSM*.