Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

# WikiReddit: Tracing Information and Attention Flows Between Online Platforms

**Patrick Gildersleve**[1], **Anna Beers**[2], **Viviane Ito**[2], **Agustin Orozco**[2], **Francesca Bolla Tripodi**[2]

[1] University of Exeter
[2] University of North Carolina at Chapel Hill
p.gildersleve@exeter.ac.uk, {albeers@, itovivi@, aorozco@, ftripodi@email.}unc.edu

## Abstract

The World Wide Web is a complex interconnected digital ecosystem, where information and attention flow between platforms and communities throughout the globe. These interactions co-construct how we understand the world, reflecting and shaping public discourse. Unfortunately, researchers often struggle to understand how information circulates and evolves across the web because platform-specific data is often siloed and restricted by linguistic barriers. To address this gap, we present a comprehensive, multilingual dataset capturing all Wikipedia mentions and links shared in posts and comments on Reddit 2020–2023, excluding those from private and NSFW subreddits. Each linked Wikipedia article is enriched with revision history, page view data, article ID, redirects, and Wikidata identifiers. Through a research agreement with Reddit, our dataset ensures user privacy while providing a query and ID mechanism that integrates with the Reddit and Wikipedia APIs. This enables extended analyses for researchers studying how information flows across platforms. For example, Reddit discussions use Wikipedia for deliberation and fact-checking which subsequently influences Wikipedia content, by driving traffic to articles or inspiring edits. By analyzing the relationship between information shared and discussed on these platforms, our dataset provides a foundation for examining the interplay between social media discourse and collaborative knowledge consumption and production.

**Code** — https://github.com/pgilders/WikiReddit
**Dataset** — https://doi.org/10.5281/zenodo.14653265

## Introduction

Researching flows of information, attention, and discussion on online platforms, and how they both reflect and shape public discourse, is a key challenge in understanding the dynamics of the modern web. There is ample work studying these effects within individual communities and platforms (Crane and Sornette 2008; Lehmann et al. 2012; Twyman, Keegan, and Shaw 2017; Kobayashi et al. 2021; Shen and Rosé 2022; Johnson et al. 2021; Matsui, Miyazaki, and Murayama 2024). While this foundational work set the stage for the importance of computational social science, the web is a complex, interconnected ecosystem. Users freely move between platforms, communities, and devices. To better understand how information propagates and spreads, we must consider interactions across platforms. Indeed, an increasing number of multi-platform studies have identified effects not observed or observable within individual platforms (McDowall, Antoniak, and Mimno 2024; Kloo, Cruickshank, and Carley 2024; Moyer et al. 2015; Dubois and Blank 2018). Unfortunately, undertaking this task in research is becoming increasingly difficult. The deterioration of the open web towards that of large centralized platforms where data is often walled off behind APIs, paywalls, or other restrictions, has made it harder to study the web as a whole (Freelon 2018; de Vreese and Tromble 2023). Research opportunities that can bridge these silos are thus of great value.

Wikipedia, as the globally recognized repository of collective knowledge, and Reddit, the foremost social news aggregation and forum site, form an intriguing pairing in this ecosystem. Links to Wikipedia articles on Reddit often serve as references, fact-checking tools, or catalysts for discussion, making them a key touchpoint for understanding how collective attention and knowledge circulates and evolves online (Moyer et al. 2015). However, foundational work on this relationship documented a "paradox of re-use;" private platforms like Reddit boost engagement and revenue by relying on Wikipedia links but the relationship is not reciprocal (Taraborelli 2015; Vincent, Johnson, and Hecht 2018). These studies documented the uneven value derived from Wikipedia volunteerism and shed light on how Wikipedia data sustains the economic interests of private platforms and satisfies informational needs. By expanding on this important work, our database will help researchers better understand the social and information dynamics of these interactions. How Wikipedia links are shared, interpreted, and acted upon within Reddit communities and the wider web—remains relatively underexplored.

We present WikiReddit, a comprehensive dataset capturing all Wikipedia mentions (including links) shared in posts and comments on Reddit from 2020 to 2023, excluding those from private and NSFW (not safe for work) subreddits. The SQL database comprises 336K total posts, 10.2M comments, 1.95M unique links, and 1.26M unique articles spanning 59 languages on Reddit and 276 Wikipedia language subdomains. Each linked Wikipedia article is enriched with its revision history and page view data within a ±10-day win-

dow of its posting, as well as article ID, redirects, and Wikidata identifiers. Supplementary anonymous metadata from Reddit posts and comments further contextualizes the links, offering a robust resource for analysing cross-platform information flows, collective attention dynamics, and the role of Wikipedia in online discourse. This dataset is distinct from those used in other works documenting similar relationships between Wikipedia and Reddit due to its scope (four years), diverse supporting data included, and approach to providing a long-term, sustainable resource, using officially licensed Reddit4Researchers API access. This approach ensures the long-term availability of the resource while adhering to ethical standards and user expectations around data use. By ensuring sustainability, our dataset not only allows for initial analysis in this paper and beyond, but also creates an opportunity for longitudinal analysis moving forward.

In initial explorations of our dataset, we study the use and success of Wikipedia links on Reddit over time, the association between Wikipedia articles being posted on Reddit and changes in page view and editing activity, and finally the distribution of languages used in Reddit posts and linked to on Wikipedia. We find Wikipedia is being referred to in Reddit posts less frequently over the period of study, but the performance of those posts remains stable. There are notable associations between activity on Reddit and Wikipedia; article page views tend to increase on the day the link posts to Reddit and continue, less markedly for a week after. However, the relationship is much weaker for editing activity. Regarding languages, English is the dominant language in the Reddit data, and this is reflected in the Wikipedia links posted. However, there is substantial cross-lingual linking to and from English.

In the following sections, we review related work, articulate how the WikiReddit dataset was developed and summarize its structure. We then undertake some exploratory analysis, before concluding with suggested applications and future research that may be undertaken with the dataset.

## Related Work

Many studies have documented the important role Wikipedia plays in how people find and validate information. Wikipedia content appears in over 80% of knowledge panels and top-linked content across three different search engines (Google, Bing, and DuckDuckGo) comprising a large portion of most user-facing knowledge graph assets (McMahon, Johnson, and Hecht 2017; Vincent and Hecht 2021; Vincent, Johnson, and Hecht 2018). Given the tight integration between search engines and Wikipedia, these studies shed light on how Wikipedia impacts human decision-making and influences other knowledge classification systems (Lerner and Lomi 2018; C. Thompson et al. 2024; Formisano et al. 2024). Corporate-owned platforms also depend on Wikipedia data. Many websites use Wikipedia hyperlinks and content to increase visitation, engagement, and revenue (Gómez-Martínez, Orden-Cruz, and Martínez-Navalón 2022; Lerner and Lomi 2018; Moyer et al. 2015; Vincent, Johnson, and Hecht 2018). Unfortunately, this economic dependency appears nonreciprocal—while Wikipedia's open licenses make it easy for corporate sites to capitalize on its content, it does not produce migratory benefits like more viewership or edits on Wikipedia itself (Vincent, Johnson, and Hecht 2018).

In addition to Wikipedia, audiences consult a range of online services for news including news websites, apps, and social media (St. Aubin and Liedke 2024; Vraga and Tully 2021). People who include, but do not exclusively rely on, social media in their news diets tend to have higher news media knowledge (Schulz, Fletcher, and Nielsen 2024). Among these social media sites is Reddit, a social media platform where community members share, vote, and comment on content and foster community-based engagement on topics or themes—colloquially referred to as subreddits. Previous research has explored the relationship between news coverage and Reddit engagement. Gozzi et al. (2020) demonstrated that COVID-19 news coverage drove users to comment on Reddit and search for information on Wikipedia, though this effect decreased over time—probably due to media saturation. Further research on Reddit has found that fact-checked information lasts longer when a post is deemed true (Bond and Garrett 2023), reinforcing that users rely on Reddit's discussions when interpreting news. However, other work has scrutinized the credibility of information posted to Reddit—without editorial oversight, users may share biased viewpoints and content from known misinformation sources (Chipidza et al. 2022). Tangled into these findings on social media and news consumption is the role Wikipedia might still play in this process. Studies have found that Reddit users regularly rely on Wikipedia hyperlinks to validate information—especially within the "Today I Learned" subreddit (Moyer et al. 2015; Vincent, Johnson, and Hecht 2018). Nonetheless, access to this previous dataset is no longer feasible—since 2023 Reddit's API is no longer available for free public use. Researchers wishing to access Reddit data must now submit an application for access to the "Reddit4Researchers" beta program—a new approach to partnering with data scientists to balance between data accessibility and user protection (Perez 2024).

Datasets like the one we present in this paper are part of a long line of research committed to making Wikipedia data accessible and available to other social scientists. Previous papers have built Wikipedia datasets to assess the quality of content on Wikipedia (Das et al. 2024); study its hyperlink structure (Consonni, Laniado, and Montresor 2019); understand how people interact with 'news events' (Gildersleve, Lambiotte, and Yasseri 2023); and document platform interdependencies (Meier 2022). The organizational structure of the site, its size, and the fact that Wikipedia is open access facilitate these dataset creations (Mitrevski, Piccardi, and West 2020). These studies are also pushing the boundaries of Anglocentrism, creating databases that leverage "language-agnostic" or multilingual techniques to identify linguistic gaps, explain the informational needs of marginalized populations, and analyze how ideas propagate across the languages (Das et al. 2024; Valentim et al. 2021; Miquel-Ribé and Laniado 2019). Creating these datasets is no simple task, Wikipedia is a massive corpus of densely interlinked content, not just a "ready-made data source" (Gildersleve, Lambiotte, and Yasseri 2023; Valentim et al. 2021).

## Dataset Development

### Overview

The dataset is shared as a SQLite3 database via Zenodo: https://doi.org/10.5281/zenodo.14653265. Replication code for collection, exploratory analysis, and demo code is provided in the project repository: https://github.com/pgilders/WikiReddit. Data is collected from the Reddit4Researchers and Wikipedia APIs (MediaWiki 2024). This data from these APIs is licensed under the "Reddit License" (Reddit 2024) and CC BY-SA 4.0 respectively. Following the drastic changes made to the old Reddit API, as well as the shutdown of secondary tools and archives such as Pushshift (Mehta 2024), Reddit has opened a new API for researchers program, which this project relies on and aims to integrate into for future data gathering (u/KeyserSosa 2024). We used the WikiToolkit (Gildersleve 2023) Python package for fast, reliable collection for a variety of Wikipedia data from their APIs and dumps. Data collection and demo code using WikiToolkit is provided in the repository for this article.

### Reddit Data

We used API access to the Reddit4Researchers program to collect data from Reddit. Data is available for 4 years (2020-2023). We collected all posts that mention Wikipedia, either in the title or post body, including content and associated metadata. We also collected all comments on Reddit that mention Wikipedia including content and associated metadata.

For the purposes of this dataset, only post, comment, and subreddit IDs, together with anonymous metadata such as timestamps, score, and extracted Wikipedia links, are shared. All IDs are securely hashed using SHA-256, and checked for uniqueness. This measure preserves individuals' privacy and also enables future researchers to collect and analyze additional data on entries of interest with access to the Reddit4Researchers API by matching against our hashes. This data is stored in the `posts` and `comments` tables.

### Parsing URLs to Articles

Not every post/comment mentioning Wikipedia includes a Wikipedia URL, and not all posted Wikipedia URLs are valid or correctly formatted such that they map to a valid Wikipedia page. Furthermore, there are a variety of ways in which a Wikipedia URL can link to an article (e.g., directly, via a redirect, to a specific revision). We developed a strategy to reliably extract these URLs and identify which articles they link to. The procedure is outlined as follows:

1. Parse the post / comment text with a markdown parser, and extract all correctly hyperlinked Wikipedia URLs.
2. For any malformed links from markdown and the full remaining text, run a split and regex for any further Wikipedia links.
3. For all extracted links, run a validation step to ensure it successfully connects to Wikipedia
4. If the link does not successfully connect, run cleaning steps with regex, removing unnecessary trailing characters.
5. Recheck validity, and resolve any http redirects as necessary for all URLs.
6. For all links, try to extract the Wikipedia language subdomain and article title with regex.
7. If the link is indirect (e.g., to a revision ID), query the Wikipedia API to get the article title.
8. Query the Wikipedia API with the extracted article titles to identify any Wikipedia article redirects. Return this as 'canonical_title'.

This data is stored in the `post_links` (all links from posts), `comment_links` (all links from comments), and `links_articles` (unique valid links that map to an article) tables. A summary of all Wikipedia mentions, links, and articles posted to Reddit is provided in Table 1.

| | Posts | Comments | Total |
|---|---|---|---|
| # mentioning Wikipedia | 335,897 | 10,264,340 | 10,600,237 |
| # with Wikipedia links | 286,359 | 9,465,316 | 9,751,675 |
| Total # of Wikipedia links | 658,493 | 11,573,36 | 12,231,860 |
| # unique Wikipedia links | 295,439 | 1,890,497 | 1,954,003 |
| # unique Wikipedia articles | 252,846 | 1,196,494 | 1,260,479 |

Table 1: A summary of all Wikipedia mentions, those including links, and those that map to Wikipedia articles by Reddit format.

### Wikipedia Data

Wikipedia data is collected via the Wikipedia APIs using WikiToolkit.

**IDs and Redirects** Wikipedia links typically link to an article name. We resolved this name, which might be outdated, or non-canonical, to the current canonical title (i.e., resolve redirects), collected the page ID, and collected all pages that redirect to the article canonical title. In cases where links are to a page ID or revision ID, we similarly gathered the appropriate page ID, canonical title, and redirects as appropriate (as previously indicated). These are stored in the `wiki_ids`, `resolved_redirects`, and `collected_redirects` tables.

**Page Views** Daily page view counts are collected for every article posted to Reddit $\pm 10$ days from initially post/comment date (`created_at`) and $\pm 10$ days from last modified date (`updated_at`, `last_modified_at`). Page views are collected for both the original posted page title and any canonical redirected title. These are stored in the `page_views` table.

**Revisions** For every article posted to Reddit all article revisions (IDs and timestamps) made $\pm 10$ days from initially post/comment date (`created_at`) and $\pm 10$ days from last modified date (`updated_at`, `last_modified_at`) are collected. In addition, the revision at the start of this time period is collected, regardless of when it was created (i.e., the state of the article -10 days from initial post/comment date). These are stored in the `revisions` table.

### Wikidata Data

For each Wikipedia article collected, we also collected their Wikidata identifier (if present), this is also stored in the `wiki_ids` table. This allows for cross-referencing the Reddit and Wikipedia data with Wikidata's structured knowledge graph, as well as interlanguage concept resolution.

### Summary

A summary of the database structure is provided in Table 2.

## Ethical and FAIR Considerations

This work was approved by the University of Exeter ethical review procedures (ID 8969882). The dataset is based on public data collected via the Reddit4Researchers and Wikipedia APIs. This research was observational, and no personally identifiable information from Reddit users was included. Data is limited to publicly accessible posts and comments containing Wikipedia links, excluding private and NSFW subreddits. The dataset conforms to the FAIR principles (Wilkinson et al. 2016) as follows.

- **Findable**: The dataset is publicly available on Zenodo and is assigned a permanent DOI: 10.5281/zenodo.14653265
- **Accessible**: Anyone with an internet connection can freely access the dataset, which is shared under a licensing agreement that ensures long-term availability and responsible use.
- **Interoperable**: The dataset is provided in SQLite3 format and includes multilingual and cross-platform identifiers to enhance compatibility and facilitate integration with other tools and datasets.
- **Re-usable**: The dataset comes with replication and demo code for data collection, and exploratory analysis, making it easy for researchers to reproduce or extend analyses. The dataset is licensed CC BY 4.0.

This dataset could be misused by researchers attempting to make claims about Wikipedia's political effects on other platforms as a basis for erroneously discrediting the Wikipedia platform, which has recently faced increased politicization (Rascouët-Paz 2024). We rely on the academic community to audit and resist bad faith usage of this dataset. Users do not explicitly consent to data collection because their posts are made on a public platform—although researchers have noted that some users, nonetheless, prefer to have their work not used for research (Fiesler and Proferes 2018). Because this dataset online includes IDs linking to usernames, posts, and comments, users can effectively remove their data from future re-collection and use by deleting their associated data on Reddit.

## Exploratory Analysis

**Reddit & Wikipedia Use Over Time** The histograms in Figure 1 and line plots in Figures 2 and 3 document the distribution of Reddit posts or comments that mention Wikipedia either by name or with an article link over time. The long tail distributions of Fig. 1 indicates that the vast majority of posts and comments have a score $\approx 0$, whereas a small number prove to be very high scoring. Unsurprisingly, since they are more readily presented to more Reddit users, posts tend to score more highly than comments.

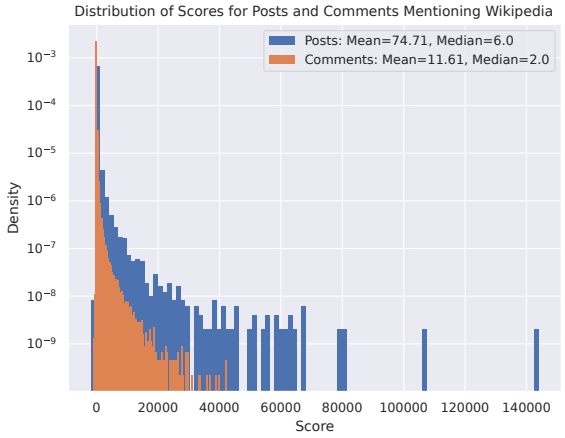


Figure 1: Histograms for the Reddit score of the posts and comments that mention Wikipedia (in text or as a link).

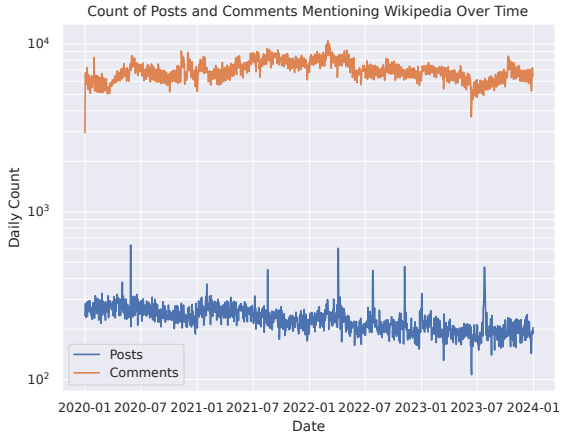


Figure 2: Plot showing the daily count of posts and comments that mention Wikipedia (in text or as a link) over 2020-2023.

When we analyze the distributions longitudinally, we find that Wikipedia is referenced about 8000 times per day in comments and that number has remained relatively stable over time. Original posts that refer to Wikipedia have decreased slightly over time going from approximately 300 posts per day to around 200 posts per day. While not fully answerable solely with our dataset, we hypothesize that these shifts might be related to changes in platform content moderation, engagement preferences with both platforms (e.g., mobile vs laptop) or other technological affordances (e.g., Reddit's transition to more image and video-based participation).

| Column Name | Type | Description |
|---|---|---|
| **Table: posts** | | |
| subreddit_id | TEXT | The unique identifier for the subreddit. |
| crosspost_parent_id | TEXT | The ID of the original Reddit post if this post is a crosspost. |
| post_id | TEXT | Unique identifier for the Reddit post. |
| created_at | TIMESTAMP | The timestamp when the post was created. |
| updated_at | TIMESTAMP | The timestamp when the post was last updated. |
| language_code | TEXT | The language code of the post. |
| score | INTEGER | The score (upvotes minus downvotes) of the post. |
| upvote_ratio | REAL | The ratio of upvotes to total votes. |
| gildings | INTEGER | Number of awards (gildings) received by the post. |
| num_comments | INTEGER | Number of comments on the post. |
| **Table: comments** | | |
| subreddit_id | TEXT | The unique identifier for the subreddit. |
| post_id | TEXT | The ID of the Reddit post the comment belongs to. |
| parent_id | TEXT | The ID of the parent comment (if a reply). |
| comment_id | TEXT | Unique identifier for the comment. |
| created_at | TIMESTAMP | The timestamp when the comment was created. |
| last_modified_at | TIMESTAMP | The timestamp when the comment was last modified. |
| score | INTEGER | The score (upvotes minus downvotes) of the comment. |
| upvote_ratio | REAL | The ratio of upvotes to total votes for the comment. |
| gilded | INTEGER | Number of awards (gildings) received by the comment. |
| **Table: postlinks** | | |
| post_id | TEXT | Unique identifier for the Reddit post. |
| end_processed_valid | INTEGER | Whether the extracted URL from the post resolves to a valid URL. |
| end_processed_url | TEXT | The extracted URL from the Reddit post. |
| final_valid | INTEGER | Whether the final URL from the post resolves to a valid URL after any redirections. |
| final_status | INTEGER | HTTP status code of the final URL. |
| final_url | TEXT | The final URL after any redirections. |
| redirected | INTEGER | Indicator of whether the posted URL was redirected (1) or not (0). |
| in_title | INTEGER | Indicator of whether the link appears in the post title (1) or post body (0). |
| **Table: commentlinks** | | |
| comment_id | TEXT | Unique identifier for the Reddit comment. |
| end_processed_valid | INTEGER | Whether the extracted URL from the comment resolves to a valid URL. |
| end_processed_url | TEXT | The extracted URL from the comment. |
| final_valid | INTEGER | Whether the final URL from the comment resolves to a valid URL after any redirections. |
| final_status | INTEGER | HTTP status code of the final URL. |
| final_url | TEXT | The final URL after any redirections. |
| redirected | INTEGER | Indicator of whether the URL was redirected (1) or not (0). |
| **Table: linkarticles** | | |
| final_url | TEXT | The final URL after any redirections. |
| lang | TEXT | The language code of the page. |
| mobile | INTEGER | Indicator of whether the link was mobile-specific (1) or not (0). |
| raw_title | TEXT | The raw, unprocessed title text extracted from the link. |
| **Table: resolved_redirects** | | |
| lang | TEXT | The language code of the Wikipedia page. |
| raw_title | TEXT | The raw title of the from the Wikipedia link before redirection. |
| norm_title | TEXT | The normalized raw title of the page. |
| canonical_title | TEXT | The canonical title after resolving the redirect. |
| **Table: collected_redirects** | | |
| lang | TEXT | The language code of the Wikipedia page. |
| canonical_title | TEXT | The canonical title of the page. |
| other_title | TEXT | Other titles associated with the page that redirect to the canonical title. |
| **Table: wiki_ids** | | |
| lang | TEXT | The language code of the Wikipedia page. |
| title | TEXT | The title of the Wikipedia page. |
| pageid | INTEGER | Unique identifier for the page in Wikipedia. |
| wikidata_id | TEXT | The Wikidata identifier for the page. |
| **Table: pageviews** | | |
| lang | TEXT | The language code of the Wikipedia page. |
| title | TEXT | The title of the Wikipedia page (not strictly the canonical title). |
| date | TIMESTAMP | The date of the page view count. |
| pageviews | INTEGER | The number of page views on the given date. |
| **Table: revisions** | | |
| lang | TEXT | The language code of the Wikipedia page. |
| canonical_title | TEXT | The canonical title of the Wikipedia page. |
| revid | INTEGER | The unique revision identifier. |
| parentid | INTEGER | The ID of the parent revision. |
| timestamp | TEXT | The timestamp of the revision. |

Table 2: SQL Database Schema with all tables. All IDs from Reddit are hashed using SHA-256 for anonymization.

**Wikipedia Activity** Information from Wikipedia being posted to Reddit may be an indicator of what kinds of information people are paying most attention to. Topics of news attention often play out on Wikipedia (Gildersleve, Lambiotte, and Yasseri 2023), but these attention patterns—sometimes together with the Wikipedia articles to help contextualize the events—also move throughout the internet. Much like the Gozzi et al. (2020) findings, our data indicates that people trying to make sense of news as it unfolds go between both platforms. This may be reflected in the

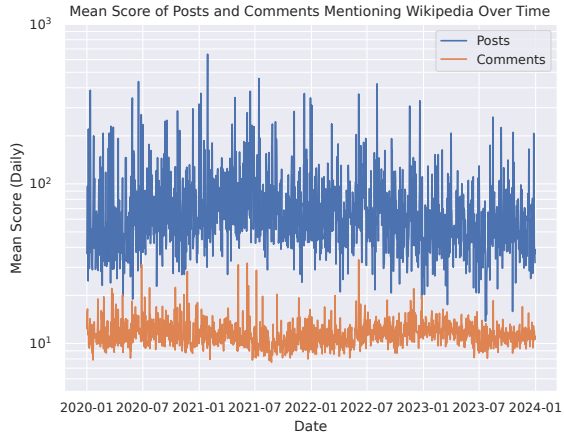


Figure 3: Plot showing the daily average Reddit score of the posts and comments that mention Wikipedia (in text or as a link) over 2020–2023.


Figure 4: Figure showing the daily page views to Wikipedia articles on the day of posting and in the week after posting relative to the week before posting. A small number of points in the extremes of the distributions are cut for visual clarity.

Wikipedia page view and editing patterns of the posted articles. The posting of articles on Reddit may itself also drive interest and editing activity towards Wikipedia.

**Page views:** As an indication of these effects, we calculate and plot (Fig. 4) the relative values of daily page views the day an article is posted and week after it is posted as compared to the week before it is posted ($\frac{\text{Views on date}}{\text{Mean views in week before}}$ and $\frac{\text{Mean views in week after}}{\text{Mean views in week before}}$, where the views on the date of posting are not included in either of the other ranges).

For page views, we first deal with 0 counts, typically indicating an article doesn't yet exist or is deleted by removing them from analysis. We then compute the geometric mean of the relative values, since the distribution is extremely long-tailed. We find that for links in Reddit posts, we observe a 45% increase in page views on the day of posting, and a 6% increase in the week after posting, as compared to the week before posting. For links in Reddit comments, we observe a 45% increase in page views on the day of posting, and a 5% increase in the week after posting, as compared to the week before posting.

The similarity between posts and comments in Figure 4 would suggest much of this activity is due to some external stimulus, as one would assume the level of attention to posts vs comments, and the subsequent spillovers to Wikipedia, would be different. This indicates that audiences are not passive news receivers, but consult different online platforms to support their comprehension of current events.

**Edits:** The picture regarding edits is less clear. This is partially due to the fact that many of the linked articles receive no edits in the days before / on / after being posted on Reddit, in line with the findings of Vincent, Johnson, and Hecht (2018). Nevertheless, we do see some indications of increased editing activity in the wake of being posted. We first remove cases where the pages receive no edits over the full time period, then, due to the smaller, discrete edit values, consider the absolute change in number of daily edits. We compare the number of daily edits on the day of posting on Reddit and in the week after posting against that in the week before posting. Considering the arithmetic mean, for links in Reddit posts, we observe an increase of 0.462 daily edits on the day of posting, but a decrease of 0.031 daily edits in the week after posting, as compared to the week before posting. For links in Reddit comments, we observe an increase of 0.127 daily edits on the day of posting, but a decrease of 0.022 daily edits in the week after posting, as compared to the week before posting. The relatively small effect size here, plus high levels of zero-inflation, and the presence of extreme outliers warrants more rigorous analysis.

To be clear, we present this analysis as a demonstration of association, rather than an investigation of any causal relationship between Reddit and Wikipedia cross-posting. When there is an association, it is likely in the majority of cases that the Reddit posting and page view / editing behavior is mostly in response to some common external stimuli (see discussion of Fig. 4). Studies interrogating any causal relationships are most welcome to be performed using the dataset, making the appropriate decisions for research design, data subsetting, and controls.

**Linking by Language** Both Reddit and Wikipedia are used in different languages. Regarding the use of Wikipedia links on the social media platform, activity is very much English language dominated. 95.8% of Reddit posts with Wikipedia links in the dataset are in English and 93.9% of all linked articles are to English Wikipedia, with the next most links being to German, French, and Spanish Wikipedias (Fig. 5). However, there is substantial cross-lingual linking. 50.85% of links to non-English Wikipedias come from English language Reddit posts. It is also notable how frequently non-English language Reddit posts link to the English language Wikipedia (Fig. 6). The most active other languages

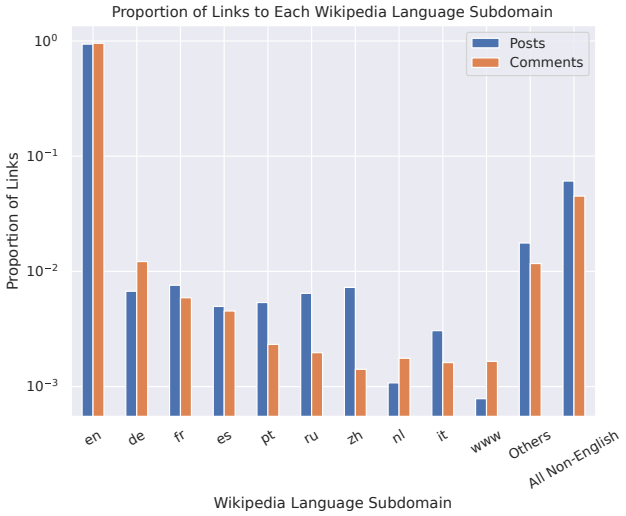


Figure 5: Figure showing the proportion of links to each Wikipedia language subdomain for the most frequently occurring languages in the dataset.

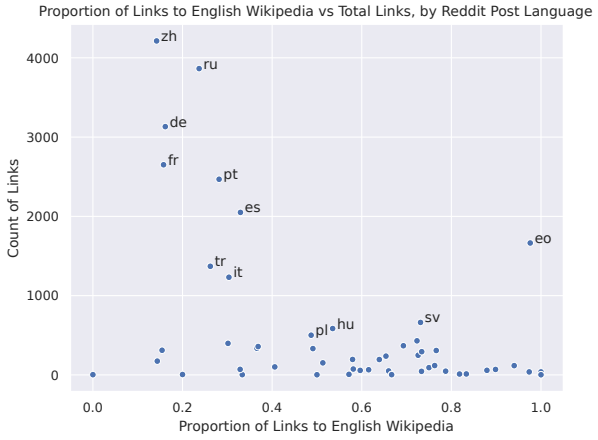


Figure 6: Figure showing the proportion of links to English Wikipedia from non-English Reddit posts vs the total number of links in that post language.

are less likely to link to English Wikipedia. An interesting dynamic is observed for Esperanto (eo), the artificially constructed international second language. Esperanto Wikipedia exists, but Esperanto users on Reddit almost exclusively use English Wikipedia—presumably their first language.

The use of non-English Wikipedia articles in English Reddit posts may highlight knowledge gaps on the English Wikipedia, or be used to compare accounts in different language editions. The extensive use of English Wikipedia in other language Reddit posts strongly highlights (perceived) knowledge gaps in these languages. Unfortunately, it is likely that this linking behavior weakens the already fragile pipeline of contributions coming from social media, lessening the likelihood of edits being made to the Wikipedia language editions in need of them. These results, and future work, can have wide-ranging implications for the multilingual state of Wikipedia and the web.

## Conclusion and Future Work

We present WikiReddit, a comprehensive dataset capturing all Wikipedia mentions (including links) shared in posts and comments on Reddit from 2020 to 2023, excluding those from private and NSFW (not safe for work) subreddits. The SQL database comprises 336K total posts, 10.2M comments, 1.95M unique links, and 1.26M unique articles spanning 59 languages on Reddit and 276 Wikipedia language subdomains. Our exploratory analysis reveals Wikipedia is being referred to in Reddit posts less frequently over the period of study, but the performance of those posts remains stable. We also find notable associations between activity on Reddit and Wikipedia; article page views tend to increase on the day the link posts to Reddit and continue, less markedly for a week after. However, the relationship is much weaker for editing activity. Finally, English is unsurprisingly the dominant language in the Reddit posts and comments, which is reflected in the Wikipedia links posted. However, there is substantial cross-lingual linking to and from English.

We foresee several applications of this dataset and preview four here. First, Reddit linking data can be used to understand how attention is driven from one platform to another. Wikipedia, somewhat uniquely among widely-used internet platforms, publishes all of its traffic data for each of its articles on an hourly basis (Wikimedia Foundation 2024), facilitating fine-grained analyses of how, for example, Reddit posts and comments direct attention to the encyclopedia. This analysis of cross-platform attention to Wikipedia is particularly worthwhile during times of crisis, where viral traffic driven by social media linking can stress the response capabilities of Wikipedia's volunteer editing community (Avieson 2019, 2022).

Second, Reddit linking data can shed light on *how* Wikipedia's archive of knowledge is used in the larger social web. While we have ample research on the uses of Wikipedia within the Wikipedia platform (Singer et al. 2017), and tentative research on the on-platform misuses of Wikipedia (Saez-Trumper 2019; Kharazian, Starbird, and Hill 2024), we know less about how Wikipedia is used or misused off-platform. Prior research gives us reason to believe that even trustworthy content can be recontextualized outside of its authors' intent, unwittingly supporting disinformation campaigns (Beers et al. 2023). Datasets like these can help us understand how the intellectual work of Wikipedia editors travel, and potentially inform how to make Wikipedia's information resilient to malicious decontextualization.

Third, our dataset could provide insights into how external attention is topically distributed across Wikipedia. Many have observed biases both who views Wikipedia and how Wikipedia is edited (Johnson et al. 2021; Tripodi 2023; Menking, Erickson, and Pratt 2019). Our dataset can help extend that analysis into the disparities in what types of external communities Wikipedia is used in, and how it is used.

Fourth, a topic analysis of our dataset could reveal how

Wikipedia usage on Reddit contributes to societal benefits and harms. Reddit has taken steps in recent years to reform its content moderation to address documentation of vitriolic discourse (see Farrell et al. (2019) or Massanari (2017) for earlier examples of sexism/racism). However, both communities still predominantly consist of white men (Gilbert 2020; Menking, Erickson, and Pratt 2019), a demographic imbalance that may contribute to larger inequities in information (Tripodi 2023). Our dataset could help examine if homogeneity within these groups shapes topic patterns and assess whether these relationships mitigate or amplify problematic engagement online.

There are some limitations to what can be done with this dataset. Some analyses of WikiReddit, such as those that recover the full text of Reddit comments, required continued access to the Reddit4Researchers API platform. While Reddit has shown a positive stance towards data access for external researchers via is Reddit4Researchers program, we must be cautious in what has been dubbed the "post-API age" of limited data access from private companies (Freelon 2018). Researchers should also note that URL-sharing is not the only method by which Wikipedia content is shared across platforms like Reddit. For example, text screenshots of external content, previously implicated in the spreading is misinformation (Matatov, Naaman, and Amir 2022), will not be captured here, and surely represent some of the external representation of Wikipedia on Reddit. We thus caution that any full census of Wikipedia's external usage is partial if accessed solely through the in-text mentions and URL-sharing included in this dataset. Nevertheless, the WikiReddit dataset represents an important step forward in understanding how Wikipedia is referenced and engaged with across a major social media platform, and provides a reliable resource for long-term study.

## References


Avieson, B. 2019. Breaking news on Wikipedia: collaborating, collating and competing. *First Monday*.

Avieson, B. 2022. Editors, sources and the 'go back' button: Wikipedia's framework for beating misinformation. *First Monday*.

Beers, A.; Nguyéễn, S.; Starbird, K.; West, J. D.; and Spiro, E. S. 2023. Selective and deceptive citation in the construction of dueling consensuses. *Science Advances*, 9(38): eadh1933. Publisher: American Association for the Advancement of Science.

Bond, R. M.; and Garrett, R. K. 2023. Engagement with fact-checked posts on Reddit. *PNAS Nexus*, 2(3): pgad018.

C. Thompson, N.; Luo, X.; McKenzie, B.; Richardson, E.; and Flanagan, B. 2024. User-Generated Content Shapes Judicial Reasoning: Evidence from a Randomized Control Trial on Wikipedia. *Information Systems Research*, 35(4): 1948–1964.

Chipidza, W.; Krewson, C.; Gatto, N.; Akbaripourdibazar, E.; and Gwanzura, T. 2022. Ideological variation in preferred content and source credibility on Reddit during the COVID-19 pandemic. *Big Data & Society*, 9(1): 20539517221076486. Publisher: SAGE Publications Ltd.

Consonni, C.; Laniado, D.; and Montresor, A. 2019. WikiLinkGraphs: a complete, longitudinal and multi-language dataset of the Wikipedia link networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 598–607.

Crane, R.; and Sornette, D. 2008. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41): 15649–15653.

Das, P.; Johnson, I.; Saez-Trumper, D.; and Aragón, P. 2024. Language-Agnostic Modeling of Wikipedia Articles for Content Quality Assessment across Languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1924–1934.

de Vreese, C.; and Tromble, R. 2023. The Data Abyss: How Lack of Data Access Leaves Research and Society in the Dark. *Political Communication*, 40(3): 356–360. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2023.2207488.

Dubois, E.; and Blank, G. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, communication & society*, 21(5): 729–745.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM conference on web science*, 87–96.

Fiesler, C.; and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1): 2056305118763366. Publisher: SAGE Publications Ltd.

Formisano, G.; Hine, E.; Juneja, P.; Laitila, J.; Novelli, C.; Chiu, E.; Dejanikus, E.; Levin, M.; Schroder, T.; West, A.; and Floridi, L. 2024. Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections. https://papers.ssrn.com/abstract=4990973. Accessed: 2025-01-14.

Freelon, D. 2018. Computational Research in the Post-API Age. *Political Communication*, 35(4): 665–668. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2018.1477506.

Gilbert, S. A. 2020. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1): 19:1–19:27.

Gildersleve, P. 2023. pgilders/WikiToolkit. https://github.com/pgilders/WikiToolkit. Accessed: 2025-01-15.

Gildersleve, P.; Lambiotte, R.; and Yasseri, T. 2023. Between news and history: identifying networked topics of collective attention on Wikipedia. *Journal of Computational Social Science*, 6(2): 845–875.

Gozzi, N.; Tizzani, M.; Starnini, M.; Ciulla, F.; Paolotti, D.; Panisson, A.; and Perra, N. 2020. Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis. *Journal of Medical Internet Research*, 22(10): e21597.

Gómez-Martínez, R.; Orden-Cruz, C.; and Martínez-Navalón, J. G. 2022. Wikipedia pageviews as investors' attention indicator for Nasdaq. *Intelligent Systems in Accounting, Finance and Management*, 29(1): 41–49.

Johnson, I.; Lemmerich, F.; Sáez-Trumper, D.; West, R.; Strohmaier, M.; and Zia, L. 2021. Global Gender Differences in Wikipedia Readership. *Proceedings of the International AAAI Conference on Web and Social Media*, 15: 254–265.

Kharazian, Z.; Starbird, K.; and Hill, B. M. 2024. Governance Capture in a Self-Governing Community: A Qualitative Comparison of the Croatian, Serbian, Bosnian, and Serbo-Croatian Wikipedias. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–26.

Kloo, I.; Cruickshank, I. J.; and Carley, K. M. 2024. A Cross-Platform Topic Analysis of the Nazi Narrative on Twitter and Telegram during the 2022 Russian Invasion of Ukraine. *Proceedings of the International AAAI Conference on Web and Social Media*, 18: 839–850.

Kobayashi, R.; Gildersleve, P.; Uno, T.; and Lambiotte, R. 2021. Modeling collective anticipation and response on Wikipedia. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 315–326.

Lehmann, J.; Gonçalves, B.; Ramasco, J. J.; and Cattuto, C. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, 251–260.

Lerner, J.; and Lomi, A. 2018. Knowledge categorization affects popularity and quality of Wikipedia articles. *PLOS ONE*, 13(1): e0190674. Publisher: Public Library of Science.

Massanari, A. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, 19(3): 329–346. Publisher: Sage Publications Sage UK: London, England.

Matatov, H.; Naaman, M.; and Amir, O. 2022. Stop the [Image] Steal: The Role and Dynamics of Visual Content in the 2020 U.S. Election Misinformation Campaign. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2): 541:1–541:24.

Matsui, A.; Miyazaki, K.; and Murayama, T. 2024. Throw Your Hat in the Ring (of Wikipedia): Exploring Urban-Rural Disparities in Local Politicians' Information Supply. *Proceedings of the International AAAI Conference on Web and Social Media*, 18: 1027–1040.

McDowall, L.; Antoniak, M.; and Mimno, D. 2024. Sensemaking about Contraceptive Methods across Online Platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, 18: 1041–1053.

McMahon, C.; Johnson, I.; and Hecht, B. 2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 142–151. Issue: 1.

MediaWiki. 2024. API:Main page - MediaWiki. https://www.mediawiki.org/wiki/API:Main_page. Accessed: 2025-01-15.

Mehta, I. 2024. Social networks are getting stingy with their data, leaving third-party developers in the lurch — TechCrunch. https://techcrunch.com/2024/02/09/social-network-api-apps-twitter-reddit-threads-mastodon-bluesky/. Accessed: 2025-01-15.

Meier, F. 2022. TWikiL–the Twitter Wikipedia Link Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1292–1301.

Menking, A.; Erickson, I.; and Pratt, W. 2019. People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5970-2.

Miquel-Ribé, M.; and Laniado, D. 2019. Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 620–629.

Mitrevski, B.; Piccardi, T.; and West, R. 2020. WikiHist. html: English Wikipedia's full revision history in HTML format. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 878–884.

Moyer, D.; Carson, S.; Dye, T.; Carson, R.; and Goldbaum, D. 2015. Determining the Influence of Reddit Posts on Wikipedia Pageviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(5): 75–82. Number: 5.

Perez, S. 2024. Reddit locks down its public data in new content policy, says use now requires a contract. https://techcrunch.com/2024/05/09/reddit-locks-down-its-public-data-in-new-content-policy-says-use-now-requires-a-contract/. Accessed: 2025-01-14.

Rascouët-Paz, A. 2024. Elon Musk Urged People to Stop Donating to Wikipedia. Here's Why. https://www.snopes.com//fact-check/elon-musk-stop-donating-wikipedia/. Accessed: 2025-01-15.

Reddit. 2024. Developer Terms. https://redditinc.com/policies/developer-terms. Accessed: 2025-01-15.

Saez-Trumper, D. 2019. Online Disinformation and the Role of Wikipedia. http://arxiv.org/abs/1910.12596. Accessed: 2025-01-13.

Schulz, A.; Fletcher, R.; and Nielsen, R. K. 2024. The role of news media knowledge for how people use social media for news in five countries. *New Media & Society*, 26(7): 4056–4077. Publisher: SAGE Publications.

Shen, Q.; and Rosé, C. P. 2022. A Tale of Two Subreddits: Measuring the Impacts of Quarantines on Political Engagement on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 932–943.

Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1591–1600. ISBN 978-1-4503-4913-0.

St. Aubin, C.; and Liedke, J. 2024. News Platform Fact Sheet. https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/. Accessed: 2025-01-14.

Taraborelli, D. 2015. The sum of all human knowledge in the age of machines: a new research agenda for Wikimedia. In *ICWSM-15 Workshop on Wikipedia*.

Tripodi, F. 2023. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 25(7): 1687–1707.

Twyman, M.; Keegan, B. C.; and Shaw, A. 2017. Black Lives Matter in Wikipedia: Collective Memory and Collaboration around Online Social Movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1400–1412. Association for Computing Machinery. ISBN 978-1-4503-4335-0.

u/KeyserSosa. 2024. Our plans for Researchers on Reddit. https://www.reddit.com/r/reddit4researchers/comments/1co0mqa/our_plans_for_researchers_on_reddit/. Accessed: 2025-01-15.

Valentim, R. V.; Comarela, G.; Park, S.; and Sáez-Trumper, D. 2021. Tracking knowledge propagation across wikipedia languages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 1046–1052.

Vincent, N.; and Hecht, B. 2021. A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–15.

Vincent, N.; Johnson, I.; and Hecht, B. 2018. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. ACM. ISBN 978-1-4503-5620-6.

Vraga, E. K.; and Tully, M. 2021. News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society*, 24(2): 150–166.

Wikimedia Foundation. 2024. Page view. https://meta.wikimedia.org/wiki/Research:Page_view. Accessed: 2025-01-13.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; and Bourne, P. E. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9. Publisher: Nature Publishing Group.

## Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. The data published in this dataset is anonymous and free from any personally id entifiable information.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. We also provide prompts for future work to more robustly analyze the initial exploratory findings.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Methodological choices are explained in the dataset development and exploratory analysis sections.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. In particular we interrogate the linguistic distribution of the dataset.
   (e) Did you describe the limitations of your work? Yes. We note limitations in the concluding section
   (f) Did you discuss any potential negative societal impacts of your work? Yes. In the Ethical and FAIR Considerations section.
   (g) Did you discuss any potential misuse of your work? Yes. In the Ethical and FAIR Considerations section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. Any potentially personally identifiable information in the raw API data is removed for the publication of this dataset.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes. We have made sure to add a section on Ethics and FAIR considerations, as well as follow suggestions in the Datasheets for Datasets article
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA - Dataset paper with no hypothesis tests
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA - Dataset paper with no theoretical results
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA - Dataset paper with no ML experiments
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA - Dataset paper with no ML experiments
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA - Dataset paper with no ML experiments
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA - Dataset paper with no ML experiments
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA - Dataset paper with no ML experiments
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA - Dataset paper with no ML experiments
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes. We cite the Wikipedia and Reddit APIs.
   (b) Did you mention the license of the assets? Yes.
   (c) Did you include any new assets in the supplemental material or as a URL? No.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. In the Ethical and FAIR Considerations section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. As discussed in the Ethical and FAIR Considerations section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. FAIR considerations discussion is provided.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, this is included in the Zenodo record.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA - no human participants

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA