DATASHEET:
Data from "*Extracting Participation in Collective Action from Social Media*"

This document is based on *Datasheets for Datasets* by Gebru *et al.* [2]. Please see the most updated version here.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
Our dataset was created to assess the expression of participation in collective action from textual traces. We performed human annotation of a sample of Reddit comments published on communities centered on activism to serve as training and test sets for the development of a classifier that addresses such a research objective. We aim at filling a gap given by the lack of curated datasets on the topic, as most approaches used to address research on collective action rely on dictionary-based methods or on classifier of a limited subset of collective action types (see "Related Work" section in the paper).

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The creators of the dataset will be disclosed upon paper acceptance.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
Funding will be disclosed upon paper acceptance.

Any other comments?
No.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The data consists of

- **Training set:** One CSV file in which each data point is a Reddit comment, real or synthetic. It contains the subreddit of reference (if applicable), the ID of the comment (if applicable), the original text (if applicable), the text used for annotation and classification (focused on sentences containing high percentages of collective action terms, see "Methodological Framework" in the paper), the assigned labels, and information on the type of augmentation to which the comment belongs.
- **Test set:** One CSV file in which each data point is a Reddit comment. It contains the subreddit of reference, the ID of the comment, the original text, the text used for annotation and classification (focused on sentences containing high percentages of collective action terms, see "Methodological Framework" in the paper), and the assigned labels.

How many instances are there in total (of each type, if appropriate)?
The **Training set** CSV file contains 1817 lines, of which: 369 resulting from the crowdsourced annotated data, 764 resulting from the synthetic augmentation and 684 resulting from the augmentation through Reddit extension. The **Test set** CSV file contains 809 lines.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset is a sample from a larger set of Reddit comments. The larger set consists of 120 million comments posted on 42 subreddits identified as relevant to activism based on titles, descriptions, and manual review. These subreddits were selected from the top 40,000 ranked by subscriber count and contained terms like "activism," "activist," or "rights."

To focus on comments related to collective action, we applied filtering steps: (i) Removed *deleted* and *removed* comments. (ii) Retained only comments with at least two matches to a 47-term corpus of collective action terms [3]. (iii) Extracted sentences with the highest action term frequency, along with preceding and following sentences for context. The resulting dataset includes 2.7 million comments for the training pool and 150,000 for testing. The final training set consists of 369 comments, randomly sampled (50 per

subreddit) and crowd-annotated via MTurk. Quality-control measures reduced the sample size, leading to the under-representation of some subreddits. The test set includes 809 comments, randomly sampled (20 per subreddit, limited to one per author-discussion thread combination) for diversity. Some subreddits have fewer than 20 comments due to this constraint.

While the employed stratification to ensure representativeness of the 42 subreddits, the training set is not fully representative, as filtering prioritized annotation quality over coverage. This choice reflects a focus on data diversity and balance for model training.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Our training and test sets contain a rich set of variables corresponding to each column in the CSV file, providing clear and detailed context for each data element (i.e. comment). Below is an overview of the each column:

- **CommentID**: ID of the retrieved comment as assigned by Reddit, if available. Otherwise, set to "None"
- **Subreddit**: subreddit in which the comment has been published, if available. Otherwise, set to "None"
- **OriginalText**: original text of the Reddit comment, if available. Otherwise, set to "None"
- **ActionFocusedText**: processed text of the Reddit comment, containing the sentence containing the highest percentage of collective action words [3], one before and one after.
- **Label**: Level of participation in collective action assigned by the annotators (majority voting in the case of crowdworkers). The possible levels are: *Problem-solution*, *Call-to-action*, *Intention*, *Execution*, and *None*.
- **SimplifiedLabel**: Binary label of participation in collective action. The possible values are *None* and *Participation*. This variable is derived from **Label**, by assigning *Problem-solution*, *Call-to-action*, *Intention*, *Execution* to the *Participation* class.
- **AugmentationType**: Type of augmentation process performed to create the comment of reference. The possible values are: *Synthetic*, *RedditExtension*, and *None*. This column is only present in the training set.

Is there a label or target associated with each instance? If so, please provide a description.

The targets are the participation in collective action (**SimplifiedLabel**) and its levels (**Label**) and are assigned by the annotators (majority voting in the case of crowdworkers for the training set).

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Yes. CommentID values might be missing because not available in the files used for data collection or because the instance of reference is from the synthetic augmentation set. The synthetic augmentation set is also lacking the reference to subreddits and original text, since it is created by synthetically augmenting the action-focused text only.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

NA.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The training and test sets are provided as already separate entities. A validation set can be defined from the training set, and a sampling fraction of 10% is recommended to reproduce the paper results.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The target labels provided results from a human validation process. While quality controls have been considered, there might be some mistakes in the annotations.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No. All reported text is from comments that were publicly available on Reddit at the time of data collection.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Potentially yes. The data consists of Reddit comments and we cannot control the harmfulness of its content.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, we considered content published by users on Reddit. We only provide information about the comments text and

do not disclose user accounts.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
We acknowledge that our data sample is partial and derived from a limited population of Reddit users interacting in the context of English content.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
If the comments are still publicly available on Reddit (i.e. have not been deleted), it might be possible to derive the authors of such comments from their IDs and/or text. However, we avoid direct sharing of the authors' information.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
Yes, the data consists of Reddit comments and may include information that touches on sensitive attributes. However, it is important to note that these comments were shared voluntarily by their authors in a public forum, with the understanding that Reddit is an open platform where published content becomes accessible to others.

Any other comments?
No.

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data for this study was obtained from publicly available Pushshift [1] Reddit dumps. We conducted checks to ensure data integrity. Specifically, we: (i) Verified that the data matched Reddit's public structure (e.g., correct fields for metadata like timestamps, user IDs, and subreddit names). (ii) Cross-referenced sample data with actual Reddit comments (where accessible) to confirm consistency and authenticity. (iii) Applied preprocessing steps, such as de-duplication and format standardization, to enhance data quality for analysis.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The dataset was created in April 2024. It references Reddit comments through December 2023.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
The data used in this study was sourced from publicly available Pushshift Reddit dumps. As the data originates from Reddit's publicly available content, it includes metadata and comments from user discussions on the platform.

We conducted checks to ensure data integrity. Specifically, we: (i) Verified that the data matched Reddit's public structure (e.g., correct fields for metadata like timestamps, user IDs, and subreddit names). (ii) Cross-referenced sample data with actual Reddit comments (where accessible) to confirm consistency and authenticity. (iii) Applied preprocessing steps, such as de-duplication and format standardization, to enhance data quality for analysis.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[4] for approaches in this area.)
All resources for the extraction from dump files were deployed locally on an Apple M1 Pro machine with 8 cores and 16GB of RAM. The carbon footprint for the extraction of Reddit dump files is estimated around 46g CO^2.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
Random sampling with stratification per subreddit (see previous sections).

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Authors and crowdworkers on Amazon MTurk. Each crowdworker was paid $0.18 per annotation, corresponding to an estimated hourly rate of $7.20 based on 90 seconds per annotation.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any

supporting documentation.
No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
We collected the data using Pushshift public Reddit dumps.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
The individuals involved in the data collection were not directly notified, as the data collection was conducted using information publicly available online.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
NA.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
NA.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Any other comments?
No.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Yes. For the training and test sets, we began our analysis by considering the top 40,000 subreddits ranked by all-time subscriber count, from which we identified a subset of 73 subreddits that contained the terms “activism," “activist," or “rights" in their titles or descriptions. Following a manual review, we selected 42 subreddits pertinent to our study’s scope, containing a total of 120 million comments posted from the creation of these subreddits through December 2023 (Appendix of the paper for the exclusion criteria and the list of selected subreddits). To refine our analysis towards comments more likely related to collective action, we adopted a word-matching approach. Specifically, we used an existing LIWC-like corpus comprising 47 terms associated with collective action [3]. We retained only comments with a minimum of two term matches. We then split the comments into sentences and keep the sentence with the highest frequency of these terms, along with one preceding and one following sentence to preserve the discussion context. We remove deleted and removed comments and split this final dataset of comments into possible candidates for training set (2.7 million comments) and test set (150,000 comments). Moreover, a sampling of data has been performed to produce the final training and test sets as described in previous sections. Data labeling is performed by human annotators, crowdworkers in the case of the training set (see paper for description of the process) and the authors in the case of the test set.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.
We saved the original text of comments in addition to the preprocessed one.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Yes, as part of the code in the shared repository for the paper.

Any other comments?
No.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Yes. The datasets have been used throughout the analysis within the paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
NA.

What (other) tasks could the dataset be used for?
The dataset could be used for other computational social

science tasks centered on group dynamics on social media in the context of collective action.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset should be used responsibly and ethically, avoiding malignancy in persuasion intents.

Any other comments?
No.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is shared on the repository for the paper.

When will the dataset be distributed?
It is shared on the paper repository.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). This license allows others to distribute, remix, adapt, and build upon the work, even commercially, as long as they credit the original creation.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

Any other comments?
No.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is shared on the paper repository and will be hosted on GitHub upon acceptance. Its maintenance will be overseen by the authors of the dataset themselves.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The creators' information will be provided upon the paper's acceptance for publication.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
No.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
The dataset is self-contained and is released communicating its creation date. Thus, its composition reflects the state of Reddit data as of the creation date.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a

description.

The dataset is made available under the Creative Commons Attribution 4.0 International License (CC BY 4.0), ensuring open access. For those interested in extending, augmenting, building upon, or contributing to the dataset, the primary point of contact will be the authors. Their contact information will be provided upon the dataset's acceptance for publication, facilitating direct communication for any collaborative efforts or contributions to the dataset.

Any other comments?

No.

REFERENCES

[1] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[3] Laura GE Smith, Craig McGarty, and Emma F Thomas. After aylan kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time. *Psychological science*, 29(4):623–634, 2018.

[4] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.