DATASHEET:
Data from “*Narratives of Collective Action in YouTube’s Discourse on Veganism*”

Annonymous Submission

This document is based on *Datasheets for Datasets* by Gebru *et al.* [3]. 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 curated to assess the presence of discourse narratives in the context of vegan-related YouTube videos and to explore the ability of such narratives to elicit different levels of collective action in responses. We computed several measures both on video content and comments textual traces to address such a research objective.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Will be disclosed after 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.)

Will be disclosed after 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 one CSV file in which each data point is a YouTube video. It contains relevant descriptive data provided by YouTube API and metrics computed on video content (transcripts) and comments.

How many instances are there in total (of each type, if appropriate)?

The file contains 3,045 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 file results from the mapping of video content to theoretically-defined narratives from social science literature. As such, it contains only videos mapped to specific groups of narrative types (i.e. communal-oriented and agency-oriented) and related to the topic of interest (i.e. vegan challenges). Moreover, the dataset is the result of a series of pre-processing steps on video content, including an English language filter, to guarantee the applicability of validated tools for the operationalization of social science concepts.

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 dataset contains a rich set of metadata corresponding to each column in the CSV file, providing clear and detailed context for each data element (i.e. video). Below is an overview of the metadata for each column:

- **VideoID**: ID of the retrieved video as assigned by YouTube.
- **CommentsTotal**: number of comments for the video as retrieved from the YouTube API.
- **CommentsClean**: number of comments for the video after the pre-processing (consisting of removing urls, mentions and comments with less than five unique words).
- **CareScore**: presence of the *care* moral dimension in the video transcript. Result of scaling the 0-1 ranged scores obtained from the `mformer` model [5] with a set of baseline YouTube videos.
- **FairnessScore**: presence of the *fairness* moral dimension in the video transcript. Result of scaling the 0-1 ranged scores obtained from the `mformer` model with a set of baseline YouTube videos.
- **LoyaltyScore**: presence of the *loyalty* moral dimension

in the video transcript. Result of scaling the 0-1 ranged scores obtained from the mformer model with a set of baseline YouTube videos.

- **AuthorityScore**: presence of the *authority* moral dimension in the video transcript. Result of scaling the 0-1 ranged scores obtained from the mformer model with a set of baseline YouTube videos.
- **SanctityScore**: presence of the *sanctity* moral dimension in the video transcript. Result of scaling the 0-1 ranged scores obtained from the mformer model with a set of baseline YouTube videos.
- **CIIndex**: Collective Identity index derived from first singular and plural person pronouns extracted from video transcript.
- **NarrativeLabel**: narrative mapping to the reference social science theory.
- **SilhouetteScore**: silhouette score of the video given the clustering into narrative types, based on S-BERT [7].
- **VideoCommentAlignment**: cosine similarity between the S-BERT embedding of the video (in terms of transcript) and the centroid of its comments.
- **CollectiveAction**: average relative frequency of collective action markers in video comments, based on the dictionary defined by [8].

Is there a label or target associated with each instance? If so, please provide a description.
The label (NarrativeLabel) is obtained by annotating clusters of videos determined by moral scores and CIIndex features. Such an annotation is performed both manually and through the use of Llama-2 70B Chat [10].

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: videos that are not part of a theory-defined narrative type (i.e. Noise type in NarrativeLabel) and/or have CommentsClean equal to 0 have missing information on VideoCommentAlignment and CollectiveAction. This is because those two metrics are relevant only for videos within theory-mapped narrative types and with at least one valid comment.

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.
NA.

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

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 contains video ids of YouTube videos (VideoID). (a) We cannot guarantee that any content will not be removed or hidden from the platform in the future. (b) The dataset already constitutes a complete version containing resources as they existed at the time of creation. (c) The dataset is self-contained. To get additional metadata from the videos, users must have access to YouTube API.

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.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Pontetially yes. The data refers to YouTube videos and we cannot control the harmfulness of its content. However, we do not share transcripts and comments text to avoiding disclosing possibly offensive data.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes, we considered content published by creators and comments by users. We only provide information about the videos and aggregated measures for the comments.

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 YouTube 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.
Only video creators (i.e. public YouTube channels) can be

identified from the video IDs.

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.
No.

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.
Textual data was derived from YouTube videos, either via auto-captions (through the use of Youtube Transcript API [2]) or by audio-to-text annotation based on Whisper [6], which has been proven to be as robust as human-based transcription. We then used validated tools and measures derived from theoretically grounded concepts to operationalize social science theories.

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 September 2023. It references YouTube videos published between December 2013 and June 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?
We used YouTube API to collect information on YouTube videos and comments. We then used the `mformer` [5] model to extract moral scores from textual traces, a collective identity index derived from personal pronouns fractions, S-BERT embedding-based measures relying on cosine similarity, and a validated dictionary of collective action markers [8]. All these tools have been previously validated or are based on theoretically validated concepts. We then based our narrative mapping on a well-studied process of HDBSCAN clustering [1] based on UMAP embeddings [4] and annotated resulting clusters both manually and through the use of Llama-2 70B Chat [10].

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.*[9] for approaches in this area.)
We utilized a V100 30GB GPU for both extracting moral foundations scores through `mformer` and generating sentence embeddings via `S-BERT`. All other resources were deployed locally on an Apple M1 Pro machine with 8 cores and 16GB of RAM. It was not possible to calculate the carbon footprint due to our reliance on the Hugging Face Inference API for Llama-based annotations. In the case of other models, we did not perform any training except for UMAP and HDBSCAN, which collectively resulted in less than 5 minutes of training on CPU.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
NA.

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)?

NA.

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 textual data using the YouTube API.

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 through YouTube services and 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. We preprocessed video transcripts by excluding those with fewer than five unique words and removing text marked with music tags indicating musical pieces within a transcript. Additionally, we filtered the pre-processed set of videos to retain only English content and conducted topical modeling to ensure high precision by keeping only videos related to plant-based challenges and their impact. As for comments, we removed mentions and URLs, and filtered out texts with fewer than five unique words. Moral dimensions scores were re-scaled by considering a set of baseline videos selected within the most frequent video category for each challenge, of a size similar to the target set and within the same reference timeframe. We labeled each video with its narrative mapping based on a clustering procedure.

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.
No.

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 shared code in the Supplementary Material.

Any other comments?
No.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Yes. Subsets of the dataset 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 on the topic of veganism.

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 will be shared on the platform `figshare`, obtaining a DOI.

When will the dataset be distributed?
Upon acceptance of the submission.

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 will be 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 will be hosted on figshare. 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)?
Contact details will be disclosed upon acceptance.

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 will be released communicating its creation date. Thus, its composition will reflect the state of YouTube 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 will be 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] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[2] Jonas Depoix. Youtube transcript api. `https://github.com/jdepoix/youtube-transcript-api`, 2023.

[3] 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.

[4] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[5] Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. Measuring moral dimensions in social media with mformer. *arXiv preprint arXiv:2311.10219*, 2023.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[7] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[8] 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.

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

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.