Datasheet Template

Content Warning: Trans-antagonistic Rhetoric and Terminology

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The datasheet was created in order to 1. develop a classification pipeline to classify TikTok content as pro-trans, anti-trans, or neutral, and 2. analyze interactions between pro-trans and anti-trans communities on the platform and how they affect the structures of these communities over time.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

Yes. ¡INSERT ARXIV LINK¿

C. What (other) tasks could the dataset be used for?

It could be used to analyze particular kinds of anti-trans rhetoric and the spread of particular pieces of misinformation.

D. Who funded the creation dataset?

The dataset creation is not funded.

E. Any other comment?

No

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

TikTok videos and associated data

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

59,860 videos are in the dataset.

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance in the dataset consists of the coded username of the creator of the video, the video id, date and time of post creation, associated written description, comments, a link to the video on TikTok, a list of other users mentioned in the video, and automated transcription of any audio in the video if possible. For a subset of this data (the largest connected component of the reply network formed by mentions), there is also a label of pro-trans, anti-trans, or neutral associated for each instance in the subset.

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

For a subset of this data (the largest connected component of the reply network formed by mentions), there is also a label of pro-trans, anti-trans, or neutral associated for each instance in the subset.

E. 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.

Some instances do not contain transciptions of audio content due to either not having any audio in the video, the video having been taken down before transcription could be completed, or the audio present in the video failing to be parsed.

F. 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.

Interactions with other users via mentions in the video (i.e. tags, replies, duets, stitches) are made explicit in the dataset via a list of mentioned users for each instance.

G. 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 contains all possible instances, but only a subset is labeled.

H. 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.

There is a recommended data split of 300 instances for validation (manually labeled by annotators).

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

Some mentions associated with videos were also unable to be parsed due to being presented in the form of non-unique display names instead of distinct usernames with no whitespace.

J. 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 links to the original vidoes on TikTok, and is thus dependant on an external source. There are not official archival versions of any videos that are deleted from TikTok after our data collection.

K. Any other comments?

No

III. Collection Process

A. 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?

Data was collected via the Official TikTok Research API by requesting videos containing the following hashtags created between January 1st, 2022 and February 16th, 2024:

- #groomers
- #protectourchildren
- #nooneisborninthewrongbody
- #saveoursinglesexspaces
- #socialcontagion
- #protectwomen
- #leavekidsalone
- #parentalrights
- #leavethekidsalone
- #gendercriticalfeminism
- #gendercriticalfeminist
- #protectthechildren
- #terf
- #letwomenspeak
- #whatisaman
- #genderconfusion
- #realwomen
- #biologicalwomen
- #childrencannotconsent
- #biologicalreality
- #gendercriticalfeminist
- #transdebate
- #genderdebate
- #dylanmulvaney
- #transwomenaremen
- #transwomenaretranswomen
- #transmedicalism
- #transmedicalist
- #transmed
- #transrights
- #transgender
- #transisbeautiful
- #transwoman
- #transwomenarewomen
- #transmen
- #transman
- #transmenaremen
- #transawareness
- #tdov
- #tdor
- #nonbinary
- #nonbinaryawareness
- #transvisibility
- #nonbinaryvisibility
- #transphobia
- #transandrophobia
- #transmasc
- #transmasculine
- #transfem
- #transfeminine
- #transmisogyny
- #enbyphobia
- #protecttranskids
- #trans
- #transsexual
- #ywnbaw
- #genderideology
- #adulthumanfemale
- #sexnotgender

- #gendercritical
- #terftok
- #sexmatters
- #whatisawoman
- #savethetomboys
- #savewomenssports
- #protectwomensspaces
- #femaleerasure

B. 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.

For each instance in the dataset the following are provided via the Official TikTok Research API: the username of the creator of the video, the video id, date and time of post creation, associated written description, comments, and a link to the video on TikTok. We process the written description to generate a list of other users mentioned in the video, and utilize the third party library pyktok to download the associated video in order to extract audio and generate an automated transcription via OpenAI's Whisper. For a subset of this data (the largest connected component of the reply network formed by mentions), we utilized LlaMA 3 to label each sample as either pro-trans, anti-trans, or neutral based on the transcription and written description of the video. We tested the following prompts for this classification task, ultimately selecting the bolded prompt:

- Classify the sentiment of the following text as either pro-transgender, anti-transgender, or neutral.
- Rate the content of the following text from one of the following categories: anti-transgender, pro-transgender, or neutral.
- **Classify the following text as either anti-transgender, neutral, or pro-transgender.**
- Categorize the following text as either pro-transgender, neutral, or anti-transgender.
- Is the following text pro-transgender, anti-transgender, or neutral?
- You are a transgender person advocating for their rights. Evaluate the following text as either anti-transgender, pro-transgender, or neutral.
- Read the following text and determine if it is anti-transgender, pro-transgender, or neutral.
- Is the following text expressing anti-transgender, pro-transgender, or neutral sentiment?

We utilize Retrieval Augmented Generation in our classification pipeline, providing both labeled examples and the text of our classification codebook to the RAG database. The labeled examples consisted of 10 pro-trans, 10 anti-trans, and 10 neutral examples. The text from our codebook was provided sentence by sentence as follows:

- Examples that advocate for abortion rights using language like 'women's bodies' or 'women's health' are erasing trans men, transmasculine people, and other nonbinary people with uteruses. Thus, they should at very least be rated anti-transgender, with the anti-transmasculinity sublabel, with a confidence value of 1 or 2, with increasing levels of confidence as the exclusion becomes more explicit.
- Trans/nonbinary people affirming their own identities should be rated as pro-transgender, with the sublabel "Celebration of Trans Existence".
- Trans/nonbinary people existing and participating in hobbies while out should be rated as pro-transgender.
- Cisgender people refusing to be identified as cisgender, stating that they are not cisgender but simply "a woman" or "a man", should be rated as anti-transgender.
- Attacks on gender-neutral language like "chest-feeding" and "birthing person" should be rated as anti-transgender, with the sublabels anti-transmasculinity, exorsexism, and potentially transmisogyny if the speaker claims the people responsible for this language are "men".
- Claims that "cis is a slur" should be rated as anti-transgender.
- Any language claiming that there's only two genders, or that your gender is what you were assigned at birth, should be rated as anti-transgender with the sublabel exorsexism.
- Examples denying the womanhood of black women are rooted in transmisogynoir and should be rated as anti-transgender, with the sublabel transmisogyny.
- Content claiming that trans men are dangerous to women should be rated as anti-transgender with the sub label anti-transmasculinity
- If the example claims that trans men experience the same gender privilege that cis men do, then it should be labeled as anti-transgender with the sublabel anti-transmasculinity.
- Content that says "trans men are men" specifically in response to a trans man doing something wrong should be rated as anti-transgender with the sublabel anti-transmasculinity.
- Examples claiming to protect women's sports from "men" (referring to trans women) should be labeled as anti-transgender with the sublabel transmisogyny.
- Language attacking gender affirming care, claiming that it's dangerous or that kids shouldn't be able to access it, should be rated as anti-transgender.
- Claims that the "medical establishment" is profiting off of gender affirming care as a way to attack access to it should be rated as anti-transgender.
- Any example of anti-transgender rhetoric that comes from a speaker that explicitly self-identifies as trans in the example should be given the sublabel intracommunity.
- Rhetoric that comes from a speaker that explicitly self-identifies as trans stating that minors shouldn't have

access to gender affirming care should be labeled anti-transgender and given the sublabel intracommunity.
- Content from self-identified trans people mocking non-binary people should be labeled anti-transgender and given the sublabel intracommunity.
- If anti-transgender rhetoric is followed by a user refuting or contradicting that rhetoric, the sample should be labeled as pro-transgender with the sublabel "Refuting Anti-Trans Rhetoric"
- If the example has a transcript that contains pro-trans content, but the description and hashtags indicate that the pro-trans position is being mocked, then the example should be rated as anti-transgender.
- If the example claims to be ok with trans people, but claims that both "extreme" trans people and "extreme" transphobes are equally bad, it should be labeled as anti-transgender.
- If the example mentions how reproductive rights and trans rights are interconnected, then the sample should be labeled pro-transgender with the sublabel "Connection to Broader Liberation".
- If the example mentions how anti-blackness and racism are connected to gender and anti-trans rhetoric, then the sample should be labeled pro-transgender with the sublabel "Connection to Broader Liberation".
- If the example claims to be neutral on trans people, so long as trans people "leave them alone", it should be labeled as anti-transgender.
- Examples positively celebrating trans/nonbinary public figures should be rated as pro-transgender.
- Transmisogyny as an aspect of anti-transgender rhetoric often portrays trans women and transfeminine people as threats to children and cis women.
- Anti-transmasculinity as an aspect of anti-transgender rhetoric portrays trans men and transmasculine people as both helpless, confused, and tricked into being trans, and also predatory and threatening.
- Content that uses the hashtag "#protectthechildren" in the context of gun violence should be labeled neutral
- Examples that use the hashtag "#protectthechildren" in the context of child sexual abuse from religious institutions should be labeled neutral
- Any examples with a transcript that expresses pro-transgender sentiment but has a description that expresses anti-transgender sentiment should be rated as anti-transgender.
- Content that appears unrelated or neutral but contains multiple hashtags relating to trans and queer identites should be rated as pro-transgender, with the sublabel "Celebration of Trans Existence".

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

The subset that was labeled by our classification pipeline consisted of instances belonging to the largest connected component in the either of the reply networks generated (one composed of tags and replies, the other composed of duets and stitches).

D. 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)?

For annotators that are not internal members of the research team, we considered compensation policy in the context of their multiply marginalized status in society, paying the annotator at a rate of $40 per hour for 2.5 hours of work. In addition, we considered power dynamics by hiring any outside annotators as temporary employees as opposed to using crowdwork platforms like Amazon Mechanical Turk, which is unregulated, involves large amounts of unpaid labor in searching for tasks, and enables exploitation through rejection of completed work even as requesters retain access rights to the worker's output [1], [2]. For this dataset, all annotators were internal members of the research team with experience in trans activism as members of the trans/nonbinary community.

E. 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.

Data was collected from June 19th, 2024 through August 9th, 2024. The timeframe in which the data was created spans January 1st, 2022 through February 16th, 2024.

IV. Data Preprocessing

A. 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.

Duplicate instances were removed, and automated transcriptions of audio present in the videos were added to the dataset. Other users mentioned in each video were extracted from the provided description for each instance. Instances tagged with hashtags indicating irrelevance to trans sentiment that frequently co-occurred with the trans related hashtags used for data collection were filtered out. Said hashtags are:

- #dog
- #pet
- #puppy
- #doghair
- #guncontrol

B. 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.

Yes. A link will not be provided because the dataset is not being distributed.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

The effectiveness of this dataset is limited by the accuracy of the classification pipeline, the accuracy of the automated transcriptions, and the lack of accurate information on users mentioned in some instances.

E. Any other comments

No

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

Due to safety concerns for individuals in the dataset, the dataset will not be distributed.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

USC Information Sciences Institute

B. Will the dataset be updated? If so, how often and by whom?

The dataset may be updated for future work by Maxyn Leitner, Rebecca Dorn, Fred Morstatter, and Kristina Lerman. However, these updates will not be made available as the dataset is not being distributed.

C. How will updates be communicated? (e.g., mailing list, GitHub)

Updates will be communicated in future publications, but will not be made available as the dataset is not being distributed.

D. If the dataset becomes obsolete how will this be communicated?

It will not be communicated, as the dataset is not being distributed.

E. Is there a repository to link to any/all papers/systems that use this dataset?

No, as the dataset is not being distributed.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

No, as the dataset is not being distributed.

VII. Legal and Ethical Considerations

A. 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.

Yes, this dataset is created under USC IRB number UP-24-00931.

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

No

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

Yes, some of the transribed audio and linked TikTok videos contain incredibly trans-antagonistic, queer-antagonistic, racist, sexist, and ableist content

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

Yes

E. 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.

No

F. 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.

Individuals who show their face or voice in videos contained within the dataset may be directly identifiable. Publicly famous individuals who appear in the dataset may also be identifiable via video transcription or description. As such, we are not distributing this dataset.

G. 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.

The dataset contains self-expressed racial and ethnic origins, gender identity, sexual orientation, and religious and political beliefs for several instances via video transcriptions and/or descriptions. As such, we are not distributing this dataset.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

It was obtained via the TikTok Official Research API, along with the use of third party libraries.

I. 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.

No

J. 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.

No

K. 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).

N/A

L. 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.

https://github.com/maxynrl/TikTokDatsheet/ImpactAnalysis.pdf

M. Any other comments?

No

REFERENCES

[1] Janine Berg. Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comparative Labor Law and Policy Journal*, 04 2016.

[2] Sarah Roberts. Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: Journal of Mobile Media*, 10:1–18, 01 2016.