Datasheet:

Navigating the Gym Landscape: Analyzing Customer Reviews in Danish Facilities

Chrisanna K. Cornish
IT University of Copenhagen
ccor@itu.dk

Christian M. Hansen
IT University of Copenhagen
chmh@itu.dk

Constantin-Bogdan Craciun
IT University of Copenhagen
cocr@itu.dk

Gino F. Fazzi
IT University of Copenhagen
gifa@itu.dk

Veron Hoxha
IT University of Copenhagen
veho@itu.dk

This datasheet was created following the guidelines and questions suggested in the Datasheets for Datasets paper(Gebru et al., 2018).

1 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.

The dataset was created as a requirement for the Master in Data Science course: Data in the Wild Course at the IT University of Copenhagen, Autumn 2023. The aim was to facilitate research into gyms in the main cities in Denmark, with more detailed labels than the star system provided by platforms such as Google and Trustpilot, to see if there was a significant difference in quality across gym enterprises in the city.

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

The dataset was primarily created by the Data Wild West team: Constantin-Bogdan Craciun, Gino F. Fazzi, Christian M. Hansen, Veron Hoxha and Chrisanna K. Cornish as students in the Data in the Wild course.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

There is no funding associated with this work.

Any other comments?

None.

2 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 dataset comprises customer reviews for fitness facilities in main Danish cities, including as attributes: enterprise, author, review, rating and platform. A significant subset of the dataset was annotated and further enriched with topic level attributes. A full breakdown of the attributes can be seen in Table 1.

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

(i) There are 3,586 unannotated instances.

(ii) There are 608 annotated instances, which are a subset of previous dataset.

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 data is a sample of reviews available via the Google API, scraping Trustpilot and Copenhagen Municipality. Reviews were restricted to gyms physically located within Denmark. Beyond that, the sample is arbitrary.

Field	Description
enterprise	The gym brand
author	The review author name
review	The review text (English)
rating	The rating given (from 1 to 5)
platform	The platform where the review was posted
labels*	Sentiment in specific topics (Staff, Equipment, Hygiene, Location, Not Determined)

Table 1: Dataset Attributes. Those with * are only available in the annotated subset.

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

An instance consists of the following attributes:

- **Enterprise**: One of the collected brands for fitness facilities, namely PureGym, Sats, Vesterbronx or Other, for smaller brands.
- **Author**: The user name of the review author in the platform.
- **Review**: The full text of the review. If the review is originally in Danish, then it is translated to English using machine translation.
- **Rating**: A value from 1 to 5 given to the review.
- **Platform**: The platform the review was posted on: Google or TrustPilot.

In the annotated instances this was then followed by 5 categories: 'Not Determined', 'Staff', 'Equipment', 'Hygiene' and 'Location', which have a value of -1, 0, or 1, representing negative, neutral or positive for each area. Not Determined was used as an 'other' category when the review mentioned something not represented adequately by one of the other categories. An instance can be positive for one category, and negative in another, or any combination.

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

The intention was the 5 annotation categories along with the rating would act as labels, see above.

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.

No data is missing.

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.

Individual instances can be linked through the enterprise, platform or, author.

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 are no recommended data splits. The annotated dataset can potentially be used to automate the annotation process, but there is no strict methodology as to how it should be split for this purpose.

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

The text was translated from Danish to English using machine translation and checked against human translations. This, along with spelling mistakes present in the raw reviews, is a source of inaccuracy within the dataset. Several of these errors were discovered during manual inspection.

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 dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

None.

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.

All data was publicly available online at the time of 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.

The reviews were written by people, and it is possible that offensive, or otherwise unpleasant language/themes were able to bypass the policies of the websites they were left on. However, in the annotated dataset, there were no reviews that should cause undue distress.

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, unless reviewers self-identify in their text. It was not a focus of the data collection, and not explicitly collected. However, it might be possible to infer information about the individual reviewers by their usernames.

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.

The collected reviews contain the reviewer's user name, unless an alias, this acts as a direct identifier. In combination with a named location, this could further be used to identify an individual reviewer.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.

Likewise, this was not data that was explicitly collected, nor was any attempt made to remove. There are some exercise locations scraped from Copenhagen Municipality that cater towards specific conditions, such as dementia, reviews for these locations would potentially disclose the reviewer's condition.

Any other comments?

None.

3 Collection Process

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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Data was collected either from Google reviews via the Google API, or directly from Trustpilot using a webscraper. No data that wasn't publicly available was accessed. Users are free to leave any review they like, as long as it abides to the relevant websites policies. There is no validation as to whether they are actually users of the specific gym, or if their review is valid.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

A combination of APIs and custom software. Initial tests were manually checked for correctness. An annotation guide was produced, and 100 reviews randomly selected to be annotated by all 5 annotators so inter-annotator agreement could be checked.

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

No sampling strategy was used.

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)?

No external data collectors were used, so all data was collected by the Data Wild West team of students. No compensation was involved.

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.

The data was collected between September and December 2023. All of the available Trustpilot reviews were scraped, from 2019-2023 (with a single entry dated 2009). The review dates were not collected from the other data sources.

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.

None.

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

All data came from third-party sources, websites and services which hosted reviews or data on exercise locations.

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.

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.

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).

Not applicable.

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?

None.

4 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 remaining questions in this section.

The review text was translated from Danish to English using a HuggingFace model with no additional training. Preprocessing was conducted in relation for this specific task, but is not preserved in the dataset.

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 the raw data was saved and can be accessed via the Data Wild West - GitHub Repo.

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

The pretrained model can be found at: Helsinki-NLP.

Any other comments?

None.

5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset was used for simple analysis for the course it was developed in (see Motivation section). This has not been published.

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.

Data Wild West - GitHub Repo.

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

Uncertain at this point in time.

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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Not to our knowledge.

Any other comments?

None.

6 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)?

There is no intention to distribute the dataset.

When will the dataset be distributed?

Not applicable.

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.

Not applicable.

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.

Not applicable.

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.

Not applicable.

Any other comments?

None.

7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

There is no intention to continue supporting this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Via one of the emails at the top of this document.

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 dataset consumers (e.g., mailing list, GitHub)?

There is no intention to update this dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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 dataset consumers.

Not applicable.

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 dataset consumers? If so, please provide a description.

There exists a mechanism in the project notebook for others to independently build and/or extend their own local dataset. This is achieved by setting a collection flag to "True". There is no common repository for an individual to contribute towards, and any independently generated dataset will not be validated, as there is no long-term support for the dataset.

Any other comments?

None.

References

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.