# Datasheet for ‘Pizza Hut Ratings and Reviews’*

Chenyiteng Han

17 April 2024

The ‘Pizza Hut Ratings and Reviews’ dataset, available on Kaggle, contains 4000 unprocessed customer reviews and ratings from a specific Pizza Hut outlet in Colombo, Sri Lanka, aimed at tasks like sentiment analysis and trend identification. Created by Kanchana Gajamuthu, it offers a focused glimpse into customer satisfaction at this location but lacks broader representativeness across different branches. The dataset adheres to GDPR compliance by anonymizing personal identifiers. Its limited scope and raw format could influence its utility and interpretation in broader analytical contexts.

Extract of the questions from Gebru et al. (2021).

**Motivation**

1. *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 to analyze customer experiences at a Pizza Hut branch in Sri Lanka for tasks like sentiment analysis, trend identification, and measuring satisfaction through reviews and ratings. It aimed to fill a gap in understanding customer perceptions at this specific location.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by Kanchana Gajamuthu.
3. *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.*
    - No funding or grants were associated with the creation of the dataset.
4. *Any other comments?*

[*] The dataset is distributed via Kaggle, accessible through the following URL: https://www.kaggle.com/datasets/kanchana1990/pizza-hut-ratings-and-reviews/data.

- None.

## Composition

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

    - The instances in the dataset represent customer ratings and textual reviews of the Pizza Hut outlet at Union Place Colombo. There are no multiple types of instances; it strictly includes customer reviews (text) and ratings (stars) related to this specific Pizza Hut location.

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

    - There are 4000 instances included in total.

3. *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 (for example, 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 (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset is a sample, containing only reviews and ratings from the Pizza Hut outlet at Union Place, Colombo, one of the pioneer branches in Sri Lanka. It is not representative of the larger set of all Pizza Hut branches globally or even nationally within Sri Lanka. This sampling is due to focusing specifically on one location, capturing experiences and feedback that are unique to that branch at a specific point in time. This focus means that the dataset may not adequately represent the diverse conditions, management practices, or customer experiences across other Pizza Hut locations. Changes in service levels, management, or local conditions at different times could lead to variability in reviews that are not captured in this dataset.

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

    - Each instance in the dataset consists of "raw" data, including a customer-provided rating (stars) on a scale from 1 to 5 and the customer's written review (text). There is no mention of processed features or additional data derived from these raw instances.

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

- The dataset does not have additional labels or targets associated with each instance beyond the customer-provided ratings and textual reviews. These elements serve as the primary data points, with no further classification or annotation mentioned.

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

    - The dataset has been maintained in its raw state, and no specific measures have been taken to address or impute missing data. The datasheet does not detail why certain information might be missing, other than a general approach to preserving privacy by removing personal identifiers. There's no mention of redacted text beyond the removal of commenter names for privacy.

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

    - Each instance stands alone, with customer reviews and ratings specific to their experience, without linking or relating these instances to one another in the dataset.

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

    - No.

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

    - The dataset contains customer feedback in its original form and has been minimally processed to remove personal identifiers, which means: There might be errors such as inaccuracies in ratings or text due to the data scraping process. Noise could be present in the form of irrelevant comments or spam in the review texts. Redundancies may exist if reviews were duplicated during collection. These issues could affect the analysis, and users of the dataset might need to perform additional cleaning steps to ensure accuracy in their findings.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, 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 (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, 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.*

- The dataset is self-contained and does not explicitly link to or rely on external resources for its core content, which only consists of customer ratings and textual reviews for the Pizza Hut outlet at Union Place Colombo.

11. *Does the dataset contain data that might be considered confidential (for example, 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.*

    - The dataset has been processed to remove personal identifiers, ensuring compliance with privacy standards such as GDPR, but it does not mention containing data that would be considered confidential, such as data protected by legal privilege, doctor-patient confidentiality, or the content of individuals' non-public communications. The focus is on publicly available customer reviews and ratings.

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

    - This dataset includes customer ratings and text reviews of Pizza Hut stores, focusing on customer experience. Some customer reviews may contain potentially sensitive or harmful content.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The dataset does not include demographic information about the reviewers, such as age or gender, adhering to privacy standards. Therefore, it does not identify sub-populations or provide distributions of these within the dataset.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, it is not possible to identify individuals directly or indirectly from the dataset, as personal identifiers have been removed to ensure privacy.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, 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.*

    - The dataset does not contain data that would be considered sensitive, such as information revealing race, ethnic origins, sexual orientations, religious beliefs, political opinions, locations, financial or health data, biometric or genetic data, government identification numbers, or criminal history. It focuses on customer reviews and ratings for a Pizza Hut outlet, without including such sensitive personal information.

16. *Any other comments?*

    - None.

**Collection process**

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, 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.*

    - The data was directly acquired from customer reviews using Apify, a tool for scraping Google Reviews, making the data directly observable.

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

    - The data was collected using Apify, a software tool for scraping Google Reviews.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset focuses on reviews from a specific Pizza Hut location in Colombo due to its long-standing reputation and historical significance as one of the first branches in Sri Lanka. The selection of this branch was intentional but not based on a formal sampling strategy.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data was scraped by the dataset authors using Apify, and the collection was automated and did not involve direct human collectors such as students, crowdworkers, or contractors.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, 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 on 11th March 2024 and includes customer reviews and ratings up to that date. The timeframe of data collection matches the creation timeframe of the data instances, as they were all recent and contemporaneous up to the collection date.

6. *Were any ethical review processes conducted (for example, 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.*

   - Ethical considerations focused on privacy standards like GDPR compliance by removing personal identifiers from the data. There were no formal ethical review processes, such as those conducted by an institutional review board, nor details on outcomes or supporting documentation regarding ethical reviews.

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

   - The data was collected via a third party, specifically through the use of Apify, a tool for scraping Google Reviews. This method involved obtaining customer reviews and ratings from the website directly, rather than collecting data from individuals directly.

8. *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 were not specifically notified about the data collection for this dataset, as the data was collected using a web scraping tool from Google Reviews. Generally, users posting reviews on such platforms agree to the site's terms of service, which might include data usage by third parties.

9. *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.*

   - The individuals were not specifically notified about the data collection for this dataset, as the data was collected using a web scraping tool from Google Reviews.

10. *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).*

    - Consent from individuals for collecting their reviews was not specifically obtained for this dataset, as it was gathered using automated scraping from Google Reviews. Therefore, no mechanism to revoke consent was provided for this particular dataset collection.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, 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 analysis of the potential impact of the dataset and its use on data subjects, such as a data protection impact analysis, has been conducted.

12. *Any other comments?*

    - None.

### Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, 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.*

    - Yes, preprocessing of the data was done, which involved the removal of personal identifiers from the reviews to ensure compliance with GDPR and to maintain reviewer privacy. Beyond this, the dataset presents customer ratings and comments in their raw, unaltered form with no additional processing or labeling such as tokenization or part-of-speech tagging. There is no process for missing values either.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The "raw" data, prior to preprocessing for the removal of personal identifiers, is not preserved separately from the preprocessed data.

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

    - The preprocessing of the dataset was done using R language (R Core Team (2022)) and RStudio.R and RStudio can be accessed via the following links: R: https://cran.r-project.org/. RStudio: https://www.rstudio.com/products/rstudio/download/. These tools are publicly available for data processing tasks.

4. *Any other comments?*

    - None.

### Uses

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

    - No.

2. *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.*

    - No.

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

    - The dataset is ideal for sentiment analysis, trend identification, and developing predictive models to enhance customer service and compare performance across different locations.

4. *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 (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - This dataset, focusing solely on a single location and lacking demographic data, requires careful contextual consideration to avoid biased conclusions, with mitigation possible through combining it with broader datasets and adhering to ethical analysis guidelines.

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

    - The dataset should not be used for tasks requiring demographic insights or generalizations across the broader Pizza Hut brand or fast-food industry, due to its specific focus on one location and lack of demographic data.

6. *Any other comments?*

    - None.

**Distribution**

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - The dataset is publicly available on Kaggle under the ODC Attribution License, allowing for distribution to third parties as long as they credit the original source.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is distributed via Kaggle, accessible through the following URL: https://www.kaggle.com/datasets/kanchana1990/pizza-hut-ratings-and-reviews/data. Additionally, it can be downloaded using the Kaggle API (Kaggle (2022)) with the command: kaggle datasets download -d kanchana1990/pizza-hut-ratings-and-reviews. The dataset does not have a Digital Object Identifier (DOI).

3. *When will the dataset be distributed?*

    - The dataset is already available for access as of the last update provided in the document, which mentions the data collection date as March 11, 2024. It is currently hosted on Kaggle, where it can be accessed by interested parties

4. *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 ODC Attribution License, allowing free use and adaptation as long as proper credit is given and any modifications are noted, with no associated fees.

5. *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 third-party IP-based or other restrictions are imposed on the dataset, which is freely available under the ODC Attribution License with no associated fees.

6. *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 export controls or regulatory restrictions apply to the dataset.

7. *Any other comments?*

    - None.

**Maintenance**

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset is hosted and maintained on Kaggle, as indicated by the availability information provided in the document.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The owner of the dataset, Kanchana Gajamuthu, can be contacted through their Kaggle profile at https://www.kaggle.com/kanchana1990.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - No.

4. *Will the dataset be updated (for example, 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 (for example, mailing list, GitHub)?*

   - The dataset is not expected to be updated.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, 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.*

   - There are no specific restrictions or policies regarding the retention of the data associated with the instances in the dataset. This means the data does not have a predefined deletion or expiration timeline based on the information provided in the dataset's documentation.

6. *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.*

   - Older versions of the dataset do not exist, as the current version is the first and only version. Additionally, there are no plans for future updates to the dataset.

7. *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 is no formal process in place for extending, augmenting, or contributing to the dataset. If contributions are desired, individuals interested in making enhancements should contact the dataset author directly, likely through their Kaggle profile or other provided contact information. Any contributions would need to be coordinated personally with the dataset's curator, who would handle the integration and validation of the new data on a case-by-case basis.

8. *Any other comments?*

   - None.

# References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Kaggle. 2022. "Kaggle API." https://www.kaggle.com/docs/api.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.