Datasheet for 'MyAnimeList(MyAnimeList 2023) Dataset'*

Dataset for anime analytics

Doran Wang

Invalid Date

This dataset is a collection of data from MyAnimeList(MyAnimeList 2023), including manually compiled 'num_favorites' data, intended for analysis of anime popularity and user preferences.

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 enable analysis of anime and user preferences from MyAnimeList(MyAnimeList 2023). Specifically, the gap addressed was the unavailability of "num_favorites" data through the API, which was manually collected to enhance the dataset.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Doran Wang created the dataset.
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.*
   - The dataset creation was self-funded.
4. *Any other comments?*
   - The dataset combines automated API data and manually collected information to ensure completeness.

*Code and data are available at: https://github.com/Wang20030509/Anime_Rating.

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.*

    - Each instance represents an anime entry from MyAnimeList(MyAnimeList 2023), including details like title, genre, rating, and the manually added number of favorites.

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

    - 50 obs (I set the limit, it can be more).

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 subset of all anime entries on MyAnimeList(MyAnimeList 2023), focusing on a curated selection relevant for analysis.

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 includes features such as title, score, genre, and num_favorites (manually collected).

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

    - No specific label or target was defined; the dataset is for exploratory analysis and modeling.

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.*

    - Some attributes, like character details or specific user interactions, are not included because they are not available via the API.

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.*

    - No explicit relationships are included.

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.*

    - Splits can be created based on user-defined analysis needs, such as genre or rating distribution.

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

    - Manual entry of "num_favorites" may introduce small human errors despite best efforts for accuracy.

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 relies on MyAnimeList(MyAnimeList 2023) (MAL) as the source. Since MAL is a live platform, changes to its structure may affect reproducibility.

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 observable through the use of the MyAnimeList API(*MyAnimeList API (Beta Ver.) (2)* 2024). Information about anime titles, ratings, popularity, number of users, and other metrics were directly extracted via automated requests to the API

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?*

    - Data was collected through the MyAnimeList API (*MyAnimeList API (Beta Ver.) (2)* 2024), which provides detailed information on anime, including ratings, popularity, and other related metadata.

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

    - I sampled popular anime titles, I use a probabilistic sampling strategy based on user ratings or popularity

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)?*

    - Student: Doran Wang

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.*

    - Data was collected over November 2024. The dataset reflects the current state of information as of the last API query on November 30, 2024.

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.*

    - The dataset was collected using publicly available API data from MyAnimeList(MyAnimeList 2023), which does not involve personal information or sensitive data. As such, no formal ethical review was required. However, the dataset complies with the terms of use provided by the MyAnimeList(MyAnimeList 2023) platform.

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 obtained from a third-party source, specifically the MyAnimeList API. The dataset does not involve collecting personal 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.*

    - As the data was collected from the MyAnimeList(MyAnimeList 2023) platform, which is publicly available and adheres to the site's terms and conditions, individual notifications were not required. The platform's terms of use govern the collection of data through the API.

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 dataset does not involve direct data collection from individuals, as it relies on publicly available information from MyAnimeList(MyAnimeList 2023)'s platform. MyAnimeList(MyAnimeList 2023) users consented to the use of their data through their agreement to the platform's terms of service, which govern the use of publicly available data.

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).*

    - Since the dataset does not involve the direct collection of personal data from individuals, there was no mechanism for individuals to revoke their consent.

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 data protection impact analysis was required, as the data does not involve personally identifiable information or sensitive data.

12. *Any other comments?*

    - The dataset complies with MyAnimeList's API terms of service, and no personal information is included in the dataset. It is intended for use in analysis of anime-related trends and statistics, without any personal data related to the users of MyAnimeList(MyAnimeList 2023).

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.*

    - Not

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.*

    - None

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

    - None

4. *Any other comments?*

- None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has been used for personal analysis related to anime ratings, popularity, and trends. The analysis includes exploring the relationship between user ratings and various attributes such as genre, score, and ranking.
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.*
    - The dataset is used for personal projects and is hosted on GitHub. You can access all project content at https://github.com/Wang20030509/Anime_Rating.
3. *What (other) tasks could the dataset be used for?*
    - The dataset could be used for sentiment analysis of anime ratings, recommendation systems based on user preferences, analyzing trends in anime genres, or exploring how popularity correlates with ratings. Additionally, it could be used for modeling user behavior or predicting anime success.
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?*
    - The dataset contains only publicly available data from the MyAnimeList API, which does not include any personally identifiable information. However, users of the dataset should be aware that the dataset could reflect certain biases inherent in the MyAnimeList(MyAnimeList 2023) community, such as demographic skew (e.g., age or location of users). This could influence the generalizability of any analysis. To mitigate this, users should consider supplementing this data with more diverse sources or applying fairness-aware techniques when analyzing user behavior.
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 any task that involves personal data analysis or attempts to infer sensitive personal characteristics, as it does not contain such data. It should also not be used to make assumptions about the broader population of anime viewers, as the dataset may not be representative of all global anime fans.
6. *Any other comments?*

- The dataset is regularly updated with new data from MyAnimeList(MyAnimeList 2023) to ensure its relevance and accuracy for ongoing analysis.

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 will be made publicly available on GitHub, and it may be used by third parties for various research or personal analysis purposes. No specific third-party entities are involved in the distribution at this time.

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 will be distributed via a GitHub repository, where users can download the data directly. It does not have a DOI at this time.

3. *When will the dataset be distributed?*

    - The dataset is already available for public access on GitHub at https://github.com/Wang20030509/An and will continue to be updated as new data is collected.

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 will be distributed under a permissive open-source license, such as the MIT License, which allows anyone to use, modify, and distribute the dataset, with the condition that proper attribution is given. A link to the license will be provided in the GitHub repository.

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.*

    - The dataset is based on publicly available data from the MyAnimeList API, which may have its own terms of use. Users should review the MyAnimeList(MyAnimeList 2023) Terms of Service before using the data.

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.*

    - There are no known export controls or regulatory restrictions that apply to the dataset at this time.

7. *Any other comments?*
    - The dataset will be maintained and updated periodically to reflect new data available via the MyAnimeList API. Users are encouraged to check the GitHub repository for the latest versions.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be maintained by Doran Wang, the author of the GitHub repository.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The dataset owner can be contacted via email at doran.wang@mail.utoronto.ca.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - There is no erratum at the moment, but any necessary corrections will be updated in the GitHub repository.
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 will be updated regularly based on new data from the MyAnimeList API. Updates will be made as the API provides new information. Updates will be communicated via the GitHub repository, with details in the commit logs.
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.*
    - The dataset does not contain personal data, so retention and deletion limits do not apply.
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 may not be maintained; however, users can access the commit history on GitHub to retrieve past versions if needed.
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.*

- Contributions can be made via pull requests on the GitHub repository. All contributions will be reviewed and validated by the dataset owner before being merged. Any accepted contributions will be communicated via the repository's commit logs.

8. *Any other comments?*

   - No further comments at this time.

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.

MyAnimeList. 2023. *MyAnimeList: Anime and Manga Database and Community*. MyAnimeList Co., Ltd. https://myanimelist.net/.

MyAnimeList API (Beta Ver.) (2). 2024. https://myanimelist.net/apiconfig/references/api/v2.