Datasheet for '2014-2024 Metacritic video game data'*

Ziqi Zhu

November 29, 2024

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 factors influencing a video game's likelihood of winning the prestigious Game of the Year (GOTY) award at The Game Awards (TGA). Because I can't found a public dataset that aggregate info for video games from all platforms in recent years. While platforms like Metacritic aggregate reviews, and genre in formations of game.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset is crated by Ziqi Zhu with scarping of publicly available data sourced from Metacritic (Metacritic 2024), a well-known aggregation of reviews for video games, films, and other media.
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 is created by Ziqi Zhu, for the purpose of this study only.
4. *Any other comments?*
    - No

Composition

*Code and data are available at: https://github.com/zzq20010617/2024TGA-goty-predictions.

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 video games, with each row corresponding to a single game or downloadable content(DLC) of a game.

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

    - There are 4207 entries for analyse.

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 of video games released between 2014 and 2024 across all platforms. It was specifically curated based on the availability and visibility of games on Metacritic, which primarily lists games with a relatively large influence (e.g., games that have received sufficient attention from users or critics to warrant aggregated reviews).

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 corresponds to a single video game and includes features(e.g., name, score) that have been preprocessed and engineered from raw data sourced from Metacritic.

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

    - Yes, goty_status is a binary variable indicating whether a video game won the Game of the Year (GOTY) award at The Game Awards, 1 means game won GOTY for its release year, 0 otherwise.

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.*

    - No critical information is missing from individual instances in the dataset after the data cleaning process.

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

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.*

   - No redundancies after cleaning.

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.

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.*

    - No

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.*

    - No

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.*

    - No

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

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.

- No

16. *Any other comments?*

- No

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 associated with each instance was directly observable, as it was collected by scraping publicly available information from Metacritic.

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?*

- Python script is used to scrape data from website. The original script is created by Bruno Vieira Ribeiro and posted on Github (Ribeiro 2021) and updated by Ziqi Zhu according to use of this study and the change on Metacritic website, updated version can be find in repo (Zhu 2024).

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 is scraped from Metacritic, initial sampling is done by Metacritic.

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 process is done by creater.

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 is collected on November 26, 2024, but it encompasses video games released between 2014 and 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.*

    - Not applicable

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

    - Data were obtain from third party website Metacritic

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.*

    - Not applicable

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.*

    - Not applicable

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).*

    - Not applicable

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

12. *Any other comments?*

    - No

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, the data cleaning process can be find in appendix section in relate paper in Github repo (Zhu 2024)

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 can be found in data/01-raw_data directory from Github repo (Zhu 2024)

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

   - Data was cleaned and labeled by R.

4. *Any other comments?*

   - No

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?*

   - It could be used to analyzing trend of game reviewing system by time and study the correlation between user scores and critic scores.

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?*

   - Some modification have done based on original collected data, genres with fewer than five games were grouped into an "Other" category to simplify analysis.

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

   - Any commercial use is forbidden.

5. *Any other comments?*
    - No

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 not be distributed to third parties outside of its creator's control. It was developed independently by Ziqi Zhu for academic research purposes and is intended for use only within the scope of this study.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Not applicable
3. *When will the dataset be distributed?*
    - Not applicable
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.*
    - Not applicable
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.*
    - Not applicable
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.*
    - Not applicable
7. *Any other comments?*
    - No

Maintenance

Note that this dataset will not be updated after the completion of this study. This decision was made to avoid repetitive web scraping, which could place unnecessary load on the source platform,

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be hosted by the creator, Ziqi Zhu.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Through e-mail of author for this datasheet ziqi.zhu@mail.utoronto.ca
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)?*
    - No
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 information collected is at the game level and does not include any personally identifiable information.
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.*
    - No
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.*
    - No external contributions are accepted.
8. *Any other comments?*
    - No

References

Metacritic. 2024. "Metacritic: Browse Games by Score." https://www.metacritic.com/browse/game/.

Ribeiro, Bruno Vieira. 2021. "projectGames." *GitHub Repository*. https://github.com/BrunoBVR/projectGames; GitHub.

Zhu, Ziqi. 2024. "2024TGA Goty Prediction." *GitHub Repository*. https://github.com/zzq20010617/2024TGA-goty-predictions; GitHub.