1. Motivation

1.1 *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.*

Since 2011, live streaming has become a popular interactive form of internet-based multimedia entertainment on a global scale (Needleman, 2015, Twitch, 2017). Live streaming has grown so commonplace that, in some instances, more people watch other people perform tasks, like play video games, than actually carry out those tasks themselves (Kaytoue, Silva, Cerf, Meira Jr, & Rassi, 2012).

Twitch, one of the main destinations for gamers and game fans, has embraced this trend. With 140 million active users per month and 9.2 million monthly streams, Twitch is even in competition with YouTube. According to Statista's projections, the gaming sector will expand at a compound annual growth rate (CAGR) of more than 10.5% between 2021 and 2026. By 2026, the value of the worldwide games market is expected to be close to $256 billion. In addition to Twitch, there are other streaming services, such as YouTube gaming and Facebook gaming. Several streaming websites were investigated, and Twitch's website was ultimately chosen as the best option for scraping. Especially since the number of streamers on Twitch is rapidly increasing and there is a wealth of information available about the streamers themselves.

Twitch additionally offers a partner program for anyone who wants to make money by broadcasting on the website – all of which makes it one of the most successful platforms out there. The website's users spend more than 20 hours a week on it, giving it a great opportunity for advertisers to connect with Gen Z. Users and partnerships undoubtedly cross in some way. Since its launch in June 2011, Twitch has been a global phenomenon, with an annual increase in partners. The number of users receiving money from subscriptions and advertisements increased substantially from 3.400 in 2012 to 40.000 in 2020. The future looks bright for this quickly growing streaming website.

Our goal is to scrape data from the 'about page' of approximately 24 to 30 popular Twitch categories so that researchers are able to investigate the revenue model of those streamers, for example. The categories are based on popularity (viewer count) during the moment of scraping, but can also be adapted to the researcher's preferred categories if necessary.

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

The dataset of twitch.tv has been scraped by a project group of the course Online Data Collection and Management. This course has been given by Hannes Datta at Tilburg University and is part of the Master Marketing Analytics.

1.3 *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.*

There was no funding or grant for the development of this dataset.

2. Composition

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

Each instance of the dataset represents the 'about-page' of a live streamer in one of the popular categories on Twitch.tv. The instance gives not only information about the streamer, but also about their socials, advertisements, the amount of followers and their subscription model.

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

Each category contains, on average, 130 instances, dependent on the viewer count for each respective streamer. When combining the number of categories and the number of streamers scraped per category, the dataset will contain, approximately 3,800 instances.

2.3 *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a*

[1]

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 that is generated has been limited to an X number of viewers. This allows for a higher technical feasibility. The number of viewers that has been decided on as a cut-off value is 50 viewers. After this value, the scraper will stop scrolling, thus stopping the scraper from collecting streamerdata below the rendered streamers. A cut-off value of 50 has been chosen, as this dataset's main purpose is to generate insights on the revenue models and sponsordeals of streams, that hardly occur when a streamer has below the viewer threshold of 50 viewers. This does mean that the dataset is not entirely representative of the entire population, however, it is representative of the top 1% of streamers (bron 1%). Next to this, the categories being scraped are based on popularity for the same reason as the streamer's cut-off value. This does not result in a fully representative sample, but does include the main revenue-generating categories, as it is based on viewercount.

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

Variable	**Description**	**Type of data**
Time of collection	Time of which the data of the streamer is collected.	Metadata
Name of the streamer	The name of the streamer.	Unprocessed text
Category	The category of the live streaming.	Unprocessed text (if applicable, else metadata: 'NA')
Team name	Sometimes live streamers are also part of a team (a team with other live streamers).	Unprocessed text (if applicable, else metadata: 'NA')
Followers	Number of followers of the streamer.	Unprocessed text
Content block	Content on the 'about page' of each streamer. See below for further explanation.	URL to image
Textual content block	Information about the streamer. This can differ among streamers. See below for further explanation.	Unprocessed text

Every streamer has content on his/her 'about page'. This content contains, among other things, information about the sponsorship deals, subscription model, social media channels, donations, etc. This content is classified differently for each streamer and therefore this is seen as one variable 'content block', which includes every 'img' element from the 'about page' of the respective streamer.

The textual content block varies between streamers. Some streamers have written about themselves, such as their age and where they live. Others can write something in the textual block about the game in which they excel. Because this varies between streamers, this data is scraped under a single variable called 'textual content', which includes every textual element on the 'about page' of the respective streamer.

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

Since the goal of this project is not to apply machine learning to predict any outcome, there is no label or target assigned to each instance in it. Therefore, it is not relevant to this project.

2.6 *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.*

For some instances, live streamers, there was no information about the 'category' and the 'team name'. When live streamers stopped streaming while their information was being scraped, the category in which they were streaming could not be captured, as the tags change when this happens. Furthermore, if a streamer did not have a team name, this information was missing because it was unavailable. When a live streamer's 'category' or 'team name' information is missing, "NA" is filled in. As a result, there will be no blank observations for these variables. In the case of a category missing, the researcher can expect the category to be the same as the category of the streamers surrounding the streamer with 'category' as "NA".

2.7 *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 could only have a relationship based on the category and team name of the streamers. Streamers can thus be members of the same team or stream in the same category. It is possible to analyze, for example, which category contains the most followers or which team has the most sponsorships or similar revenue models.

2.8 *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 suggested data splits. There are, on average, 130 live streamers scraped in each category. When the dataset is divided into smaller subgroups based on categories, the dataset for each subgroup could become very small. For example, during one scraping run the lowest amount of streamers scraped in a category was 55. If the categories are scrolled through, increasing the amount of categories added to the list which will be iterated over, categories with a very small number of streamers will occur. This can be solved by reducing the viewer threshold per streamer, which is now set as 50 viewers, resulting in more streamers being scraped.

2.9 *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 ex- isted 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 is self-contained and therefore does not link to any external resources. The data within the dataset is exclusively collected from the Twitch website. The Twitch website is expected to remain the same, but there is no guarantee. Archive.org has official archived versions of the Twitch website, which allows researchers to inspect the old environment of Twitch.

2.10 *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.*

The dataset contains no information that could be considered confidential. The data will be retrieved by the publicly accessible website twitch.tv. The about page of a live streamer contains information about that live streamer, but this information is not private because anyone can search for it. Also, live streamers have a Twitch username, which is saved in the list and not their official names.There is some gray area in the 'Textual content block' because streamers sometimes post their top donators there. This could mean the username of a viewer, which could be their real name, is presented in the dataset.

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

There are no offensive, disrespectful, threatening, or anxiety-inducing items in the dataset. Twitch has strict community guidelines, which means both content blocks are not allowed to contain offensive data.

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

Yes, the dataset is related to people through their account on Twitch which gives personal information.

2.13 *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.*

There are no subpopulations in the dataset. But, segmentation has been used: the about pages of streamers of multiple different categories are scraped. Next to this, 'team names' are scraped, which indicate whether a streamer is part of a team. This can be seen as a subpopulation.

2.14 *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.*

It is possible to identify individuals from the dataset, because there is information scraped about the amount of followers, the social media accounts of the streamer and their username. But individuals could only be identified when the streamers are using their real names on Twitch and their social media accounts.

2.15 *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.*

There is no data in the dataset that could be considered sensitive.

3. Collection Process

3.1 *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (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.

The data that is collected is data that is directly observable on the Twitch website. This data is put on the 'about page' by the Twitch streamer him/herself.

3.2 *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

There is an API available for Twitch streamers. However, this API is not used because it contains no information about the 'about page' of the streamers. As a result, the API does not contain all of the data required for the research project. The API is generally used by Twitch streamers to add plug-ins to their streams, such as donation voice-over, chat moderators, etc. Therefore, it was decided to use web scraping. Web scraping allows for the extraction of all publicly available content and data from a website, which is required for research purposes.

To collect the data BeautifulSoup and Selenium are both imported. With Selenium, a webdriver can be opened and Selenium also allows for scrolling on a page. With the help of BeautifulSoup, which helps locating HTML elements, a scroll function was created which stops scrolling when the page arrives at the viewer threshold. When the webdriver arrived at the viewer threshold, with the help of BeautifulSoup a list was created of all of the streamer names. In addition, BeautifulSoup scrapes all the necessary information of the 'about' page of the streamer. In conclusion, Selenium was used to open the webdriver and to perform actions in the webdriver and BeautifulSoup was used to reformat the HTML code into scrapable text.

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

The dataset that has been generated by the scraper is a sample from the entire available data on Twitch.tv. The criteria on which the sampling strategy has been based are linked to the objective of the dataset; 'to generate a dataset that allows researchers to analyze the revenue models and sponsorships of Twitch streamers'. For this reason, categories are scraped based on popularity and streamers within these categories are scraped based on a viewer threshold of at least 50 viewers, which puts the streamer into the top 1% of Twitch streamers.

3.4 *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)?*

Only five students from the MSc Marketing Analytics were involved in the data collection process. Because no crowdfunding or employment were required for this project, monetary compensation is not a topic that is relevant here. The five students were supported by Hannes Datta, the lecturer of the course.

3.5 *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 time- frame in which the data associated with the instances was created.*

The code was created between September 5, 2022, and October 12, 2022. The code was written and improved throughout this time in order to gather the data. Because live streamers are being scraped, the code collects data in real time. On the 10th of October from 19:06 to 22:33, the specific dataset of 3833 streamers from 30 different categories was collected. Within the

dataset, for each instance, a separate timestamp has been generated.

3.6 *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.*

There were no ethical review processes conducted, so this question is not applicable.

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

The dataset is related to people, as Twitch streamers are human beings.

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

The collected data is obtained indirectly from the website Twitch.tv, which can be considered a third party. However, the data on the 'about page' of the live streamers, is provided by the streamers themselves.

3.9 *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 other- wise reproduce, the exact language of the notification itself.*

Individuals were not informed that their information was being collected.

3.10 *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.*

Individuals were not asked for consent to scrape their data. However, they agreed that the information is displayed on their 'about page' on Twitch.tv, which is publically available.

3.11 *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).*

Individuals did not provide their consent, as all of the information is publicly available.

3.12 *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.*

The potential impact of the dataset was not examined.

4. Preprocessing, cleaning, labeling

4.1 *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 remain- der of the questions in this section.*

To obtain the final dataset, we used a variety of pre-processing techniques. A number of steps were required before extracting information from the "about pages" of live Twitch streamers. First, after scraping all categories, all category names were adjusted to fit into an URL. That means spaces had to be substituted for

'%20' and colon had to be substituted for '%3A'. To access the 'about page' of live streamers in that category on Twitch, a base URL with each live streamer's username is required. The live streamers' URLs are as follows: url = "https://www.twitch.tv/" + "/about" + streamer name. This list is then used to loop through the list and scrape each streamer's "about page."

The asian alphabet is not included when printing the list of usernames for live streamers because this is not required and is not used in the links of streamers. The variables are labeled as a final stage, and care was taken to ensure that the labels were as clear and descriptive as possible so that it was instantly clear what the variable was about (for example, "name of streamer" or "category"). In addition, the "team name" of the live streamer is being collected. This "team name" was not present with every streamer. When this variable was missing, " NA" was used in place of the "team name". This also applies to the information about the category. When the category could not be captured because live streamers stopped streaming while being scraped (as mentioned in paragraph 2.6), "NA" was filled in.

4.2 *Was the "raw" data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

Other than the completed dataset generated in this project, no raw data was saved.

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

The Python program was used to write and process the code in a Jupyter Notebook. Selenium and BeautifulSoup are used to make website requests and scrape the required information. The following link will redirect the reader to the location where the source code is stored: https://github.com/bodr101/webscraper_twitch.tv.

5. Uses

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

Other than the project team that scraped the dataset's information, the dataset has not yet been used by other people or researchers.

5.2 *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.*

There is currently no repository for papers and systems that use the dataset scraped in this project.

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

The dataset obtained by scraping the 'about page' of live Twitch streamers can be used to determine, among other things, the revenue model of the streamers, sponshipdeals per category and advertising opportunities for businesses. Twitch, as mentioned in paragraph 1.1, has a partner program for anyone interested in generating revenue by broadcasting on the website. Streamers on Twitch can have a large number of followers, making advertising on Twitch an excellent way for businesses to connect with their target audience. Companies can benefit greatly from knowing, for example, what types of advertisements are most effective on Twitch and what time to advertise is most effective. These are some examples of what can be done with the data, but many more study directions can be based on this data.

5.4 *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 future user 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 undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The Twitch website is a dynamic website. This means that there are, for example, new streamers on Twitch on a regular basis, new companies advertising, or that something on the 'about page' can be changed. This project's dataset is derived from live streamers at a specific point in time. The code used to generate the dataset will continue to function in the future, but it will generate a different dataset. There is no way to solve this issue. The dataset may not be the same, but this does not necessarily affect the data's quality.

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

This information is gathered in order to analyze it, gain insights, and ultimately draw conclusions about the analysis's findings. The data should only be used for academic studies and other forms of research. This dataset may not be used for plagiarism or other similar activities. Furthermore, the dataset should not be used to contact live streamers for commercial or other purposes.

Total scrapes	Timing per scrape	Hours of Scraping
3,833	3 seconds	3,20 hours

Bronnen:

https://www.sciencedirect.com/science/article/pii/S0747563218300712?casa_token=5qszTfyhy30AAAAA:oOTuYmFwjCi5Cq9tMgSacUPZLqMX1xEyC2MCkU0wMBna3XZt7oDKA_6wvpIIlZ1-s5wBdh6RMw

https://earthweb.com/twitch-statistics/#:~:text=Twitch%20has%20140%20million%20monthly%20active%20users%20in%202022.&text=More%20than%202.2%20million%20users%20make%20the%20most%20of%20the,of%20Twitch%20users%20are%20male

https://www.dexerto.com/entertainment/new-twitch-stats-reveal-how-few-viewers-are-needed-to-be-a-top-streamer-1527638/