Datasheets for Datasets

Adapted from: Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.*

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.*

Previous research suggests that YouTube tutorials have a positive impact on the skill development of youth. However, the creative segment of YouTube such as tutorials for creating music was not covered. The purpose of this study is to fill this gap so that researchers can perform studies on video attributes within the creative segment of YouTube. Tutorials on FL Studio were chosen since many famous artists create their music with this software.

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

This dataset was created by Quinten de Putter, Jeroen Maagdenberg, Sam van de Ven and Tayfun Ozcan. They formed group 15 during the Online Data Collection Management course at Tilburg University.

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.*

The YouTube API used for this dataset is free of charge. Therefore, there is no associated grant needed.

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.*

The dataset consists of unique YouTube videos about the topic FL studio tutorials. These datasets contain the search results for the topic and provide information and statistics about each video such as date and time of publishment, channel name, view count, like-count, dislike-count and comment count.

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

For this research, the only instances that are collected are YouTube videos and YouTube channels.

2.3 *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 (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 is limited and does not contain all possible instances since the API is bound to a certain quota limit. The larger set would be all possible instances that are on YouTube. The dataset is representative since the API is from YouTube itself and the search results are not discriminated on language or geographical location of upload.

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 description.*

1* https://arxiv.org/abs/1803.09010

The instances that are gathered for the research consist of raw data. This means that the dataset contains unprocessed text and images. As mentioned in 2.1, information and statistics about each instance, or here video, are retrieved. Examples of the information and statistics are view count, like-count, channel name, channel id, thumbnail images, descriptions and similar results that one can find on a YouTube page.

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

Each instance is uploaded by a YouTube channel. The target of the instance is to gain interaction in the form of views, likes etc.

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.*

There is no missing information from individual instances from YouTube videos and YouTube channels.

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.*

The dataset used for this research consists of YouTube videos and YouTube channels. First, we scraped the most relevant videos and afterwards, we scrape the corresponding channels.

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 recommend data splits used in this research.

2.9 *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., website*

s, 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 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.

This dataset is linked to the external resource YouTube and the data extracted from the YouTube API are snapshots. The data will exist, but will not stay constant over time since the algorithm could change. There are no official archival versions of the complete dataset. Each API key allows the researcher to use 10.000 quotas each day to gather data for free. There are no license fees for this API key. However, this quota limit can be increased by buying certain packages/subscriptions on the Google Cloud Platform.

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 raw data collected for this research is publicly available. Therefore, the dataset is not considered confidential. People that upload a video can choose their own channel name and thus a large number of channel names in our dataset are nicknames or company names.

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.*

The dataset contains only raw data about YouTube videos and provides information and statistics about these videos. These statistics are not offensive, insulting, threatening and will not cause anxiety since

each video has to comply to the YouTube guidelines and policies.

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

The dataset does not relate to people. Therefore, the questions from 2.12 till 2.15 will not be answered.

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.*

N/A

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.*

N/A

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.*

N/A

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

The collected data consists of raw descriptive data and statistics for each instance. A large part of the data was not directly observable, since it was raw text. However, some information such as titles, descriptions, view count, like- and dislike-count are directly observable. The data was not reported by subjects and not derived from other data sources.

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?*

As mentioned before, for this research, a YouTube API was used to gather data. This API was developed by Google itself and, for access, it is required to use an API key that is requested through the Google Cloud Platform.

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)?*

YouTube allows researchers a maximum of 10.000 quotas to gather raw video data. Since there is an endless number of videos about FL studio tutorials, it obligates the researchers to work with a sample that fits the quota limit. Also, since the purpose of this research is to find out which FL studio tutorial videos are the most popular, the data consist of the most relevant videos.

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)?*

The people who were involved are four students from Tilburg University (Quinten de Putter, Jeroen Maagdenberg, Sam van de Ven and Tayfun Ozcan). They were not financially compensated for the collection process. The process was initiated and regularly reviewed by professor, dr. Hannes Datta from Tilburg University.

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.*

As mentioned in section 2.9, the data is a snapshot of the most relevant results for the search query “FL tutorial” at a given moment. Therefore, there is not a

specific timeframe wherein the data is collected.

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.

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

N/A

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)?*

N/A

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.*

N/A

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.*

N/A

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).*

N/A

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.*

N/A

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.*

Within this project, there is no preprocessing, cleaning or labeling done. Therefore, the remainder of the section will be skipped.

On a side note, YouTube adjusts the date and time of a video upload to the local time of a person that accesses a video on the website or the API.

4.2 *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.*

N/A

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

N/A

5. Uses

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

No, for this research only data has been collected. Therefore, no tasks were performed.

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 no repository for this research.

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

The dataset allows other researchers to combine different instances and statistics to run multiple linear regressions with different independent- and dependent variables. One could use the videos in this dataset to perform a research on the comment section of the videos. However, the data of the comment section was not gathered during this collection process.

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?*

A possible limitation for this research is that the algorithm of Google and/or YouTube changes constantly and it may provide different results per user.

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

The data might contain real and fake first and last names. Therefore, it is not advised to use the names gathered as a reference for contacting certain people.