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.*

The COVID-19 pandemic has changed the world tremendously. Many places in the world are currently in lockdown and therefore stores, museums, sports clubs and restaurants have been, or are, closed. The Netherlands has experienced several 'corona waves' in which the extent of the measures was different. For a while, in the first wave, professional group sports as well as amateur group sports were cancelled completely. Currently, the Netherlands allows group sports without competition for all ages under 27 years old. Next to this, professional players are allowed to play again without public in the stadiums.

One of the sports that is played the most and attracts the biggest crowds is soccer or football (In Dutch: voetbal. It translates to football, but the game is very different from American football. The game is played like British soccer) (Boen, Vanbeselaers & Feys, 2002). In general, Dutch soccer fans are known as very dedicated and enthusiastic (Lechner, 2007). Many soccer fans hold a season ticket and are present at almost every game of their favorite soccer club. The soccer stadiums in the Netherlands can hold up to 56.000 people and are, when not in lockdown, often fully booked.

As mentioned above, the current COVID-19 pandemic has forced fans to stay home and watch soccer on the television. This has created a situation that has almost never occurred in the past and it is therefore very interesting to see how this has impacted soccer fans and their interaction. Normally, the most dedicated fans can express themselves within the stadium but when watching from home, it can well be that they converted some of their interaction to social media and thus the amount of online interaction has gone up. Next to this, we are interested in whether the sentiment in the online interactions changed due to the changed situation. It is intuitive that the sentiment of fans at home is different than when standing in an arena with dozens of enthusiastic and hyped up people. As this research aims to be explorative, no clear hypotheses were generated beforehand.

As it are hard financial times for football clubs because there are no fans allowed in their stadiums (Grix, Brannagan, Grimes & Neville, 2020), it is vital for soccer clubs to keep their fans engaged with their team. As COVID-19 has shifted face to face interactions to digital interactions, social media is now also the main source of interaction of football clubs with their fans.

The social media platform that is scrapped is Twitter because there is a lot of soccer fan interaction on this platform. Other social media platforms such as Facebook or Instagram are more focused on sharing content whereas Twitter is for interacting with one another and giving opinions (Kassing, J. W., & Sanderson, 2010). Twitter also provides several tools for additional interaction because the tweets can be 'retweeted', and Tweeters can directly reply on Tweets of other Tweeters. 'Retweeting' entails sharing the Tweet someone else created with your own network.

In general, this dataset was thus created to explore the effects of the COVID-19 measures on the social media interaction of soccer fans. The research question is: To

---

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

Datasheets for Datasets

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

what extent did the Twitter conversation of Dutch Football fans change due to COVID-19 (no fans in the stadiums)? The data set contains multiple variables. Next to this, the scraper that was built for this project can easily be adapted to scrape a broader time period than the one that was chosen now. As historical Twitter data cannot be collected from the Application Protocol Interface (API), the scraper that is used is fairly unique in being able to collect historical tweets for free. Because of this, similar data sets are not publicly available.

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 has been created by Stan Wiggers, Kevin Stekelenburg, Ruben Custers, Eric Volten and Anne van Veenendaal. The dataset has been created on behalf of Tilburg University for the courses Online Data Collections (oDCM) and Data Preparation and Workflow Management (DPrep).

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 creation of this dataset was not funded nor were there any costs incurred.

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.*

That data was scraped from twitter.com with hashtags of the 18 Eredivisie clubs. The following hashtags were used: #AdoDenHaag #AFCAjax #AZalkmaar #FCEmmen #FCGroningen #FCTwente #FCUtrecht #Feyenoord #FortunaSittard #Heracles #PECZwolle #PSV #RKCWaalwijk #SCHeerenveen #SpartaRotterdam #Vitesse #VVVVenlo #WillemII. Some hashtags (e.g. #ajax) return foreign tweets that are containing this hashtag but are unrelated to the football club. Therefore, hashtags which resulted in the most related tweets (by observing a sample of tweets) were sought and proved to be the full names of the soccer clubs.

As the research objective is to make a comparison between before COVID-19 and during COVID-19, different weekends were selected to scrape. The weekends selected are:

Season 2019/2020 (Before COVID-19)

Period 1: Round 14 - 22 23 24 November 2019

Period 2: Round 20 - 24 25 26 January 2020

Season 2020/2021 (During COVID-19)

Period 1: Round 10 - 27 28 29 November 2020

Period 2: Round 18 - 22 23 24 January 2021

These specific weekends were selected because they have the same amount of games with a similar degree of hypothesized 'buzz'. With buzz we refer to the degree of rivalry between the soccer clubs and degree of exciting results.

The entities or instances that are scraped are tweets and the different variables that the tweets contain. The tweets can be from: private individuals, organizations and the soccer teams. For the analysis the variables *date, content, unique id* and *usernames* are needed. Next to this, the variables *URL, reply count, retweet count, like count, location, user followers count, user friends count,* and *tweet*

source were scraped. The data is available through a CSV file.

The variable content contains all the content in a specific tweet. The variable *unique id* contains the unique id every object in Twitter gets assigned. The variable *username* contains the username of the person that posted the tweet. The variable *location* displays the town or place the user has as location on its profile. This does not have to be the same as the actual location of the user; the user is free to adapt this. The variable *source* reveals where tweets are posted from (e.g. Android). For the variables the *reply count, retweet count, like count, user followers count, user friends count* the count of each object is included.

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

From the weekends that were scraped, 5849 instances were selected. 2871 of these instances are from during COVID-19 and 2978 from before COVID-19. All of these instances contain at least one of the hashtags specified in section 2.1. Some instances can be linked to each other through the variable retweet.

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 contains all possible instances that match the requirements specified in section 2.1 (e.g. they contain at least one of the hashtags and were tweeted in a selected weekend). Considering the requirements specified in 2.1, all tweets were scraped for representative reasons. Foreign tweets were also kept in the data set because the soccer teams could possibly have foreign fans as well.

However, one must note that not the whole season was scraped; specific weekends were selected to make the amount of data more manageable. As mentioned before, careful consideration was put into the selection of the weekends, but no validation could be done because of the newness of the topic. The larger set would contain all soccer matches of season 19/20 and all soccer matches of season 20/21.

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.*

The dataset consists of raw data. Tweets from Twitter can include the following objects: text, numbers, emoticons, special characters, images, GIF-images and Tweeters can create polls. The tweets were collected by using a scraper called snscrape[2] so historical tweets could be scraped. The scraping was not monitored or formatted in any way; all tweets that fit the criteria (of timeframe and hashtag) were kept in the data set. Part of the text can also be the signs used to decode smileys used in the tweets. These signs are the Unicode of the smileys which can make the overall text look strange in some occasions.

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

The overarching goal of this study is to evaluate whether the COVID-19 influenced the online interaction about Dutch soccer teams. This analysis is important because soccer is played without fans in the stadiums due to the COVID-19 measures and this could have many implications. For soccer clubs it is essential that their fans remain engaged. The analysis will be done by comparing whether the online interactions have gone up and by analyzing the sentiment within the tweets. Thus, the target associated with each

[2] The scraper was retrieved from:
https://github.com/JustAnotherArchivist/snscrape

instance is evaluating the sentiment of the tweet and the total amount of tweets per period.

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.*

In the variable location some missing information is observed. This is caused by the privacy settings of Twitter. The code includes measures to control for this missing data in the sense that even when values are missing, in any variable, they will still be present in our dataset. As the missing information is primarily located in a variable that is not essential for our research purposes, it was decided to keep all instances in the data set.

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 tweets can be clustered on their hashtags, there are for example some tweets that are about ajax (#AFCajax) whilst others are about PSV (#PSV). This means that tweets can be about different teams and also about different games which can cause differences. During the data collection process all tweets that contained one of the hashtags and were tweeted in the selected time frame were retained.

There are also relationships between instances because tweets can be retweeted. From researcher's perspective this type of interaction between Tweeters can be seen as 'relationship'. Capturing retweets was not the main goal of this data collection and dataset but there is a variable that counts the amount of retweets and retweets are also retained in the dataset Retweets could give possibly give valuable information but these research objectives go beyond the scope of this study.

The replies on a tweet in the dataset that are not containing the hashtags specified in section 2.1 are not retained in this dataset. Nevertheless, the variable reply count does give an indication on the amount of interaction a tweet generates. For future research, it would be advised to include replies without the hashtag in the data set to make the analysis of interaction more well-rounded.

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 recommended data splits. The research objective is to get an overview that is the most inclusive and since only a limited amount of the total season time is scraped already, it is not advised to reduce the data set further.

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 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.*

The dataset relies on an external source, namely Twitter. The data is captured directly through the website through a scraper called snscrape. The reason that the API of Twitter was not used to scrape the tweets is that it did not allow us to scrape far enough in the past. The data will exist and will be constant over time if the Twitter user, or Twitter itself does not delete the tweet. Tweets cannot be posted in the past so it not likely that anything will be added to the tweets in the data set.

One should note that if Twitter decides to change its web structure, the scraper used for this data collection might not work anymore. In this case, the data is stored on a private location and can thus not be accessed without permission because of privacy reasons. In this case, it is advised that one looks out for another scraper or one creates a similar scraper.

Twitter uses MySQL and Manhattan NoSQL as their primary databases for storing user data and input. The official archival dataset of Twitter is not public and

when scraping one should adhere to strict rules. By not taking the Twitter API, this risk of being blocked was taken as using the API would ensure adhering to Twitter regulations. In Twitter regulations it is stated that ".. scraping the Services without the prior consent of Twitter is expressly prohibited".

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.*

With Twitter, users have the option to make their profile either private or public. The dataset presented only entails tweets that were posted from a public account. So, the dataset does not contain confidential data as the tweets were all posted from a public account. However, one should note that data entries were not asked for their informed consent to let their tweets be used for research purposes. As some users have usernames similar to their own name, this might be raising some anonymity issues. Next to this, the location of the users was also scraped which some might consider confidential as well. However, as it is not the actual location of the user but the location they put on their profile, confidentiality issues can be disregarded. Next to this, it would be against the privacy rules of Twitter to share the private information (such as username and location) online without permission.

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.*

Even though, most tweets are about cheering on the soccer teams and discussing the games, some tweets can be seen as offensive or insulting. Soccer fans are often very dedicated to their favorite teams and this can result in a groundless hatred against the other teams (Johnes, 2008). This can lead to offensive tweets about a team or a specific player. Players that made mistakes are also called out and are called things like 'miskoop' (dutch for bad buy). However, one should note that Twitter also has a policy to stop tweets that are offensive or insulting so a good share of tweets of this nature are already selected out by Twitter.

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

The dataset does relate to people because the data contains tweets which are posted by people or organizations. Thus, the dataset focuses on the interaction of people on the platform Twitter.

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 based on socio-demographics identified because the socio-demographics are not included in the dataset. The reason that this is not included in the dataset is that this is private confidential information and Twitter does not share this publicly. One could create subpopulations in the dataset based on the hashtags that were used. This would entail grouping tweets based on what hashtag or hashtags were used and thus based on what team(s) their tweet was about. The data can also be clustered based on the period they were posted in.

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.*

There are no subpopulations based on socio-demographics identified because the socio-demographics are not included in the dataset. The reason that this is not included in the dataset is that this is private confidential information and Twitter does not share this publicly. One could create subpopulations in the dataset based on the hashtags that were used. This would entail grouping tweets based on what hashtag or hashtags were used and thus based on what team(s) their tweet was about. The data can also be clustered based on the period they were posted in.

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.

As Twitter reviews the tweets and the tweets are posted on public accounts, the overall nature of the tweets is not sensitive. Soccer is not a topic where sensitive information on topics such as politics or religious beliefs is commonly discussed so one can assume that most data in the dataset is not sensitive.

However, locations, if they were disclosed by the user, are included in the dataset which can be classified as sensitive. As mentioned before, for the analysis the locations will not be used so the sensitivity issues that rise with scraping the locations can be disregarded.

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 instances that compromise the data set are tweets. All data associated with each instance is acquired by scraping Twitter using the package snscrape in Jupyter Notebook. After creating a data frame from the tweets list, the raw data text is directly observable in a CSV file.

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 previously mentioned, we use a tweet-scraping library from Martin Beck, called snscrape, which can also be used for other social networking sites. This scraper allows us to scrape historical tweets with a text search.

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 is not a sample from a larger set. Each tweet that is posted within one of the time frames containing at least one of the 18 hashtags is included in the final data set. In total, the data set consists of 5849 instances.

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)?*

There were five Tilburg University students involved in the data collection process: Stan Wiggers, Kevin Stekelenburg, Ruben Custers, Eric Volten and Anne van Veenendaal. Data is collected on behalf of Tilburg University for the courses Online Data Collection Management (oDCM) and Data Preparation and Workflow Management (DPrep). The students are not compensated with a financial reward for their work. However, by participating in both courses they gained vital knowledge and experience.

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 data collection process found place in March 2021. This timeframe does not match the creation timeframe of the data associated with the instances, since the data was created in two weekends (periods) in the Eredivisie season 2019/2020 and two weekends (periods) in the Eredivisie season 2020/2021.

Season 2019/2020 (Before COVID-19)

Period 1: Round 14 - 22 23 24 November 2019

Period 2: Round 20 - 24 25 26 January 2020

Season 2020/2021 (During COVID-19)

Period 1: Round 10 - 27 28 29 November 2020

Period 2: Round 18 - 22 23 24 January 2021

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.

Scraping tweets with the Twitter API from up to seven days ago is not enough to fulfill the research goals. Consequently, we explored several third-party scrapers and eventually found the scraper from Martin Beck. This scraper allows to scrape further in the past without using an API. However, this is not according to the rules and terms described on Twitter's website, which states the following (Twitter, n.d.):

(iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by Twitter (and only pursuant to the applicable terms and conditions), unless you have been specifically allowed to do so in a separate agreement with Twitter (NOTE: crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, **scraping the Services without the prior consent of Twitter is expressly prohibited)**

Besides, even though tweets are publicly available, people on Twitter have not given their informed consent on doing analysis with their tweets. Therefore, scraping tweets is ethically questionable. However, as Twitter is a data source that is often used for researches, we decided that we could scrape the tweets considering some ethical concerns and measures. Our reasoning on this is twofold: first, we will not connect the analysis to individual users and therefore their identity remains private and anonymous. Next to this, we argue that tweets about football contain relatively little sensitive or personal information as compared with many other topics such as politics for example.

3.7 *Does the dataset relate to people?*

The dataset does relate to people because the data contains tweets which are posted by people or organizations. Thus, the dataset focuses on the interaction of people on the platform Twitter.

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 data is not directly collected from the individuals in question. The data is obtained via social networking site Twitter and all tweets are publicly available.

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.*

The individuals in question were not notified about the data collection.

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.*

As mentioned in 3.6, the individuals or organizations in question have not given their informed consent on doing analysis with their tweets. However, all tweets derive from public Twitter accounts. This means that anyone can view and interact with the collected tweets. If someone wanted to hide their tweets from the general public, he or she could have chosen for a private Twitter account.

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).*

Consent was not obtained.

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.*

An analysis of the potential impact of the dataset and its use on data subjects has not been conducted.

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.*

No preprocessing or cleaning has taken place. Labeling is done in Jupyter Notebook by adding more meaningful and clear column names to the data frame. In total there are 14 columns, named as follows: 'Season', 'Period', 'URL', 'Datetime', 'Tweet ID', 'Text', 'Username', 'Replies', 'Retweets', 'Likes', 'Location', 'Followers', 'Friends', and 'Source'.

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.*

Because we did not preprocess and clean the data, the ''raw'' data is not saved in addition to the labeled data. It is expected that any unanticipated future study that wants to make use of the data, is likely to reuse the above-mentioned column names. The data set is provided in four CSV files, one for each period. These CSV files are merged into one final CSV file and subsequently exported to our local directory.

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

As mentioned in 4.1, Jupyter Notebook is used to label the data. Jupyter Notebook is an open-source web application, which can be installed using the following link: https://jupyter.org/install

5. Uses

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

The dataset was used for the courses Online Data Collections (oDCM) and Data Preparation and Workflow Management (DPrep) Tilburg University. In the DPrep course, the data was explored, and a sentiment analysis has been performed. This sentiment analysis counted the words that contained the following emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Next to the emotions, tweets were evaluated on negativity and positivity. The result showed that there are more negative words in the tweets during corona than before corona. Surprisingly, in the period before corona more tweets were posted than during corona.

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.*

The repository that links to this dataset and its analysis can be found on GitHub. The repository can be found by using this link: https://github.com/kevinStekelenburg/conversation-change-covid19

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

This data can be used for many types of analyses that compare the pre-corona tweets to during corona tweets. Next to this, one could look further into the content and examine what is commonly tweeted and which type of words or smileys are often used.

It is also advised that the other variables included in the data set such as retweets, followers count, and friends count are explored further. There are tremendous ways one could analyze this. One could for example look into whether location corresponds to the soccer club people tweet about.

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?*

One should note that the first variable corresponds to the season the tweet was posted. The second variable corresponds to the period the tweet was scraped. Period 1 corresponds to Round 14 and date 22 23 24 November 2019. Period 2 corresponds to Round 20 on 24 25 26 January 2020. The second Period 1 (of season 20/21) corresponds to Round 10 on 27 28 29 November 2020. The second period 2 corresponds to Round 18 on 22 23 24 January 2021.

Next to this, keep in mind that scraper includes data entries regardless of missing values.

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

The data set was not created to focus on individual persons, and it is therefore strongly discouraged to link the data to discover the identity of the specific users. As the users did not post the tweets to participate in a study, the data should be treated as anonymous and unlinked to specific people. Furthermore, one should not contact the users for further questions.

6. Distribution

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

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

6.3 *When will the dataset be distributed?*

6.4 *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.*

6.5 *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.*

6.6 *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.*

7. Maintenance

7.1 *Who will be supporting/hosting/maintaining the dataset?*

7.2 *How can the owner/curator/manager of the dataset be contacted (e.g., email address)?*

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

7.4 *Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

7.5 *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were 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.*

7.6 *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 users.*

7.7 *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 other users? If so, please provide a description.*