MOTIVATION FOR DATASHEET CREATION

[question1] Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)
[answer1]
[question2] Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?
[answer2]
[question3] What (other) tasks could the dataset be used for?
[answer3]
[question4] Who funded the creation dataset?
[answer4]
[question5] Any other comment?
[answer5]

DATASHEET COMPOSITION

[question6] What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)
[answer6]
[question7] How many instances are there in total (of each type, if appropriate)?
[answer7]
[question8] What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the in-stances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?
[answer8]
[question9] Is there a label or target associated with each instance? If so, please provide a description.
[answer9]
[question10] 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.
[answer10]
[question11] 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.
[answer11]
[question12] 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).
[answer12]
[question13] 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.
[answer13]

[question14] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
[answer14]
[question15] 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.
[answer15]
[question16] Any other comments?
[answer16]

COLLECTION PROCESS

[question17] What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
[answer17]
[question18] How was the data associated with each instance ac-quired? 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.
[answer18]
[question19] If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
[answer19]
[question20] 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)?
[answer20]
[question21] 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 timeframe in which the data associated with the instances was created.
[answer21]

DATA PREPROCESSING

[question22] 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, pro-cessing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
[answer22]
[question23] 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.
[answer23]

[question24] Is the software used to preprocess/clean/label the in-stances available? If so, please provide a link or other access point.
[answer24]
[question25] Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?
[answer25]
[question26] Any other comments
[answer26]

DATASET DISTRIBUTION

[question27] How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)
[answer27]
[question28] When will the dataset be released/first distributed? What license (if any) is it distributed under?
[answer28]
[question29] Are there any copyrights on the data?
[answer29]
[question30] Are there any fees or access/export restrictions?
[answer30]
[question31] Any other comments?
[answer31]

DATASET MAINTENANCE

[question32] Who is supporting/hosting/maintaining the dataset?
[answer32]
[question33] Will the dataset be updated? If so, how often and by whom?
[answer33]
[question34] How will updates be communicated? (e.g., mailing list, GitHub)
[answer34]
[question35] If the dataset becomes obsolete how will this be communicated?
[answer35]
[question36] Is there a repository to link to any/all papers/systems that use this dataset?
[answer36]
[question37] If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?
[answer37]

LEGAL AND ETHICAL CONSIDERATIONS

[question38] 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.
[answer38]

[question39] Does the dataset contain data that might be considered confidential (e.g., 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.
[answer39]
[question40] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why
[answer40]
[question41] Does the dataset relate to people? If not, you may skip the remaining questions in this section.
[answer41]
[question42] 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.
[answer42]
[question43] 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.
[answer43]
[question44] 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.
[answer44]
[question45] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
[answer45]
[question46] 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.
[answer46]
[question47] 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.
[answer47]
[question48] 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).
[answer48]
[question49] 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.
[answer49]
[question50] Any other comments?
[answer50]