MOTIVATION.

[question1] 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.
[answer1] We collected the datasets in this benchmark to evaluate supervised machine learning (classification/regression) algorithms designed to jointly operate on text and tabular features. The original versions of these data were also initially created primarily for a similar purpose.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] The authors of this paper, all scientists employed by Amazon, curated this benchmark. Curating the benchmark did not cost significant money, and the benchmark data are currently hosted on cloud servers (S3) provided by Amazon.

[question3] 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.
[answer3] The original data sources were created/curated/funded by various companies/individuals, please refer to each individual source for more details.

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] 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.
[answer5]

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6]

[question7] 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).
[answer7] Each dataset is a sample of instances from a larger set. We caution these samples may not be at all representative of the larger set, and thus the benchmark should not be used to draw domain-specific conclusions/insights through scientific data analysis of individual datasets.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.
[answer8] The 18 datasets in our benchmark represent all

of the public text/tabular datasets we could find that do not violate our exclusion criteria and satisfy our main desiderata: the dataset must entail a meaningful prediction problem with real enterprise data (as opposed to contrived toy task without real-world application). Note that we only consider tabular datasets that contain text fields, which is a small fraction of publicly available tabular datasets (even though such data are ubiquitous in private enterprises). Our dataset search was conducted over the following sources: Kaggle, MachineHack, UCI ML Repository; the first two are the best sources of publicly available enterprise datasets (with meaningful prediction problems) that we are aware of. Within each source, we searched for datasets matching the keyword "text" in their metadata/ descriptions for consideration in our benchmark (although the majority such datasets either had no tabular features or failed to provide the original raw text presenting only a featurized version such as bag-of-words). We also conducted some dataset searches via Google, but did not find serious candidates for the benchmark via this avenue. Beyond the primary requirement that data must stem from a real enterprise application with a meaningful classification/regression task, our other exclusion criteria ensured each dataset in the benchmark has: IID examples, non-prohibitive licensing, some text fields beyond just 1-2 words and in the English language (for simplicity), sample size of at least 1000, and predictive signal across both text and tabular (numeric+categorical) modalities (meaning one modality does not appear entirely useless for the prediction problem, evaluated via preliminary AG-Stack+Ngram runs without each modality).

[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 (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] Yes there are many missing fields in certain datasets. It is unclear why they are missing or if the missingness mechanism satisfies the missing at random assumption.

[question11] 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.

[answer11] For evaluating ML performance, we simply assume the data are IID. However this may be violated by certain datasets. For example, product datasets may contain near duplicate products and products may be related (reviewed by the same users, price of a product can affect price of others, etc.). We do not explicitly know the relationships between instances in these data.

[question12] 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.

[answer12] Yes the benchmark provides a recommended training/test split, but ML systems are free to split validation data from the training set as they see fit. The split was done randomly (stratified based on labels for classification) to best reflect an IID setting for which supervised learning methods are primarily intended.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13]

[question14] 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.

[answer14]

[question15] 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.

[answer15] Not to our knowledge, but it is possible that a person entered confidential information into the text fields (although they knew these would be publicized).

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] The data are mostly non-offensive data used for business purposes. Exceptions are the text fields in the jigsaw dataset, which contain toxic online comments, and the channel/pop datasets, which contain news article titles that may be anxiety-inducing. Furthermore, some of the user reviews of products may be offensive to certain people, although we did not spot any.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] 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.

[answer17] Yes some datasets contain information from people. These all stem from commercial sources where people upload their data intentionally to share it with the world (e.g. user reviews, Kickstarter fundraising, public questions, etc.). There is no sensitive/personal information in these data, beyond what a person intended to publicize.

[question18] 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.

[answer18] Yes it may be possible as some datasets contain text fields where an individual may have entered arbitrary information (although they knew the information would appear publicly).

[question19] 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.
[answer19] Not to our knowledge given all this data was already publicly available, but it is possible given the nature of free form text fields.
[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] 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.
[answer21] We processed each dataset from the original source using the publicly available scripts in the scripts/data_processing/ folder of our benchmark GitHub repository. To create versions for our benchmark, we omitted certain features (columns), badly formatted or duplicated rows and subsampled overly large datasets.
[question22] 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?
[answer22]
[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
[answer23]
[question24] 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)?
[answer24]
[question25] 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.
[answer25]
[question26] 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.

[answer26]
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipB] NO
[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
[answer27]
[question28] 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.
[answer28]
[question29] 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.
[answer29]
[question30] 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).
[answer30]
[question31] 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.
[answer31]
[question32] Any other comments?
[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] 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.
[answer33]
[question34] 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.

[answer34]

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35]

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] Yes many of the datasets have been used to evaluate ML systems, some through formal prediction competitions. Other datasets have been used to demonstrate data analysis techniques. For the datasets originally stemming from Kaggle, one can find some of the previously considered tasks in the discussion forum or notebooks associated with the original dataset.

[question38] 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.

[answer38]

[question39] What (other) tasks could the dataset be used for?

[answer39] We recommend these datasets only be used for evaluation of machine learning algorithms. One could select different target variables in each dataset to create new prediction tasks to evaluate, but these will likely be less practically meaningful (i.e. representative of a real application) than the target variable we have selected for each dataset. Also note that none of the datasets has extremely large sample-size (say over a million), so modeling conclusions drawn based on this benchmark may not translate to applications with massive datasets.

[question40] 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?

[answer40]

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41]

[question42] Any other comments?

[answer42] No

DISTRIBUTION.

[question43] 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.
[answer43] Yes the benchmark is made publicly available.
[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
[answer44]
[question45] When will the dataset be distributed?
[answer45]
[question46] 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.
[answer46]
[question47] 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.
[answer47] Yes please refer
to the licenses corresponding to each original data source (linked from our repository) for more details.
[question48] 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.
[answer48] Not to our knowledge.
[question49] Any other comments?
[answer49] No

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?
[answer50]
[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?
[answer51] You can open a GitHub issue at the
benchmark repository, or email the authors of this paper.
[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52]
[question53] 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)?
[answer53] Yes updates will be done via GitHub and publicly announced there.
[question54] 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.
[answer54]
[question55] 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.
[answer55]
[question56] 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.
[answer56] Yes anybody may open Pull Request with desired changes on GitHub.
[question57] Any other comments?
[answer57] No.