Datasheets for Datasets

Timnit Gebru*[1], Jamie Morgenstern[2], Briana Vecchione[3], Jennifer Wortman Vaughan[4], Hanna Wallach[4], Hal Daumé III[4,5], and Kate Crawford[4,6]

[1]Google
[2]Georgia Institute of Technology
[3]Cornell University
[4]Microsoft Research
[5]University of Maryland
[6]AI Now Institute

April 16, 2019

Abstract

The machine learning community currently has no standardized process for documenting datasets. To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using datasets, and the characteristics of these datasets will fundamentally influence a model's behavior. A model is unlikely to perform well in the wild if its deployment context doesn't match its training or evaluation datasets, or if these datasets reflect unwanted biases. Mismatches like this can have especially severe consequences when machine learning is used in high-stakes domains such as criminal justice [2, 20, 44], hiring [29], critical infrastructure [10, 35], or finance [28]. And even in other domains, mismatches may lead to loss of revenue or public relations setbacks.

Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in datasets. Much like a faulty capacitor in a circuit,

*Much of this research was conducted while Gebru, Morgenstern, and Vecchione were at Microsoft.

the effects of these biases can propagate throughout an entire machine learning system. For example, Buolamwini and Gebru [7] found that three commercial gender classifiers had near-perfect performance for lighter-skinned men while error rates for darker-skinned women were as high as 33%; Amazon canceled the development of an automated hiring system because the system amplified gender biases in the tech industry [12]; and Bolukbasi et al. [5] showed that low-dimensional embeddings of English words inferred from news articles reproduce gender biases by, for example, completing the analogy "man is to computer programmer as woman is to X" with the stereotypical "homemaker." Holstein et al. [24] found that many industry practitioners turn first to datasets when attempting to address issues like these, but that the sources of such issues can be difficult to identify.

The risk of unintentional dataset misuse increases when developers are not experts, either in machine learning or in the domain where machine learning will be used. This concern is is particularly salient due to the increased prevalence of tools that "democratize AI" by providing easy access to datasets and models for general use. As these tools are made available, it is important that the entities who provide them enable developers to understand their characteristics and limitations.

For these and other reasons, the World Economic Forum suggests that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes [45]. Although data provenance has been studied extensively in the databases community [4, 8], it is rarely discussed in the machine learning community, and documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is no standardized process for documenting machine learning datasets.

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks.

After outlining our objectives below, we begin with a discussion of parallels in other contexts, including the role of datasheets in the electronics industry. We then provide a set of questions covering the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. Next we describe the process by which we developed the questions and workflow. Finally, we conclude with a summary of the impact of datasheets for datasets and a discussion of implementation challenges and avenues for future work.

1.1 Objectives

Datasheets for datasets are intended for two key stakeholder groups: dataset creators and dataset consumers. For dataset creators, the primary objective is to encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions,

potential risks or harms, and implications of use. Answering the questions in section 3 can guide such reflection. Dataset creators should read through all questions prior to any data collection and then provide answers during the creation, distribution, and maintenance process. For dataset consumers, the primary objective is to ensure they have the information they need to make informed decisions about using a dataset, including information about its composition, collection process, recommended uses, and restrictions, as well as any underlying assumptions, potential risks or harms, etc. Transparency on the part of dataset creators is necessary for dataset consumers to be sufficiently well informed that they can select appropriate datasets and avoid unintentional misuse.

Beyond these key stakeholder groups, datasheets for datasets may be valuable to policy makers, consumer advocates, individuals whose data is included in those datasets, and individuals who may be impacted by models trained or evaluated on those datasets. They also serve a secondary objective of facilitating greater reproducibility of machine learning results: without access to a dataset, researchers and practitioners can use the information in a datasheet to reconstruct the dataset.

Although we provide a set of questions covering the information that a datasheet for a dataset might contain, they are not intended to be prescriptive. Indeed, we expect that datasheets will vary depending on factors such as the domain or existing organizational infrastructure and workflows.

We emphasize that the process of creating a datasheet is not intended to be automated. Although automated documentation processes are convenient, they run counter to our objective of encouraging dataset creators to carefully reflect on the process of creating, distributing, and maintaining a dataset. Moreover, manually creating a datasheet provides dataset creators with an opportunity to alter their creation, distribution, and maintenance process in response to their reflection.

Similarly, datasheets are not intended to replace schemas for taxonomizing datasets. Nor are they intended for data provenance or version tracking, although updated versions of a dataset should be accompanied with updated datasheets. Finally, datasheets are not intended to serve as verification or proof of a dataset's composition or to serve as a demonstration of construct validity (although a datasheet may contain sufficient information to highlight a lack of construct validity).

2 Parallels in Other Contexts

To contextualize and motivate our proposal, we first discuss parallels in two other contexts.

Electronics. Like datasets, electronic components are incorporated into systems whose ultimate behavior may be far removed from the roles of the components themselves. Small variations that seem insignificant when studying a component in isolation can have serious consequences for a system as a whole. For example, although all types of capacitors can be abstracted into an idealized mathematical model, different non-idealities may be more or less detrimental depending on the context of use [43]. As a result, organizations like the International Electrotechnical Commission (IEC) have developed international standards specifying manufacturing conditions, operating characteristics, and tests for components. According to the IEC, "Close to 20,000 experts from industry, commerce, government, test and research labs, academia and consumer groups participate

in IEC Standardization work" [26]. To complement these standards, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, and recommended uses. Datasheets are prominently displayed on components' product websites, allowing users to better understand their characteristics and limitations. In other words, datasheets facilitate communication between component manufacturers and component users.

Medicine. Just as data is critical to machine learning, clinical trials play an important role in drug development. When the US justice system stopped viewing clinical trials as a form of medical malpractice [14], standards for clinical trials were put in place, spurred by the gross mistreatment of non-consenting participants [11, 16, 31]. Now, prior to the start of a clinical trial for a drug, participants must be informed that the drug is experimental and not proven to be effective, and they must provide consent. An institutional review board and the Food and Drug Administration (FDA) must review evidence of the drug's relative safety (including its chemical composition and the results of animal testing) and the design of the trial (including participant demographics) [19]. Historically, lack of diversity in clinical trials has led to the development of drugs that are more dangerous and less efficacious for subpopulations by age [18], sex [34], and race [38]. In 2014, the FDA therefore promoted an action plan to make available the results of clinical trials broken down by subpopulation [17]. These progressions parallel recent concerns in machine learning regarding the use of individuals' data without their consent and the prevalence of performance disparities between subpopulations. Machine learning's closest legal analog to these medical safeguards is the EU's General Data Protection Regulation, which targets individuals' data protection and privacy.

3 Questions and Workflow

In this section, we provide a set of questions covering the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions. Specifically, the questions are grouped into sections that roughly match the key stages of the dataset creation, maintenance, and distribution process: motivation, composition, collection process, pre-processing/cleaning/labeling, uses, distribution, and maintenance. By grouping the questions in this way, dataset creators are encouraged to reflect on the process of creating, distributing, and maintaining a dataset, and can even alter this process in response to their reflection. We recommend that dataset creators read through all questions prior to any data collection so as to flag potential issues early on and then provide answers to the questions in each section during the relevant stage of the process. We note that not all questions will be applicable to all datasets, and dataset creators should omit those that do not apply. More generally, we again emphasize that these questions and workflow are not intended to be prescriptive. We expect that dataset creators will need to modify them based on factors such as the domain or existing organizational infrastructure and workflows.

To prompt dataset creators to provide sufficient information, all questions are worded so as to discourage yes/no answers. The questions not intended to serve as a checklist, and dataset creators must be as transparent and forthcoming as possible for datasheets to be useful to dataset consumers.

To illustrate how these questions might be answered in practice, we provide in the appendix examples of datasheets for two well-known datasets: Labeled Faces in the Wild [25] and Pang and Lee's polarity dataset [37]. We chose these datasets in large part because their creators provided exemplary documentation, allowing us to easily find the answers to many of our questions.

3.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

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

3.2 Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset. For example, if the dataset relates to people, dataset creators are asked to identify subpopulations so that dataset consumers can determine whether the dataset is appropriate for their chosen tasks. Some of the questions in this section are intended to encourage reflection and transparency around confidential or sensitive data and around individuals' data protection and privacy. The answers to these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Several of the questions in this section may additionally help others to reconstruct the dataset without access to it.

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

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
- **Is there a label or target associated with each instance?** If so, please provide a description.
- **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.
- **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.
- **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.
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
- **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.
- **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.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
- **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.
- **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.
- **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.
- **Any other comments?**

3.3 Collection Process

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. Again, the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset, to encourage reflection and transparency around individuals' data protection and privacy (specifically around notice and consent), and to provide others with information that may help them to reconstruct the dataset without access to it.

- **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.
- **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?
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
- **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)?**
- **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.
- **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.
- **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
- **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.
- **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.
- **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).
- **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.
- **Any other comments?**

3.4 Preprocessing/cleaning/labeling

Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

- **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 remainder of the questions in this section.
- **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.
- **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.
- **Any other comments?**

3.5 Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- **Has the dataset been used for any tasks already?** If so, please provide a description.
- **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.
- **What (other) tasks could the dataset be used for?**
- **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?
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
- **Any other comments?**

3.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

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

3.7 Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and to be transparent with dataset consumers about their maintenance plans.

- **Who is supporting/hosting/maintaining the dataset?**
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
- **Is there an erratum?** If so, please provide a link or other access point.
- **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)?

- **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.
- **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.
- **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.
- **Any other comments?**

4 Development Process

We refined the questions and workflow over a period of roughly a year. During this period, we incorporated feedback from dozens of researchers, practitioners, civil servants, and lawyers.

We developed an initial set of questions based on our own experiences as researchers with diverse backgrounds working in different domains at different institutions. We drew on our knowledge of dataset characteristics, unintentional misuse, unwanted biases, and other issues to produce a set of questions that spanned these topics, and then "tested" the questions by creating example datasheets for two well-known datasets: Labeled Faces in the Wild [25] and Pang and Lee's polarity dataset [37]. While creating these datasheets, we noted gaps in our questions, as well as redundancies and lack of clarity. Following these notes, we then refined our initial set of questions.

After these refinements, we solicited feedback from a wide range of stakeholders. We distributed the questions to product teams in two major US-based technology companies, in some cases helping them create datasheets for their own datasets and observing where the questions did not achieve their intended objective. Contemporaneously, we circulated an initial draft of this paper to colleagues directly, through social media, and on arXiv (draft posted 23 March 2018). Via these channels we received extensive comments from dozens of researchers, practitioners, and civil servants. We also worked with external counsel to review the questions from a legal perspective.

We incorporated this feedback to yield the questions and workflow in the previous section. We refined the content of the questions, added missing questions, deleted redundant or similar questions, and reordered the questions to better match the key stages of the dataset creation, maintenance, and distribution process. Based on our experiences with product teams, we reworded the questions to discourage yes/no answers. The biggest changes were the addition of a section on "Uses" and the deletion of a section on "Legal and Ethical Considerations." We found that product teams were likely to avoid answering questions about legal and ethical considerations if these questions were grouped together, as opposed to integrated into sections about the relevant stages of the dataset creation process. Following feedback from external counsel, we removed questions explic-

itly asking about compliance with regulations, and introduced factual questions intended to elicit relevant information about compliance without requiring dataset creators to make legal judgments.

We do not view the questions or workflow as fixed, and expect that they will evolve as datasheets for datasets become more widely used. We also expect that dataset creators will need to modify them to better reflect their domain or existing organizational infrastructure and workflows.

5 Impact and Challenges

Since circulating an initial draft of this paper in March 2018, datasheets for datasets have already gained traction in a number of settings. Academic researchers have adopted our proposal and released datasets with accompanying datasheets [e.g., 9, 15, 41]. Microsoft, Google, and IBM have begun to pilot datasheets for datasets internally within product teams. Researchers at Google published follow-up work on *model cards* that document machine learning models [30] and released a *data card* (a lightweight version of a datasheet) with the Open Images dataset [27]. Researchers at IBM proposed *factsheets* [22] that document various characteristics of AI services, including whether the datasets used to develop the services are accompanied with datasheets. Finally, the Partnership on AI, a multistakeholder organization focused on sharing best practices for developing and deploying responsible AI, is actively working on industry-wide documentation guidance that builds on datasheets, model cards, and factsheets. Datasheets for datasets have therefore already begun to increase transparency and accountability within the machine learning community.

These initial successes have also revealed a number of implementation challenges that may need to be addressed to support wider adoption. Chief among them is the need for dataset creators to modify the questions and workflow in section 3 based on their existing organizational infrastructure and workflows. We again emphasize that our questions are not intended to be prescriptive, and there is no "one-size-fits-all" approach. Many teams have sophisticated processes and tools for creating, distributing, and maintaining datasets, and for documenting compliance with regulations. Dataset creators will likely only create datasheets if they are integrated into their existing workflows. In some cases, this may mean documenting less information, as with Google's data cards [27] or the dataset nutrition label framework [23]. In other cases, this may mean implementing new tools to create datasheets on the platforms where datasets are being created or distributed.

We expect that dataset creators in different domains will need to modify our questions and workflow. The most salient information about a dataset may be domain-specific and not captured by our questions. For example, Bender and Friedman [3] note that for language-related datasets, it may be valuable to document the native language or socioeconomic status of the individuals involved in labeling. Some domains, including geoscience, medicine, and information science, have existing practices relating to metadata [1, 6, 13, 21, 32, 33, 36, 39, 40], which may be integrated into datasheets. We also note that our questions and workflow may pose challenges for dynamic datasets. If a dataset changes only infrequently, we recommend accompanying updated versions with updated datasheets. However, streaming data or other datasets that change very frequently will likely require modifications to both the questions and the workflow (e.g., dataset creators may

wish to document the number of instances per second rather than the total number of instances).

We emphasize that datasheets for datasets do not provide a complete solution to mitigating unwanted biases or potential risks or harms. In general, it is not possible for dataset creators to anticipate every possible use (or misuse) of a dataset, and identifying unwanted biases often requires additional labels indicating demographic information for individuals, which may not be available to dataset creators for reasons including those individuals' data protection and privacy [24]. We also note that many datasets are not created from scratch, and are instead assembled from other sources (e.g., scraped from websites or social media) with no way of acquiring demographic information from individuals or requesting their consent. In some of these cases, aggregate demographic information may still be available. For example, although per-employee demographic information is not available for the Enron email dataset [42], aggregate information about the employee population is.

When creating datasheets for datasets that relate to people, it may be necessary for dataset creators to work with experts in other domains. For example, researchers in anthropology are well-versed in collecting demographic information. Similarly, there are complex and contextual social, historical, and geographical factors that influence how best to collect a dataset in a manner that is respectful of individuals and their data protection and privacy. The questions and workflow in section 3 should be modified as appropriate to encourage dataset creators to follow best practices when creating their datasets, without discouraging them from being transparent about their process.

Finally, creating datasheets for datasets will necessarily impose overhead on dataset creators. Although datasheets may reduce the amount of time that dataset creators spend answering one-off questions about datasets, the process of creating a datasheet will always take time, and organizational infrastructure and workflows will need to be modified to accommodate this investment. In addition, the information in a datasheet may expose the entity on behalf of which the dataset was created to legal risks, including inadvertent release of proprietary information. Despite these challenges, there are many benefits to creating datasheets for datasets. In addition to facilitating better communication between dataset creators and dataset consumers, datasheets provide the opportunity for dataset creators to distinguish themselves as prioritizing transparency and accountability. Ultimately, we believe that the benefits to the machine learning community outweigh the costs.

Acknowledgments

We thank Peter Bailey, Emily Bender, Yoshua Bengio, Sarah Brown, Steven Bowles, Joy Buolamwini, Amanda Casari, Eric Charran, Alain Couillault, Lukas Dauterman, Leigh Dodds, Miroslav Dudík, Michael Ekstrand, Noémie Elhadad, Michael Golebiewski, Nick Gonsalves, Martin Hansen, Andy Hickl, Michael Hoffman, Scott Hoogerwerf, Eric Horvitz, Mingjing Huang, Surya Kallumadi, Ece Kamar, Krishnaram Kenthapadi, Emre Kiciman, Jacquelyn Krones, Erik Learned-Miller, Lillian Lee, Jochen Leidner, Rob Mauceri, Brian Mcfee, Emily McReynolds, Bogdan Micu, Margaret Mitchell, Brendan O'Connor, Thomas Padilla, Bo Pang, Anjali Parikh, Lisa Peets, Alessandro Perina, Michael Philips, Barton Place, Sudha Rao, David Van Riper, Anna Roth, Cynthia Rudin, Ben Shneiderman, Biplav Srivastava, Ankur Teredesai, Rachel Thomas, Martin Tomko, Panagiotis Tziachris, Meredith Whittaker, Hans Wolters, Ashly Yeo, Lu Zhang, and the

attendees of the Partnership on AI's April 2019 ABOUT ML workshop for valuable feedback.

References

[1] C A. Manduca, Sean Fox, and H Rissler. 2006. DataSheets: Making Geoscience Data Easier to Find and Use. *AGU Fall Meeting Abstracts* (12 2006).

[2] Don A Andrews, James Bonta, and J Stephen Wormith. 2006. The recent past and near future of risk and/or need assessment. *Crime & Delinquency* 52, 1 (2006), 7–27.

[3] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[4] Anant P. Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J. Elmore, Samuel Madden, and Aditya G. Parameswaran. 2014. DataHub: Collaborative Data Science & Dataset Version Management at Scale. *CoRR* abs/1409.0798 (2014).

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 4349–4357.

[6] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, et al. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics* 29, 4 (2001), 365.

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM, New York, NY, USA, 77–91.

[8] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases* 1, 4 (2009), 379–474.

[9] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. *CoRR* abs/1808.07036 (2018).

[10] Glennda Chui. 2017. Project will use AI to prevent or minimize electric grid failures. [Online; accessed 14-March-2018].

[11] William J Curran. 1973. The Tuskegee syphilis study. *The New England Journal of Medicine* 289, 14 (1973).

[12] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUS

[13] Liping Di, Yuanzheng Shao, and Lingjun Kang. 2013. Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 19115-2 lineage model. *IEEE Transactions on Geoscience and Remote Sensing* 51, 11 (2013), 5082–5089.

[14] Harry F Dowling. 1975. The emergence of the cooperative clinical trial. *Transactions & Studies of the College of Physicians of Philadelphia* 43, 1 (1975), 20–29.

[15] Erkut Erdem. 2018. Datasheet for RecipeQA.

[16] Ruth R Faden, Susan E Lederer, and Jonathan D Moreno. 1996. US medical researchers, the Nuremberg Doctors Trial, and the Nuremberg Code. A review of findings of the Advisory Committee on Human Radiation Experiments. *JAMA* 276, 20 (1996), 1667–1671.

[17] Food and Drug Administration. 1985. Content and Format of a New Drug Application (21 CFR 314.50 (d)(5)(v)).

[18] Food and Drug Administration. 1989. Guidance for the Study of Drugs Likely to Be Used in the Elderly.

[19] Food and Drug Administration. 2018. FDA Clinical Trials Guidance Documents.

[20] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. 2016. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, New Jersey Ave NW, Washington, DC.

[21] Tracy D Gunter and Nicolas P Terry. 2005. The emergence of national electronic health record architectures in the United States and Australia: Models, costs, and questions. *Journal of Medical Internet research* 7, 1 (2005).

[22] Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R. Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *CoRR* abs/1808.07261 (2018).

[23] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). http://arxiv.org/abs/1805.03677

[24] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do

Industry Practitioners Need?. In *2019 ACM CHI Conference on Human Factors in Computing Systems*.

[25] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Technical Report 07-49. University of Massachusetts Amherst.

[26] International Electrotechnical Commission. 2017. About the IEC: Overview.

[27] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. OpenImages: A public dataset for large-scale multi-label and multi-class image classification.

[28] Tom CW Lin. 2012. The new investor. *UCLA Law Review* 60 (2012), 678.

[29] G Mann and C O'Neil. 2016. Hiring Algorithms Are Not Neutral. `https://hbr.org/2016/12/hiring-algorithms-are-not-neutral`.

[30] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 220–229. `https://doi.org/10.1145/3287560.3287596`

[31] Jonathan D Moreno. 2013. *Undue Risk: Secret State Experiments on Humans*. Routledge, New York, NY, USA.

[32] National Electrical Manufacturers Association. 2018. Digital Imaging and Communications in Medicine. `https://www.dicomstandard.org/`.

[33] National Library of Medicine. 2018. National Library of Medicine. `https://www.nlm.nih.gov/`.

[34] Martha R Nolan and Thuy-Linh Nguyen. 2013. Analysis and Reporting of Sex Differences in Phase III Medical Device Clinical Trials—How Are We Doing? *Journal of Women's Health* 22, 5 (2013), 399–401.

[35] Mary Catherine O'Connor. 2017. How AI Could Smarten Up Our Water System. [Online; accessed 14-March-2018].

[36] Thomas Padilla. 2016. Humanities data in the library: Integrity, form, access. *D-Lib Magazine* 22, 3 (2016), 1.

[37] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 271.

[38] Katrina L. Pariera, Sheila T. Murphy, Jingbo Meng, and Margaret L. McLaughlin. 2017. Exploring Willingness to Participate in Clinical Trials by Ethnicity. *Journal of Racial and Ethnic Health Disparities* 4, 763 (2017).

[39] Andreas Rauber, Ari Asmi, Dieter van Uytvanck, and Stefan Proell. 2016. Identification of reproducible subsets for data citation, sharing and re-use. *Bulletin of IEEE Technical Committee on Digital Libraries* 12, 1 (2016), 6–15.

[40] SDMX. 2018. Statistical Data and Metadata eXchange. `https://sdmx.org/?page_id=3425`.

[41] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a datasheet for the Cerema AWP dataset. *CoRR* abs/1806.04016 (2018). `http://arxiv.org/abs/1806.04016`

[42] Jitesh Shetty and Jafar Adibi. 2004. The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California* 4, 1 (2004), 120–128.

[43] Larry D Smith, Raymond E Anderson, Douglas W Forehand, Thomas J Pelc, and Tanmoy Roy. 1999. Power distribution system design methodology and capacitor selection for modern CMOS technology. *IEEE Transactions on Advanced Packaging* 22, 3 (1999), 284–291.

[44] Doha Suppy Systems. 2017. Facial Recognition. [Online; accessed 14-March-2018].

[45] World Economic Forum Global Future Council on Human Rights 2016–2018. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. `https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning`.

A Appendix

In this appendix, we provide examples of datasheets for two well-known datasets: Labeled Faces in the Wild [25] (figure 1 to figure 6) and Pang and Lee's polarity dataset [37] (figure 7 to figure 10).

Motivation

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.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

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

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

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 construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

Composition

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

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

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

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

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 does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

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

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

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

Each image is accompanied by a label indicating the name of the person in the image.

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.

Everything is included in the dataset.

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.

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

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.

The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy,** $\hat{\mu}$ and the **standard error of the mean:** S_E for View 2.

$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \tag{1}$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \tag{2}$$

[1] All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.
Original paper: http://www.cs.cornell.edu/people/pabo/movie-review-data/; LFW survey: http://vis-www.cs.umass.edu/lfw/lfw.pdf; Paper measuring LFW demographic characteristics : http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf; LFW website: http://vis-www.cs.umass.edu/lfw/.

Figure 1: Example datasheet for Labeled Faces in the Wild [25], page 1.

Where $\hat{\sigma}$ is the estimate of the standard deviation, given by:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10}(p_i - \hat{\mu})^2}{9}} \quad (3)$$

The multiple-view approach is used instead of a traditional train/validation/test split in order to maximize the amount of data available for training and testing.

Training Paradigms: There are two training paradigms that can be used with our dataset. Practitioners should specify the training paradigm they used while reporting results.

- **Image-Restricted Training** This setting prevents the experimenter from using the name associated with each image during training and testing. That is, the only available information is whether or not a pair of images consist of the same person, not who that person is. This means that there would be no simple way of knowing if there are multiple pairs of images in the train/test set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than comparing names). Thus, to form training pairs of matched and mismatched images for the same person, one can use image equivalence to add images that consist of the same person.

  The files pairsDevTrain.txt and pairsDevTest.txt support image-restricted uses of train/test data. The file pairs.txt in View 2 supports the image-restricted use of training data.

- **Unrestricted Training** In this setting, one can use the names associated with images to form pairs of matched and mismatched images for the same person. The file people.txt in View 2 of the dataset contains subsets of people along with images for each subset. To use this paradigm, matched and mismatched pairs of images should be formed from images in the same subset. In View 1, the files peopleDevTrain.txt and peopleDevTest.txt can be used to create arbitrary pairs of matched/mismatched images for each person. The unrestricted paradigm should only be used to create training data and not for performance reporting. The test data, which is detailed in the file pairs.txt, should be used to report performance. We recommend that experimenters first use the image-restricted paradigm and move to the unrestricted paradigm if they believe that their algorithm's performance would significantly improve with more training data. While reporting performance, it should be made clear which of these two training paradigms were used for a particular test result.

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

http://vis-www.cs.umass.edu/lfw/#download lists a small number of errors including a few incorrect matched pairs in the dataset and other known labeling errors. Errors could also have been introduced while determining the name of each individual in the dataset if the original caption associated with each person's photograph is incorrect. Some additional potential limitations and sources of bias are also listed at the end of the datasheet.

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 is self-contained.

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.

No. All data was derived from publicly available news sources.

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

No. The dataset only consists of faces and associated names.

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

Yes. The dataset contains people's faces.

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.

While subpopulation data was not available at the initial release of the dataset, a subsequent paper[2] reports the distribution of images by age, race and gender. Table 2 lists these results. The age, perceived gender and race of each individual in the dataset was collected using Amazon Mechanical Turk, with 3 crowd workers labeling each image. After exact age estimation, the ages were binned into groups of 0-10, 21-40, 41-60 and 60+.

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.

Each image is annotated with the name of the person that appears in the image.

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.

The dataset does not contain confidential information since all information was scraped from news stories.

Any other comments?

[2] http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

Figure 2: Example datasheet for Labeled Faces in the Wild [25], page 2.

Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images.* Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

Collection Process

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 names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

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?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[3]. The images in this database were gathered from news articles on the web using software to crawl news articles.

[3]Faces in the Wild: http://tamaraberg.com/faceDataset/

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

The original Faces in the Wild dataset is a sample of pictures of people appearing in the news on the web. Labeled Faces in the Wild is thus also a sample of images of people found on the news on line. While the intention of the dataset is to have a wide range of demographic (e.g. age, race, ethnicity) and image (e.g. pose, illumination, lighting) characteristics, there are many groups that have few instances (e.g. only 1.57% of the dataset consists of individuals under 20 years old).

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)?

Subsequent gender, age and race annotations listed in http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf were performed by crowd workers found through Amazon Mechanical Turk.

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.

Unknown

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.

Unknown

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

Yes. Each instance is an image of a person.

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 was crawled from public web sources.

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.

Unknown

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.

No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

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).

Figure 3: Example datasheet for Labeled Faces in the Wild [25], page 3.

No. The data was crawled from public web sources, and the individuals appeared in news stories. But there was no explicit informing of these individuals that their images were being assembled into a dataset.

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.

Unknown

Any other comments?

Preprocessing/cleaning/labeling

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 remainder of the questions in this section.

The following steps were taken to process the data:

1. **Gathering raw images:** First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web.
2. **Running the Viola-Jones face detector**[4] The OpenCV version 1.0.0 release 1 implementation of Viola-Jones face detector was used to detect faces in each of these images, using the function cvHaarDetectObjects, with the provided Haar classifier—cascadehaarcascadefrontalfacedefault.xml. The scale factor was set to 1.2, min neighbors was set to 2, and the flag was set to CV HAAR DO CANNY PRUNING.
3. **Manually eliminating false positives:** If a face was detected and the specified region was determined not to be a face (by the operator), or the name of the person with the detected face could not be identified (using step 5 below), the face was omitted from the dataset.
4. **Eliminating duplicate images:** If images were determined to have a common original source photograph, they are defined to be duplicates of each other. An attempt was made to remove all duplicates but a very small number (that were not initially found) might still exist in the dataset. The number of remaining duplicates should be small enough so as not to significantly impact training/testing. The dataset contains distinct images that are not defined to be duplicates but are extremely similar. For example, there are pictures of celebrities that appear to be taken almost at the same time by different photographers from slightly different angles. These images were not removed.
5. **Labeling (naming) the detected people:** The name associated with each person was extracted from the associated news caption. This can be a source of error if the original news caption was incorrect. Photos of the same person were combined into a single group associated with one name. This was a challenging process as photos of some people were associated with multiple names in the news captions (e.g."Bob McNamara" and "Robert McNamara"). In this scenario, an attempt was made to use the most common name. Some people have a single name (e.g. "Madonna" or "Abdullah"). For Chinese and some other Asian names, the common Chinese ordering (family name followed by given name) was used (e.g. "Hu Jintao").
6. **Cropping and rescaling the detected faces:** Each detected region denoting a face was first expanded by 2.2 in each dimension. If the expanded region falls outside of the image, a new image was created by padding the original pixels with black pixels to fill the area outside of the original image. This expanded region was then resized to 250 pixels by 250 pixels using the function cvResize, and cvSetImageROI as necessary. Images were saved in JPEG 2.0 format.
7. **Forming pairs of training and testing pairs for View 1 and View 2 of the dataset:** Each person in the dataset was randomly assigned to a set (with 0.7 probability of being in a training set in View 1 and uniform probability of being in any set in View 2). Matched pairs were formed by picking a person uniformly at random from the set of people who had two or more images in the dataset. Then, two images were drawn uniformly at random from the set of images of each chosen person, repeating the process if the images are identical or if they were already chosen as a matched pair). Mismatched pairs were formed by first choosing two people uniformly at random, repeating the sampling process if the same person was chosen twice. For each chosen person, one image was picked uniformly at random from their set of images. The process is repeated if both images are already contained in a mismatched pair.

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.

The raw unprocessed data (consisting of images of faces and names of the corresponding people in the images) is saved.

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

While a script running a sequence of commands is not available, all software used to process the data is open source and has been specified above.

Any other comments?

[4] Paul Viola and Michael Jones. *Robust real-time face detection*. IJCV, 2004

Figure 4: Example datasheet for Labeled Faces in the Wild [25], page 4.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

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.
Papers using this dataset and the specified training/evaluation protocols are listed under "Methods" section of http://vis-www.cs.umass.edu/lfw/results.html

What (other) tasks could the dataset be used for?
The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[5]

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?
There is minimal risk for harm: the data was already public.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset should not be used for tasks that are high stakes (e.g. law enforcement).

Any other comments?

Distribution

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.
Yes. The dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset can be downloaded from http://vis-www.cs.umass.edu/lfw/index.html#download. The images can be downloaded as a gzipped tar file.

When will the dataset be distributed?
The dataset was released in October, 2007.

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.
The crawled data copyright belongs to the news papers that the data originally appeared in. There is no license, but there is a request to cite the corresponding paper if the dataset is used: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

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.
There are no fees or restrictions.

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.
Unknown

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset is hosted at the University of Massachusetts.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
All questions and comments can be sent to Gary Huang: gbhuang@cs.umass.edu.

Is there an erratum? If so, please provide a link or other access point.
All changes to the dataset will be announced through the LFW mailing list. Those who would like to sign up should send an email to lfw-subscribe@cs.umass.edu. Errata are listed under the "Errata" section of http://vis-www.cs.umass.edu/lfw/index.html

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)?
All changes to the dataset will be announced through the LFW mailing list.

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.
No.

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.
They will continue to be supported with all information on http://vis-www.cs.umass.edu/lfw/index.html unless otherwise communicated on the LFW mailing list.

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.

[5] Unconstrained face recognition: Identifying a person of interest from a media collection: http://biometrics.cse.msu.edu/Publications/Face/BestRowdenetal_UnconstrainedFaceRecognition_TechReport_MSU-CSE-14-1.pdf

Figure 5: Example datasheet for Labeled Faces in the Wild [25], page 5.

Unknown

Any other comments?

There some potential limitations in the dataset which might bias the data towards a particular demographic, pose, image characteristics etc.

- The Viola-Jones detector can have systematic errors by race, gender, age or other categories
- Due to the Viola-Jones detector, there are only a small number of side views of faces, and only a few views from either above or below
- The dataset does not contain many images that occur under extreme (or very low) lighting conditions
- The original images were collected from news paper articles. These articles could cover subjects in limited geographical locations, specific genders, age, race, etc. The dataset does not provide information on the types of garments worn by the individuals, whether they have glasses on, etc.
- The majority of the dataset consists of people whose perceived gender has been labeled as male, and race as White.
- There are very few images of people who under 20 years old.
- The proposed train/test protocol allows reuse of data between View 1 and View 2 in the dataset. This could potentially introduce very small biases into the results

Figure 1. Examples of images from our dataset (matched pairs)

Figure 6: Example datasheet for Labeled Faces in the Wild [25], page 6.

Motivation

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 dataset was created to enable research on predicting sentiment polarity: given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

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

The dataset was created by Bo Pang and Lillian Lee at Cornell University.

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.

Funding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

Composition

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

The instances are movie reviews extracted from newsgroup postings, together with a sentiment rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The polarity rating is binary {positive,negative}. An example instance is shown in Figure 1.

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

There are 1400 instances in total in the original (v1.x versions) and 2000 instances in total in v2.0 (from 2014).

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 a sample of instances. It is (presumably) intended to be a random sample of instances of movie revies from newsgroup postings. No tests were run to determine representativeness.

[1] Information in this datasheet is taken from one of five sources; any errors that were introduced are our fault. http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

> these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

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

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and alter fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

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

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.

Everything is included. No data is missing.

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.

None explicitly, though the original newsgroup postings include poster name and email address, so some information could be extracted if needed.

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.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

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

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.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient con-

Figure 7: Example datasheet for Pang and Lee's polarity dataset [37], page 1.

fidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

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

Some movie reviews might contain moderately inappropriate or offensive language, but we do not expect this to be the norm.

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

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.

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.

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.

The raw form of the dataset contains names and email addresses, but these are already public on the internet newsgroup.

Any other comments?

Collection Process

Similar to Composition, this section should be read during the initial planning phase, and filled out during the collection of data. Again, these questions provide general transparency into the makeup of the data help both the dataset creator and dataset consumer uncover risks and potential harms, for example by questioning whether those whose information is contained in the dataset have control over usage of their data or the ability to remove their information from the dataset entirely.

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 data was mostly observable as raw text, except the labels were extracted by the process described below. The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at http://reviews.imdb.com/Reviews.

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?

Unknown.

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

The sample of instances collected is English movie reviews from the `rec.arts.movies.reviews` newsgroup, from which a "number of stars" rating could be extracted. The sample is limited to forty reviews per unique author in order to achieve broader coverage by authorship. Beyond that, the sample is arbitrary.

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)?

Unknown

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.

Unknown

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.

Unknown

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

The dataset relates to people in that the reviews themselves are authored by people. Personally identifying information (e.g., email addresses) was removed.

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 was collected from newsgroups.

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.

No. The data was crawled from public web sources, and the authors of the posts presumably knew that their posts would be public, but there was no explicit informing of these authors that their posts were to be used in this way.

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.

No (see previous question).

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.

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.

Any other comments?

Figure 8: Example datasheet for Pang and Lee's polarity dataset [37], page 2.

Preprocessing/cleaning/labeling

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 remainder of the questions in this section.

Instances for which an explicit rating could not be found were discarded. Also only instances with strongly-positive or strongly-negative ratings were retained. Star ratings were extracted by automatically looking for text like "`**** out of *****`" in the review, using that as a label, and then removing the corresponding text. When the star rating was out of five stars, anything at least four was considered positive and anything at most two negative; when out of four, three and up is considered positive, and one or less is considered negative. Occasionally half stars are missed which affects the labeling of negative examples. Everything in the middle was discarded. In order to ensure that sufficiently many authors are represented, at most 20 reviews (per positive/negative label) per author are included.

In a later version of the dataset (v1.1), non-English reviews were also removed.

Some preprocessing errors were caught in later versions. The following fixes were made: (1) Some reviews had rating information in several places that was missed by the initial filters; these are removed. (2) Some reviews had unexpected/unparsed ranges and these were fixed. (3) Sometimes the boilerplate removal removed too much of the text.

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.

Yes. The dataset itself contains all the raw dta.

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

No.

Any other comments?

Uses

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

At the time of publication, only the original paper http://xxx.lanl.gov/pdf/cs/0409058v1. Between then and 2012, a collection of papers that used this dataset was maintained at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

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 a repository, maintained by Pang/Lee through April 2012, at http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/otherexperiments.html.

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

The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrases that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

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?

There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were removed.

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

This data is collected solely in the movie review domain, so systems trained on it may or may not generalize to other sentiment prediction tasks. Consequently, such systems should not—without additional verification—be used to make consequential decisions about people.

Any other comments?

Distribution

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.

Yes, the dataset is publicly available on the internet.

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

The dataset is distributed on Bo Pang's webpage at Cornell: http://www.cs.cornell.edu/people/pabo/movie-review-data. The dataset does not have a DOI and there is no redundant archive.

When will the dataset be distributed?

The dataset was first released in 2002.

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.

The crawled data copyright belongs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: *Thumbs up? Sentiment classification using machine learning techniques.* Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Proceedings of EMNLP, 2002.

Figure 9: Example datasheet for Pang and Lee's polarity dataset [37], page 3.

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.

No.

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.

Unknown

Any other comments?

Maintenance

This section should be completed once the dataset has been constructed, before it is distributed. These questions help the dataset creator think through their plans for updating, adding to, or fixing errors in the dataset, and expose these plans to dataset consumers.

Who is supporting/hosting/maintaining the dataset?

Bo Pang is supporting/maintaining the dataset.

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

Unknown

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

Since its initial release (v0.9) there have been three later releases (v1.0, v1.1 and v2.0). There is not an explicit erratum, but updates and known errors are specified in higher version `README` and `diff` files. There are several versions of these: v1.0: http://www.cs.cornell.edu/people/pabo/movie-review-data/README; v1.1: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/README.1.1 and http://www.cs.cornell.edu/people/pabo/movie-review-data/diff.txt; v2.0: http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/poldata.README.2.0.txt. Updates are listed on the dataset web page. (This datasheet largely summarizes these sources.)

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)?

This will be posted on the dataset webpage.

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.

N/A.

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.

The dataset has already been updated; older versions are kept around for consistency.

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.

Others may do so and should contact the original authors about incorporating fixes/extensions.

Any other comments?

Figure 10: Example datasheet for Pang and Lee's polarity dataset [37], page 4.