Datasheets for Datasets

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

Abstract

The machine learning community has no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns. To address this gap, we propose the concept of datasheets for datasets. In the electronics industry, it is standard to accompany every component with a datasheet providing standard operating characteristics, test results, recommended usage, and other information. Similarly, we recommend that every dataset be accompanied with a datasheet documenting its creation, composition, intended uses, maintenance, and other properties. Datasheets for datasets will facilitate better communication between dataset creators and users, and encourage the machine learning community to prioritize transparency and accountability.

1. Introduction

Machine learning is no longer a purely academic discipline. Domains such as criminal justice (Garvie et al., 2016; Systems, 2017; Andrews et al., 2006), hiring and employment (Mann & O'Neil, 2016), critical infrastructure (O'Connor, 2017; Chui, 2017), and finance (Lin, 2012) all increasingly depend on machine learning methods.

By definition, machine learning models are trained using data; the choice of data fundamentally influences a model's behavior. However, there is no standardized way to document how and why a dataset was created, what information it contains, what tasks it should and shouldn't be used for, and whether it might raise any ethical or legal concerns. This lack of documentation is especially problematic when datasets are used to train models for high-stakes applications.

[1]Microsoft Research, New York, NY [2]Georgia Institute of Technology, Atlanta, GA [3]Cornell University, Ithaca, NY [4]University of Maryland, College Park, MD [5]AI Now Institute, New York, NY. Correspondence to: Timnit Gebru <tgebru@gmail.com>.

Proceedings of the 5^{th} Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

We therefore propose the concept of datasheets for datasets. In the electronics industry, every component is accompanied by a datasheet describing standard operating characteristics, test results, and recommended usage. By analogy, we recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information. We anticipate that such datasheets will increase transparency and accountability in the machine learning community.

Section 2 provides context for our proposal. Section 3 discusses the evolution of safety standards in other industries, and outlines the concept of datasheets in electronics. We give examples of questions that should be answered in datasheets for datasets in Section 4, and discuss challenges and future work in Section 5. The appendix includes a more complete proposal along with prototype datasheets for two well-known datasets: Labeled Faces in the Wild (Huang et al., 2007) and Pang and Lee's polarity dataset (2004).

2. Context

A foundational challenge in the use of machine learning is the risk of deploying systems in unsuitable environments. A model's behavior on some benchmark may say very little about its performance in the wild. Of particular concern are recent examples showing that machine learning systems can amplify existing societal biases. For example, Buolamwini & Gebru (2018) showed that commercial gender classification APIs have near perfect performance for lighter-skinned males, while error rates for darker-skinned females can be as high as 33%.[1] Bolukbasi et al. (2016) showed that word embeddings trained on news articles exhibit gender biases, finishing the analogy "man is to computer programmer as woman is to X" with "homemaker," a stereotypical role for women. Caliskan et al. (2017) showed these embeddings also contain racial biases: traditional European-American names are closer to positive words like "joy," while African-American names are closer to words like "agony."

These biases can have dire consequences that might not be easily discovered. Much like a faulty resistor or a capacitor in a circuit, the effects of a biased machine learning

[1]The evaluated APIs also provided the labels of female and male, failing to address the complexities of gender beyond binary.

component, such as a dataset, can propagate throughout a system making them difficult to track down. For example, biases in word embeddings can result in hiring discrimination (Bolukbasi et al., 2016). For these and other reasons, the World Economic Forum lists tracking the provenance, development, and use of training datasets as a best practice that all companies should follow in order to prevent discriminatory outcomes (World Economic Forum Global Future Council on Human Rights 2016–2018, 2018). But while provenance has been extensively studied in the database literature (Cheney et al., 2009; Bhardwaj et al., 2014), it has received relatively little attention in machine learning.

The risk of unintentional misuse of datasets can increase when developers are not domain experts. This concern is particularly important with the movement toward "democratizing AI" and toolboxes that provide publicly available datasets and off-the-shelf models to be trained by those with little-to-no domain knowledge or machine learning expertise. As these powerful tools become available to a broader set of developers, it is increasingly important to enable these developers to understand the implications of their work.

We believe this problem can be partially mitigated by accompanying datasets with datasheets that describe their creation, their strengths, and their limitations. While this is not the same as making everyone an expert, it gives an opportunity for domain experts to communicate what they know about a dataset and its limitations to the developers who might use it.

The use of datasheets will be more effective if coupled with educational efforts around interpreting and applying machine learning models. Such efforts are happening both within traditional "ivory tower" institutions (e.g., the new ethics in computing course at Harvard) and in new educational organizations. For instance, one of Fast.ai's missions is "to get deep learning into the hands of as many people as possible, from as many diverse backgrounds as possible" (Fast.ai, 2017); their educational program includes explicit training in dataset biases and ethics. Combining better education and datasheets will more quickly enable progress by both domain experts and machine learning experts.

3. Safety Standards in Other Industries

To put our proposal into context, we discuss the evolution of safety standards for automobiles, drugs, and electronics. Lessons learned from the historical dangers of new technologies, and the safety measures put in place to combat them, can help define a path forward for machine learning.

3.1. The Automobile Industry

Similar to current hopes that machine learning will positively transform society, the introduction of automobiles promised to expand mobility and provide additional recreational, social, and economic opportunities. However, much like current machine learning technology, automobiles were introduced with few safety checks or regulations. When cars first became available in the US, there were no speed limits, stop signs, traffic lights, driver education, or regulations pertaining to seat belts or drunk driving (Canis, 2017). This resulted in many deaths and injuries due to collisions, speeding, and reckless driving (Hingson et al., 1988). Reminiscent of current debates about machine learning, courtrooms and newspaper editorials argued the possibility that the automobile was inherently evil (Lewis v. Amorous, 1907).

The US and the rest of the world have gradually enacted driver education, drivers licenses (Department of Transportation Federal Highway Administration, 1997), and safety systems like four-wheel hydraulic brakes, shatter-resistant windshields, all-steel bodies (McShane, 2018), padded dashboards, and seat belts (Peltzman, 1975). Motorists' slow adoption of seat belts spurred safety campaigns promoting their adoption. By analogy, machine learning will likely to require laws and regulations (especially in high-stakes environments), as well as social campaigns to promote best practices. The automobile industry routinely uses crash-test dummies to develop and test safety systems. This practice led to problems similar to the "biased dataset" problems currently faced by the machine learning community: almost all crash-test dummies were designed with prototypical male physiology; only in 2011 did US safety standards require frontal crash tests with "female" crash-test dummies (National Highway Traffic Safety Administration, 2006), following evidence that women sustained more serious injuries than men in similar accidents (Bose et al., 2011).

3.2. Clinical Trials in Medicine

Like data collection and experimentation for 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 (Dowling, 1975), standards for clinical trials were put in place, often spurred by gross mistreatment, committed in the name of science. For example, the US government ran experiments on citizens without their consent, including a study of patients with syphilis who were not told they were sick (Curran, 1973) and radiation experiments (Faden et al., 1996; Moreno, 2013). The poor, the imprisoned, minority groups, pregnant women, and children comprised a majority of these study groups.

The US now requires drug trials to inform participants that drugs are experimental and not proven to be effective (and participants must consent). Prior to the start of a clinical trial, an Institutional Review Board and the Food and Drug Administration (FDA) review evidence of the drug's relative safety (including the drug's chemical composition and results of animal testing) and the trial design (including

participant demographics) (Food and Drug Administration, 2018). Machine learning's closest legal analog to these safeguards is the EU's General Data Protection Regulation (GDPR), which aims to ensure that users' personal data are not used without their explicit consent. Standards for data collection, storage, and sharing are now a central topic of concern for scientific research in general, and clinical trials are no exception (National Institutes of Health, 2018).

Finally, the lack of diversity in clinical trial participants has led to the development of drugs with more danger and less efficacy for many groups. In the late 1980s, the FDA mandated that clinical trial participants should be composed of populations from different age groups (Food and Drug Administration, 1989). Regulation stating that safety and efficacy data be broken down by sex, race, and age was only passed in 1998. As late as 2013, a majority of federally funded clinical trials still did not break down their results by sex (Nolan & Nguyen, 2013). In 2014, the FDA promoted an action plan to make results of clinical trials broken down by subpopulation more easily available (Food and Drug Administration, 1985). These regulations and policies followed evidence of high risk-to-reward trade-offs for drugs treating these populations. For example, eight out of ten drugs recalled between 1997 and 2001 had more adverse effects for women (Liu & Dipietro Mager, 2016). These progressions parallel recent examples showing that various machine learning systems exhibit accuracy disparities between subpopulations, and calls for more diverse datasets, inclusive testing, and standards to address these disparities.

3.3. Electrical and Electronic Technologies

Like datasets, electronic components are incorporated into a system whose larger goal may be far removed from the tasks of specific components. Thus, small deviations that may seem insignificant while studying a component in isolation can have serious consequences for the system as a whole. For instance, while all types of capacitors can be abstracted into an idealized mathematical model, different non-idealities are significant depending on the context. As an example, having a low equivalent series resistance is important in certain power supply and radio frequency applications, while this parameter is lower priority in most other designs (Smith et al., 1999). Thus, the electronics industry has developed standards specifying ideal operating characteristics, tests, and manufacturing conditions for components manufactured with different tasks in mind.

Many of these standards are specified by the International Electrotechnical Commission (IEC). 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" (International Electrotechnical Commission, 2017). In addition to international standards, all electronic components, ranging from the cheapest and most ubiquitous resistor to highly complex integrated circuits like CPUs, are accompanied by datasheets that are prominently displayed on the components' product webpages. A component's datasheet contains a description of the component's function, features, and other specifications (such as absolute maximum and minimum operating voltages); physical details of the component (such as size and pin connections) and lists of available packages; liability disclaimers to protect the manufacturer (e.g., in case the component is used in high-stakes environments like nuclear power plants or life-support systems); and compliance with relevant (e.g., IEC) standards.

For example, the datasheet for a miniature aluminum electrolytic capacitor (Passive Components, 2005) contains

- a description of the component's function;
- notable features like the component's compliance with the Restriction of Hazardous Substances Directive (European Parliament and the Council of the EU, 2003);
- standard operating characteristics, including operating temperature range and capacitance tolerance;
- a diagram of its dimensions and pin connections;
- plots of time vs. temperature and frequency.

4. Datasheets for Datasets

In the appendix, we propose a set of questions that a datasheet for a dataset should arguably contain. We also include prototype datasheets for two well-known datasets that illustrate how these questions might be answered in practice, one for Labeled Faces in the Wild (Huang et al., 2007) and one for Pang and Lee's polarity dataset (2004). We chose these two datasets in part because their authors provided exemplary documentation, allowing us to easily find the answers to many of our proposed questions.

Our development of these questions was driven by several fundamental objectives. First, a practitioner should be able to decide, from reading a datasheet, how appropriate the corresponding dataset is for a task, what its strengths and limitations are, and how it fits into the broader ecosystem. Second, the creators of a dataset should be able to use our proposed questions to prompt thought about aspects of dataset creation that may not have otherwise occurred to them.

The questions are divided into seven categories: motivation for dataset creation; dataset composition; data collection process; data preprocessing; dataset distribution; dataset maintenance; and legal and ethical considerations. Not all questions will be applicable to all datasets, in which case they can be omitted. The questions are not intended to be definitive. Instead, we hope that they will initiate a broader conversation about how data provenance, ethics, privacy, and documentation might be handled by the machine learn-

ing community. Below are a few examples of the questions:

- **Why was the dataset created?** (e.g., was there a specific intended task gap that needed to be filled?)
- **Who funded the creation of the dataset?**
- **What preprocessing/cleaning was done?** (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances)
- **If it relates to people, were they told what the dataset would be used for and did they consent?** If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?
- **Will the dataset be updated?** How often, by whom?

5. Challenges and Future Work

Our proposal faces a number of implementation challenges: how to converge on standard formats and content for datasheets, how to identify incentives that will encourage datasheet creation and overcome inertia, and how to communicate with outside experts to properly address the complex ethical considerations that relate to data about people. We describe these challenges below, and urge the machine learning community to make progress on them in future work.

Researchers, practitioners, and other enthusiasts that create and use datasets must come to a consensus about what information should be included in a datasheet, and how best to gather and share that information. Just as different categories of electronic components have different characteristics, the most relevant information about a dataset will likely be context-specific. A dataset of photographs of human faces will have different relevant information than datasets about health or weather. Some domains, including geoscience, medicine, and information science, have existing standards and methods for gathering metadata (National Electrical Manufacturers Association, 2018; Gunter & Terry, 2005; Brazma et al., 2001; National Library of Medicine, 2018; SDMX, 2018; Manduca et al., 2006; Di et al., 2013; Padilla, 2016; Rauber et al., 2016). Because this line of work is new to machine learning, we should not expect consensus to happen easily. Experts in a particular domain might first agree on a small number of critical domain-specific attributes for their own datasets. For example, recent work focuses on some of these attributes in natural language processing (Anonymous, 2018). There will also be questions that are relevant to all datasets (e.g., Why was the dataset created? Who funded the creation of the dataset?) about which the field should eventually come to some agreement. We are heartened to see other groups working on projects similar to datasheets for datasets (Holland et al., 2018).

When designing datasheets, it will be necessary to communicate with experts in other areas. For example, researchers in fields like anthropology are well-versed in the collection of demographic information about people. There are additional complex social, historical, and geographical contextual questions regarding how best to address ethical issues such as biases and privacy. Questions should be framed to encourage practitioners to follow ethical procedures while gathering data, without discouraging them from providing relevant information about the process (or creating a datasheet).

Although this paper proposes the concept of datasheets for datasets, datasheets are also needed for pretrained models and their APIs. What questions should be asked about the behavior of models and APIs, and how should the answers be measured and communicated given that models are often developed using multiple datasets, expert knowledge, and other sources of input? Organizations that produce pretrained models and APIs should iterate with developers and customers to arrive at appropriate questions and guidelines for "datasheets for models" that would parallel our proposal.

There will necessarily be overhead in creating datasheets, some of which we are working to mitigate by designing an interactive datasheet creation tool. Although carefully crafted datasheets might, in the long run, reduce the amount of time that dataset creators need to spend answering one-off questions about their datasets, organizations can face hurdles when creating datasheets. For instance, details in a datasheet may result in exposure to legal or PR risks or the inadvertent release of proprietary information. Developers may delay releasing *any* datasheet—even a useful but incomplete one—in order to "perfect" it. Although the overhead in creating datasheets might be more costly for small organizations, they have an opportunity to differentiate themselves as more transparent than larger, more established, slower-to-change organizations. Publication venues can incentivize academics to release datasheets along with their datasets, while negative media attention for datasets without datasheets might drive companies to adopt the concept. Ultimately, we believe that the work involved in creating a dataset far exceeds the work involved in creating a datasheet. Moreover, a datasheet can dramatically improve the utility of a dataset for others and even mitigate potential harms.

Finally, it is important to note that machine learning datasets are rarely created "from the ground up" in a way that makes it possible to gather additional information about them. Instead, they are often scraped from some source, with no way of acquiring demographic information, consent, or other features. Some contextual information might still be available, however. For the Enron email dataset, for example, per-employee demographic information is not available, but demographic information about Enron as a whole is available. Ultimately, we see efforts aimed at more detailed annotation of datasets as a key step in strengthening the fairness, accountability, and transparency of machine learning systems.

6. Acknowledgements

We thank Peter Bailey, 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, Michael Hoffman, Eric Horvitz, Mingjing Huang, Surya Kallumadi, Ece Kamar, Krishnaram Kenthapadi, Emre Kiciman, Jacquelyn Krones, Lillian Lee, Jochen Leidner, Rob Mauceri, Brian Mcfee, Erik Learned-Miller, Bogdan Micu, Margaret Mitchell, Brendan O'Connor, Thomas Padilla, Bo Pang, Anjali Parikh, Alessandro Perina, Michael Philips, Barton Place, Sudha Rao, David Van Riper, Cynthia Rudin, Ben Schneiderman, Biplav Srivastava, Ankur Teredesai, Rachel Thomas, Martin Tomko, Panagiotis Tziachris, Meredith Whittaker, Hans Wolters, and Lu Zhang for valuable discussion and feedback.

References

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

Anonymous. Data statements for NLP: Toward mitigating system bias and enabling better science. https://openreview.net/forum?id=By4oPeX9f, 2018.

Bhardwaj, Anant, Bhattacherjee, Souvik, Chavan, Amit, Deshpande, Amol, Elmore, Aaron J, Madden, Samuel, and Parameswaran, Aditya G. DataHub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798*, 2014.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y, Saligrama, Venkatesh, and Kalai, Adam T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

Bose, Dipan, Segui-Gomez, Maria, and Crandall, Jeff R. Vulnerability of female drivers involved in motor vehicle crashes: An analysis of US population at risk. *American Journal of Public Health*, 2011.

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

Buolamwini, Joy and Gebru, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 77–91, 2018.

Caliskan, Aylin, Bryson, Joanna J, and Narayanan, Arvind. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Canis, Bill. Issues with federal motor vehicle safety standards, 2017. [Online; accessed 18-March-2018].

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

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

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

Department of Transportation Federal Highway Administration. Year of first state driver license law and first driver examination, 1997. [Online; accessed 18-March-2018].

Di, Liping, Shao, Yuanzheng, and Kang, Lingjun. 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): 5082–5089, 2013.

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

European Parliament and the Council of the EU. Directive 2002/95/EC of the European Parliament and of the Council of 27 January 2003 on the restriction of the use of certain hazardous substances in electrical and electronic equipment, 2003.

Faden, Ruth R, Lederer, Susan E, and Moreno, Jonathan D. 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):1667–1671, 1996.

Fast.ai. Fast.ai. http://www.fast.ai/, 2017.

Food and Drug Administration. Content and format of a new drug application (21 CFR 314.50 (d)(5)(v)), 1985.

Food and Drug Administration. Guidance for the study of drugs likely to be used in the elderly, 1989.

Food and Drug Administration. FDA clinical trials guidance documents, 2018.

Garvie, Clare, Bedoya, Alvaro, and Frankle, Jonathan. *The Perpetual Line-Up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

Gunter, Tracy D and Terry, Nicolas P. 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.

Hingson, Ralph, Howland, Jonathan, and Levenson, Suzette. Effects of legislative reform to reduce drunken driving and alcohol-related traffic fatalities. *Public Health Reports*, 1988.

Holland, Sarah, Hosny, Ahmed, Newman, Sarah, Joseph, Joshua, and Chmielinski, Kasia. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

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

International Electrotechnical Commission. About the IEC: Overview, 2017.

Lewis v. Amorous. 3 Ga.App. 50, 59 S.E. 338, 1907. [Online; accessed 18-March-2018].

Lin, Tom CW. The new investor. *UCLA Law Review*, 60:678, 2012.

Liu, Katherine A and Dipietro Mager, Natalie A. Women's involvement in clinical trials: Historical perspective and future implications. *Pharmacy Practice (Granada)*, 14(1), 2016.

Manduca, CA, Fox, S, and Rissler, H. Datasheets: Making geoscience data easier to find and use. In *AGU Fall Meeting Abstracts*, 2006.

Mann, G and O'Neil, C. Hiring algorithms are not neutral. *Harvard Business Review*, 2016.

McShane, Clay. *The Automobile*. Routledge, 2018.

Moreno, Jonathan D. *Undue Risk: Secret State Experiments on Humans*. Routledge, 2013.

National Electrical Manufacturers Association. Digital imaging and communications in medicine. https://www.dicomstandard.org/, 2018.

National Highway Traffic Safety Administration. Final Regulatory Evaluation, Amendment to Federal Motor Vehicle Safety Standards 208, 2006. [Online; accessed 18-March-2018].

National Institutes of Health. NIH sharing policies and related guidance on NIH-funded research resources, 2018.

National Library of Medicine. National library of medicine. https://www.nlm.nih.gov/, 2018.

Nolan, Martha R and Nguyen, Thuy-Linh. Analysis and reporting of sex differences in phase III medical device clinical trials—how are we doing? *Journal of Women's Health*, 22(5):399–401, 2013.

O'Connor, Mary Catherine. How AI could smarten up our water system, 2017. [Online; accessed 14-March-2018].

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

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

Passive Components, XICON. Miniature aluminum electrolytic capacitors: XRL series, 2005. [Online; accessed 14-March-2018].

Peltzman, Sam. The effects of automobile safety regulation. *Journal of Political Economy*, 1975.

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

SDMX. Statistical data and metadata exchange. https://sdmx.org/?page_id=3425, 2018.

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

Systems, Doha Suppy. Facial recognition, 2017. [Online; accessed 14-March-2018].

World Economic Forum Global Future Council on Human Rights 2016–2018. How to prevent discriminatory outcomes in machine learning. https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning, 2018.

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

Any other comments?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

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 the latter two, were they validated/verified and if so how?

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Are there any known errors, sources of noise, or redundancies in the data?

Any other comments?

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

Was the "raw" data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)

Is the preprocessing software available?

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

Any other comments?

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

What license (if any) is it distributed under? Are there any copyrights on the data?

Are there any fees or access/export restrictions?

Any other comments?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

If the dataset becomes obsolete how will this be communicated?

Is there a repository to link to any/all papers/systems that use this dataset?

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Any other comments?

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

Does the dataset contain information that might be considered inappropriate or offensive?

Any other comments?

Prototypes of Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

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]

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

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.[2]

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

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

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

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".

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

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.

How many instances of each type are there?

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.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format. Each image is accompanied by a label indicating the name of the person in the image. While subpopulation data was not available at the initial release of the dataset, a subsequent paper[3] reports the distribution of images by age, race and gender. Table 2 lists these results.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

Everything is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

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:

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

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

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

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.

[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/.

[2] 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

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

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 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 particular test result.

What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

The dataset was originally released without reported experimental results but many experiments have been run on it since then.

Any other comments?

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).

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[4]. The images in this database were gathered from news articles on the web using software to crawl news articles.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Unknown

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

Unknown

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

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 the latter two, were they validated/verified and if so 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. Some people could have given incorrect names particularly if the original caption was incorrect.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

The dataset does not contain all possible instances.

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

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).

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Unknown

Are there any known errors, sources of noise, or redundancies in the data?

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

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**[5] 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

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

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 data? (e.g., to support unanticipated future uses)
The raw unprocessed data (consisting of images of faces and names of the corresponding people in the images) is saved.

Is the preprocessing software available?
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.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
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 White males
- 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

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)
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 released/first distributed? (Is there a canonical paper/reference for this dataset?)
The dataset was released in October, 2007.

What license (if any) is it distributed under? Are there any copyrights on the data?
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.

What license (if any) is it distributed under? Are there any copyrights on the data?

Are there any fees or access/export restrictions?
There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?
The dataset is hosted at the University of Massachusetts and all and comments can be sent to: Gary Huang - gbhuang@cs.umass.edu.

Will the dataset be updated? How often and by whom? How will updates/revisions be documented (e.g., mailing list, GitHub)? Is there an erratum?
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

If the dataset becomes obsolete how will this be communicated?
All changes to the dataset will be announced through the LFW mailing list.

Is there a repository to link to any/all papers/systems that use this dataset?
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

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?
Unknown

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)
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.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)
Not applicable

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)
Unknown

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?
No (see first question).

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?
There is minimal risk for harm: the data was already public.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?
Unknown

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?
No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?
The dataset does not comply with GDPR because subjects were not asked for their consent.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)
The dataset does not contain confidential information since all information was scraped from news stories.

Does the dataset contain information that might be considered inappropriate or offensive?
No. The dataset only consists of faces and associated names.

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

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

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]

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

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.

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

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.

Who funded the creation of the dataset? If there is an associated grant, provide the grant 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.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

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.

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

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

How many instances of each type are there?

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

[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)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

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 everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

Everything is included.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

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

What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

Several experiments are reported in the README for baselines on this data, both on the original dataset (Table 1) and the cleaned version (Table 2). In these results, NB=Naive Bayes, ME=Maximum Entropy and SVM=Support Vector Machine. The feature sets include unigrams (with and without counts), bigrams, part of speech features, and adjectives-only.

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

The data was collected by downloading reviews from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, at http://reviews.imdb.com/Reviews.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Features	corrected NB	in paper NB	in paper ME	in paper SVM
unigrams (freq.)	79.0	78.7	n/a	72.8
unigrams	81.5	81.0	80.4	82.9
unigrams+bigrams	80.5	80.6	80.8	82.7
bigrams	77.3	77.3	77.4	77.1
unigrams+POS	81.5	81.5	80.4	81.9
adjectives	76.8	77.0	77.7	75.1
top 2633 unigrams	80.2	80.3	81.0	81.4
unigrams+position	80.8	81.0	80.1	81.6

Table 1. Results on the original dataset (first column is after data repair specified in the erratum, later).

Features	# features	NB	ME	SVM
unigrams (freq.)	16162	79.0	n/a	73.0
unigrams	16162	81.0	80.2	82.9
unigrams+bigrams	32324	80.7	80.7	82.8
bigrams	16162	77.3	77.5	76.5
unigrams+POS	16688	81.3	80.3	82.0
adjectives	2631	76.6	77.6	75.3
top 2631 unigrams	2631	80.9	81.3	81.2
unigrams+position	22407	80.8	79.8	81.8

Table 2. Results on the cleaned dataset (first column is the number of unique features).

Unknown

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

Unknown

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 the latter two, were they validated/verified and if so how?

The data was mostly observable as raw text, except the labels were extracted by the process described below.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

The dataset is a sample of instances.

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

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.

Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

No data is missing.

Are there any known errors, sources of noise, or redundancies in the data?

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

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 data? (e.g., to support unanticipated future uses)

Yes.

Is the preprocessing software available?

No.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

The overarching goal of this dataset is to study the task of sentiment analysis. From this perspective, the current dataset represents a highly biased sample of all texts that express affect. In particular: the genre is movie reviews (as opposed to other affective texts), the reviews are all in English, they are all from the IMDb archive of the `rec.arts.movies.reviews` newsgroup, and all from a limited time frame. As mentioned above, at most forty reviews were retained per author to ensure better coverage of authors. Due to all these sampling biases, it is unclear whether models trained on this dataset should be expected to generalize to other review domains (e.g., books, hotels, etc.) or to domains where affect may be present but where affect is not the main point of the text (e.g., personal emails).

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

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 released/first distributed? (Is there a canonical paper/reference for this dataset?)

The dataset was first released in 2002.

What license (if any) is it distributed under? Are there any copyrights on the data?

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.

Are there any fees or access/export restrictions?

No.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

Bo Pang is supporting/maintaining the dataset.

Will the dataset be updated? How often and by whom? How will updates/revisions be documented (e.g., mailing list, GitHub)? Is there an erratum?

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.)

If the dataset becomes obsolete how will this be communicated?

This will be posted on the dataset webpage.

Is there a repository to link to any/all papers/systems that use this dataset?

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

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

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

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

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.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

Not applicable.

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

Unknown

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

No (see first question).

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

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

If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?

Unknown

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

No; however, while most names have been removed from the preprocessed/tokenized versions of the data, the original data includes names and email addresses, which were also present on the IMDb archive.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

The preprocessed dataset may comply with GDPR; the raw data does not because it contains personally identifying information.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

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

Does the dataset contain information that might be considered inappropriate or offensive?

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