Working Draft: This is work in progress!

Datasheets for Datasets*

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

[1]Microsoft Research: {timnit.gebru, hal3, jenn, wallach, kate}@microsoft.com
[2]Georgia Tech: jamiemmt.cs@gatech.edu
[3]briana.vecchione@gmail.com
[4]University of Maryland
[5]AI Now Institute

March 23, 2018

Abstract

Currently there is no standard way to identify how a dataset was created, and what characteristics, motivations, and potential skews it represents. To begin to address this issue, we propose the concept of a datasheet for datasets, a short document to accompany public datasets, commercial APIs, and pretrained models. The goal of this proposal is to enable better communication between dataset creators and users, and help the AI community move toward greater transparency and accountability. By analogy, in computer hardware, it has become industry standard to accompany everything from the simplest components (e.g., resistors), to the most complex microprocessor chips, with datasheets detailing standard operating characteristics, test results, recommended usage, and other information. We outline some of the questions a datasheet for datasets should answer. These questions focus on when, where, and how the training data was gathered, its recommended use cases, and, in the case of human-centric datasets, information regarding the subjects' demographics and consent as applicable. We develop prototypes of datasheets for two well-known datasets: Labeled Faces in The Wild [33] and the Pang & Lee Polarity Dataset [45].

1 Introduction

Artificial Intelligence is rapidly moving from a purely academic discipline to a technology that is embedded in everyday products. Our cars, homes, and computers are regularly controlled by machine learning algorithms. These algorithms are also used by state and federal agencies and corporations to predict our behavior: law enforcement uses facial recognition software to catch suspects [30, 52], the US criminal justice system uses risk assessment scores to estimate a person's likelihood of committing a crime [5], and companies use models to filter job applications before requesting interviews [37, 32]. Critical components of our world's infrastructure rely on machine learning models, for example to

*Working Draft: This is work in progress, so writing may be less polished than usual and is subject to revision. New versions forthcoming. We'd love to incorporate your feedback, so if you have comments for us, please let us know!

monitor and manage our water systems [41] and power grids [11]. Since 2011, a majority of financial trades made in US on our stock exchanges are made automatically [35].

By definition, all such machine learning models are trained from some data: indeed, the datasets on which they are trained are an integral in determining their functionality. The ubiquity with which data is gathered and models are trained and tested on that data allows for incredible innovation, but also poses a number of risks and questions: Under what circumstances do we expect such models to perform well? In different circumstances, how will the model's performance degrade? Is this dataset fixed, or is more data gathered over time? Is it possible to observe and ameliorate any bias perpetuated and exacerbated by these models when making choices affecting people? All of these questions depend upon how the data was gathered, cleaned and processed, and how a model incorporates that data. Without this information, even experts trained in machine learning cannot hope to say with any certainty when a model might fail badly. Moreover, there is currently no standard practice for carefully testing or describing these datasets, APIs, or freely available models. This is of particular concern when we consider high-stakes uses of these models or datasets.

In this paper, we argue that the first step towards transparency must involve attaching significant, standardized information about any dataset, API, or pretrained model. We concentrate specifically on datasets, and recommend a standardized description format with standard questions. Even if the eventual goal is to understand pretrained models or APIs, understanding datasets is an important first step: regardless of how they are trained, models and APIs are typically evaluated against fixed datasets, and understanding the characteristics of that evaluation is of paramount importance. We draw inspiration from standardized forms of sharing information and rigorous testing conducted in the much more mature field of hardware. This document, called henceforth a **datasheet**, would be a short (typically 3-5 page) document detailing any tests that have been conducted on a dataset, its recommended usage, its collection procedure, any regulations governing its use, and so on. A datasheet for a pretrained model would contain information pertaining to the model's training data and tested behavior.

We structure this paper as follows. Section 2 discusses the motivation and context for our work, Section 3 briefly discusses the evolution of safety standards in other industries to draw a comparison with AI. Section 4 discusses the concept of datasheets in hardware, and Section 5 outlines the most important questions that should be answered by a datasheet. Two prototypes of datasheets for datasets can be found in Appendix B. The paper ends with a discussion of challenges and future work in Section 6.

2 Context

Many datasets have little or no documentation describing the settings under which they were gathered, the characteristics of the training sets, how representative the dataset is of a particular population, or recommended usage. This lack of information makes it difficult to assess what contexts models and APIs trained on these datasets can be used for and what scenarios should be avoided. Most of these APIs are not accompanied by detailed information describing recommended usage, standard operating characteristics, and tests performed to verify these conditions.

Of particular concern is the recently discovered extent to which AI systems exhibit and amplify biases. Buolamwini and Gebru [8] showed that commercial gender classification APIs have near perfect performance for light skinned males while error rates for darker skinned females can be as high as 33%.[1] In a different setting, Bolukbasi et al. [6] showed that word embeddings trained on news articles exhibit gender bias, finishing the analogy "Man is to computer programmer as woman is to X" with "home maker," a stereotypical role for women. Caliskan et al. [9] showed that racial biases also exist in word embeddings: traditional European-American names, for example, are more closely related to words that are considered pleasant (e.g., joy); those associated with African-American names however, are closer to words such as agony. These biases can have dire consequences that might not be easily discovered. For example, Bolukbasi et al. [6] argue that bias in word embeddings can result in hiring discrimination. Much like a faulty resistor or a capacitor in a circuit, the effects of a biased AI-based component can propagate throughout a system making them difficult to track down.

One of the biggest challenges in the deployment of AI technology is deploying systems, built using datasets or pretrained models, in unsuitable environments. Such systems can exhibit poor or unpredictable performance when deployed: the models' behavior on some benchmark may say very little about their behavior in a different setting. Such unpredictability is of concern when these systems make high-stake real-time decisions, such as controlling a large amount of trading volume on the stock exchanges or allocating valuable resources to different markets. Unpredictability is of even greater concern when the system is used to interact with or make decisions directly influencing humans.

This problem of unintentional misuse of datasets is often exacerbated when the users are not domain experts. We believe that this problem can be mitigated, at least partially, by accompanying datasets with datasheets that describe their creation, strengths and limitations. While this is not the same as making everyone a domain expert, it gives an opportunity for domain experts to easily communicate what they know about the dataset and its limitations to whomever may be using it. This is particularly important today, when companies such as Google, Amazon, Microsoft, and Facebook are moving toward "democratizing AI" by creating toolboxes such as Cloud AutoML that can be trained by those with little-to-no machine learning expertise and domain knowledge [12]. Pre-trained models and standard datasets are freely available for use with open source tools such as Caffe [22], Tensorflow [2], and PyTorch [48]. As powerful machine learning tools become available to a much broader set of users, it becomes even more important to enable those users to understand all the implications of their work.

Part of the educational gap, outside traditional "ivory tower" education, is being covered by organizations like Coursera [20], Udacity [53], Fast.ai [26], and others, which offer online AI courses to those with no prior experience. A goal of these educational platforms is to enable people around the world to use AI to solve problems in their communities. 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" [26]. These readily available AI courses and toolkits belong to a movement whose laudable goal is to enable many people to integrate AI into their systems. Coupling this educational strategy, which often (in particular in the case of Fast.ai) includes explicit training in dataset bias and ethics (topics sometimes lacking even in the "ivory tower"), with datasheets that can explain the context and biases in existing datasets, can much more quickly enable progress by both domain experts and AI experts.

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

3 The Evolution of Safety Standards in Other Industries

We briefly discuss the evolution of safety standards in 3 different industries: automobiles, health care and electronics. Understanding the dangers that were posed by the proliferation of new technology in these industries, and the safety measures that were put in place to combat them, can help us carve out a path forward for AI.

3.1 The Automobile Industry

Similar to our current hopes for AI to positively transform society, the introduction of automobiles promised to expand mobility and provide additional recreational, social, and economic opportunities. However, much like current AI technology, the automobile was introduced with few safety checks or regulations in place. When cars first became available in the United States, there were no speed limits, stop signs, traffic lights, driver education, or regulations pertaining to seat belts or drunk driving [10]. Thus, the early 1900s saw many deaths and injuries due to collisions, speeding, and reckless driving [31]. Much like current debates regarding the future of AI, there were courtroom discussions and news paper editorials outlining the possibility that the automobile was inherently evil [1].

It took many years in different parts of the world for laws, traffic signals, driver education and fines to be enacted. In the United States, for instance, driver's licenses were not fully implemented across all states until 1954 [43]. By the 1930s, a variety of technical safety responses were introduced such as four-wheel hydraulic brakes, shatter-resistant windshields, and all-steel bodies [38]. Despite these safety standards, the automobile industry fell victim to similar "bad dataset" problems faced by AI technology, in particular in the construction of the crash-test dummies used to evaluate vehicle safety. Although crash-test dummies became a mandated part of U.S. safety standards in 1973, almost all the crash-test dummies in use were modeled after prototypical male physiology [4]. It wasn't until almost four decades later that, in 2011, using "female" crash-test dummies became mandatory for frontal crash tests [3]. Subsequent studies suggest that male-centric engineering design is responsible for disparate rates of vehicle injuries by sex: A safety study of automobiles manufactured between 1998 and 2008 concluded that women wearing seat belts were 47% more likely to be seriously injured than males in similar accidents [7].

Automobile safety standards in the US have continued to evolve since their introduction in the early 1920s, and there are many countries without adequate road safety measures. Road accidents are still the biggest global killer of teenagers [44]. In the case of seat belts in particular, it was only in 1968 that features like padded dashboards and seat belts were made mandatory in the United States [49]. Still, most motorists were reluctant to use seat belts, and safety campaigns had to be sponsored to promote adoption. By analogy, an "AI solution" is likely to require both laws and regulations (in particular in high-stakes environments) and also social campaigns to promote best practices. This underscores the need to aggressively work towards standardization of best practices for the creation and proliferation of datasets and APIs in AI, and have a mechanism by which these practices can continue to evolve.

3.2 Clinical Trials in Medicine

We can draw lessons from studying the evolution of safety measures for clinical trials, and some of the harms that were caused by inadequate standards and unethical procedures. Like the need for large scale data collection and experimentation before the deployment of an AI system, clinical trials are an important step in any drug development. However, the US legal system viewed clinical trials as a form of medical malpractice until well into the 20th century, making standardized large-scale evaluations difficult [23]. After the acceptance of certain clinical trials came various standards which were to be followed, most of which were spurred by some atrocity or another committed in the name of science. For example, the US government ran a number of experiments on its citizens without their consent, from a public health study on syphilis where participants were not informed of their disease [21], to radiation experiments [25, 39]. The poor, the imprisoned, minority groups, pregnant women, and children comprised a majority of these study groups.

Currently, in the US, participants in a drug trial must be informed that the drug is experimental and not proven to be effective, and subjects must be willing participants. Prior to a drug being tested in a clinical trial, an Institutional Review Board and the Food and Drug Administration must approve an application which shows evidence of the drug's relative safety (in terms of basic chemistry and animal testing results) and lays out the design of the trial (including who will participate in the trial) [29].

The closest legal analog in AI are laws like the European Union's General Data Protection Regulation (GDPR), which are starting to be deployed to ensure that people consent to having their data used in the training of AI based models [47]. Data collection standards are now a central topic of concern for scientific research broadly, and clinical trials are no exception. The US National Institute of Health has a body of guidelines for how their funded projects are to gather, store, and share human subject data [42].

Finally, the lack of diversity in clinical trial participants has led to the development of drugs that do not work well for many groups of people. For example, eight out of ten drugs pulled from circulation between 1997 and 2001 had more adverse effects for women, suggesting clinical trials without representative samples did not accurately display risks for those drugs for women [36]. As late as 2013, a majority of federally-funded clinical trials still did not break down their results by sex [40]. In 2014, the FDA promoted an action plan to make results of clinical trials broken down by subpopulation more easily available [27]. In the late 1980s, the FDA moved to require different age groups participate in clinical trials [28]. Not until 1998 was a regulation stating that safety and efficacy data be provided broken down by sex, race, and age. These progressions parallel some recent results showing disparities in accuracy of various AI based models by subpopulation (e.g., [8]), and calls for more diverse datasets, inclusive testing, and standards in place to reduce these disparities.

3.3 Electrical and Electronic Technologies

Similar to the current proliferation of AI within everyday products, electrical and electronic components are used in devices that are incredibly widespread. They are designed into devices ranging from those used in communication (TV, radio, phones), transportation (automobiles, trains, planes), healthcare, military equipment, energy production, and transmission. With the move towards smart homes and the internet of things [54], soon one may be hard-pressed to find a synthetic object without electronic components.

Like datasets and the models trained on them, electronic components, such as resistors or capacitors, are designed into a system whose larger goal may be far removed from the task of that component. Thus, small deviations that may seem insignificant while studying a component in isolation can result in dire consequences due to its interactions with the rest of the system.

For example, while all types of resistors can be abstracted into an idealized mathematical model, different non-idealities are important depending on the context. The operating temperature range of a power resistor meant for operation under high voltage conditions is much more crucial than that for low power thin film resistors in a motherboard [55]. Even still, within the power resistor family, the safety considerations for those used in power plants for instance are different from those in consumer electronics [55]. Thus, the electronic component community has developed standards that specify ideal operation characteristics, tests and manufacturing conditions for components manufactured with different tasks in mind.

Many of these standards are specified by the International Electrotechnical Comission (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” [13]. After the International System of Electrical and Magnetic Units was agreed to at the first International Electrical Congress in 1881, it became clear that many other questions of standardization would arise [34]. The IEC was founded in 1906 and in 1938, it published an international vocabulary to unify terminology relating to electrical, electronic and related technologies [13, 17]. There are currently $9,000$ IEC standards in use today, with over 10 standards pertaining to different types of resistors alone [18]. For example IEC 60195:2016 describes the recommended procedures to assess the magnitude of current noise in fixed resistors of any type, IEC 60115-2:2014 provides standards for leaded fixed low-power film resistors for use in electronic equipment, and IEC 60322 states operating conditions and testing methodology for power resistors used in railway applications [14, 15, 16]. We argue that standards with similar detail and scope can be set in place for datasets and pre-trained AI models used in different scenarios.

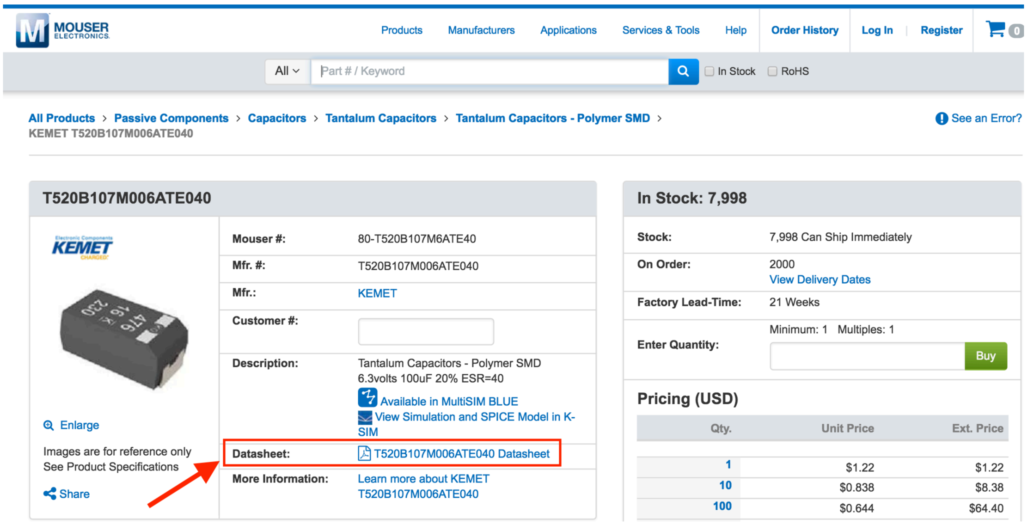
4 Datasheets for Electronic Components

We take our inspiration from the standardization of datasheets for electronic components. All electronic components, ranging from the cheapest and most ubiquitous resistors, to highly complex integrated circuits (like CPUs), are accompanied by datasheets characterizing their recommended operating conditions and other detailed technical characteristics. While the information contained in any given datasheet depends on the specific product, there are several aspects that are common to all datasheets (such as the manufacturer's name, and the product name and number).

Most datasheets start with a description of the component's function and features, and contain other specifications like the absolute maximum and minimum operating voltages. In addition to the technical characteristics, datasheets contain physical details of the component (such as size and pin connections), and list of available packages. Datasheets can also contain 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). If the component design, manufacturing and testing adheres to some standard (e.g., an IEC standard), this is typically also stated in the datasheet.

When one navigates to a product webpage, there is a datasheet associated with each product. Figure 4 shows a screenshot of the the product page for a KEMET Corporation tantalum capacitor [24]. The datasheet for this product (indicated by the red arrow), is prominently featured along with other documentation. An example of a datasheet for a miniature aluminum electrolytic capacitor is shown in Appendix A. Like many datasheets for similar components, this one contains:

- A short description of the component's function and notable features (including the component's compliance with the RoHS directive adopted by the European Union restricting the use of certain hazardous materials [46])



KEMET T520B107M006ATE040

T520B107M006ATE040

Mouser #:	80-T520B107M6ATE40
Mfr. #:	T520B107M006ATE040
Mfr.:	KEMET
Customer #:	
Description:	Tantalum Capacitors - Polymer SMD 6.3volts 100uF 20% ESR=40 Available in MultiSIM BLUE View Simulation and SPICE Model in K-SIM
Datasheet:	T520B107M006ATE040 Datasheet
More Information:	Learn more about KEMET T520B107M006ATE040

Images are for reference only
See Product Specifications

In Stock: 7,998

Stock:	7,998 Can Ship Immediately
On Order:	2000 View Delivery Dates
Factory Lead-Time:	21 Weeks
Enter Quantity:	Minimum: 1 Multiples: 1

Pricing (USD)

Qty.	Unit Price	Ext. Price
1	$1.22	$1.22
10	$0.838	$8.38
100	$0.644	$64.40

Figure 1: A screenshot of the product webpage for a KEMET Corporation tantalum capacitor with the datasheet for the component highlighted in red.

- Various standard operating characteristics, such as its operating temperature range and capacitance tolerance

- A diagram of the component showing its dimensions and pin connections

- Plots showing the change in various characteristics vs. time, temperature and frequency; for example, it is well known that capacitance decreases over time, and the first graph measures this change across 1000 hours

Some examples of other datasheets are those for semiconductors [51] (56 pages), resistors [50] (2 pages), and other components [19] (18 pages).

4.1 What has driven the use of datasheets in hardware?

To better understand how we might hope for proliferation of datasheets for datasets, it is useful to note some of the potential reasons for their standardization in hardware.

De facto industry standard. The “highest order bits” (one to two phrase summary describing a component, such as conductance or resistance) are nowhere near enough to understand in what settings a component should/could be used, how it will behave in a variety of settings, how robust it is, what size it is, and so forth. In order to make informed purchasing decisions, one needs to know additional information as provided in a datasheet. *In the AI setting, many dataset characteristics need to be outlined to understand their use cases, potential biases, and limitations.*

Product Liability. The seller of a hardware component wants to clearly outline the appropriate settings for using their component. If the component fails during some use and causes damage or injury, they could be liable if they have not clearly specified that type of use as outside the component’s operating characteristics. *In the AI setting, many failure modes are not as clearly visible to all, and the liability may not be easily traced back to datasets. This could make the adoption of semi-standardized datasheets more difficult.*

Market forces. A hardware builder would never buy a component that does not have a datasheet since they would have no way of knowing how it would operate. *In the AI setting, practitioners are training custom models with specific datasets without understanding the limitations of these datasets. The introduction of datasheets can encourage the AI community to perform a thorough analysis of dataset characteristics before applying them to specific settings.*

Understanding out-of-spec behavior. Any of the testing done on components can be described in the datasheet, and different tests are performed on different types of components. However, there are a set of minimal tests that are usually included in all datasheets. *In the AI setting, datasets containing the same types of instances (examples) meant for use in different settings will need different specifications. For example, a dataset of faces used to train a model recognizing people from around the world should contain a representative sample of people around the world.*

5 Datasheets for Datasets

In the context of artificial intelligence and data science, datasets play a central role in both training and evaluation. This is the case regardless of whether the dataset is used to build a predictor that will be deployed as part of a system, or used to ask scientific questions and reach scientific conclusions. In both cases, the specific properties of a dataset can have profound impact on the quality of a learned predictor, or the quality of scientific conclusions. Akin to how it is important to understand the operating characteristics of a resistor when "debugging" a microcontroller, it is also important to understand the specific properties of a dataset to understand how it fits into the larger data ecosystem.

Below we have proposed sample questions that a datasheet should arguably contain. The prototypes in the appendix of this paper are provided as examples of how these might be answered in practice. Several fundamental objectives drove our formation of these questions. First, a practitioner should be able to decide, from reading this datasheet, how appropriate this dataset is for their task, what its strengths and limitations are, and how it fits into the broader dataset ecosystem. Second, the creators of a dataset should be able to use the questions on a datasheet to help them think about aspects of data creation that may not have otherwise occurred to them. Third, users should be able to understand—based on the performance of a model or API on a dataset—what that performance measure actually means, and when to be comfortable using such models.

The set of questions we provide here is not intended to be definitive. Instead, we hope it will initiate a larger conversation about how data provenance, ethics, privacy, and documentation might be handled by the community of data curators. Below are our proposed questions, which include details about the gathering, cleaning, testing, and releasing of a dataset. Not all questions will be applicable to all datasets, in which case they can simply be left out.

Appendix B includes prototypes of datasheets for two datasets: Labeled Faces in The Wild [33] and the Pang & Lee Polarity Dataset [45]. (In the creation of these datasheets, sometimes information was unknown; this is marked in red text.)

Motivation for Dataset Creation

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

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

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?

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 are there? (of each type, if appropriate)?

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 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; c) are there access restrictions or fees?

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

What experiments were initially run on this dataset? Have a summary of those results.

Any other comments?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API)

Who was involved in the data collection process? (e.g., students, crowdworkers) and 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 of the instances?

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 a sample (not necessarily random) of instances from a larger set?

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?

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)

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? If not, what are the limitations?

Any other comments?

Dataset Distribution

How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

When will the dataset be released/first distributed?

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?

Will the dataset be updated? If so, how often and by whom?

How will updates be 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 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?

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?

6 Challenges and Future Work

Our proposal for a set of standardized datasheets faces several challenges in implementation; we outline the most pressing of these below and urge the machine learning community to make progress on these in future work. These challenges fall into several categories: how to converge on the format and content of the datasheet, the incentives required to encourage datasheet production and the need to overcome inertia, and the communication with outside experts that is necessary to properly address ethical and demographic considerations of datasets containing data about people.

As a community, we will need to come to some consensus about what should be included in a datasheet, and how that data can be most effectively solicited and communicated. Just as in the context of hardware datasheets (Section 4), the most relevant information regarding each dataset will likely be context-specific; just as hardware has different categories of components with differing relevant characteristics, datasets comprised of photographs of human faces will have different relevant documentation needs than datasets of health or weather records. We should not expect this consensus to come easily; researchers and practitioners who work in an individual domain might first want to agree upon on a small number of critical domain-specific attributes for their own datasets. However, there are also universal questions relevant to all datasets (e.g., who paid for the collection of the data, or whether there were human subjects generating the dataset), upon which the broader community should perhaps come to an agreement. This will likely be part of a larger conversation, both at a high level (datasheets for all types of datasets), as well as in a domain-specific manner. Along the lines of the former, a group at the MIT Media Lab recently publicized a "Data Nutrition Label" idea[2], which, at the time of writing, has similar goals to our proposal, though the details are not yet available. Along the lines of the latter, an anonymous paper was published contemporaneously with this paper in the natural language processing domain[3].

It is also unclear to what extent a datasheet should delve into ethical questions such as bias or privacy. Questions regarding ethical considerations should be framed in a manner that encourages practitioners to use ethical procedures to gather data, without discouraging them from providing as much information as possible about the process.

While this paper outlines questions that a datasheet for datasets would ideally answer, a similar endeavor needs to be undertaken for datasheets pertaining to models that have been pre-trained and their APIs. In particular, what are the important questions to ask about the behavior of these models, and how should these be measured and communicated, especially when the models are built based on multiple datasets, together with expert knowledge and other sources of input? Institutions that produce such models should iterate with customers and developers to arrive at the right set of questions and guidelines in a "datasheet for models" that would parallel our proposal for datasets.

There will be overhead in creating datasheets, some of which we can mitigate by carefully designing an interactive survey that would automatically produce a datasheet based on answers to questions. Moreover, hopefully a carefully crafted datasheet will, in the long run, reduce the amount of time the dataset creators will need to spend answering one-off questions about their data. Both large and small organizations will face hurdles in producing these datasheets. For instance, extra details in a datasheet may result in an organization being exposed to legal or PR risks, or such details might contain proprietary information which give the organization a competitive edge. Organizations may also delay releasing *any* datasheet—even an imperfect one—in order to "complete" it. Small organizations might consider the overhead in preparing a datasheet more onerous than large organizations. On the other hand, datasheets also provide an opportunity for smaller organizations to differentiate themselves as more transparent than larger, more established players. Ultimately, we believe the work involved in collecting and preparing a model or dataset for public use far exceeds the cost of creating a datasheet, which can improve the usefulness of this work for other users.

Finally, a large chunk of the work which remains in implementing datasheets for machine learning systems will be to communicate with experts in other areas. One example of this need comes from considering demographic information for datasets related to people (how the data is collected, collated, and analyzed). Other fields (such as anthropology) are well-versed in the difficulties arising from demography, and we should avail ourselves of that resource. It will also be important to consider that datasets are rarely gathered "from the ground up" in such a way that it would be theoretically possible to gather all sorts of additional information about the elements of the dataset. Instead, in many settings datasets are scraped from some source without the ability to gather additional features,

[2] See `http://datanutrition.media.mit.edu/`.

[3] See `https://openreview.net/forum?id=By4oPeX9f`.

demographic information, or consent. Some contextual information might still be available (e.g., for the Enron email dataset, we might not have demographic information on a per-employee basis, but some demographic information about the employees of the company as a whole may be available). Again, other industries have considered these difficulties (as discussed in Section 3) and we should learn from their best practices.

7 Acknowledgements

We would like to thank Meredith Whittaker, Erik Learned-Miller, Yoshua Bengio, Joy Buolamwini, Margaret Mitchell, Miroslav Dudík, and Rachel Thomas for valuable discussions and feedback.

References

[1] Lewis v. amorous. `https://groups.google.com/forum/#!topic/alt.lawyers/I-sU32WypvQ`, 1907. [3 Ga.App. 50, 59 S.E. 338 (1907). Online accessed 18-March-2018].

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[3] National Highway Traffic Safety Administration. Final regulatory evaluation, amendment to federal motor vehicle safety standards 208. `https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/dummy_milestones_812189.pdf`, 2006. [Online; accessed 18-March-2018].

[4] National Highway Traffic Safety Administration. Milestones for nhtsa's crash test dummies. `https://www.gpo.gov/fdsys/pkg/FR-2006-08-31/pdf/06-7225.pdf`, 2015. [Online; accessed 18-March-2018].

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

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

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

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

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

[10] Bill Canis. Issues with federal motor vehicle safety standards. `https://fas.org/sgp/crs/misc/R44800.pdf`, 2017. [Online; accessed 18-March-2018].

[11] Glennda Chui. Project will use ai to prevent or minimize electric grid failures. `https://phys.org/news/2017-09-ai-minimize-electric-grid-failures.html`, 2017. [Online; accessed 14-March-2018].

[12] Google Cloud. Cloud automl. `https://cloud.google.com/automl/`, 2018. [Online; accessed 14-March-2018].

[13] International Electrotechnical Commission. About the iec: Overview. https://basecamp.iec.ch/download/welcome-to-the-iec/, 2017.

[14] International Electrotechnical Commission. Iec 60195:2016. https://webstore.iec.ch/publication/24478, 2018.

[15] International Electrotechnical Commission. Iec 60322:2001. https://webstore.iec.ch/publication/768, 2018.

[16] International Electrotechnical Commission. Iec 60322:2001. https://webstore.iec.ch/publication/1462, 2018.

[17] International Electrotechnical Commission. Welcome to the iec international electrotechnical commission. http://www.iec.ch/about/history/overview/, 2018.

[18] International Electrotechnical Commission. Iec webstore. https://webstore.iec.ch/searchform&q=resistor, 2018.

[19] KEMET Electronic Components. Surface mount multilayer ceramic chip capacitors (smd mlccs. http://www.mouser.com/ds/2/212/KEM_C1035_COG_PULSE_SMD-1103961.pdf, 2018.

[20] Coursera. Coursera. http://www.coursera.org/, 2017.

[21] William J Curran. The tuskegee syphilis study, 1973.

[22] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

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

[24] Mouser Electronics. Kemet t520b107m006ate040. https://www.mouser.com/ProductDetail/KEMET/T520B107M006ATE040/?qs=A%2FUypIgi4aDLVIJ%2FABrGAw==, 2018.

[25] Ruth R Faden, Susan E Lederer, and Jonathan D Moreno. 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.

[26] Fast.ai. Fast.ai. http://www.fast.ai/, 2017.

[27] Food and Drug Administration. Content and format of a new drug application (21 cfr 314.50 (d)(5)(v)). https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/WomensHealthResearch/UCM557761.pdf, 1985.

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

[29] Food and Drug Administration. Fda clinical trials guidance documents. https://www.fda.gov/RegulatoryInformation/Guidances/ucm122046.htm, 2018.

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

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

[32] HireVue. Hirevue. https://www.hirevue.com/, 2018.

[33] 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. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[34] Randall K Kirschman. *High temperature electronics.* IEEE Press, 1999.

[35] Tom CW Lin. The new investor. *UCLA L. Rev.*, 60:678, 2012.

[36] Katherine A Liu and Natalie A Dipietro Mager. Women's involvement in clinical trials: historical perspective and future implications. *Pharmacy Practice (Granada)*, 14(1):0–0, 2016.

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

[38] Clay McShane. *The Automobile.* Routledge, 2018.

[39] Jonathan D Moreno. *Undue risk: secret state experiments on humans.* Routledge, 2013.

[40] Martha R Nolan and Thuy-Linh Nguyen. 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.

[41] Mary Catherine O'Connor. How ai could smarten up our water system. `https://medium.com/s/ai-for-good/how-ai-could-smarten-up-our-water-system-f965b87f355a`, 2017. [Online; accessed 14-March-2018].

[42] National Institute of Health. Nih sharing policies and related guidance on nih-funded research resources. `https://grants.nih.gov/policy/sharing.htm`, 2018.

[43] U.S. Department of Transportation Federal Highway Administration. Year of first state driver license law and first driver examination. `https://www.fhwa.dot.gov/ohim/summary95/dl230.pdf`, 1997. [Online; accessed 18-March-2018].

[44] World Health Organization. Global status report on road safety 2015. `Fillin`, 2015.

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

[46] The European Parliament and the Council of the European Union. 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. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:037:0019:0023:EN:PDF`, 2003.

[47] THE EUROPEAN PARLIAMENT and THE COUNCIL OF THE EUROPEAN UNION. General data protection regulation - european council. `http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf`, 2016.

[48] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch, 2017.

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

[50] Arcol Resistors. Hs aluminium housed resistors. `http://www.arcolresistors.com/wp-content/uploads/2014/03/HS-Datasheet.pdf`, 2008.

[51] Freescale Semiconductor. Mac7100 microcontroller family hardware specifications. `https://www.nxp.com/docs/en/data-sheet/MAC7100EC.pdf`, 2006.

[52] Doha Suppy Systems. Facial recognition. `http://doha.co.za/facialrecognition.html`, 2017. [Online; accessed 14-March-2018].

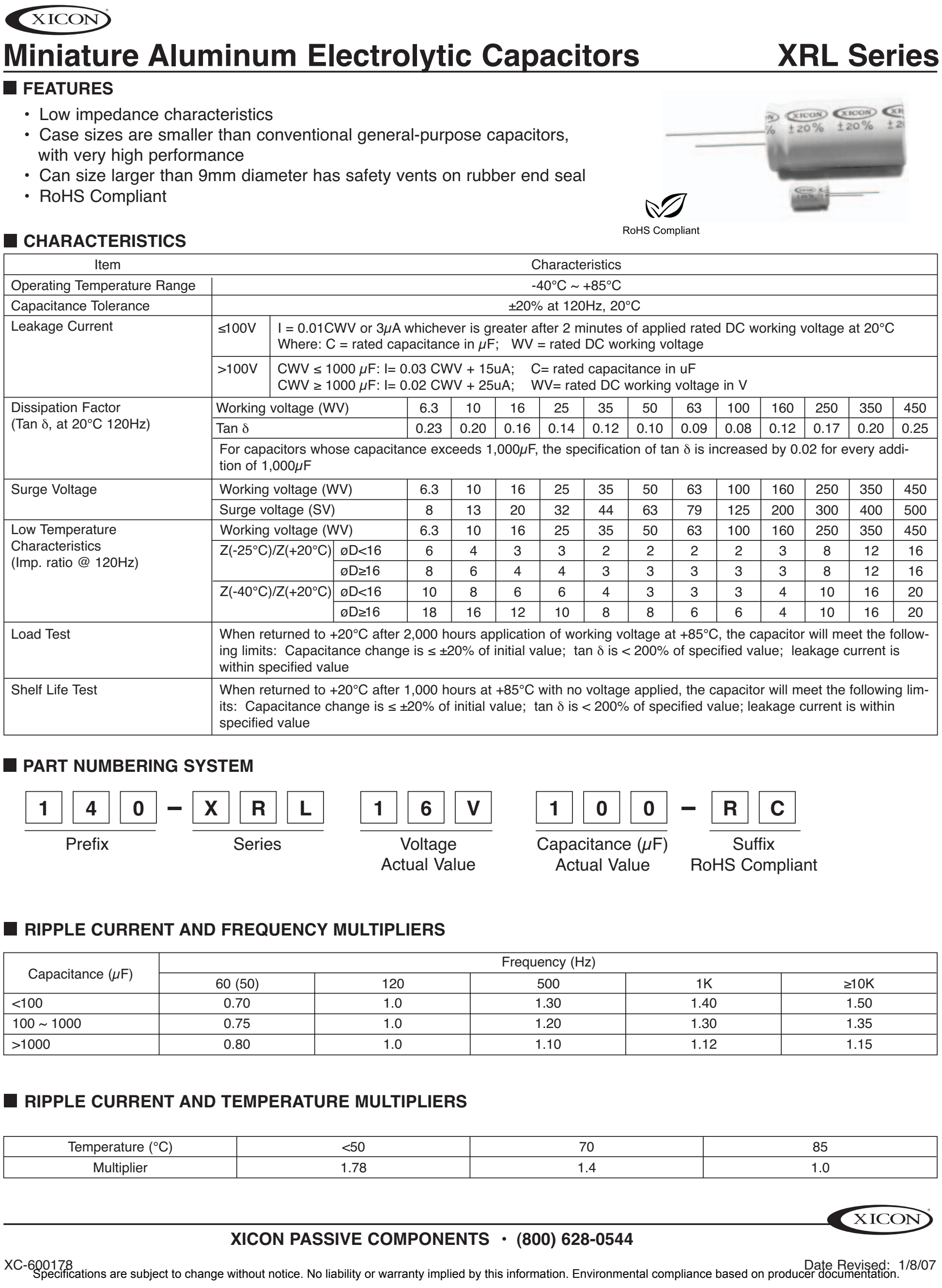
[53] Udacity. Udacity. `http://www.udacity.org/`, 2017.

[54] Rolf H Weber and Romana Weber. *Internet of things*, volume 12. Springer, 2010.

[55] Frank A Wolff. The so-called international electrical units. *Journal of the Institution of Electrical Engineers*, 34(170):190–207, 1905.

A An Example of a Datasheet for a Hardware Component

‘



Miniature Aluminum Electrolytic Capacitors XRL Series

FEATURES

- Low impedance characteristics
- Case sizes are smaller than conventional general-purpose capacitors, with very high performance
- Can size larger than 9mm diameter has safety vents on rubber end seal
- RoHS Compliant

RoHS Compliant

CHARACTERISTICS

Item	Characteristics													
Operating Temperature Range	-40°C ~ +85°C													
Capacitance Tolerance	±20% at 120Hz, 20°C													
Leakage Current	≤100V	I = 0.01CWV or 3μA whichever is greater after 2 minutes of applied rated DC working voltage at 20°C Where: C = rated capacitance in μF; WV = rated DC working voltage												
	>100V	CWV ≤ 1000 μF: I= 0.03 CWV + 15uA; C= rated capacitance in uF CWV ≥ 1000 μF: I= 0.02 CWV + 25uA; WV= rated DC working voltage in V												
Dissipation Factor (Tan δ, at 20°C 120Hz)	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Tan δ		0.23	0.20	0.16	0.14	0.12	0.10	0.09	0.08	0.12	0.17	0.20	0.25
	For capacitors whose capacitance exceeds 1,000μF, the specification of tan δ is increased by 0.02 for every addition of 1,000μF													
Surge Voltage	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Surge voltage (SV)		8	13	20	32	44	63	79	125	200	300	400	500
Low Temperature Characteristics (Imp. ratio @ 120Hz)	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Z(-25°C)/Z(+20°C)	øD<16	6	4	3	3	2	2	2	2	3	8	12	16
		øD≥16	8	6	4	4	3	3	3	3	3	8	12	16
	Z(-40°C)/Z(+20°C)	øD<16	10	8	6	6	4	3	3	3	4	10	16	20
		øD≥16	18	16	12	10	8	8	6	6	4	10	16	20
Load Test	When returned to +20°C after 2,000 hours application of working voltage at +85°C, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value													
Shelf Life Test	When returned to +20°C after 1,000 hours at +85°C with no voltage applied, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value													

PART NUMBERING SYSTEM

1 4 0 –	X R L	1 6 V	1 0 0 –	R C
Prefix	Series	Voltage Actual Value	Capacitance (μF) Actual Value	Suffix RoHS Compliant

RIPPLE CURRENT AND FREQUENCY MULTIPLIERS

Capacitance (μF)	Frequency (Hz)				
	60 (50)	120	500	1K	≥10K
<100	0.70	1.0	1.30	1.40	1.50
100 ~ 1000	0.75	1.0	1.20	1.30	1.35
>1000	0.80	1.0	1.10	1.12	1.15

RIPPLE CURRENT AND TEMPERATURE MULTIPLIERS

Temperature (°C)	<50	70	85
Multiplier	1.78	1.4	1.0

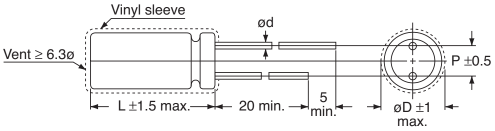
XICON PASSIVE COMPONENTS • (800) 628-0544

XC-600178 Date Revised: 1/8/07

Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

Miniature Aluminum Electrolytic Capacitors XRL Series

DIMENSIONS AND PERMISSIBLE RIPPLE CURRENT



Lead Spacing and Diameter (mm)

øD	5	6.3	8	10	13	16	18	22	25
P	2.0	2.5	3.5	5.0	5.0	7.5	7.5	10	12.5
ød	0.5	0.5	0.6	0.6	0.6	0.8	0.8	1.0	1.0

Tape and box is 5.0mm lead space.

Value (μF)	Working Voltage (WV); Dimensions: øD x L (mm); Ripple Current: mA/RMS @ 120Hz, 85°C													
	10		16		25		35		50		63		100	
	øD x L	mA	øD x L	mA	øD x L	mA	øD x L	mA	øD x L	mA	øD x L	mA	øD x L	mA
.10									5 x 11	1.5	5 x 11	3.0	5 x 11	3.0
.22									5 x 11	3.5	5 x 11	4.5	5 x 11	5.8
.33									5 x 11	5.0	5 x 11	7.5	5 x 11	8.8
.47					5 x 11	6.0			5 x 11	7.0	5 x 11	9.5	5 x 11	12
1.0					5 x 11	10			5 x 11	15	5 x 11	17	5 x 11	22
2.2			5 x 11	20	5 x 11	16			5 x 11	29	5 x 11	28	5 x 11	33
3.3			5 x 11	30	5 x 11	25			5 x 11	35	5 x 11	34	5 x 11	40
4.7			5 x 11	41	5 x 11	31	5 x 11	40	5 x 11	42	5 x 11	45	5 x 11	48
10	5 x 11	54	5 x 11	49	5 x 11	54	5 x 11	58	5 x 11	65	5 x 11	70	6.3 x 11	80
22	5 x 11	70	5 x 11	75	5 x 11	80	5 x 11	87	5 x 11	95	6.3 x 11	115	8 x 11.5	135
33	5 x 11	84	5 x 11	90	5 x 11	97	6.3 x 11	115	6.3 x 11	136	8 x 11.5	150	10 x 16	195
47	5 x 11	100	5 x 11	110	5 x 11	115	6.3 x 11	145	6.3 x 11	165	8 x 11.5	190	10 x 16	255
100	5 x 11	145	6.3 x 11	180	6.3 x 11	190	8 x 11.5	240	8 x 11.5	260	10 x 12	320	10 x 20	370
220	6.3 x 11	250	8 x 11.5	300	8 x 11.5	320	10 x 12	420	10 x 16	490	10 x 20	565	13 x 25	675
330	8 x 11.5	350	8 x 11.5	370	10 x 12.5	470	10 x 16	570	13 x 20	635	13 x 20	765	16 x 32	972
470	8 x 11.5	415	10 x 12.5	520	10 x 16	620	10 x 16	740	13 x 20	860	16 x 25	1050	18 x 36	1135
1000	10 x 12.5	650	10 x 16	785	13 x 20	1090	13 x 20	1145	16 x 25	1530	16 x 25	1700	22 x 40	2600
2200	13 x 20	1240	13 x 20	1295	16 x 25	1660	16 x 32	1890	18 x 40	2231	18 x 40	2385		
3300	13 x 20	1420	16 x 25	1840	16 x 32	2070	18 x 36	2430	22 x 40	2785	22x 40	3000		
4700	16 x 25	1980	16 x 32	2260	18 x 36	2520	18 x 36	2700	25 x 40	3300	25 x 40	3560		
6800	16 x 25	2220	16 x 32	2520	18 x 36	2880	22 x 41	2900						
10000	18 x 36	2880	18 x 36	3080	22 x 40	3440								

Value (μF)	Working Voltage (WV); Dimensions: øD x L (mm); Ripple Current: mA/RMS @ 120Hz, 85°C							
	160		250		350		450	
	øD x L	mA	øD x L	mA	øD x L	mA	øD x L	mA
.47	5 x 11	13	8 x 11.5	21	8 x 11.5	21	10 x 12.5	26
1.0	5 x 11	20	8 x 11.5	32	8 x 11.5	32	10 x 12.5	38
2.2	6.3 x 11	34	8 x 11.5	49	10 x 16	63	10 x 16	63
3.3	8 x 11.5	50	10 x 12.5	70	10 x 16	78	10 x 20	86
4.7	8 x 11.5	60	10 x 16	93	10 x 20	103	13 x 20	120
10	10 x 16	115	10 x 20	150	13 x 20	174	13 x 25	192
22	13 x 20	216	13 x 20	255	13 x 25	282	16 x 25	354
33	13 x 20	270	13 x 25	348	16 x 32	438	18 x 36	426
47	13 x 25	354	16 x 25	468	16 x 36	500	18 x 40	555
100	16 x 25	582	18 x 40	822	18 x 40	685	22 x 45	750
220	18 x 36	900	22 x 40	1134				
330	18 x 40	1010						

Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

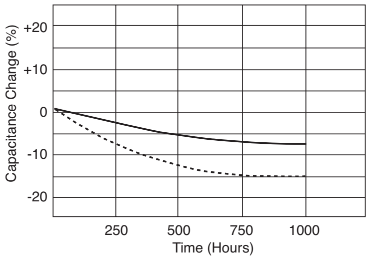
Miniature Aluminum Electrolytic Capacitors XRL Series

■ TYPICAL PERFORMANCE CHARACTERISTICS

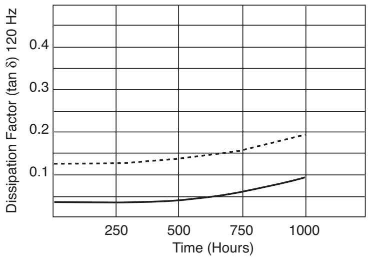
- - - - - - - - - 1000μF 16V
———— 1μF 50V

Life Test

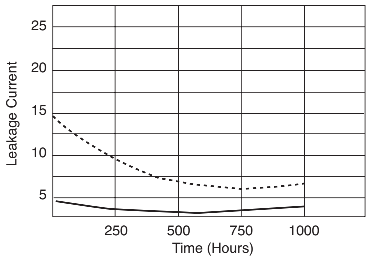
Capacitance Change vs. Time (at +85°C)



Dissipation Factor vs. Time (at +85°C)

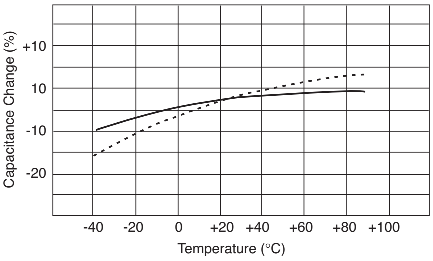


Leakage Current vs. Time (at +85°C)

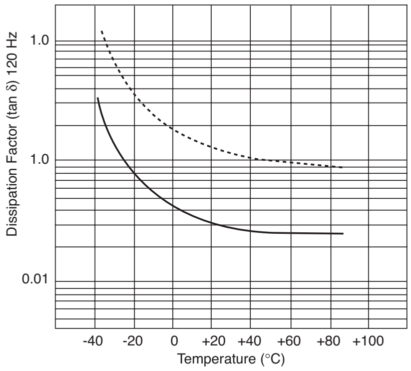


Temperature Characteristics

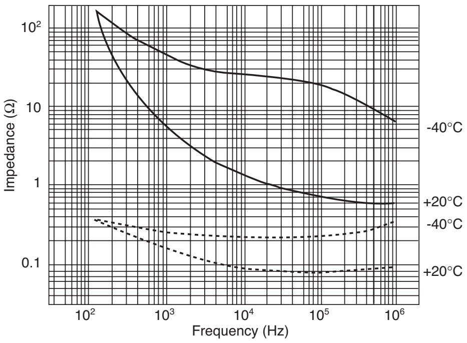
Capacitance Change vs. Temperature



Dissipation Factor vs. Temperature

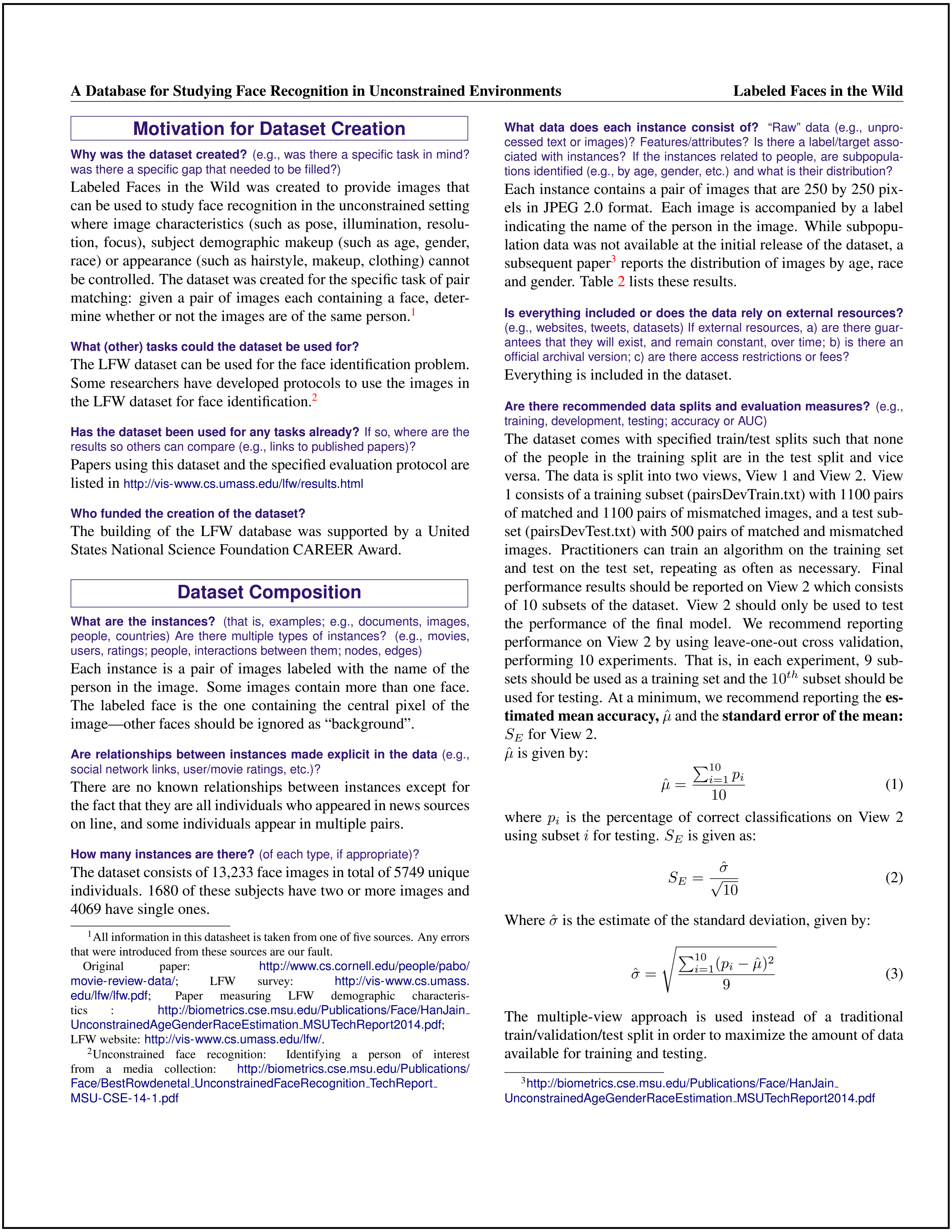


Impedance vs. Frequency



Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

B Prototypes of Datasheets for Datasets



A Database for Studying Face Recognition in Unconstrained Environments **Labeled Faces in the Wild**

Motivation for Dataset Creation

Why was the dataset created? (e.g., was there a specific task in mind? was there 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?

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?

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 are there? (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.

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 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; c) are there access restrictions or fees?

Everything is included in the dataset.

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

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)

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) and 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 of the instances?

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 a sample (not necessarily random) of instances from a larger set?

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

Data Preprocessing

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

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 images. The process is repeated if both images are already contained in a mismatched pair.

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

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? If not, what are the limitations?

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 will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and 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?

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.

Are there any fees or access/export restrictions?

There are no fees or restrictions.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

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? If so, how often and by whom?

Unknown

How will updates be communicated? (e.g., mailing list, GitHub)

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.

Is there an erratum?

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 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?

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.

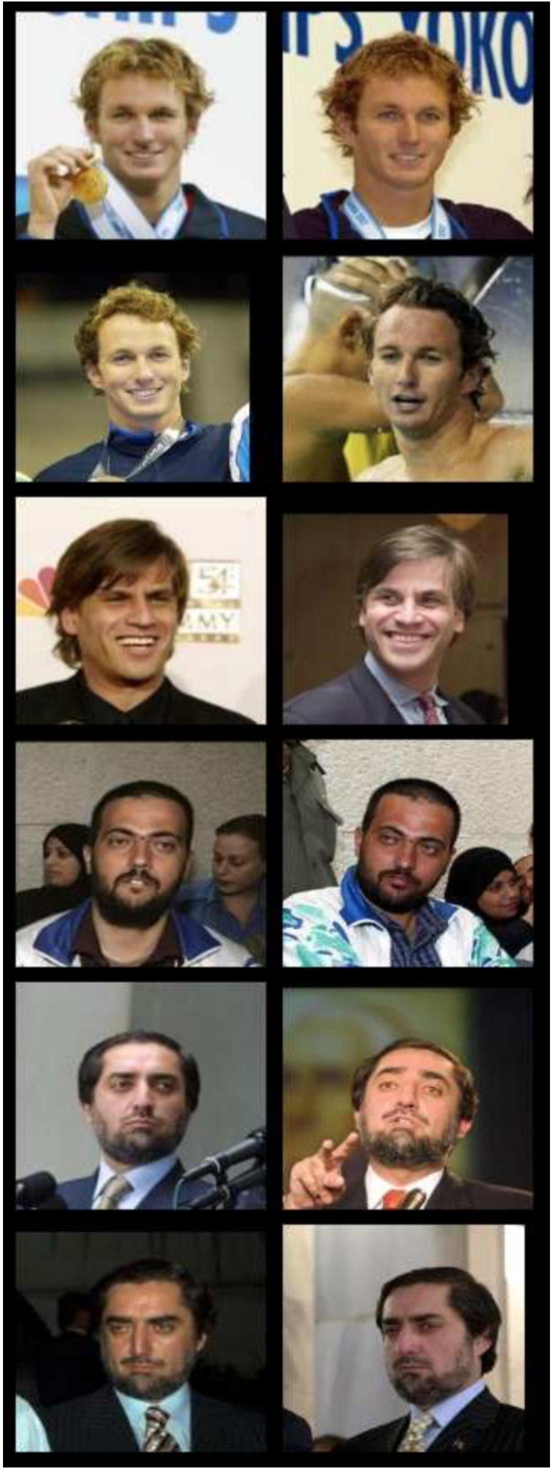
Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

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., was there a specific task in mind? was there 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?

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?

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 are there? (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).

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 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").

> 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`.

	corrected	——in paper——		
Features	NB	NB	ME	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).

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; c) are there access restrictions or fees?

Everything is included.

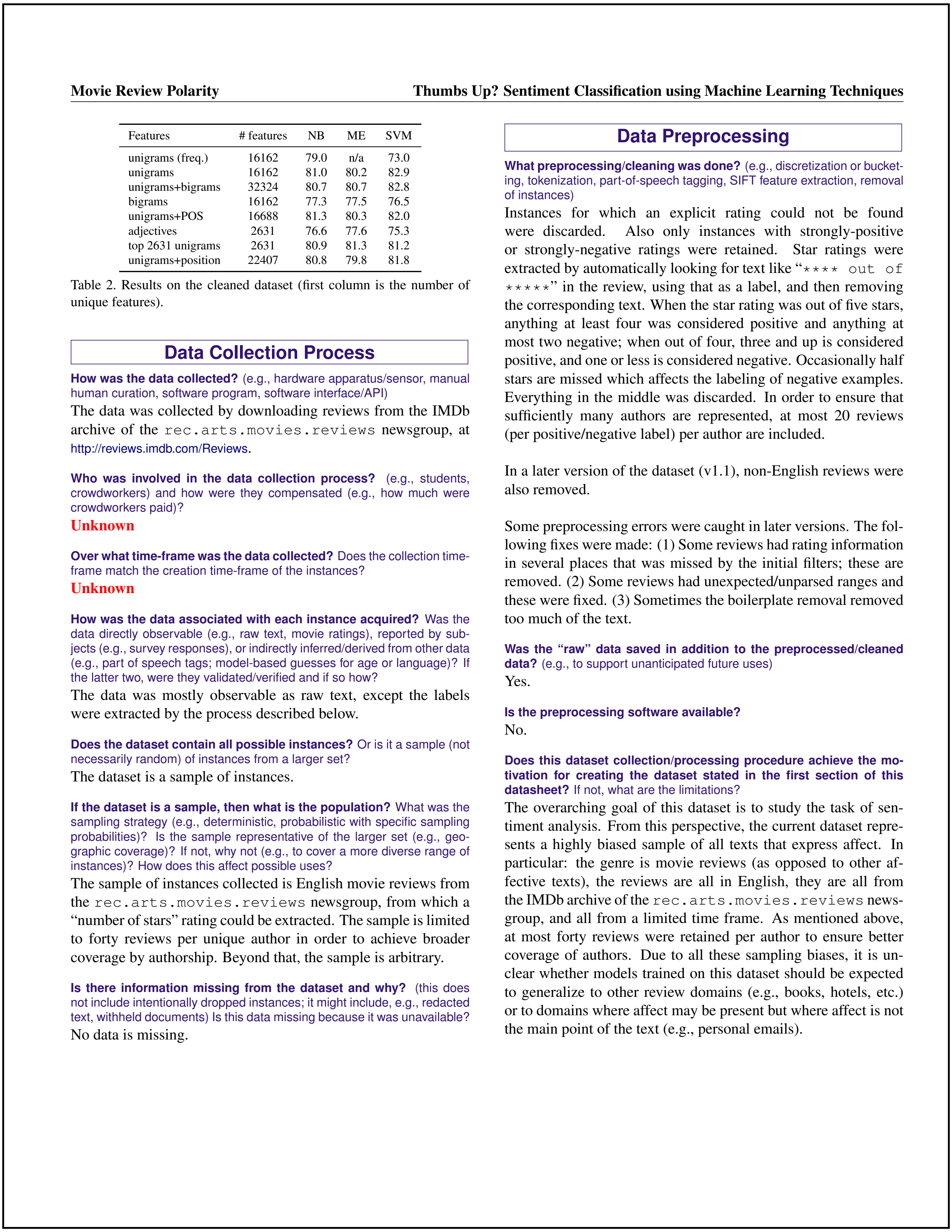
Are there recommended data splits and evaluation measures? (e.g., training, development, testing; accuracy or AUC)

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.

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.

[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.



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).

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API)

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) and 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 of the instances?

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 a sample (not necessarily random) of instances from a larger set?

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.

Data Preprocessing

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

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? If not, what are the limitations?

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 will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and 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?

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?

Bo Pang is supporting/maintaining the dataset.

Will the dataset be updated? If so, how often and by whom?

Since its initial release (v0.9) there have been three later releases (v1.0, v1.1 and v2.0).

How will updates be communicated? (e.g., mailing list, GitHub)

Updates are listed on the dataset web page.

Is there an erratum?

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. (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 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?

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.