Model Cards

(Group 5)
Kazi Noshin, Nina Chinnam, Yanxi Liu, Chaitanya Shahane

Why We Need Model Cards?

— ML models

- ML models are used in high-impact areas
- But there is no standardized way to document them
- Lack of transparency leads to misuse and harm

Real-World Harms from Missing Documentation

- Joy Buolamwini's story: face recognition failed to detect her:

  Joy's story

- Biased toxicity detection models
- Systems disproportionately fail on marginalized groups

What Are Model Cards?

- A model card = structured document about a trained ML model
- One to two page documents
- Describes performance across groups
- Includes: intended use, performance breakdowns, limitations
- Similar to nutrition labels or hardware datasheets

Who Benefits from Model/Data Cards?

- Developers → compare models
- Policymakers → assess risks
- End-users → understand potential harms
- Companies → support responsible AI practices

Datasheets vs Data Cards vs Model Cards (Purpose)

Datasheets [2018, 1]	Data Cards [2022]	Models Cards [2018, 2]
Standardize dataset documentation for transparency and accountability.	Summarize datasets to support responsible AI in real-world use.	Explain model use, performance, and ethical concerns for stakeholders.
Capture data origin, makeup, and use to assess fit for ML tasks.	Document both visible and contextual dataset info across its lifecycle.	Show performance by demographic and phenotypic subgroups.

Datasheets vs Data Cards vs Model Cards (Key Features)

Aspect	Datasheets	Data Cards	Models Cards
Format	Long-form Q&A format	Modular block format (title, Q, input)	Short (1–2 pages)
Documentation Style	Manual, reflective—not automated	Covers fairness, purpose, provenance	Shows model data, evaluation, ethics

Example Datasheet

Example datasheet for Pang and Lee's polarity dataset, page 1

This datasheet contains 4 pages

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

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

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

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

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

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

Any other comments?

None.

Composition

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

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity 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 sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

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

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

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

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

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

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

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later 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) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

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

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

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

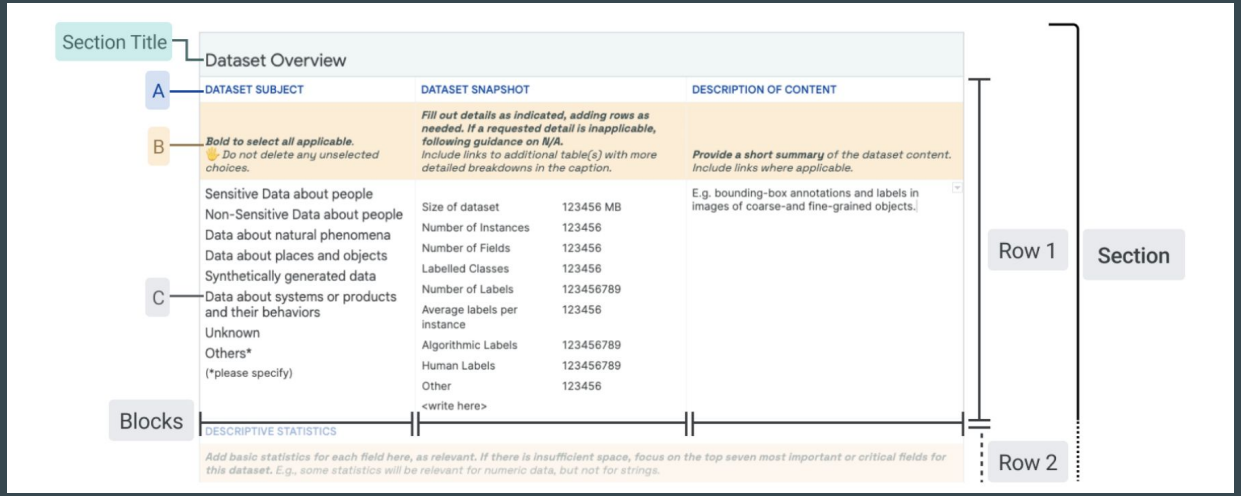
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

Example Data Card

A Data Card Template Section: This section is titled "Dataset Overview".



The section contains two rows:

- The first row has three blocks.
- The second row spans the entire width of the section.

Blocks contain
(A) A Title,
(B) A prompting question, and
(C) An answer input space with predetermined choices or suggested answer structures.

Example Model Card

Example Model Card for two versions of Perspective API's toxicity detector

Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because of privacy considerations, the model does not take into account user history when making judgments about toxicity.

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

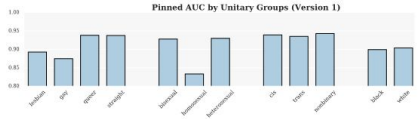
Evaluation Data

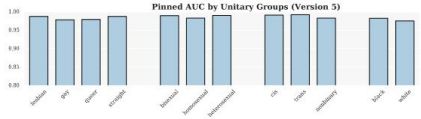
- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

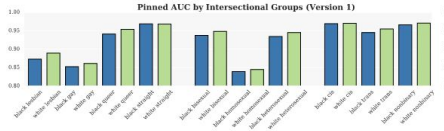
Caveats and Recommendations

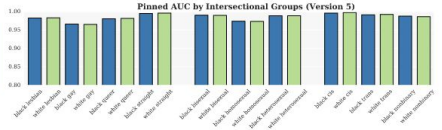
- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Quantitative Analyses









Datasheets vs Data Cards vs Model Cards (Key Features) [Cont.]

Aspect	Datasheets	Data Cards	Models Cards
Benefit	Supports reproducibility & bias checks	Serves as a shared “boundary object”	Ensures transparent, fair use
Use	Created during dataset build & updated over time	User-focused, works across platforms	Complements to datasheets

Datasheets vs Data Cards vs Model Cards (Target Audience)

Datasheets	Data Cards	Models Cards
Dataset creators and consumers (especially in academia or research).	Diverse stakeholders (developers, auditors, policymakers), not necessarily dataset experts.	ML practitioners, developers, policymakers, affected users.

Values Encoded in Machine Learning Research

Uplifted Values	Neglected Values
performance generalization efficiency novelty	social concerns justice inclusion societal need

Relation between Values and Cards

Datasheets	Data Cards	Models Cards
Tackle dataset opacity by requiring reflection on data origin, consent, use, and limits.	Challenge the assumptions by documenting context, labeling, fairness, and impact.	Address ethics and bias by detailing use cases, subgroup performance, and risks.

Discussion

1. Would it be feasible for datasets to be required to have a standardized documentation before being publicly released? What are the trade-offs?

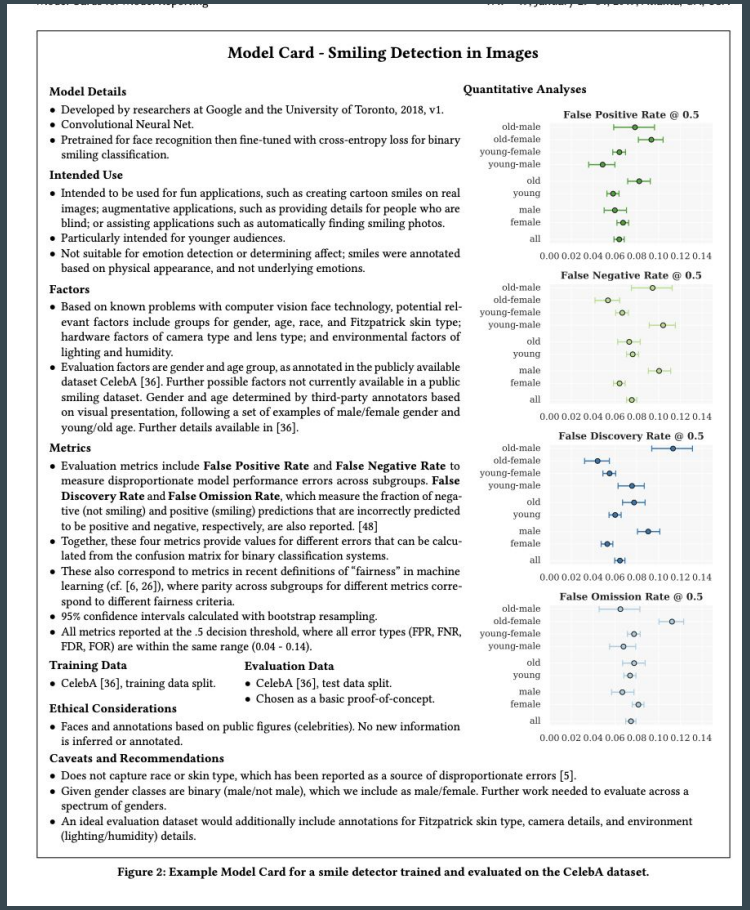
2. Can you think of a scenario where a lack of consistency in data/model documentation could lead to confusion or misinformation?

3. It is inevitable that values will influence AI design. Who should decide which values in AI design should be prioritized?

Methodologies

Model Cards for Model Reporting

- Inspired by system failures
- Built from cross-domain analogies
- Structure:
  - use case→ metrics → caveats
- Intersectional group evaluation required



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

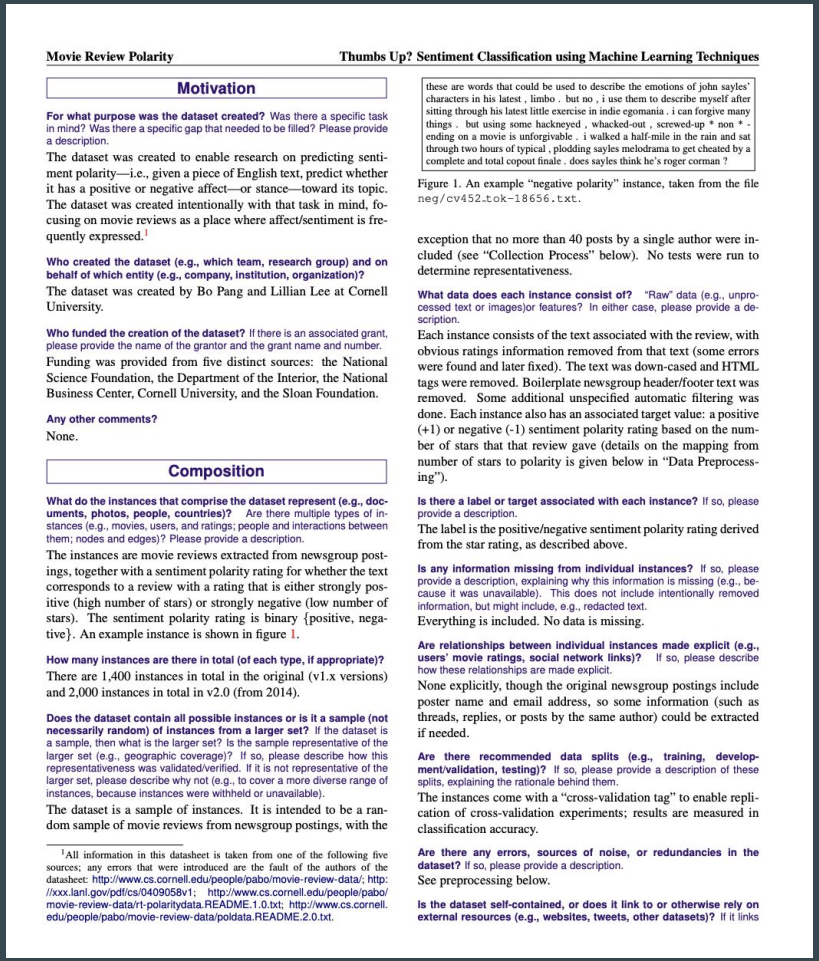
Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Datasheets for Datasets

- Drafting Questions
- Legal Review
- Pilot Testing
- Refinement
- Industry Feedback
- Structuring



Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

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

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

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

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

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

Any other comments?

None.

Composition

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

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity 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 sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

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

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

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

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

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

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

Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later 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) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

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

The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

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

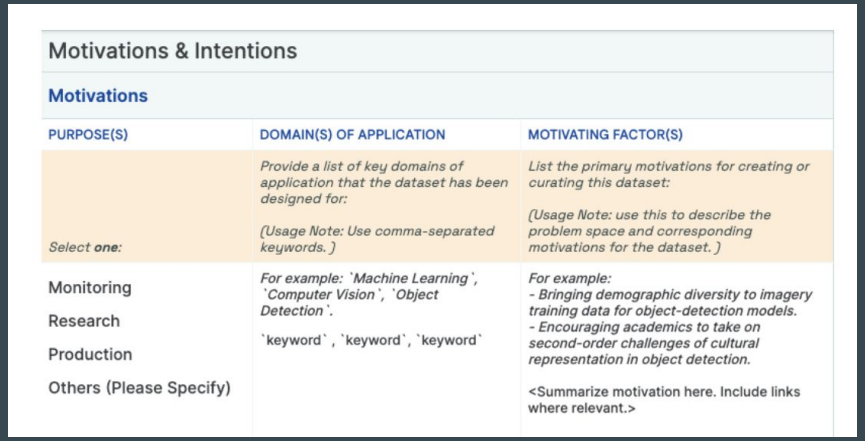
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

Data Cards

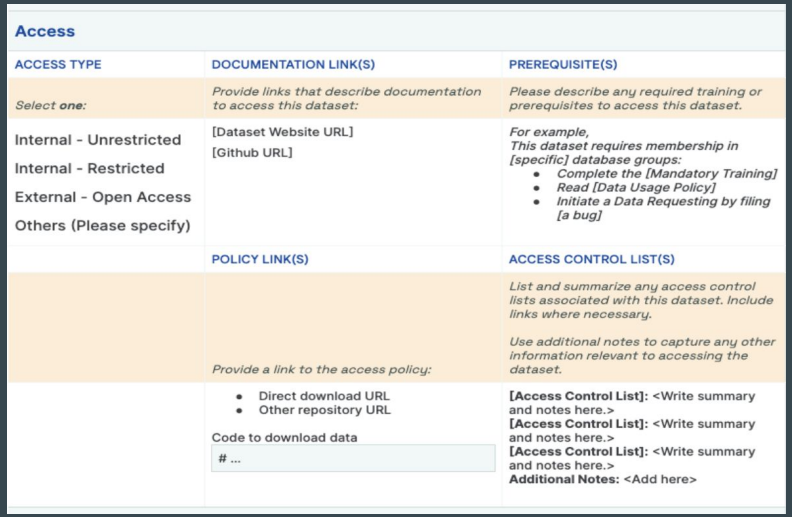
1. Participatory Design
   a. 24 month co-development with 12 teams
   b. 22 real world Data Cards
2. Surveys & Iteration
   a. MaxDiff Survey (n = 191)
   b. Identified 31 essential elements
3. OFTEn Framework
   a. Reflects on Observable vs Unobservable dataset traits



Motivations & Intentions

Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
*Select **one**:*	*Provide a list of key domains of application that the dataset has been designed for:* *(Usage Note: Use comma-separated keywords.)*	*List the primary motivations for creating or curating this dataset:* *(Usage Note: use this to describe the problem space and corresponding motivations for the dataset.)*
Monitoring Research Production Others (Please Specify)	*For example: `Machine Learning`, `Computer Vision`, `Object Detection`.* `keyword` , `keyword`, `keyword`	*For example:* *- Bringing demographic diversity to imagery training data for object-detection models.* *- Encouraging academics to take on second-order challenges of cultural representation in object detection.* <Summarize motivation here. Include links where relevant.>



Access

ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
*Select **one**:*	*Provide links that describe documentation to access this dataset:*	*Please describe any required training or prerequisites to access this dataset.*
Internal - Unrestricted Internal - Restricted External - Open Access Others (Please specify)	[Dataset Website URL] [Github URL]	*For example,* *This dataset requires membership in [specific] database groups:* • *Complete the [Mandatory Training]* • *Read [Data Usage Policy]* • *Initiate a Data Requesting by filing [a bug]*
	POLICY LINK(S)	**ACCESS CONTROL LIST(S)**
	Provide a link to the access policy:	*List and summarize any access control lists associated with this dataset. Include links where necessary.* *Use additional notes to capture any other information relevant to accessing the dataset.*
	• Direct download URL • Other repository URL Code to download data # ...	**[Access Control List]:** <Write summary and notes here.> **[Access Control List]:** <Write summary and notes here.> **[Access Control List]:** <Write summary and notes here.> **Additional Notes:** <Add here>

Values Encoded in ML Research

Sampling
100 top NeurIPS/ICML Papers

Coding
3.5k+ sentences, manually annotated

Values
Inductive + Deductive identification

Reliability
Fleiss' Kappa (0.45 - 0.79), dual coding

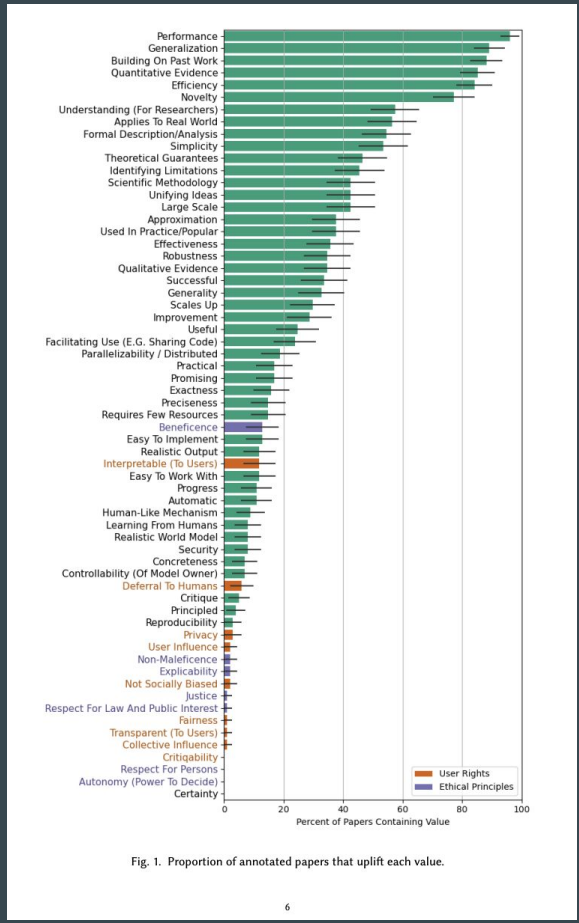


Fig. 1. Proportion of annotated papers that uplift each value.

6

Discussion

1. The methodology behind these papers emphasizes manual, reflective documentation rather than automation. What are the trade-offs of requiring dataset creators to manually complete extensive documentation, and how might this impact adoption in fast-paced industry settings?

Results and Analysis

Analysis - Data Cards

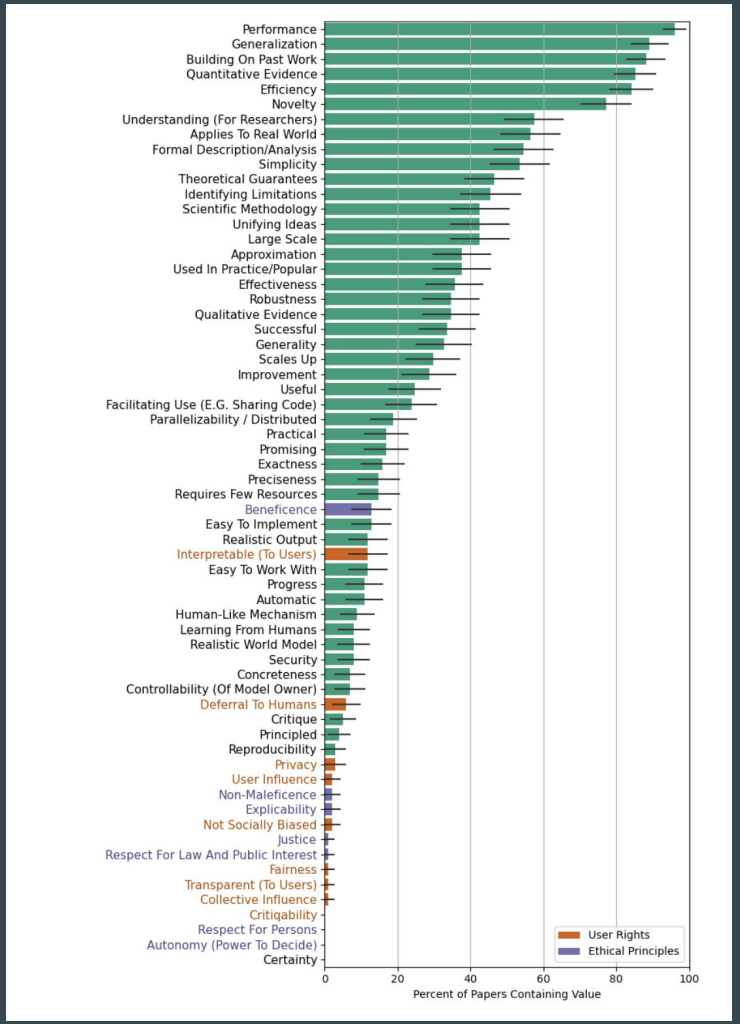
- This paper by Google Research also talks about data cards for the purpose and have interesting insights
- A single Data Card can support tasks such as conducting reviews and audits, determining use in AI systems or research, comparison of multiple datasets
- Also helps with inclusion of multiple perspectives (engineering, research, user experience, legal and ethical) to enhance the readability and relevance of documentation
- A centralized approach would definitely pave a path to hassle free future research

Analysis - Datasheets for Datasets

1. Since circulating the draft for this paper in 2018, the practice of maintaining datasheets for datasets has gained traction.
2. Academic Researchers and Researchers at organizations like IBM, Google, Microsoft have started maintaining data cards for the purpose.
3. The main common challenge observed in the process is need for dataset creators to modify the questions and workflow based on their existing organizational infrastructure and workflows.
4. Although it might help in mitigating social bias to some extent, it will also create an overhead for dataset creators.

Analysis: The Values Encoded in Machine Learning Research

- Out of the plotted prevalence values in 100 annotated papers, the top values were observed to be performance (96% of papers), generalization (89%), building on past work (88%).
- Most papers only justify how they achieve their internal, technical goal; 68% make no mention of societal need or impact, and only 4% make a rigorous attempt to present links connecting their research to societal needs.
- 98% of papers contained no reference to potential negative impacts.


Performance
Generalization
Building On Past Work
Quantitative Evidence
Efficiency
Novelty
Understanding (For Researchers)
Applies To Real World
Formal Description/Analysis
Simplicity
Theoretical Guarantees
Identifying Limitations
Scientific Methodology
Unifying Ideas
Large Scale
Approximation
Used In Practice/Popular
Effectiveness
Robustness
Qualitative Evidence
Successful
Generality
Scales Up
Improvement
Useful
Facilitating Use (E.G. Sharing Code)
Parallelizability / Distributed
Practical
Promising
Exactness
Preciseness
Requires Few Resources
Beneficence
Easy To Implement
Realistic Output
Interpretable (To Users)
Easy To Work With
Progress
Automatic
Human-Like Mechanism
Learning From Humans
Realistic World Model
Security
Concreteness
Controllability (Of Model Owner)
Deferral To Humans
Critique
Principled
Reproducibility
Privacy
User Influence
Non-Maleficence
Explicability
Not Socially Biased
Justice
Respect For Law And Public Interest
Fairness
Transparent (To Users)
Collective Influence
Critiqability
Respect For Persons
Autonomy (Power To Decide)
Certainty
User Rights
Ethical Principles
0
20
40
60
80
100
Percent of Papers Containing Value

Analysis: The Values Encoded in Machine Learning Research

- Another interesting statistic from the study was:

  Comparing papers written in 2008/2009 to those written in 2018/2019, ties to corporations nearly doubled to 79% of all annotated papers,

  ties to big tech more than tripled, to 66%, while ties to universities declined to 81%, putting the presence of corporations nearly on par with universities

Analysis: The Values Encoded in Machine Learning Research

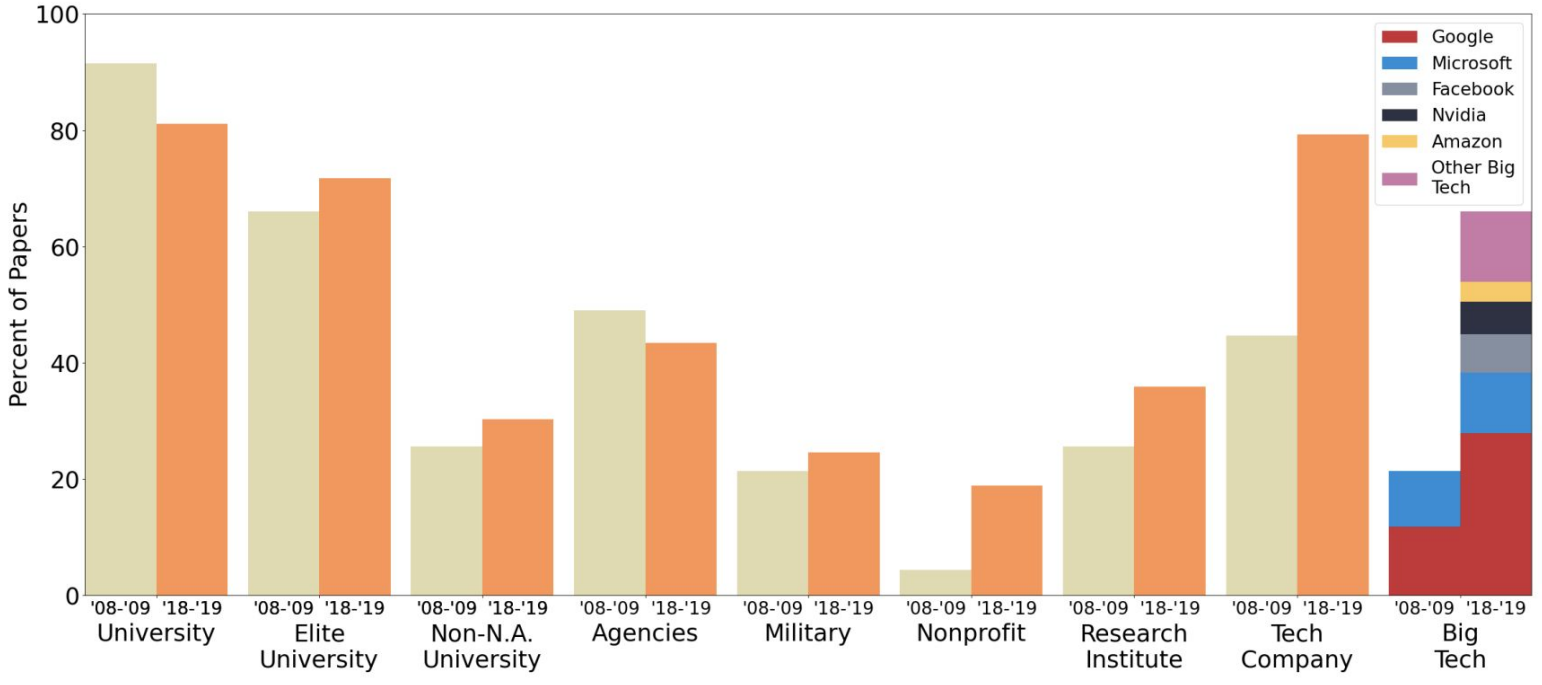


Fig. 3. Affiliations and funding ties.

Discussion

- The entry of Big corporations raises concerns for ethical research and values. But these corporations also bring a lot of money in terms of research funds, something that is vital in today's fast paced research world. What are your thoughts on this ?

Scenario Discussion

A tech company has developed an AI tool to automate hiring by analyzing resumes. The tool is trained on a dataset primarily from a specific demographic and uses performances metrics based on homogeneous job markets. As the company prepares to release the tool, it faces challenges related to potential hiring bias, the need for transparency in decision-making, and ensuring fairness across diverse groups. To address these issues, the company plans to create Model Cards, Datasheets and Data cards to document the AI systems and its datasets.

1. What steps should the company take to document potential biases in the data?
2. What ethical considerations should the company take into account when releasing an AI-powered hiring tool?
3. Should there be regulatory standards for the use of such tools, particularly when they involve sensitive decisions?

Thank you