February 8th, 2019
11:00 am PST
Online via Zoom

AI Research Discussion:
Model Cards for Model Reporting

LXAI

Accel.AI

AIDLE

AI & Deep Learning Enthusiasts

AIDLE

Meetup

https://www.meetup.com/Artificial-Intelligence-Deep-Learning-Enthusiasts-Meetup/

Github

https://github.com/AccelAI/AI-DL-Enthusiasts-Meetup

Accel.AI

https://www.accel.ai/

LatinX in AI Coalition

http://www.latinxinai.org/

Accel.AI

AI & Deep Learning Enthusiasts

AIDLE

Attendee Guidelines

- Try to read the paper and find a few good questions and/or insights.
- Be on time, latecomers can be disruptive for online discussions
- Use a headset in a quiet location if possible, and have a good internet connection: good audio quality is essential
- Have the discussion paper on your desktop ready to show via screen sharing in the event you need to point to something in a figure e.g.
- There are no silly questions!
- **AIDLE** has strict guidelines regarding discrimination based on gender, ethnicity, sexual orientation, class, and prior experience or educational background. We are all here to learn and grow! Violators will be removed from the group immediately.

Accel.AI OpenCollective

https://opencollective.com/accel-ai

Accel.AI

Resources & Links

Andrew Zaldivar

Linkedin

Google Citations

Interview with Google

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

PDF: https://arxiv.org/abs/1810.03993

Publicity

Increasing Transparency in Perspective’s Machine Learning Models by Jigsaw

Affiliated Research

Datasheets for Datasets

Videos

Joy Buolamwini - How I'm fighting bias in algorithms

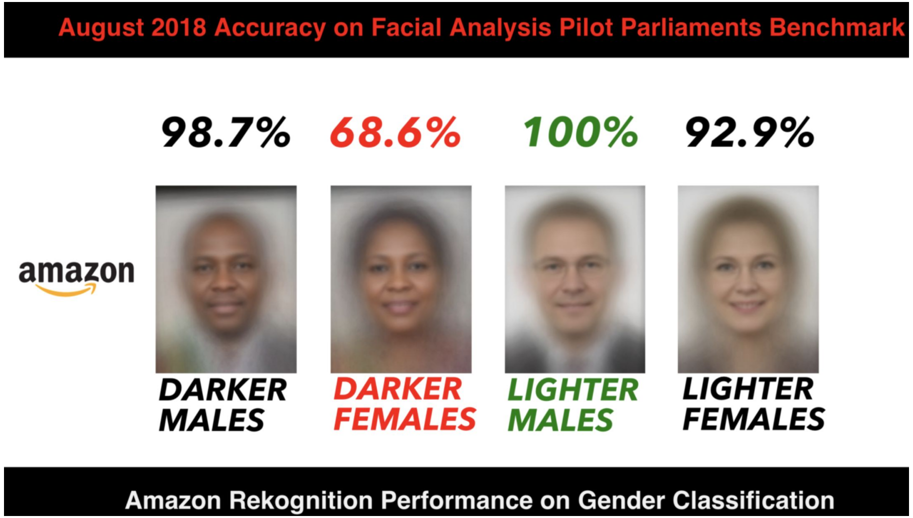
Rachel Thomas - Some Healthy Principles About Ethics & Bias In AI

Axios - Biases are being baked into artificial intelligence

Introduction - Why?

- ➔ No standardized documentation procedures for ML & AI models in research or industry today

- ➔ Lack of standardization can perpetuate systemic bias, misrepresentation and other errors



August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

	DARKER MALES	DARKER FEMALES	LIGHTER MALES	LIGHTER FEMALES
amazon	98.7%	68.6%	100%	92.9%

Amazon Rekognition Performance on Gender Classification

Introduction - Why?

- Model Cards for Model Reporting
  - Provide transparency and accountability in models
  - Information on capabilities and limitations of models
  - Detail motivation behind performance metrics
  - Benchmarked evaluation of cultural, demographic, phenotypic, and intersectional groups

- Improve impact in life saving applications
  - Healthcare
  - Employment
  - Education
  - Law enforcement

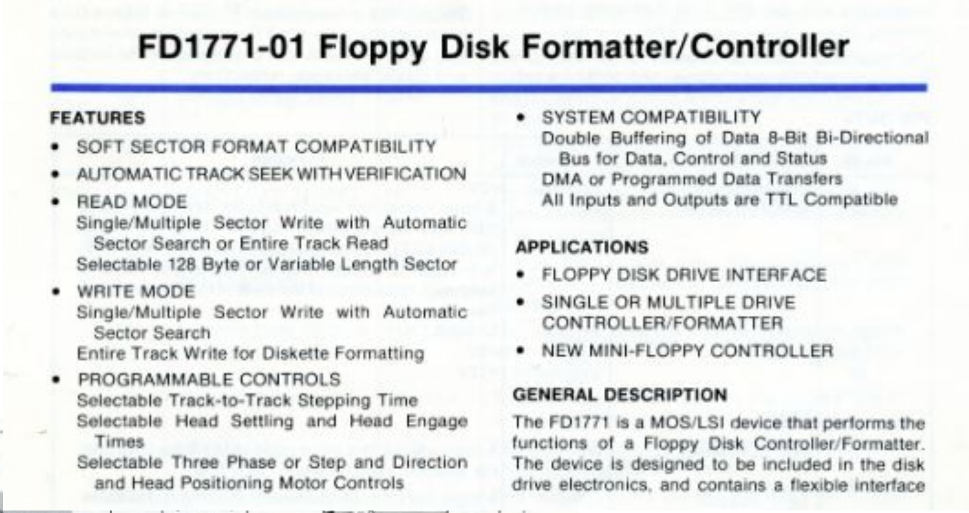
Introduction - Why?

- ➔ Can you think of specific applications that could benefit from models with model cards?
- ➔ What other possible benefits could this standardization provide?
- ➔ What other complimentary information could supplement these cards?

Accel.AI

Background

- ➔ Most mature industries have standardized methods for benchmarking
  - ◆ Electronic Hardware
  - ◆ Auto Industry
  - ◆ Pharmaceutical



FD1771-01 Floppy Disk Formatter/Controller

FEATURES

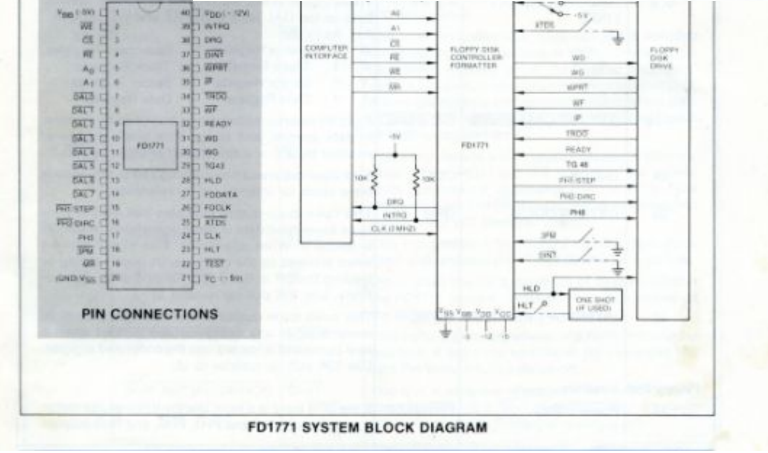
- SOFT SECTOR FORMAT COMPATIBILITY
- AUTOMATIC TRACK SEEK WITH VERIFICATION
- READ MODE
  Single/Multiple Sector Write with Automatic Sector Search or Entire Track Read
  Selectable 128 Byte or Variable Length Sector
- WRITE MODE
  Single/Multiple Sector Write with Automatic Sector Search
  Entire Track Write for Diskette Formatting
- PROGRAMMABLE CONTROLS
  Selectable Track-to-Track Stepping Time
  Selectable Head Settling and Head Engage Times
  Selectable Three Phase or Step and Direction and Head Positioning Motor Controls
- SYSTEM COMPATIBILITY
  Double Buffering of Data 8-Bit Bi-Directional Bus for Data, Control and Status
  DMA or Programmed Data Transfers
  All Inputs and Outputs are TTL Compatible

APPLICATIONS

- FLOPPY DISK DRIVE INTERFACE
- SINGLE OR MULTIPLE DRIVE CONTROLLER/FORMATTER
- NEW MINI-FLOPPY CONTROLLER

GENERAL DESCRIPTION

The FD1771 is a MOS/LSI device that performs the functions of a Floppy Disk Controller/Formatter. The device is designed to be included in the disk drive electronics, and contains a flexible interface



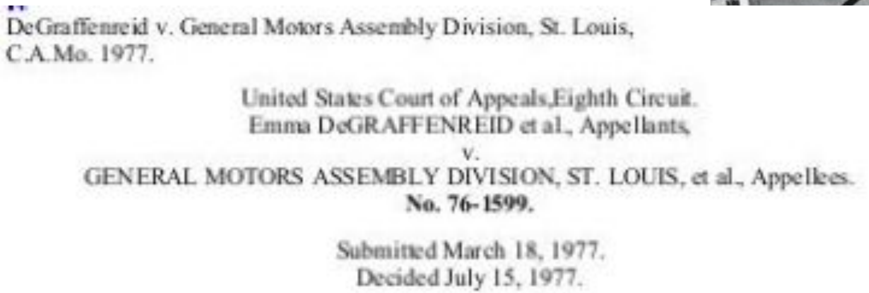
PIN CONNECTIONS

FD1771 SYSTEM BLOCK DIAGRAM

Background

Intersectional Analysis:

➔ Empirical analysis emphasizing interaction between various demographic categories



DeGraffenreid v. General Motors Assembly Division, St. Louis,
C.A.Mo. 1977.

United States Court of Appeals,Eighth Circuit.
Emma DeGRAFFENREID et al., Appellants,
v.
GENERAL MOTORS ASSEMBLY DIVISION, ST. LOUIS, et al., Appellees.
No. 76-1599.

Submitted March 18, 1977.
Decided July 15, 1977.

DeGraffenreid v. General Motors

by Skylr Smith

Background

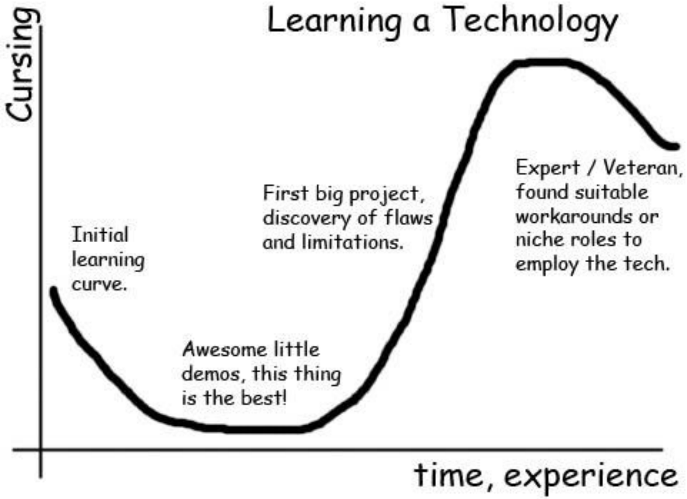
- ➔ What risks are involved with labeling gender and race in datasets?
- ➔ Should they be included at all?
- ➔ Should the difference between gender and sex be distinguished?
- ➔ Should analysis only be done on self identified individuals?
- ➔ Would classifying individuals with the "perceived" tag be sufficient for labeling?

Motivation

Increased ML use =

Increase errors and failures =

Increased negative systemic impact without explanation...

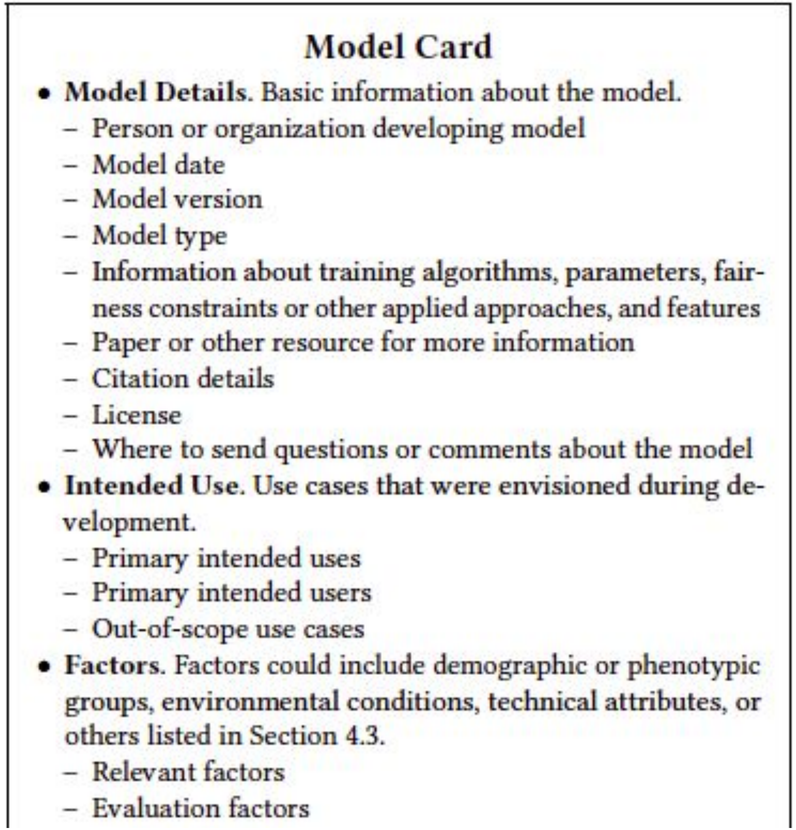

Learning a Technology
Cursing
time, experience
Initial learning curve.
Awesome little demos, this thing is the best!
First big project, discovery of flaws and limitations.
Expert / Veteran, found suitable workarounds or niche roles to employ the tech.

Motivation - Use Cases

- ➔ ML & AI Practitioners
- ➔ Model Developers
- ➔ Software Developers
- ➔ Policy Makers
- ➔ Organizations
- ➔ ML-Knowledgeable Individuals
- ➔ Impacted Individuals

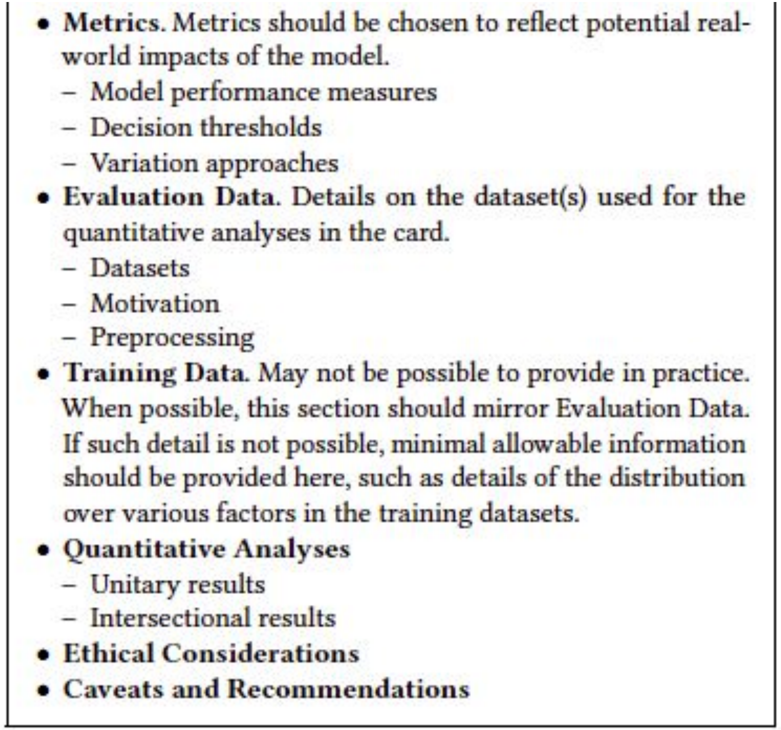
Accel.AI

Model Card Sections



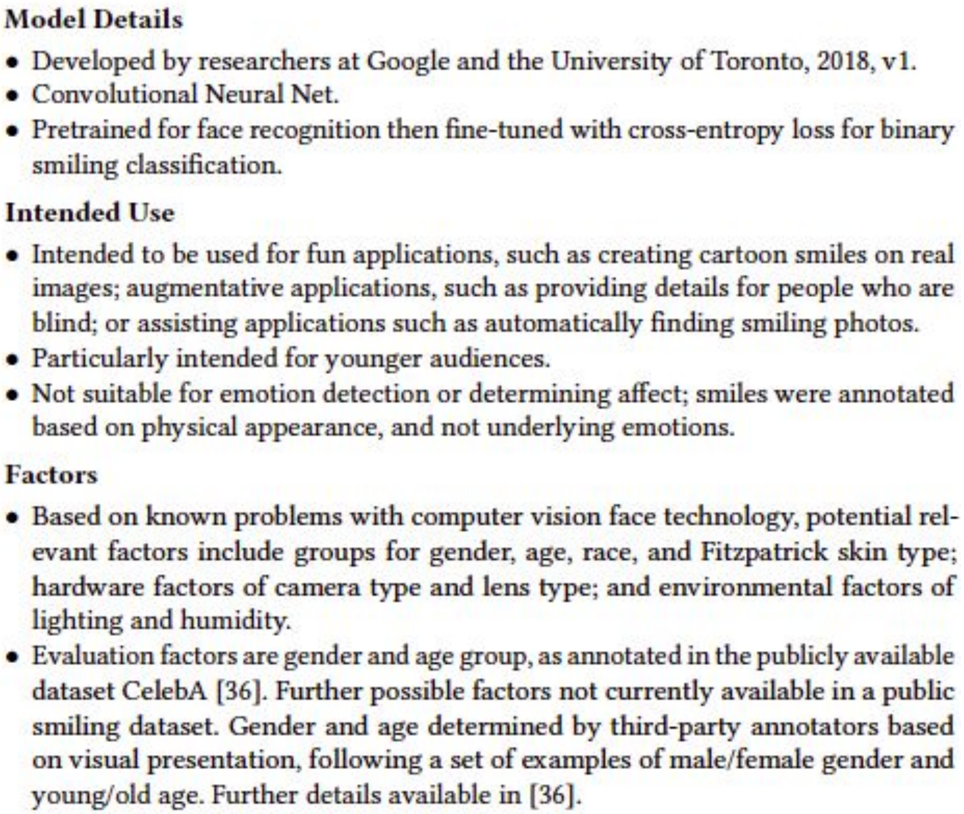
Model Card

- **Model Details.** Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use.** Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors



- **Metrics.** Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data.** Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data.** May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

V. Example - Smiling Classifier



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.

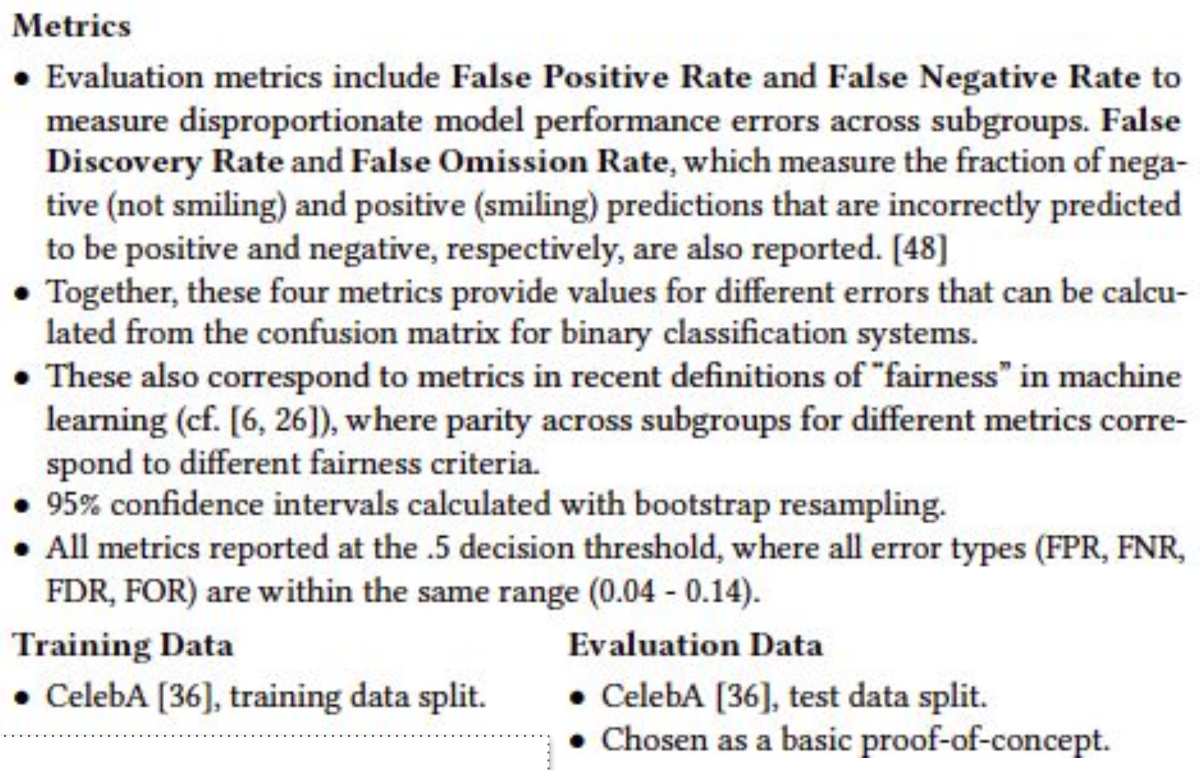
Intended Use

- 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].

V. Example - Smiling Classifier



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

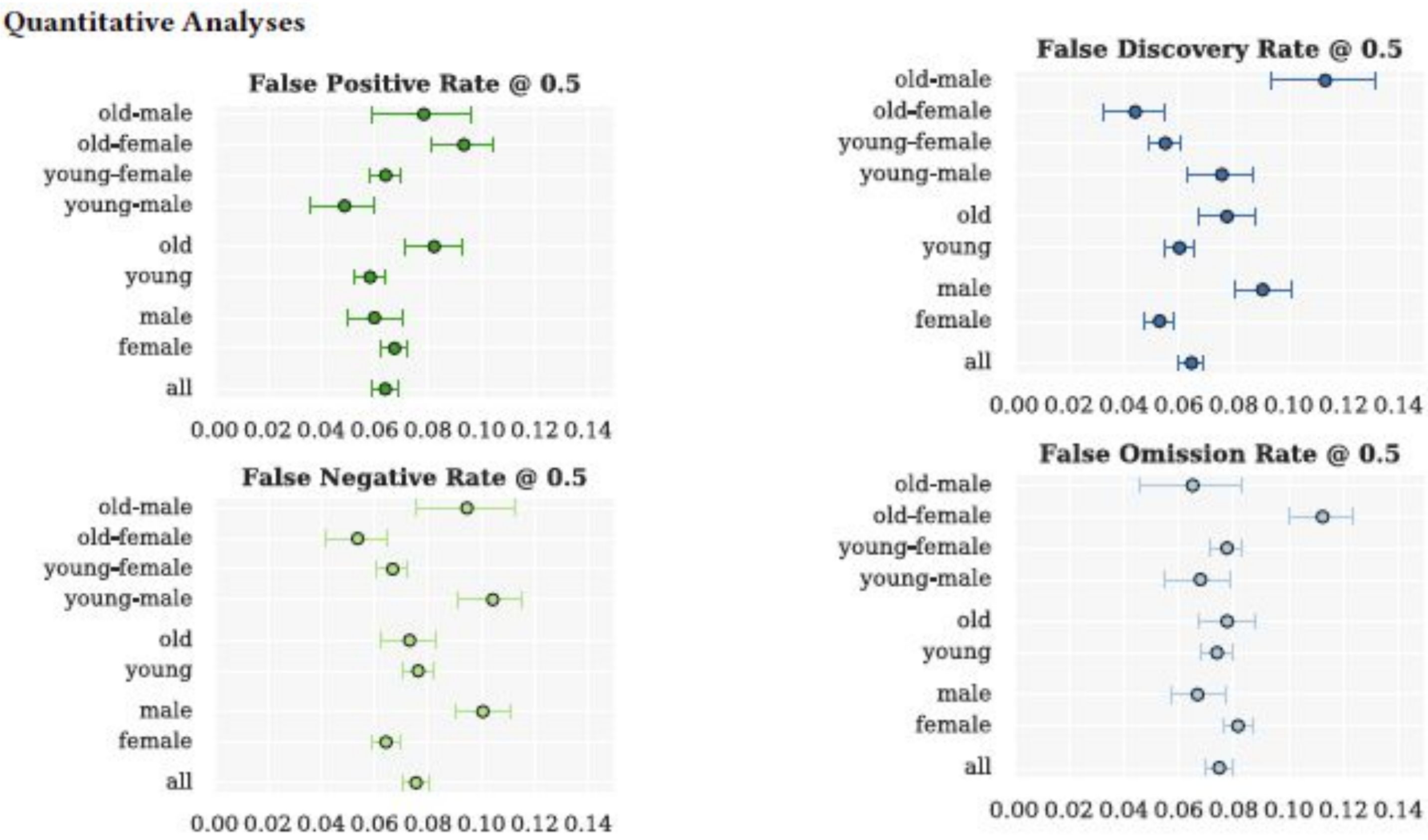
- CelebA [36], training data split.

Evaluation Data

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

Accel.AI

V. Example - Smiling Classifier



Quantitative Analyses

False Positive Rate @ 0.5

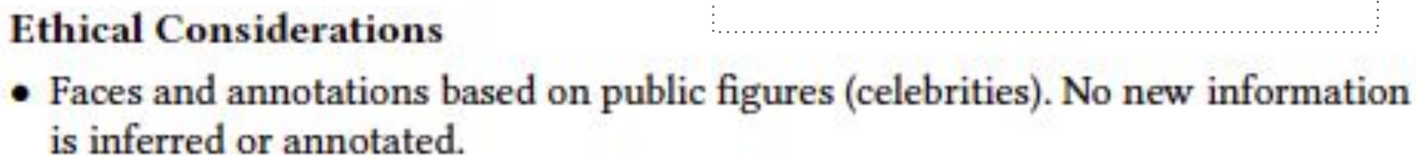
False Negative Rate @ 0.5

False Discovery Rate @ 0.5

False Omission Rate @ 0.5

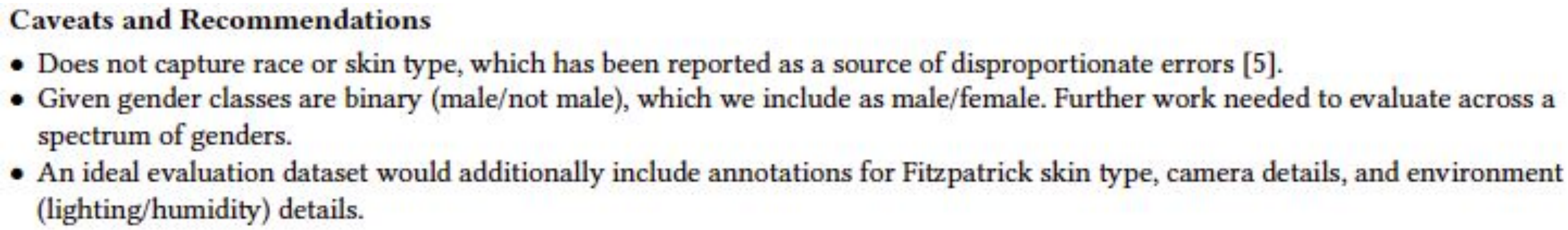
Accel.AI

V. Example - Smiling Classifier



Ethical Considerations

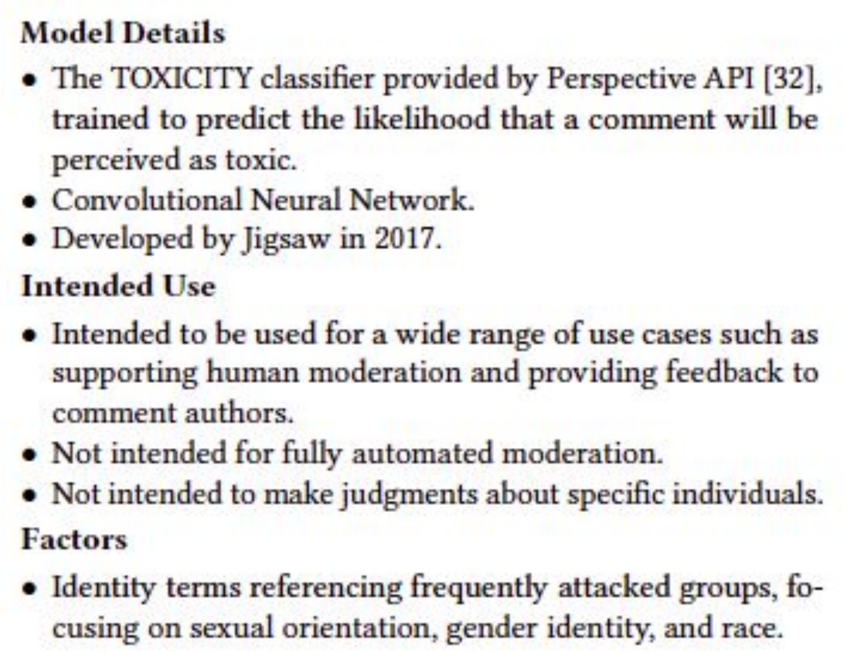
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.



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.

V. Example - Toxicity in Text



Model Details

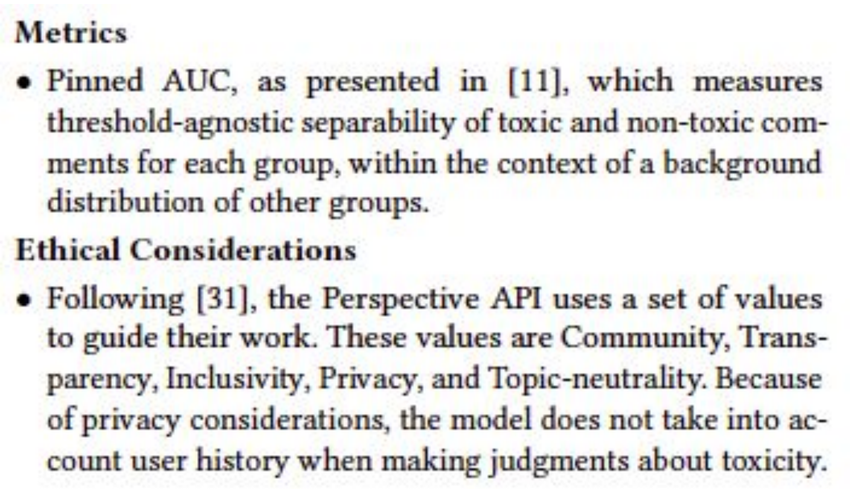
- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

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.



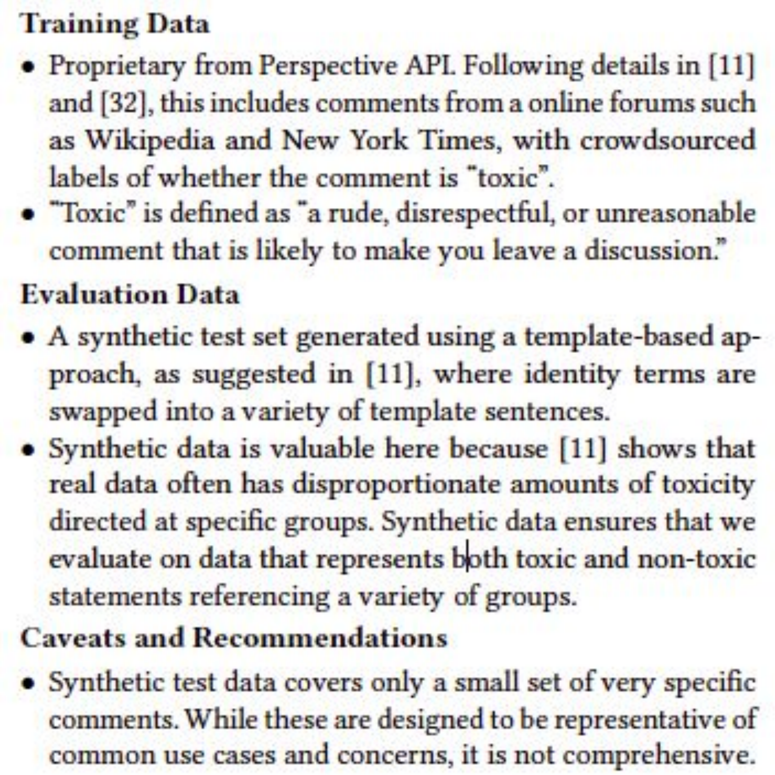
Metrics

- 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.

V. Example - Toxicity in Text



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.”

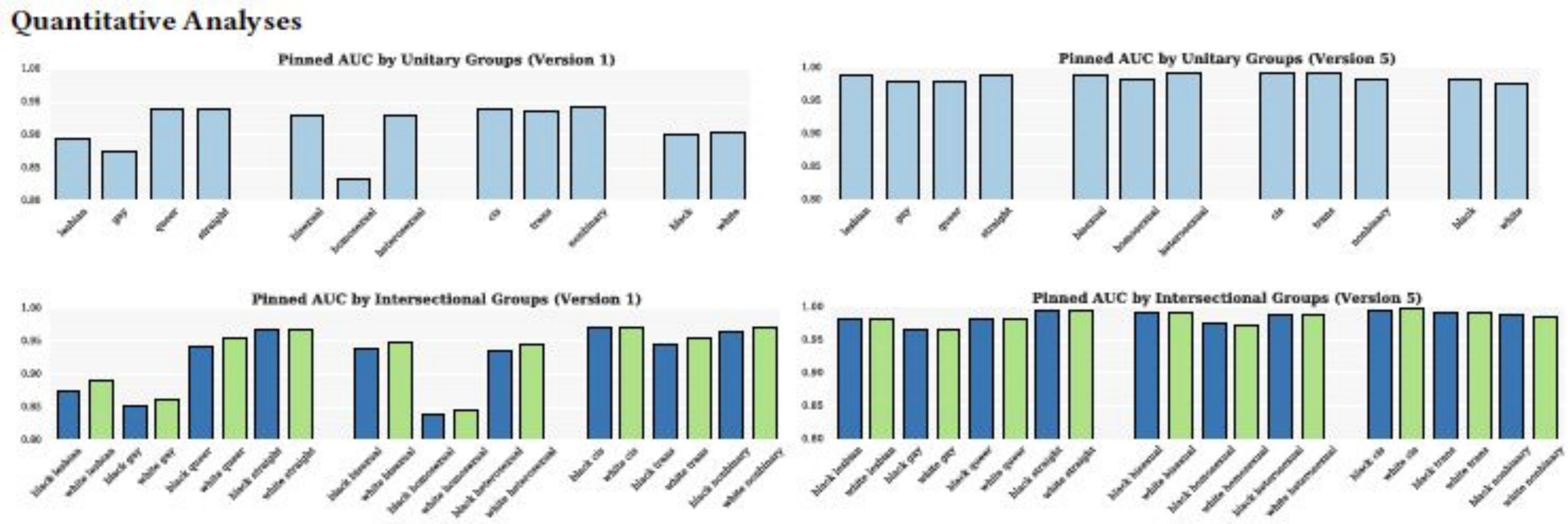
Evaluation Data

- 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.

V. Example - Toxicity in Text



Quantitative Analyses

AI & Deep Learning Enthusiasts

AIDLE

Meetup

https://www.meetup.com/Artificial-Intelligence-Deep-Learning-Enthusiasts-Meetup/

Github

https://github.com/AccelAI/AI-DL-Enthusiasts-Meetup

Accel.AI

https://www.accel.ai/

LatinX in AI Coalition

http://www.latinxinai.org/

Accel.AI OpenCollective

https://opencollective.com/accel-ai

Accel.AI

Accel.AI

Join Us!

Accel AI Institute 501c3 Non Profit

@ info@accel.ai

@AccelerateAI

/accelai/

/accel.ai/

/company/accel.ai

www.accel.ai
www.latinxinai.org