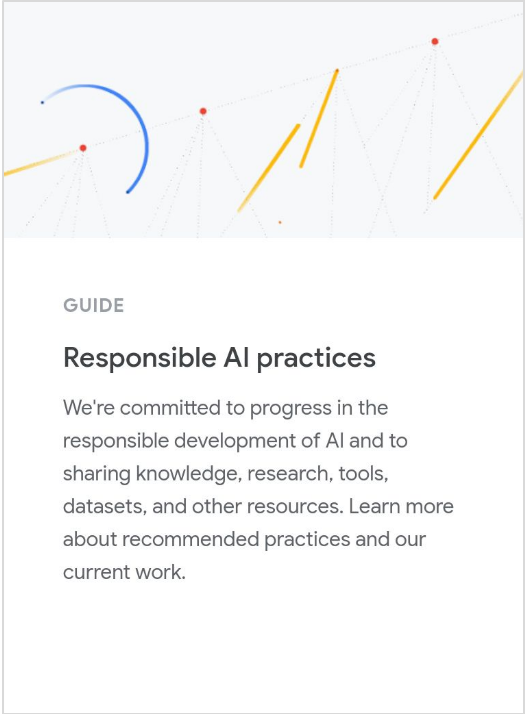
Model Cards for Model Reporting

Andrew Zaldivar, Ph.D.

In Collaboration with Margaret Mitchell, Simone Wu, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji & Timnit Gebru

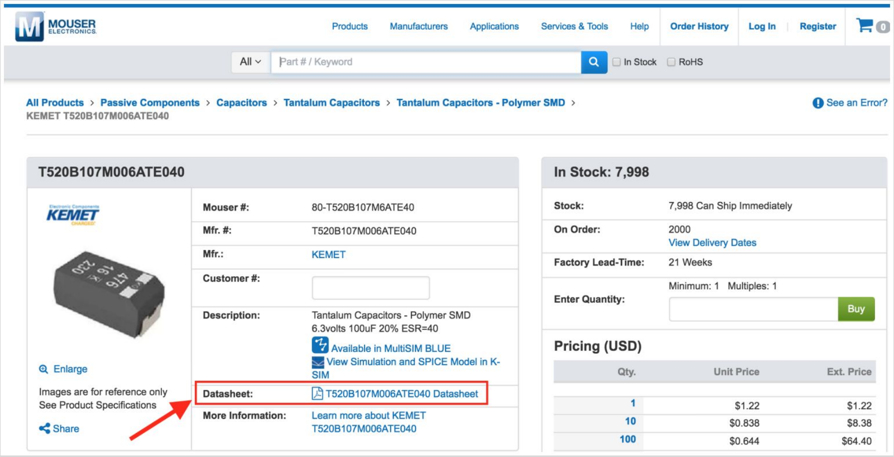
Background



GUIDE

Responsible AI practices

We're committed to progress in the responsible development of AI and to sharing knowledge, research, tools, datasets, and other resources. Learn more about recommended practices and our current work.



All Products › Passive Components › Capacitors › Tantalum Capacitors › Tantalum Capacitors - Polymer SMD ›
KEMET T520B107M006ATE040

T520B107M006ATE040

Images are for reference only
See Product Specifications

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

In Stock: 7,998

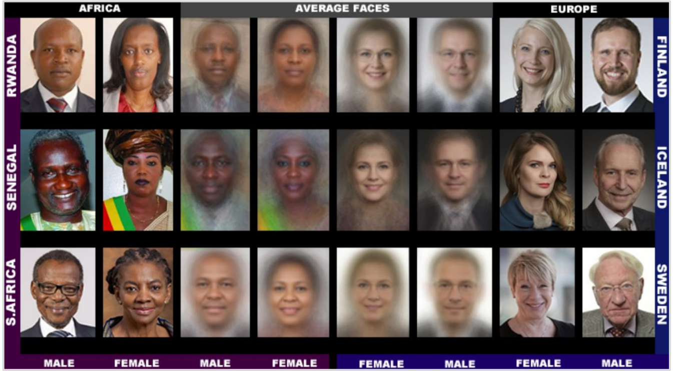
Stock:	7,998 Can Ship Immediately
On Order:	2000 View Delivery Dates
Factory Lead-Time:	21 Weeks
Enter Quantity:	Minimum: 1 Multiples: 1

Pricing (USD)

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

Timnit Gebru, et al. "Datasheets for Datasets." arXiv preprint arXiv:1803.09010 (2018)

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Conference on Fairness, Accountability and Transparency. 2018.



ai.google/education/responsible-ai-practices

Motivation

Disaggregated Evaluation

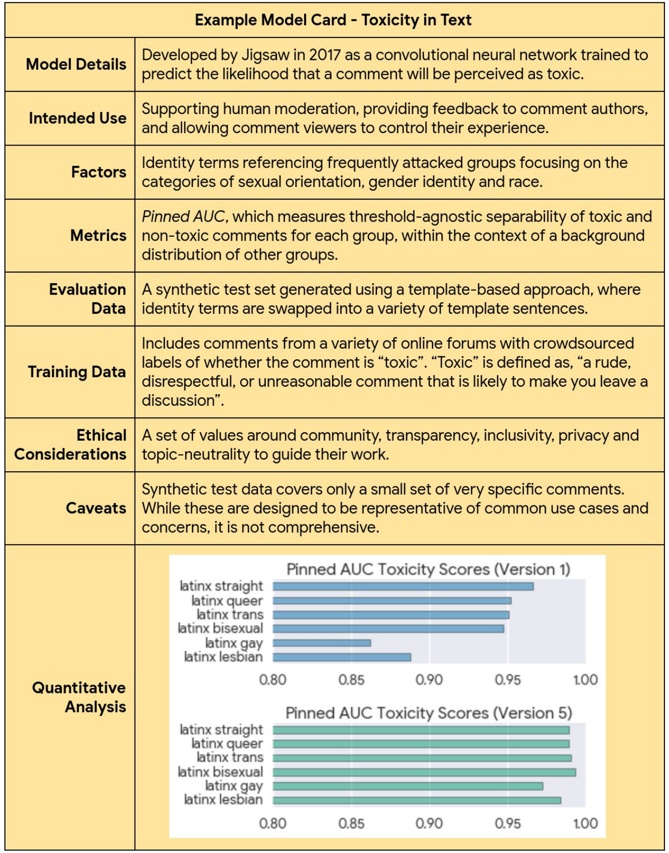
Transparency

Unitary / Intersectional Analyses

Context Around Model

Model Cards: A Proposal

Model Cards is a framework that serve to disclose information about a trained machine learning model.



Example Model Card - Toxicity in Text	
Model Details	Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.
Intended Use	Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.
Factors	Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.
Metrics	*Pinned AUC*, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.
Evaluation Data	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
Training Data	Includes comments from a variety of online forums 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”.
Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats	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 Analysis	Pinned AUC Toxicity Scores (Version 1) Pinned AUC Toxicity Scores (Version 5)

Model Details, Intended Use & Factors

Example Model Card - Toxicity in Text	
Model Details	Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.
Intended Use	Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.
Factors	Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.

Metrics, Evaluation Data & Training Data

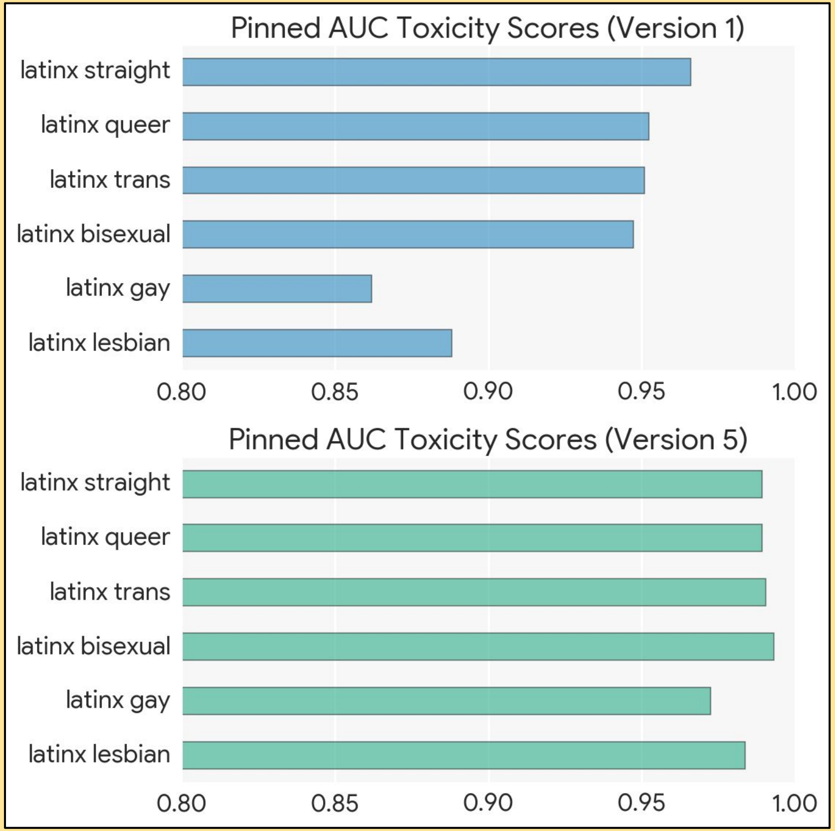
Metrics	*Pinned AUC*, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.
Evaluation Data	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
Training Data	Includes comments from a variety of online forums 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”.

Ethical Consideration & Caveats

Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats & 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 Analysis

Quantitative Analysis



Discussion

Responsible AI

We propose *Model Cards* as a step towards the responsible democratization of machine learning and related AI technology, intended to be applicable across different institutions, contexts, and stakeholders.

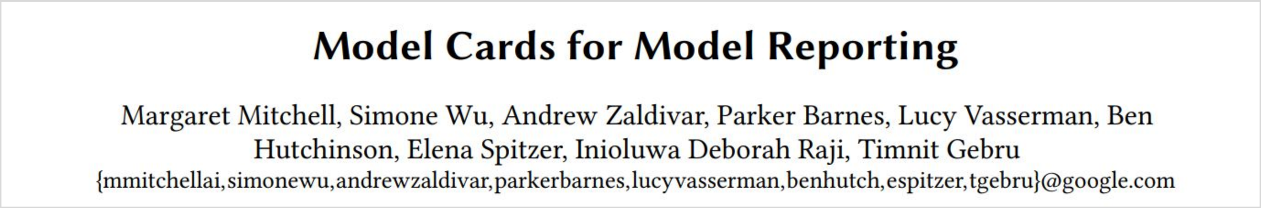
Refine Framework

Usefulness and accuracy of a model card relies on the integrity of the card creator(s). Future work will aim to refine this framework by studying how model information is interpreted and used by different stakeholders.

Other Transparency Methods

Similar work has begun for datasets and AI services (e.g., Datasheets for Datasets, Nutrition Labels for ML Datasets, IBM's Factsheets for AI Services). Worth exploring how *Model Cards* can strengthen and complement other transparency methods.

What’s Next?



Model Cards for Model Reporting

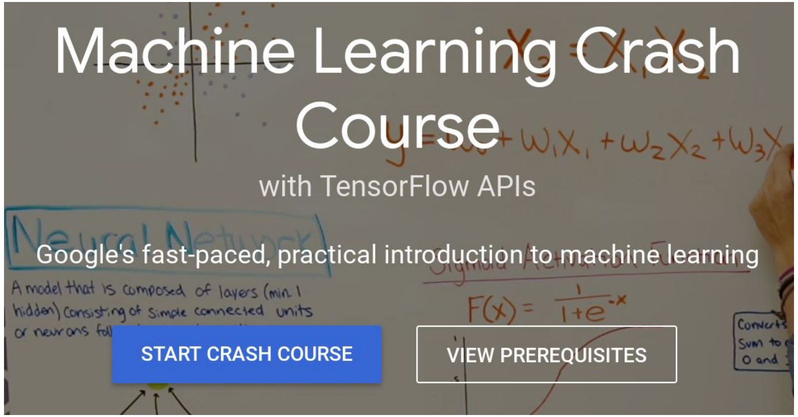
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com

Accepted at ACM Conference on Fairness, Accountability, and Transparency

Poster Presentations at Women in Machine Learning, Black in AI & LatinX in AI Workshops, NeurIPS 2018

https://arxiv.org/abs/1810.03993

Learn More!

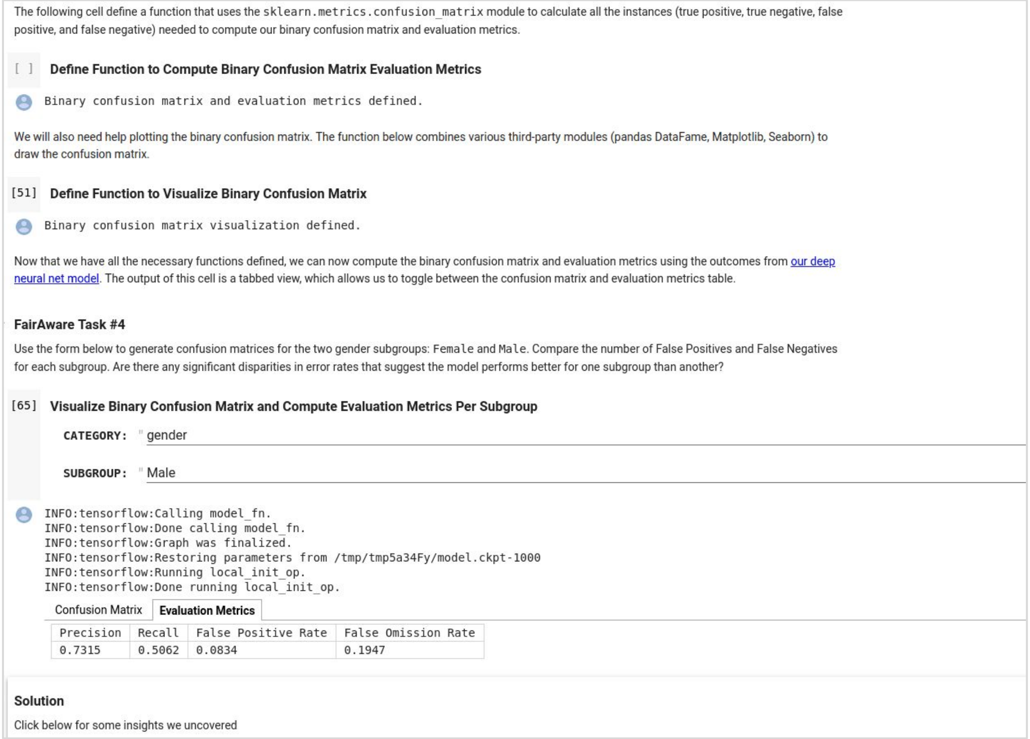


Machine Learning Crash Course

with TensorFlow APIs

Google's fast-paced, practical introduction to machine learning

START CRASH COURSE
VIEW PREREQUISITES
Neural Network
A model that is composed of layers (min 1 hidden) consisting of simple connected units or neurons fol
$F(x) = \frac{1}{1+e^{-x}}$
$w_1x_1 + w_2x_2 + w_3x$
Converts Sum to a 0 and



The following cell define a function that uses the `sklearn.metrics.confusion_matrix` module to calculate all the instances (true positive, true negative, false positive, and false negative) needed to compute our binary confusion matrix and evaluation metrics.

[ ] **Define Function to Compute Binary Confusion Matrix Evaluation Metrics**

```
Binary confusion matrix and evaluation metrics defined.
```

We will also need help plotting the binary confusion matrix. The function below combines various third-party modules (pandas DataFame, Matplotlib, Seaborn) to draw the confusion matrix.

[51] **Define Function to Visualize Binary Confusion Matrix**

```
Binary confusion matrix visualization defined.
```

Now that we have all the necessary functions defined, we can now compute the binary confusion matrix and evaluation metrics using the outcomes from our deep neural net model. The output of this cell is a tabbed view, which allows us to toggle between the confusion matrix and evaluation metrics table.

FairAware Task #4

Use the form below to generate confusion matrices for the two gender subgroups: `Female` and `Male`. Compare the number of False Positives and False Negatives for each subgroup. Are there any significant disparities in error rates that suggest the model performs better for one subgroup than another?

[65] **Visualize Binary Confusion Matrix and Compute Evaluation Metrics Per Subgroup**

CATEGORY: " gender

SUBGROUP: " Male

```
INFO:tensorflow:Calling model_fn.
INFO:tensorflow:Done calling model_fn.
INFO:tensorflow:Graph was finalized.
INFO:tensorflow:Restoring parameters from /tmp/tmp5a34Fy/model.ckpt-1000
INFO:tensorflow:Running local_init_op.
INFO:tensorflow:Done running local_init_op.
```

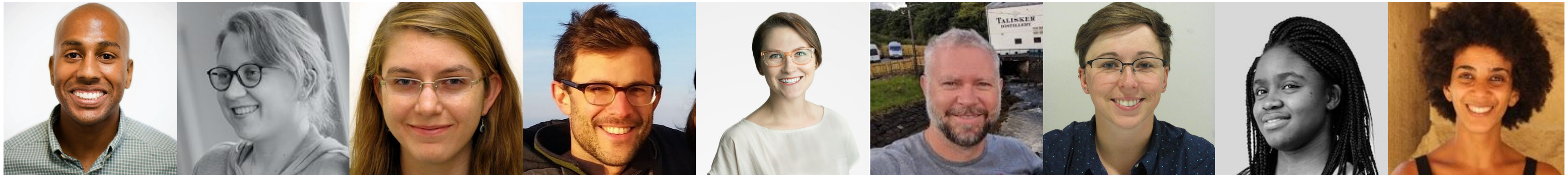
Precision	Recall	False Positive Rate	False Omission Rate
0.7315	0.5062	0.0834	0.1947

Solution

Click below for some insights we uncovered

developers.google.com/machine-learning/crash-course/fairness/

Thank You & Acknowledgements



andrewzaldivar@google.com, mmitchellai@google.com, simonewu@google.com,
parkerbarnes@google.com, lucyvasserman@google.com, benhutch@google.com,
espitzer@google.com, deborah.raji@mail.utoronto.ca, tgebru@google.com

joyab@mit.edu, @shalinianandа1 (Twitter),
ljames@google.com, @shiraamitchell (Twitter)