Critique: Model Cards for Model Reporting

Presenters: Ronghao Zhang
Wenting Song

Key Takeaways

- Proposed template for appropriate documentation
    - Spotlight importance of documentation accompanying released machine learning models
    - Handle mismatch between a model’s intended purpose and the way it is actually used.
- **Model Cards** are short documents that go alongside publicly available machine learning models to share information model users should know
    - intended usage, model training processes, evaluation results
- Evaluation across different demographic groups and communities to measure unintended bias
    - performance metrics are computed and published independently for different identity-based subsets of the evaluation data

Critique Zero

- More work to do with the simple framework…
    - Standardizing or formalizing model cards to prevent misleading representations
    - Refining the methodology of preparing model cards by considering different stakeholders.
    - Adding more criteria with the growing research in machine learning systems, e.g **reliability criteria** (likely shifts in environment, certificates of robustness, model verification). [1]

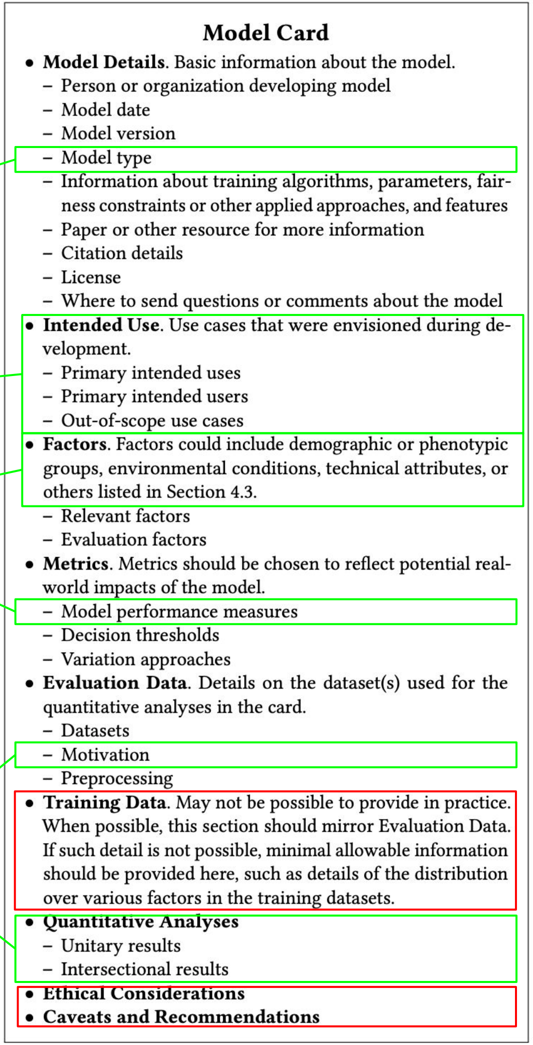
- Model Card for Perspective API's TOXICITY model
    - Continuing work on how to identify, measure, and mitigate unintended biases in models
    - Long term-work, lack of metrics or test data

[1]. Saria, Suchi, and Adarsh Subbaswamy. "Tutorial: Safe and Reliable Machine Learning." arXiv preprint arXiv:1904.07204 (2019).

Critique One - Messy Intro

Datasheets highlight characteristics of the data feeding into the model, we focus on trained model characteristics such as:

- Type of model
- Intended use cases
- Information about attributes for which model performance may vary
- Measures of model performance
  - Cultural
  - Demographic
  - Phenotypic groups
  - Domain-relevant conditions
  - Intersectional analysis combining two (or more) groups and conditions
- Motivation behind chosen performance metrics, group definitions, and other relevant factors.

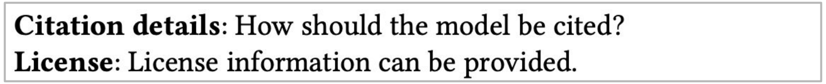


Model Card

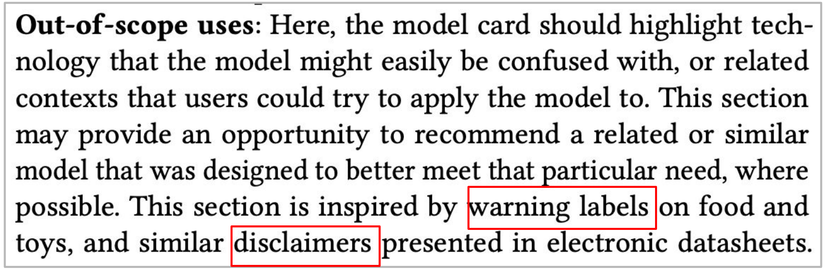
- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more 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**

Critique Two - Missing the Legal Sections

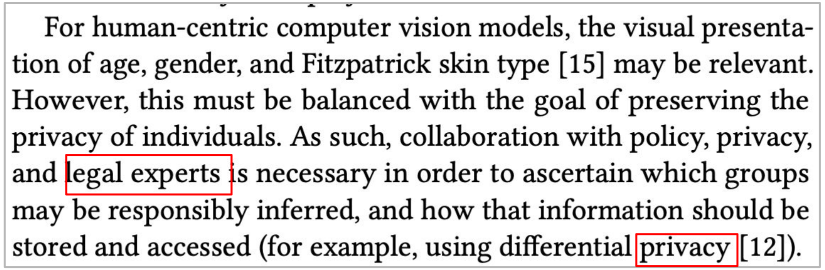
Multiple legal concerns were proposed in the model card sections:



Citation details: How should the model be cited?
License: License information can be provided.

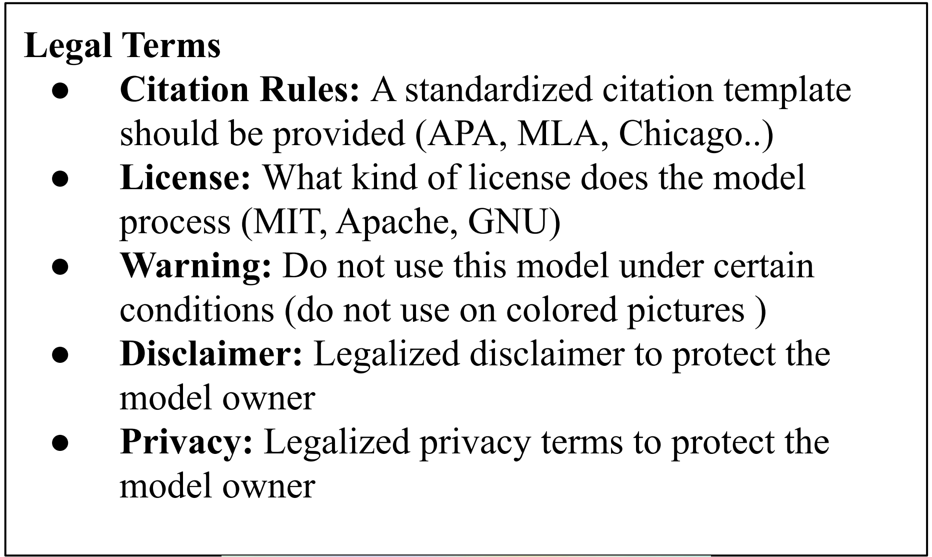


Out-of-scope uses: Here, the model card should highlight technology that the model might easily be confused with, or related contexts that users could try to apply the model to. This section may provide an opportunity to recommend a related or similar model that was designed to better meet that particular need, where possible. This section is inspired by warning labels on food and toys, and similar disclaimers presented in electronic datasheets.



For human-centric computer vision models, the visual presentation of age, gender, and Fitzpatrick skin type [15] may be relevant. However, this must be balanced with the goal of preserving the privacy of individuals. As such, collaboration with policy, privacy, and legal experts is necessary in order to ascertain which groups may be responsibly inferred, and how that information should be stored and accessed (for example, using differential privacy [12]).

It seems necessary to have a seperate section dedicated to legal components:



Legal Terms

- **Citation Rules:** A standardized citation template should be provided (APA, MLA, Chicago..)
- **License:** What kind of license does the model process (MIT, Apache, GNU)
- **Warning:** Do not use this model under certain conditions (do not use on colored pictures)
- **Disclaimer:** Legalized disclaimer to protect the model owner
- **Privacy:** Legalized privacy terms to protect the model owner

Critique Four - Moral Dilemmas

Model cards are just one approach to **increasing transparency** between developers, users, and stakeholders of machine learning models and systems.

The usefulness and accuracy of a model card relies on the **integrity of the creator(s)** of the card itself.

Homer, tell the truth in model cards!
A little unfairness isn't going to kill anyone!

Thank You!