DEVELOPING A CONCEPTUAL FRAMEWORK FOR ANALYZING PEOPLE IN UNSTRUCTURED DATA

Mark Díaz, Sunipa Dev, Emily Reif, Remi Denton, & Vinodkumar Prabhakaran
Google Research
{markdiaz,sunipadev,ereif,dentone,vinodkpg}@google.com

ABSTRACT

Unstructured data used in foundation model development is a challenge for systematic analyses to make data use and documentation decisions. From a Responsible AI perspective, these decisions often rely upon understanding how people are represented in data. We propose a framework to guide analysis of human representation in unstructured data and identify downstream risks.

1 INTRODUCTION

Insights from systematic analysis of datasets can identify potential harms and inform interventions to mitigate risk. At the same time, the large, unstructured nature of these datasets poses significant challenges for conducting analyses required to make development, documentation, and use decisions. The open-ended potential for downstream use, means that risks are wide-ranging and sometimes lack clear methods of evaluation (Weidinger et al., 2021). Prior systematic fairness audits have often focused on data labels and utilized aggregated and disaggregated analyses to identify class imbalances (e.g., (Saleiro et al., 2018; Kearns et al., 2018; Kleinberg et al., 2016; Friedler et al., 2019)). Despite increased scrutiny of large unstructured datasets (Birhane et al., 2021; Dodge et al., 2021), methods of analysis remain less robust and there is little guidance for applying them in practice to fairness workflows. As a result, practitioners apply analyses ad hoc, continue to use what they have used before, or miss relevant analyses (Madaio et al., 2022; Heger et al., 2022).

We offer a conceptual framework to standardize workflows for analyzing unstructured data. The framework (shown in Appendix A) focuses on social representation of people in data, including the data features that indicate social identity and influence the representation of different social groups. The framework is grouped according to *who* is in the data, *what* is in the data, and associations between the two. Thus, the structure and core analytical questions are modality-agnostic and extensible to new modality combinations. The framework does not strictly prescribe analysis implementations– rather, it guides responsible AI (RAI) workflows for data evaluation, documentation, and risk mitigation.

2 BACKGROUND

2.1 DATASET TRANSPARENCY AND DOCUMENTATION

A growing body of scholarship in RAI focuses on increasing AI transparency for a variety of stakeholders. These range from developers who build on pre-trained models to system end-users who may be subject to algorithmic decision making (Lima et al., 2022; Wagner et al., 2020). At the dataset level, transparency highlights critical information about the contents of a dataset as well as the processes that underpin how a dataset was created. To this end, a range of work brings structured approaches to documenting both dataset content and development processes (Bender & Friedman, 2018; Gebru et al., 2021; Dodge et al., 2021; Díaz et al., 2022; Hutchinson et al., 2021; Rostamzadeh et al., 2022; Srinivasan et al., 2021; Pushkarna et al., 2022). At the same time, datasets underpinning training and testing have been at the center of various tensions connected to privacy, consent, unfair system performance, representational harms, and harmful applications (Paullada et al., 2021). Thus, prominent ML datasets have been subject to close scrutiny, with empirical examinations and audits uncovering a range of problematic content that itself is harmful (e.g., copyright violations; representational harms such as misgendering) or that can lead to downstream

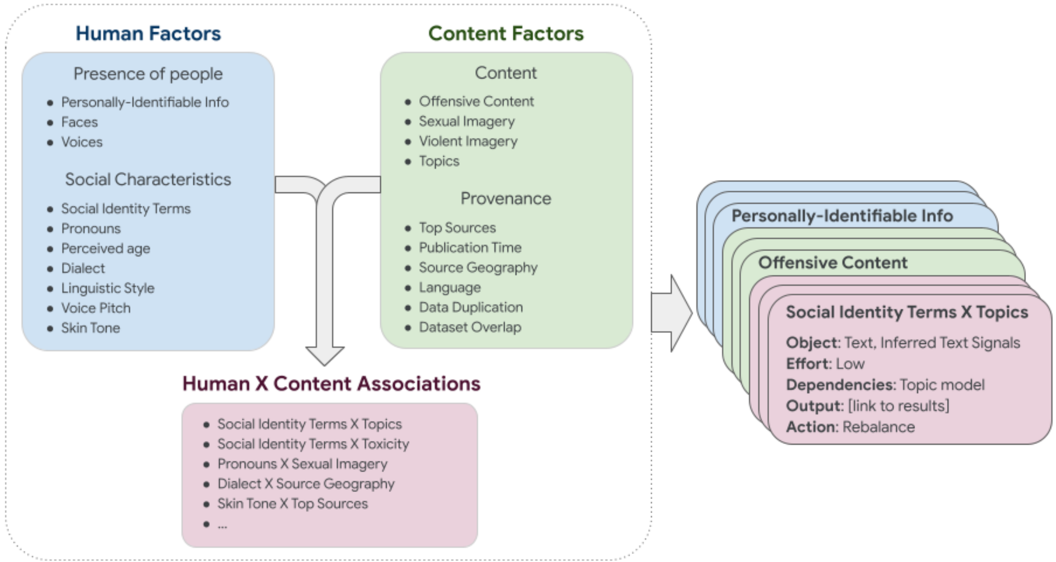

Figure 1: Framework conceptual structure. The combination of Human Factor and Content Factor analyses provided disaggregated associations. Analyses documentation and results are on the right.

harms. Dataset audits can close documentation gaps (Dodge et al., 2021), inform data filtering or re-balancing decisions (Russakovsky et al., 2014), and even lead to the deprecation of datasets (e.g., MegaFace (Kemelmacher-Shlizerman et al., 2016); Tiny Images (Torralba et al., 2008)).

2.2 Standardizing Responsible AI Workflows

Evaluating data in a structured way and communicating results to stakeholders remains an important challenge for RAI. Data work often takes a backseat to work focused on developing state of the art models and algorithms (Sambasivan et al., 2021b). In addition, current approaches to data documentation are "largely ad hoc and myopic in nature" (Heger et al., 2022) and practitioners face difficulty in understanding why documentation is needed, how best to document, and, ultimately, what to document (Chang & Custis, 2022). A range of development toolkits and checklists have been proposed to address these challenges, including documentation frameworks such as Data and Model Cards (Gebru et al., 2021; Mitchell et al., 2019; Pushkarna et al., 2022), internal auditing frameworks (Raji et al., 2020b), and impact assessment frameworks (Schiff et al., 2020). We configure our framework around downstream risks to support practitioners to evaluate for documented risks.

3 Framework

Organizing prior individual audits, we present a principled framework (see Appendix C) of RAI data evaluations. It supports goals ranging from dataset development to third-party audits. Our framework identifies components that guide the operationalization of analyses and the interpretation of results.

3.1 Framework analyses

We organize the framework around high-level, human-centered considerations in data: namely, *Who* is in the data, *What* is in the data, and *How* are the two associated? This structure also allows the analyses to focus on data questions at different levels of complexity with respect to corresponding downstream harms, as well as to prevent an over-focus on optimizing isolated analyses or metrics.

Our framework is designed to be general-purpose and extensible, thus it is not exhaustive of every single possible analysis within each section. While we use text, image, and image-text datasets as references for developing and describing the framework, it can be adapted to other modalities with appropriate changes. For instance, the dialect analysis in text could be adjusted to include elements of (or complemented with) analysis of accent in speech data. Analyses can also be modified, added, or removed as the field's collective sociotechnical understanding about relevant social biases evolve over time, while preserving the overall framework structure. For example, salient social identity term lists may iteratively change as best practices respond to social shifts, or as global socio-cultural contexts are increasingly integrated into RAI considerations. Analyses can also be updated alongside our understanding of salient social risks, the human social characteristics they are connected to, and our technical means of analyzing them. However, the motivating questions remain stable.

3.1.1 WHO IS IN THE DATA?

In asking who is in the data, we consider several human factors of data that include measuring the presence of people in data along with social characteristics.

Presence of People: These analyses tally whether individuals or identifying information appear in data. This includes calculations of personally-identifiable information and face or person detection. Extending to new modalities, the analyses implicitly ask which data characteristics can indicate the presence of a person, such as faces or bodies in visual data or voice in audio data. Results guide more focused, follow-up analyses that assess depictions of social groups.

Social Characteristics: These analyses center on data characteristics that are often associated with social identity and may be used as proxies for social identity. Some proxies appear directly in data, such as pronouns, while others, such as perceived age or gender expression in images, must be inferred, frequently using predictive methods (e.g., Lanitis et al. (2004)). These include analyses of dialect, linguistic style, skin tone, and voice pitch. Social characteristic analyses provides insight into the over- and under-representation of specific social groups, which has been associated with disparities in performance (Wilson et al., 2019; Buolamwini & Gebru, 2018) and general problems for class prediction (Johnson & Khoshgoftaar, 2019). Because these characteristics are social in nature, their measurement must be adapted to local context and time. For example, social identity terms vary across social and cultural contexts, meaning static identity term list cannot be exhaustive.

3.1.2 WHAT IS IN THE DATA?

The second grouping of analyses focuses on content that may influence human representation.

Content: This group of analyses is focused on content characteristics that relate to harmful or undesirable outcomes that are independent of specific people or social groups. Analyses include calculating the distribution of topics in text, as well as sexual content in images. Topic distribution provides a bird's-eye view of the composition of the data and can give an indication of sexually explicit or sensitive topics contained in a dataset. Topic distributions can give clues to subtle downstream biases. For example, models trained primarily on news data have been shown to exhibit biases against particular country names and professions (Huang et al., 2019).

Provenance: Data sources and metadata can indicate the values and norms likely to be contained in data, such as the geographic distribution of sources and publication dates. Geographic, cultural, and social representation in data can have implications for downstream models. For example, image classifiers trained on datasets sourced predominantly from western countries have lower rates of accuracy when applied to images from non-western countries (Shankar et al., 2017). Data recency can have particular impacts on models supporting low-resource languages, which can disproportionately rely on religious or historical texts due to data scarcity (e.g., (Ahmadi & Masoud, 2020)).

3.1.3 HUMAN $\times$ CONTENT ASSOCIATIONS

The final section focuses on *how* people are depicted. Associations disaggregate analyses within and across modalities, such as social identity terms and topics in text or occurrences between image objects and text tokens in multimodal datasets. Associations can reveal stereotype-aligned correlations, which can propagate exclusionary norms (Dev et al., 2021; Weidinger et al., 2021; Zhao et al., 2018; Hendricks et al., 2018). While highly specific analyses can be run (e.g., queer depictions in Spanish-language medical literature from a specific year), the structure of the framework facilitates beginning with the most general question (i.e., are people depicted?) followed by more specific inquiries (e.g., with which other data do people most often occur?) to provide clearer starting points.

Analysis Goals	
Dataset Development	Developing a dataset for training or evaluation through new data collection and/or adaptation of existing datasets
Use Decisions	Making decisions regarding appropriate use of a dataset, whether for training or evaluative purposes
Model Understanding	Investigating potential roots of or explanations for model behavior
Auditing	Auditing a dataset to fill documentation gaps, ensure legal or institutional compliance, or to foster greater public awareness

Table 1: A non-exhaustive list of data analysis goals.

3.2 Framework Components

Next, we outline additional framework components that structrure analysis planning and results reporting. The **Output** and **Action** fields are provided to capture the results of a given analysis and any mitigation actions. The Taking Action section discusses in more depth the process of making mitigation decisions. Each analysis includes additional fields to support planning:

Analysis Object indicates whether an analysis is calculated on data directly (e.g., tokens in text) or if it applies to an inference produced by an intermediate classifier (e.g., inferred document topic). This highlights which analyses are dependent on predictive models and susceptible to biases that those models may themselves exhibit. The distinction between "Image" and "Inferred image signals" is particularly important since few analyses in the framework are applied to image data directly.

Effort indicates rough time and cost of an analysis based on current techniques and tooling.

Dependencies indicates intermediate resources needed to conduct an analysis, classifiers which produce inferred signals. While the framework does not dictate a single, required implementation for any analysis, we point to example classifiers and term lists that may be used. Moreover, some dependencies, such as term lists, should ideally draw from qualitative insights to localize evaluations.

4 Taking Action

While this framework can be used to discover new biases, it is intended to help practitioners overcome challenges in pursuing evaluations and mitigations motivated by institutional policy, previously-documented data biases (e.g., gender bias), and regulatory concerns. These challenges include identifying risks and harms AI systems can generate, and a systematic approach for applying known fairness evaluations across a broad range of products and systems (Madaio et al., 2022; Heger et al., 2022). The framework is not meant to be exhaustively implemented for every use case, since all possible association analyses would produce an intractably large number of results. Rather analyses should be rooted in high priority institutional policy concerns and regionally-specific protected groups. Moreover, which mitigation actions to take depends on context. An exhaustive review of mitigation actions is beyond the scope of this work, however we describe considerations for follow-up action. Key questions narrow the scope of actions to be considered, making planning more tractable:

- What are the planned deliverables of the data effort (e.g., training or evaluation data)?
- What are the primary goals the analyses will support (e.g., making use decisions)?
- To what extent can development steps or data be intervened upon (e.g., data collection)?

Dataset Purpose: The framework can be applied to datasets (including pre-training or fine-tuning data) as well as model-generated data. Mitigation actions depend on the purpose of the dataset. For example, for pre-training data, it may be unclear how changes to data distributions will impact model performance, potentially making other mitigations more desirable. In contrast, benchmark data stands to be used as a standard measure, thus additional resources may be more easily justified.

Analysis Goals: Different analysis goals bring attention to different actions (shown in Table 1). For example, developing a dataset from web-scraped data raises decisions to collect additional data or adjust filtering criteria, whereas auditing a third-party dataset raises documentation and use decisions.

Development Phase: Development phases afford different mitigation actions (shown in Table 2). For example, during data collection, toxic content biased across social groups might be addressed

Dev. Phase	Actions	Description
Data Collection/ Processing	Addition	Rebalancing distributions within or across categories
	Removal	Filtering data to remove unwanted content
	Augmentation	Augmenting data (e.g., data tagging to tune model behavior)
	Flagging	Flagging results for further evaluation or documentation
	Non-Use	Not using the dataset
Model Evaluation	Addt'l Benchmarking	Selection of additional evaluation benchmarks
	Benchmark Creation	Development of benchmarks to evaluate new concerns
Documentation	Warning	Documentation of general or use case-specific limitations
	Non-Use	Documentation of cases where the data should not be used
Packaging and Release	Licensing	Development of licensing and terms of use specifications
	Access	Development of limited access policies

Table 2: A non-exhaustive list of actions that may be taken to address social risks identified in data.

by modifying the dataset or by additional model evaluations. Alternatively, documentation can flag concerns for public consumption, or data release planning can be adjusted.

5 Assessing Harm in Data

Characterizing how people are represented in data is a necessary part of identifying risks. Yet, RAI lacks systematized guidance to do so, and analyzing how people are depicted is challenging because "good" social representation changes with context. Notably, our framework has no canonical list of social identities nor an exhaustive list of evaluations for a given social identity. This is because social identity is unstable in nature (Hanna et al., 2020), and the axes along which discrimination occurs are culturally specific (Sambasivan et al., 2021a). As Chasalow & Levy (2021) posit, representativeness is both time and place specific; the social categories we attend to are shaped by normative assumptions about what should be measured, whether and how those categories map to signals in data, and the existence of language for a social category. For example, Andrews et al. (2022) suggest that terms used today to analyze disability representation would likely differ from terms 30 years ago.

We echo others advocating for more attention to data work (Sambasivan et al., 2021b), including the growing focus on data-centric AI (DCAI) (Jarrahi et al., 2022). Building from DCAI's focus on understanding the data throughout ML development, our framework sets a foundation for systematically analyzing social risks in data. At the same time, dataset evaluations are just one component of RAI evaluation. As Hooker (2021) argues, algorithm design choices, such as optimization for privacy guarantees, compression techniques, and even learning rate can contribute to model biases. As a result, dataset evaluations must be considered alongside other approaches to mitigating risk and harm.

5.1 Supporting RAI Goals

RAI development requires data analyses that complement existing RAI processes while adapting to sociotechnical risks that are contextually determined. In response, the framework eschews automated test beds or fixed implementations (e.g., specific term lists or classifiers) and, instead, aims to standardize workflow planning. This includes structured guidance to repeat and localize analyses of human depictions in data. A primary motivation for this framework is to analyze data used for foundation models. High model training costs limit opportunities to run comprehensive studies to identify which mitigation strategies best support fairness and model performance. For RAI, this means making mitigation decisions with limited information about specific impacts. This challenge is exacerbated by data cascades, which can compound to produce out-sized, negative outcomes (Sambasivan et al., 2021b). Yet, the range of potential downstream risks warrants proactive decision making. While intervening on training is difficult when downstream applications are unclear, the framework can also be used for multimodal evaluations of fine-tuning data or model-generated data.

6 Conclusion

RAI practitioners must identify risks and take appropriate action to mitigate them, often with limited information about downstream model use. In response, we build from critical dataset audits and frameworks to offer a conceptual framework to evaluate human representation in unstructured data.

REFERENCES

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Sina Ahmadi and Mariam Masoud. Towards machine translation for the kurdish language. *arXiv preprint arXiv:2010.06041*, 2020.

Miltiadis Allamanis. The adverse effects of code duplication in machine learning models of code. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pp. 143–153, 2019.

Erin E Andrews, Robyn M Powell, and Kara Ayers. The evolution of disability language: Choosing terms to describe disability. *Disability and Health Journal*, 15(3):101328, 2022.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

Jack Bandy and Nicholas Vincent. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. SIGCIS, 2017. URL `http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf`.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

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. *Advances in neural information processing systems*, 29, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

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

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

Jiyoo Chang and Christine Custis. Understanding implementation challenges in machine learning documentation. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–8. 2022.

Kyla Chasalow and Karen Levy. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 77–89, 2021.

Samantha Cole. These nudes do not exist and i don't know why this startup does either. In *Vice*, 2020. https://www.vice.com/en/article/qjdx7w/these-nudes-do-not-exist-and-i-dont-know-why-this-startup-does-either.

Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5034–5050, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.411. URL `https://aclanthology.org/2021.emnlp-main.411`.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2342–2351, 2022.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 501–512, 2020.

Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29, 2022.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 771–787, 2018.

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 560–575, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445918. URL `https://doi.org/10.1145/3442188.3445918`.

Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The principles of data-centric ai (dcai). *arXiv preprint arXiv:2211.14611*, 2022.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4873–4882, 2016.

Mahrukh Khan, Muhammad Atif Tahir, and Zeeshan Ahmed. Detection of violent content in cartoon videos using multimedia content detection techniques. In *2018 IEEE 21st International Multi-Topic Conference (INMIC)*, pp. 1–5. IEEE, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Andreas Lanitis, Chrisina Draganova, and Chris Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2103–2113, 2022.

Alexandra Sasha Luccioni and Joseph D Viviano. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*, 2021.

Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–26, 2022.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL `https://doi.org/10.1145/3287560.3287596`.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100336. URL https://www.sciencedirect.com/science/article/pii/S2666389921001847.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 145–151, 2020a.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 33–44, 2020b.

Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. Healthsheet: development of a transparency artifact for health datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1943–1961, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 315–328, 2021a.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021b.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. Principles to practices for responsible ai: Closing the gap. *arXiv preprint arXiv:2006.04707*, 2020.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. Artsheets for art datasets. 2021.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

Ben Wagner, Krisztina Rozgonyi, Marie-Therese Sekwenz, Jennifer Cobbe, and Jatinder Singh. Regulating transparency? facebook, twitter and the german network enforcement act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 261–271, 2020.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

Appendix

Table of Contents

A Framework Introduction **11**
A.1 Guiding Considerations 11

B Analysis Overview **13**
B.1 Human Factors 13
B.2 Content Factors 13
B.3 Human X Content Associations 13
B.4 Analysis Dependencies 14

C List of Framework Analyses **15**
C.1 Human Factors 15
C.2 Content Factors 18
C.3 Human X Content Associations 21

A Framework Introduction

Here we present a framework overview, detailed for text, image, and text-image application. Rather than an exhaustive list of every possible analysis, the Human-Content Associations section shows only social identity analyses for each modality. At the end of the overview we provide a list of example tools that can be used to meet the dependency requirements of the analyses. The full framework features recommended risk mitigation actions, however, in practice, mitigation steps should depend on development context and goals. The below tables list common analysis goals and mitigation actions and the following section on Guiding Considerations lists questions and prompts to guide framework use.

Analysis Goals	
Dataset Development	Developing a dataset for training or evaluation through new data collection and/or adaptation of existing datasets
Use Decisions	Making decisions regarding appropriate use of a dataset, whether for training or evaluative purposes
Model Understanding	Investigating potential roots of or explanations for model behavior
Auditing	Auditing a dataset to fill documentation gaps, ensure legal or institutional compliance, or to foster greater public awareness

Table 3: A non-exhaustive list of data analysis goals.

A.1 Guiding Considerations

- **What are the planned deliverables of the data being analyzed?**
  The planned outputs help to inform decisions regarding resource investment in analyses and potential mitigations. For example, a novel benchmark dataset may warrant more extensive analysis than a pre-training dataset because it will be repeatedly use to judge the performance of many models. Similarly, a fine-tuning dataset may be the subject of more modification than a pre-training dataset due to its smaller size and direct role in shaping model compliance with a range of fairness and policy constraints.
- **What are the primary goals the analyses will support?**
  Compliance audits and use decisions may only warrant documentation, or packaging and release actions; model understanding may simply be used to help diagnose model behavior

Dev. Phase	Actions	Description
Data Collection/ Processing	Addition	Rebalancing distributions across an entire dataset or within specified categories with additional (potentially synthetic) data
	Removal	Filtering data to remove unwanted content
	Augmentation	Augmenting data, such as through data tagging (Anil et al., 2023) to allow a model to learn undesirable content while controlling its production downstream
	Flagging	Flagging analysis results for further downstream evaluation or documentation
	Non-Use	Not using the dataset, for example if applying analyses to different candidate datasets to decide which to use
Model Evaluation	Addt'l Benchmarking	Selection of additional evaluation benchmarks
	Benchmark Creation	Development of benchmarks to evaluate new concerns
Documentation	Warning	Documentation of general or use case-specific limitations
	Non-Use	Documentation of cases where the data should not be used
Packaging and Release	Licensing	Development of licensing and terms of use specifications
	Access	Development of limited access policies

Table 4: A non-exhaustive list of actions that may be taken to address social risks identified in data.

(i.e., no mitigation action on data); whereas the development of a new dataset can involve many mitigation actions.

- **To what extent can completed or planned development steps be revisited or modified**
  Additional analyses and mitigations require time and resources, as does experimentation to compare which actions are most effective for a given project. For example, budget may be limited for additional data collection, while data filtering or tagging may still be possible.
- **To what extent is the dataset mutable? (i.e., can data be added, removed, or changed?)**
  The mutability of a dataset narrows the available mitigation actions and the degree to which data can be filtered or rebalanced.
- **Can new or additional model benchmarks be run?**
  When developing a new model, an ultimate consideration is whether biases identified in data will persist through training and finetuning steps. This is known to be true for a number of biases discovered in certain datasets. For novel or recently discovered biases, relevant benchmarks may not exist or are not feasible to create to test downstream models for the bias. This may place emphasis on other mitigation actions, depending on the risk associated with the identified bias.
- **Will the data or a model trained on the data be released or made available for external use? If so, can the release terms (e.g., terms of use, licensing, policies) be modified?**
  The use of data and models released externally cannot be as carefully controlled as data and models used strictly within a single institution. Safeguards against risks may be handled through terms of use or licensing agreements in order to mitigate data or model limitations.
- **Are there institutional or legal requirements regarding what should be analyzed in the data?**
  Legal and institutional compliance are a first step for planning analyses and reporting results. These can be supplemented with parallel fairness analyses of additional social groups, or analyzed more in depth to make mitigation decisions (e.g., analyzing which data sources feature the strongest identified biases in order to make filtering decisions)
- **Are there additional fairness or bias concerns a priori?**
  Beyond compliance concerns, consider analyses that involve groups or data subjects that have been previously impacted by similar systems or within which the context the data or model will be released.
- **What signals are available for analyzing social characteristics in the data?**
  The feasibility of running a given analysis depends on the available data features. Aside from straightforward considerations, such as whether a dataset contains faces that can be detected, varied social characteristics may map to a single data feature (e.g., common surnames across disparate racial and ethnic groups) or may not easily map to any data features (e.g., religion and disability are two examples of social characteristics that are not always 'visible' in images).

B ANALYSIS OVERVIEW

B.1 HUMAN FACTORS

B.1.1 PRESENCE OF PEOPLE

- Personally-Identifiable Information
- Faces

B.1.2 SOCIAL CHARACTERISTICS

- Social Identity Terms
- Pronouns
- Hateful Terms in Text
- Dialect
- Perceived Social Identity in Images
- Hateful Symbols in Images

B.2 CONTENT FACTORS

B.2.1 CONTENT

- Offensive Speech
- Topics
- Sexual Imagery
- Violent Imagery

B.2.2 PROVENANCE

- Top Sources
- Source Geography
- Source Publication Time
- Language
- Data Duplication
- Dataset Overlap

B.3 HUMAN X CONTENT ASSOCIATIONS

Text

- Social Identity Terms X Top Tokens
- Social Identity Terms X Offensive Speech
- Social Identity Terms X Topic

Image

- Perceived Social Identity Features X Top Image Features
- Perceived Social Identity Features X Sexual Imagery
- Perceived Social Identity Features X Violent Imagery

Text-Image

- Social Identity Terms X Sexual Imagery
- Social Identity Terms X Violent Imagery

- Perceived Social Identity Features X Text-Image
- Perceived Social Identity Features X Sexual Imagery
- Perceived Social Identity Features X Violent Imagery
- Perceived Social Identity Features X Offensive Speech
- Perceived Social Identity Features X Topic

B.4 Analysis Dependencies

Below is a list of technical dependencies required for framework analyses along with example classifiers that can be used.

PII Detection: Google Data Loss Prevention API (InfoType Detector)

Face Detection: Google Cloud Vision API (Image Annotation)

Sexual Imagery: Google Cloud Vision API (Safe Search Annotation)

Violent Imagery: Google Cloud Vision API (Safe Search Annotation)

Object Classifier: Google Cloud Vision API (Object Localization)

Topic Classifier: Google Cloud Natural Language API (Content Categories)

Offensive Speech: Perspective API (Note: Not offensive speech, but Toxicity)

Language Classifier: Google Cloud Natural Language API (Text Annotation

Dialect Classifier: No comprehensive classification approaches.

Hateful Symbols: No currently reliable method

Social Identity Features: A number of published audits point to important limitations of social identity classification systems (e.g., Buolamwini & Gebru (2018) and Scheuerman et al. (2019) audits of gender classification for image, and Raji et al. (2020a) audit of gender and age classification for image), therefore we encourage research and results interpretation with individual usage contexts in mind.

C List of Framework Analyses

C.1 Human Factors

Analyses focused on identifying individuals, social identity groups, and cultural groups in data.

Presence of People

C.1.1 Personally Identifiable Information (PII)

Personally identifiable information allows for an individual's identity to be potentially inferred through direct or indirect means. It includes, but is not limited to, names, email addresses, addresses, and images or video with identifying characteristics. The presence of personally identifiable information in a dataset is not inherently problematic. However, even if a dataset is not intended for public use, publishing a model trained on a dataset containing personally identifiable information risks "leaking" that information through attack methods designed to recover specific training instances from a model Carlini et al. (2021). Generative models that have overfit to the training distribution may also risk generating outputs that constitute a breach in privacy.

Task: Identify data containing PII

Analysis Object: Text | Inferred Text Signals

Effort: Medium

Dependencies: PII Detection

Output: Proportion and list of data points containing PII

Action: Flag PII to assess potential privacy violations that arise for model tasks or as a result of dataset publication

C.1.2 People in Images

As a subcategory of personally-identifiable information, faces in datasets are at risk of "leaking" through attack methods designed to recover specific training instances from a model Carlini et al. (2021). The presence of PII is not inherently problematic, however privacy risks must be evaluated in the context of specific tasks.

Task: Calculate the proportion of images that depict people (e.g., via face detection)

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Face Detection

Output: Proportion of images that depict people

Action: Flag images of people

Social Characteristics

C.1.3 Social Identity Terms

Under-representation in datasets can contribute to representational harms Barocas et al. (2017), such as strong under-representation of darker-skinned individuals in datasets linked to under-performance in pedestrian detection Wilson et al. (2019) and data imbalances generally causing issues for class prediction Johnson & Khoshgoftaar (2019). Attention to social identity representation in datasets also helps to support intersectional analyses which inherently focus on relatively smaller intersections of data but require sufficient data points to conduct.

Task: Calculate proportion of text referencing different social identity groups, considering unitary and intersectional groups

Analysis Object: Text

Effort: Low

Dependencies: Social Identity Term List

Output: Frequency of text depicting unitary and intersectional groups

Action: Flag social identity representation

C.1.4 Pronoun Distribution

Pronoun distributions help provide a high-level snapshot of gender representation in datasets to evaluate their connection to model performance on gender bias benchmarks. Gender stereotypic associations have been well documented in NLP Bolukbasi et al. (2016); Caliskan et al. (2017) and mitigating imbalances in gendered terms in datasets has been shown to mitigate gender bias in benchmark performance Zhao et al. (2018).

Task: Count number of gender pronouns that occur

Analysis Object: Text

Effort: Low

Dependencies: Pronoun List

Output: Distribution of pronouns

Action: Flag pronoun distribution

C.1.5 Hateful Terms in Text

When training datasets contain derogatory language, slurs, or offensive language, there is a risk models trained on the data will apply these terms in an inappropriate manner, e.g. describing people with such terms. Several recent audits of image classification datasets and image-to-text datasets have uncovered the presence racial and ethnic slurs and other derogatory and hateful language Birhane et al. (2021); Birhane & Prabhu (2021); Crawford & Paglen (2021).

Task: Count number of hateful terms that occur

Analysis Object: Text

Effort: Low

Dependencies: Hateful Term List

Output: Distribution of hateful terms, the groups they reference, and their occurrences

Action: Flag hateful terms

C.1.6 DIALECT

Dialects vary across regions and cultures. Data that lacks manners of speaking or dialects prominently used by subgroups can lead to erasure of those dialects in resultant models, or categorization of non-dominant dialects as incorrect, low quality, or even offensive Sap et al. (2019).

Task: Calculate proportion of documents representing languages and dialects aimed to be supported in downstream applications

Analysis Object: Inferred Text Signals

Effort: High

Dependencies: Dialect Classifier

Output: Proportion of utterances representing different languages and dialects

Action: Flag dialect representation. Consider rebalancing strongly underrepresented dialects or qualifying downstream model capabilities.

C.1.7 SOCIAL IDENTITY IN IMAGES

Training datasets that under-represent certain social identity groups can produce models that exhibit poor performance when presented with novel depictions of those groups. For example, image classifiers trained on datasets sourced predominantly from western countries have lower accuracy when applied to images from non-western countries Shankar et al. (2017), and facial analysis systems trained on datasets that skew heavily towards lighter skinned subjects have higher error rates when applied to images depicting faces with darker skin tones Buolamwini & Gebru (2018).

Task: Calculate proportion of images depicting social identity groups, considering unitary and intersectional groups

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Perceived Social Identity Classifiers

Output: Proportion of images depicting unitary and intersectional groups

Action: Flag social identity representation

C.1.8 HATEFUL SYMBOLS IN IMAGES

Visual imagery associated with hate groups (e.g., swastikas) may go undetected or undocumented in datasets. Identifying hateful content across modalities can help mitigate the risk of unintentionally generating hateful or offensive content. It is important to note that offensiveness of imagery is often dependent on cultural context, and hence any action to address this issue should account for this.

Task: Calculate proportion of images that depict known hateful symbols or text

Analysis Object: Inferred Image Signals

Effort: Not yet possible by automated methods

Dependencies: Hateful Symbol Classifier

Output: Proportion of images depicting known hateful symbols or text

Action: Flag images with hateful symbols

C.2 Content Factors

Analyses focused on identifying content that may heighten the sensitivity of human depictions.

Content

C.2.1 Offensive Speech Distribution

Identifying the distribution of offensive content in training data can help to understand the relationship between offensive speech in training data and model outputs as well as help mitigate the risk of unintentionally generating toxic content. Analyses of OWTC and OpenAI-WT suggest that toxic language generation has links to toxic language in web text data and toxic training examples that may be harder for a model to "forget" Gehman et al. (2020).

Task: Calculate the distribution of toxicity

Analysis Object: Inferred Text Signals

Effort: Low

Dependencies: Offensive Speech Classifier or similar (e.g., Perspective API[a])

Output: Histogram of offensiveness across documents

Action: Flag toxic content

[a] www.perspectiveapi.com

C.2.2 Topic Distribution

Topic distribution can provide a birds-eye view of the composition of the dataset across various topics such as sports, games, finance, etc. More importantly, it can also give an indication of sexually explicit or sensitive topics they contain, and potential biases they bring. For example, models trained primarily on news data have been shown to exhibit biases against particular country names and professions Huang et al. (2019). Assessments of data distributions across languages may highlight potential weaknesses across languages in generative tasks.

Task: Calculate the distribution of topics

Analysis Object: Inferred Text Signals

Effort: Low

Dependencies: Topic Classifier (e.g., Google Cloud Content Categories[a])

Output: Histogram of topic distribution

Action: Flag dominant topics

[a] https://cloud.google.com/natural-language/docs/categories

C.2.3 SEXUAL IMAGERY

Many datasets have been identified as unintentionally containing sexually explicit content. For example, audits of image datasets such as Imagenet, TinyImage, and LAION-400M found pornographic and explicit imagery, predominantly depicting women Birhane et al. (2021); Birhane & Prabhu (2021). While some datasets may be explicitly curated to contain sexually explicit content, unintentional inclusion of such content can risk accidental generation of sexually explicit content by models trained on the data. Sexually explicit image content also raises potential ethical concerns regarding data sourcing since much publically available pornographic content has been stolen from sex workers or captured in a non-consentual manner Cole (2020).

Task: Calculate the proportion of images that depict sexual content

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Visual Content Classifier (e.g., Cloud Vision API[a])

Output: Proportion of images depicting sexual content

Action: Flag sexual imagery. Consider filtering sexual imagery depending on downstream tasks and to avoid problematically-sourced dataset content

[a] `https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation`

C.2.4 VIOLENT IMAGERY

Like sexual imagery, violent content may feature in datasets as a part of efforts to moderate content, such as in the detection of cartoon content that may be inappropriate for children Khan et al. (2018). For general purposes, inadvertent generation of violent content may be undesirable or counter to institutional or platform policies.

Task: Calculate the proportion of images that depict violent content

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Visual Content Classifier (e.g., Cloud Vision API[a])

Output: Proportion of images depicting violent content

Action: Flag violent imagery. Consider filtering violent imagery depending on downstream tasks

[a] `https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation`

Provenance

C.2.5 TOP SOURCES

The top source URL domains provide a high-level indication of the range of content most represented in web-scraped data. This, in turn, provides insight into document content, such as language and cultural content represented, as well as the prevalence of machine-generated text Dodge et al. (2021).

Task: Count the top websites from which tokens were collected

Analysis Object: Metadata

Effort: Low

Dependencies: None

Output: List of the top websites ordered by tokens collected

Action: Flag top URLs

C.2.6 Source Geographic Spread

Source URL domains provide an indication of geographic, cultural, and social representation. When training datasets have strong cultural skews, a resulting model can exhibit poor performance when presented with culturally or regionally specific depictions. For example, image classifiers trained on datasets sourced predominantly from western countries have lower rates of accuracy when applied to images from non-western countries Shankar et al. (2017).

Task: Calculate the known proportions of total tokens from different geographic regions (e.g., as determined through country-specific top-level domains such as, .uk, .de, .gh, etc.)

Analysis Object: Metadata

Effort: Low

Dependencies: None

Output: List of the top country-level domains with the proportion of data sourced from each

Action: Flag geographic spread. Consider rebalancing if regions relatively underrepresented in the dataset are prioritized for downstream uses.

C.2.7 Source Data Publication Time

The relevance of some data can change over time. For example, assessments of truth or offensiveness of a statement can change in relation to when it is evaluated. Data recency may have impacts on models supporting low-resource languages, which can disproportionately rely on religious or historical texts due to data scarcity Ahmadi & Masoud (2020).

Task: Identify the dates (years) of publication for data obtained

Analysis Object: Metadata

Effort: Low

Dependencies: None

Output: Histogram of data published per year

Action: Flag publication time. Consider rebalancing if prioritizing tasks likely to be relatively sensitive to data publication time (e.g., Q&A factuality; use of contemporary slang in generative language tasks)

C.2.8 Language

Languages vary across regions and cultures, however work in NLP frequently focuses on English, even when there is data available for other languages Hovy & Prabhumoye (2021). As with under-representation of data and classes in other contexts, poorly represented languages in a language dataset may lead to worsened performance of a downstream model when applied to that langauge.

Task: Calculate proportion of documents representing languages aimed to be supported in downstream applications

Analysis Object: Text | Inferred Text Signals

Effort: Low

Dependencies: Language Classifier (e.g., Google Cloud Natural Language API)

Output: Proportion of utterances representing different languages

Action: Flag language representation. Consider rebalancing strongly underrepresented languages or qualifying downstream model capabilities.

C.2.9 Data Duplication

Duplicate text has been identified in a number of datasets (e.g., Bandy & Vincent (2021)) and has been shown to worsen model performance on some tasks Allamanis (2019) as well as induce model memorization Lee et al. (2021), which can carry privacy risk Carlini et al. (2021).

Task: Calculate the proportion of duplicated data points

Analysis Object: Text | Image

Effort: Low

Dependencies: None

Output: Proportions of duplicated data points

Action: Filter duplicate data to reduce memorization and privacy risks and improve model performance

C.2.10 Dataset Overlap

There is growing awareness and concern of test set contamination, i.e. presence of training instances in the test set, especially with large internet sourced datasets. Explicitly examining and mitigating any overlap in training and evaluation splits is increasingly commonplace (e.g., Trinh & Le (2018); Brown et al. (2020).

Task: Identify overlaps between the present dataset and relevant datasets and benchmarks

Analysis Object: Text | Image

Effort: Medium [Dependencies:] None

Output: For benchmark datasets identified, list of percent overlap with the current dataset based on exact matches of target text

Action: Filter dataset overlaps to preserve the validity of benchmark results

C.3 Human X Content Associations

Disaggregated analyses of how content is associated with human factors. Listed below are a selection of all possible associations between Human Factors and Content Factors. Disaggregated analyses enable evaluation of stereotypical or otherwise harmful associations between depictions of people and dataset content. For example, prior research has identified stereotype-aligned associations between gender and activities in image datasets Zhao et al. (2018); Hendricks et al. (2018), associations between muslims and violence in language models Abid et al. (2021), and associations between women of color and explicit content in text datasets Luccioni & Viviano (2021) and image-text datasets Birhane et al. (2021).

Text

C.3.1 Social Identity Terms X Top Tokens

Analysis Object: Text

Task: For social identity terms in text, calculate top word co-occurrences for each

Effort: Low

Dependencies: None

Output: List of most frequent social identity terms and their respective top co-occurrences, demarcating identity terms (e.g., woman, Muslim, trans)

Action: Flag associations. Consider rebalancing harmful associations that emerge in model evaluations

C.3.2 SOCIAL IDENTITY TERMS X TOPIC

Task: Calculate the distribution of topics within text, disaggregated by social identity terms or inferred social identity signals

Analysis Object: Inferred Text Signals + Social Identity Term List

Effort: Low

Dependencies: Topic Classifier (e.g., Google Cloud Content Categories[a])

Output: Distribution of topics disaggregated by social identity

Action: Flag disproportionate associations between social identity terms and stereotypical or sensitive topics

[a] https://cloud.google.com/natural-language/docs/categories

C.3.3 SOCIAL IDENTITY TERMS X OFFENSIVE SPEECH

Task: Calculate the distribution of toxicity within text, disaggregated by social identity terms or inferred social identity signals

Analysis Object: Inferred Text Signals + Social Identity Term List

Effort: Low

Dependencies: Offensive Speech Classifier or similar (e.g., Perspective API[a]

Output: Distribution of toxicity, disaggregated by social identity

Action: Flag high toxicity. Consider rebalancing harmful associations that emerge in model evaluations

[a] www.perspectiveapi.com

Image

C.3.4 PERCEIVED SOCIAL IDENTITY FEATURES X TOP IMAGE FEATURES

Task: Calculate co-occurrences between perceived social identity features and image features, such as objects and other people.

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Object Classifier (e.g., Google Cloud Vision API[a])

Output: List of perceived social identity features and the top co-occurrences for each.

Action: Flag stereotypical associations. Consider rebalancing harmful associations that emerge in model evaluations.

[a] https://cloud.google.com/vision/docs/object-localizer

C.3.5 Perceived Social Identity Features X Sexual Imagery

Task: Calculate co-occurrences between perceived social identity features and sexual imagery.

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Visual Content Classifier (e.g., Google Cloud Vision API[a]

Output: Proportion of images depicting sexual content disaggregated by social identity features

Action: Flag sexual content. Consider filtering sexual content depending on downstream tasks and to avoid problematically-sourced dataset content.

[a] https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation

C.3.6 Perceived Social Identity Features X Violent Imagery

Task: Calculate co-occurrences between perceived social identity features and violent imagery.

Analysis Object: Inferred Image Signals

Effort: Low

Dependencies: Visual Content Classifier (e.g., Cloud Vision API[a]

Output: Proportion of images depicting violent content disaggregated by social identity

Action: Flag violent content. Consider filtering sexual content depending on downstream tasks and to avoid problematically-sourced dataset content

[a] https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation

C.3.7 Perceived Social Identity Features X Hateful Symbols

Task: Calculate co-occurrences between perceived social identity features and hateful symbols detected in images.

Analysis Object: Inferred Image Signals

Effort: Not currently possible

Dependencies: Hateful Symbol Classifier

Output: Proportion of images depicting sexual content disaggregated by social identity

Action: Flag hateful content. Consider filtering hateful content depending on downstream tasks.

Text-Image

C.3.8 Social Identity Terms X Sexual Imagery

Analysis Object: Social Identity Term List + Inferred Image Signals

Task: Calculate co-occurrences between social identity terms and sexual imagery.

Effort: Low

Dependencies: Visual Content Classifier (e.g., Cloud Vision API[a]

Output: Proportion of images depicting sexual content disaggregated by social identity term

Action: Flag associations. Consider rebalancing harmful associations that emerge in downstream model evaluations

[a] https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation

C.3.9 Social Identity Terms X Violent Imagery

Analysis Object: Social Identity Term List + Inferred Image Signals

Task: Calculate co-occurrences between social identity terms and violent imagery.

Dependencies: Visual Content Classifier (e.g., Cloud Vision API[a]

Output: Proportion of images depicting violent content disaggregated by social identity term

Action: Flag associations. Consider rebalancing harmful associations that emerge in downstream model evaluations

Effort: Low

[a]https://cloud.google.com/vision/docs/reference/rest/v1/images/annotate#safesearchannotation

C.3.10 Perceived Social Identity Features X Top Text Tokens

Analysis Object: Inferred Image Signals + Text

Task: Calculate co-occurrences between perceived social identity signals and tokens in associated text.

Effort: Low

Dependencies: Perceived Social Identity Classifiers

Output: List of most frequent co-occurrences for each social identity signal

Action: Flag associations. Consider rebalancing harmful associations that emerge in downstream model evaluations

C.3.11 Perceived Social Identity Features X Offensive Speech

Analysis Object: Inferred Image Signals + Inferred Text Signals

Task: Calculate co-occurrences between perceived social identity signals and tokens in associated text.

Effort: Low

Dependencies: Offensive Speech Classifier or similar (e.g., Perspective API[a]

Output: Distribution of toxicity, disaggregated by social identity

Action: Flag high toxicity. Consider rebalancing harmful associations that emerge in model evaluations

[a]www.perspectiveapi.com

C.3.12 Perceived Social Identity Features X Topic

Analysis Object: Inferred Image Signals + Inferred Text Signals

Task: Calculate co-occurrences between perceived social identity signals and tokens in associated text.

Effort: Low

Dependencies: Topic Classifier (e.g., Google Cloud Content Categories[a])

Output: Distribution of toxicity, disaggregated by social identity

Action: Flag high toxicity. Consider rebalancing harmful associations that emerge in model evaluations

[a]https://cloud.google.com/natural-language/docs/categories