Experiences with Improving the Transparency of AI Models and Services

Michael Hind[*]
Stephanie Houde[†]
Jacquelyn Martino[*]
Aleksandra Mojsilović[*]
David Piorkowski[*]
John Richards[*]
Kush R. Varshney[*]

IBM Research
[*]Yorktown Heights, NY 10598
[†]Cambridge, MA 02142
USA
hindm@us.ibm.com
stephanie.houde@ibm.com
jmartino@us.ibm.com
aleksand@us.ibm.com
djp@ibm.com
ajtr@us.ibm.com
krvarshn@us.ibm.com

Abstract

AI models and services are used in a growing number of high-stakes areas, resulting in a need for increased transparency. Consistent with this, several proposals for higher quality and more consistent documentation of AI data, models, and systems have emerged. Little is known, however, about the needs of those who would produce or consume these new forms of documentation. Through semi-structured developer interviews, and two document-creation exercises, we have assembled a clearer picture of these needs and the various challenges faced in creating accurate and useful AI documentation. Based on the observations from this work, supplemented by feedback received during multiple design explorations and stakeholder conversations, we make recommendations for easing the collection and flexible presentation of AI facts to promote transparency.

Author Keywords

AI transparency; AI governance; documentation; FactSheets.

CCS Concepts

•Computing methodologies → Artificial intelligence;
•Software and its engineering → Software creation and management;

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the author/owner(s).
ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.3383051

Introduction

AI models and services are being used in a growing number of high-stakes areas such as financial risk assessment [9], medical diagnosis and treatment planning [18], hiring and promotion decisions [1], social services eligibility determination [6], predictive policing [5], and probation and sentencing recommendations [12].

Recent work has outlined the need for increased transparency in AI for data sets [8, 4, 10], models [14], and services [3]. Proposals in support of ethical and trusted AI are also emerging [19, 15, 11]. Although details differ, all are driving towards a common set of attributes that capture essential “facts" about a model. We are not yet aware of developers adopting these ideas for regular use or any published work describing developers’ needs or the difficulties they face when producing or consuming AI documentation.

In this paper we discuss formative research with developers of AI models to better understand their documentation needs. We also report on a study in which developers in several application areas created AI documentation in the form of a *FactSheet*.

A FactSheet, as proposed by [3], is a collection of information about an AI model or service that is generated throughout the machine learning life cycle. It includes information from the business owner (e.g., intended use and business justification), from the data gathering/feature selection/data cleaning phase (e.g., data sets, features used or created, cleaning operations), from the model training phase (e.g., bias and robustness information), and from the model validation and deployment phase (e.g., key performance indicators). A FactSheet is associated with a model (or service) and is meant to be write once, i.e., an update to a model would trigger a new FactSheet for the updated model. FactSheets will be consumed by different users for different purposes including reviewing and selecting models from a library or catalog, validating models prior to deployment, certifying model compliance with regulations, and monitoring models with respect to key business indicators once deployed. Of course, the diversity of model types, and the range of possible application domains, makes the specification of a common FactSheet schema difficult. We hope the work reported here provides useful guidance going forward.

The contributions of this paper include:

- summaries of semi-structured interviews with AI developers on their documentation needs and practices
- observations on documentation requirements and difficulties of AI developers in creating FactSheets
- recommendations for supporting mechanisms and new research to improve FactSheet creation.

Related Work

The challenge of creating useful and usable documentation is not new. Software engineering research has identified quality issues in existing documentation for conventional systems [7, 16, 17] and identified problems such as missing rationales for design decisions, too few examples to understand how to use a module or package, lack of overviews to illustrate how a system’s component parts work as a whole, and insufficient guidance on how to map usage scenarios to elements of an API. Exacerbating these problems is the fact that documentation tends to be costly to create and maintain, and is often left as a low-priority item during software development [20]. Perhaps unsurprisingly, developers may ignore documentation altogether and prefer to read source code [13] or inspect the outputs returned to various inputs, as they view these techniques to be a more reliable reflection of reality.

1. What is this model for?
2. What domain was it designed for?
3. Information about the training data (if appropriate)
4. Information about the model (if appropriate)
5. What are the model's inputs and outputs?
6. What are the model's performance metrics? (accuracy, bias, robustness, domain shift, other metrics that you think are appropriate for this model)
7. Information about the test set
8. Can a user get an explanation of how your model makes it decisions?
9. In what circumstances does the model do particularly well (within expected use cases of the model)? (e.g., inputs that work well)
10. Based on your experience, in what circumstances does the model perform poorly? (e.g. domain shift, specific kinds of input, observations from experience)

Shortened FactSheet questions.

Given the reliance of AI models on training data, and the often probabilistic behavior of AI systems with respect to test data, we believe that AI documentation will include features not found in documentation of general software. The emergence of data sheets [8, 4, 10], model cards [14], and FactSheets [3] attest to these differing needs. We turn now to how we explored the requirements and possible benefits of AI FactSheets and the potential difficulties faced by developers in creating and consuming them.

Methodology

Our formative research into potential FactSheet uses, requirements, and challenges consisted of two primary threads. First, we conducted semi-structured interviews with data scientists to investigate potential FactSheet use cases. Second, we led two FactSheet creation sessions with additional AI developers to understand their views on what content to include in FactSheets and the difficulties they faced in actually creating that content.

Data Scientist Interviews

We conducted semi-structured interviews focused on current AI documentation practices and potential FactSheet use cases. We recruited six participants (one female and five males), half from within our research and development organization and half from outside. Participants' experience in their current role ranged from four months to twenty years. Their self-reported background and experience included applied mathematics, data and business analytics, computer science, cognitive science, engineering, and machine learning. Participants suggested several use cases where FactSheets could be valuable.

Participants noted how FactSheets might be used to improve model comprehension. They also reported an ongoing need to facilitate discussions about models with others. Some envisioned use cases where FactSheets were more closely integrated within their development environments. Taken together, these discussions helped frame FactSheet requirements.

FactSheet Creation Session 1

To gather more detailed requirements, and to understand the difficulties developers might face in generating useful facts, we recruited nine additional AI model or service developers from within our larger organization. Their models varied both in type and in application domain.

Each developer was first introduced to the idea of FactSheets as the primary source of information about a model. Each was then given a sample FactSheet consisting of 39 questions and answers based on [3]. They were then asked to create a FactSheet to document their model, the stated intent of this documentation to be for use in a model marketplace where users could search for, compare, and select models appropriate for their tasks. Developers were encouraged to add or remove any facts as they deemed necessary.

FactSheet Creation Session 2

For the second FactSheet creation session, we reengaged with six of the nine developers and provided ten high-level questions to be answered during one-hour co-development sessions with one of the authors. The questions, shown to the left (with sample answers to the first two questions below), were selected as a commonly recurring subset from the first session. We asked participants to again fill out a FactSheet for a potential user of a model marketplace. After they completed this task, one of us walked them through their filled-out FactSheet to extract rationales for why they provided the information.

Purpose
This model is able to detect whether an English-language text fragment leans towards a positive or a negative sentiment. The underlying neural network is based on the pre-trained BERT-Base, English Uncased model and was fine-tuned on the IBM Claim Stance Dataset.

Domain
The primary domain of this model is Natural Language Processing/Understanding. This model was trained on Wikipedia article text in the context of debates and likely performs best with similar-styled inputs. The model performs best on "argument" styled sentences as the training data was part of Project Debater.

Sample answers to two FactSheet questions.

Observations

In this section we report our major findings from these two FactSheet creation exercises, and the comments gathered during the first interviews described above.

Perceived FactSheet Value

All the developers in our interviews and creation sessions viewed FactSheets as valuable, with the exception of one of the initial interviewees (a senior data scientist with well-established work practices). The idea of capturing key facts about an AI model or service in a form that is useful to a broad range of stakeholders was appealing. AI developers stated that FactSheets could provide useful guidance on how best to document their work. Those who have tried to consume models developed by others noted the importance of understanding how the models were structured, what data was used to train them, how features were engineered, and why a particular model or class of model was selected as fit for purpose. FactSheets could and should include these elements.

Observed FactSheet Challenges

What particular difficulties did developers face when trying to create content for a FactSheet? Some developers simply forgot important details such as how they transformed training data or which hyper parameters they explored along the path to a final model. Another area of concern for our participants was documenting facts about a model that might reflect poorly on data provenance, testing protocols, or the possibility of various biases in either the training data or model output. Several participants noted that data or model details may be proprietary. It is often not clear where to draw the line between providing enough information for a model to be adopted while not revealing information that threatens competitive advantage.

The reason that FactSheets or related forms of documentation will be produced is so others can benefit from consuming them. In many cases, however, we observed that FactSheet producers were unsure what information might be needed by potential FactSheet consumers. This seemed to be especially true in the case of models packaged for reuse in shared catalogs (which was the case we explored with our participants). How is it possible, short of extensive experience in a domain, to anticipate the ways that models might be used, or misused? More generally, how is it possible to know whether a FactSheet will be viewed from the perspective of say a testing team as opposed to industry regulators? Each consumer will have different levels of understanding and will be performing quite different tasks.

Recommendations

We have found a range of needs for those creating and consuming AI documentation. We have also found considerable diversity across domains and model types as to which facts in an AI FactSheet are likely to be useful. Despite this diversity we have learned enough to suggest some directions for future work that could lead to better, more consumable, FactSheets and more widespread FactSheet adoption.

Fact Collection

Nascent “facts", in the form of discreet events, are already arising throughout the AI life cycle. Most of these potentially interesting facts go unnoticed. Others are noticed but quickly forgotten. A few are noticed and recorded (with varying fidelity) in unconnected systems that prevent coherent documentation from being produced reliably or efficiently. If this is to change we must first acknowledge that facts will arise from the actions of different stakeholders (ranging from line of business managers, to data scientists, to risk and compliance officers) and that they will arise through the use of many different tools (from business modeling software, to Jupyter notebooks, to risk management and reporting systems). Consider a few examples. Intended use descriptions and required performance targets might be specified by a line of business owner. Data provenance and licensing details might be documented by a data steward. Accuracy, bias, and robustness metrics might be generated by a data scientist or quality assurance team. Some facts will be generated automatically as a side effect of a data transformation or a training cycle. Others will arise in discussions between high-level managers in operations meetings. Importantly, whatever we can do to make fact collection easy, either by making it more automatic or simply making it possible to write (and post) a bit of descriptive text in the moment of tool use, will decrease the incidence of forgetting the fact or forgetting important details about the fact.

It is certainly *possible* to create a single integrated system including all the tools used throughout the AI life cycle, each tool equipped with a common mechanism to collect facts for later use. But it is likely that such a system would impose unacceptable constraints on the tools that organizations want to use and the way that organizations want to work. We believe a more realistic approach is to define an open API for registering models, for posting facts about them, and for retrieving them for monitoring and reporting. This API could define the end points for a pub-sub architecture, such as Pulsar [2], enabling the creation of a FactSheet Repository for a diverse and essentially unbounded set of tools.

Fact Authoring

Some important AI facts can be captured automatically. Others, perhaps those that have coalesced around a stable set of practices within a development community, can be easily authored. Some facts, however, will require careful thought about what should be documented. Our developers found it difficult, for example, to specify the boundary conditions beyond which model use was inadvisable. In another example, developers were often unsure about the level of detail to include in descriptions of a model’s structure. We have found through our interviews and FactSheet creation exercises that this sort of human fact “authoring" is challenging and the quality of authored facts is quite variable. Both productivity and quality can be improved, however. For example, if user testing indicates that a particular FactSheet question is hard to understand, the question can be clarified through a cycle of user testing and refinement. Alternatively, if the question is understandable but answers are often incomplete or of poor quality, hints or examples of well-formed answers can be offered.

Even in our somewhat limited testing we have found that FactSheets will often include questions for which there is no relevant answer. “N/A" may be a perfectly acceptable fact value in this case, but it should be applied thoughtfully. Some facts may be known but proprietary (and could be redacted for those with insufficient access rights). In other cases, model builders will assume that some kinds of facts are not applicable even if, on further reflection, they are. We have seen, for example, that model bias is often considered inapplicable. To get better answers in cases such as this we may need to create elicitation techniques that go beyond mere form filling. For example, if a question is frequently marked as not applicable a wizard could walk the user through a process that will lead, in the end, to a suitable answer.

We have also noted that different ways of recording the same information can make model comparisons difficult. Just as answer quality can be improved through hints and examples, so too can answer consistency. More generally, open fact schemata can be created, perhaps even standardized, to promote greater consistency within particular domains or industries.

Fact Retrieval and Reporting

At some point, facts about a model that are collected automatically, along with facts that are manually authored, will generally (but not necessarily) be assembled and rendered into coherent documents. Of course, not all facts about a model need to be seen by all stakeholders. And in some situations, not all facts about a model *should* be seen by all stakeholders.

Beyond the need for facts to be excluded because of limited access rights, there is a need for facts to be assembled in different ways when they are rendered as FactSheets for different stakeholders. One way to manage this diversity of views is through the use of what we could call *FactSheet Templates*. FactSheet templates could be created through a template builder drawing on an inventory of all fact types that an organization wanted to collect and report. A particular template would determine the content and layout of the associated FactSheet instance generated for a particular stakeholder class.

In addition to excluding or including particular facts in a generated report, we have found at least one case where different stakeholders needed to see the *same* fact at different levels of detail. Not surprisingly, the prospect of creating a set of different fact forms for the same fact was viewed as unattractive. To support this case, information within a fact could be more finely structured such that portions of it were individually addressable. Alternatively, facts consisting of lengthy unstructured text might be automatically summarized for those needing only an overview. Future research might profitably focus on meeting this need.

Concluding Thoughts

Accurate and understandable facts about a model throughout its full life cycle — from requirements specification, to data curation and feature engineering, to training and testing, to deployment and monitoring — will provide a range of benefits, some of which we can only speculate about now. For most models, there are currently no such well-assembled facts.

This is not just a pain point for developers and a cost driver for organizations. The absence of useful information diminishes the perceived trustworthiness of the models we create. When important model facts are easily collected, authored, and reported, trust and responsible AI use will grow.

The focus of this paper is on the requirements and challenges of creating FactSheets, a necessary step to improve the governance and transparency of AI. Another important dimension is the level of trust in the facts themselves. Is it sufficient to have an enterprise self-report their facts or do standards bodies or third-party certification agencies conduct or validate this reporting?

We have discussed the experiences of developers in creating and consuming current AI documentation. We have also worked with developers as they have tried to create the form of documentation proposed as FactSheets. The problems they faced have led to a series of recommendations for improving system support for automatic fact collection, human fact authoring, and flexible fact reporting. Future explorations of these ideas may lead to a new, open ecosystem improving our collective understanding of the AI models, services, and systems we create.

REFERENCES

[1] Jennifer Alsever. 2017. How AI Is Changing Your Job Hunt. https://fortune.com/2017/05/19/ai-changing-jobs-hiring-recruiting/.
(2017).

[2] Apache Foundation. 2019. https://pulsar.apache.org. (2019). Last accessed 1 November 2019.

[3] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. NatesanRamamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM Journal of Research & Development* 63, 4/5 (Sept. 2019).

[4] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association of Computational Linguistics* (2018).

[5] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. (2017).

[6] Tiffany Dovey Fishman, William D. Eggers, and Pankaj Kishnani. 2019. Using cognitive technologies to transform program delivery. https://www2.deloitte.com/us/en/insights/industry/public-sector/artificial-intelligence-technologies-human-services-programs.html.
(2019).

[7] Golara Garousi, Vahid Garousi, Mahmoud Moussavi, Guenther Ruhe, and Brian Smith. 2013. Evaluating usage and quality of technical software documentation: An empirical study. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 24–35.

[8] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. 2018. Datasheets for Datasets. In *Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Workshop*. Stockholm, Sweden.

[9] Shruti Goyal. 2018. Credit Risk Prediction Using Artificial Neural Network Algorithm. https://www.datasciencecentral.com/profiles/blogs/credit-risk-prediction-using-artificial-neural-network-algorithm. (March
2018).

[10] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677.

[11] IEEE. 2017. P7006 - Standard for Personal Data Artificial Intelligence (AI) Agent. https://standards.ieee.org/project/7006.html. (2017).

[12] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwiny. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm,
(2016).

[13] Walid Maalej, Rebecca Tiarks, Tobias Roehm, and Rainer Koschke. 2014. On the comprehension of program comprehension. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 23, 4 (2014), 31.

[14] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Atlanta, USA.

[15] Partnership On AI. 2019. https://www.partnershiponai.org/about-ml/. (2019). Last accessed 3 November 2019.

[16] Martin P Robillard and Robert Deline. 2011. A field study of API learning obstacles. *Empirical Software Engineering* 16, 6 (2011), 703–732.

[17] SM Sohan, Frank Maurer, Craig Anslow, and Martin P Robillard. 2017. A study of the effectiveness of usage examples in REST API documentation. In *2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 53–61.

[18] Eliza Strickland. 2019. Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients. *IEEE Spectrum* (Oct. 2019).

[19] The European Commission's High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. Brussels, Belgium.

[20] Gias Uddin and Martin P Robillard. 2015. How API documentation fails. *IEEE Software* 32, 4 (2015), 68–75.