Disembodied Machine Learning: On the Illusion of Objectivity in NLP

Anonymized

Abstract

Machine Learning (ML) seeks to identify and encode bodies of knowledge within provided datasets. However, data encodes subjective content, which determines the possible outcomes of the models trained on it. Because such subjectivity potentially enables marginalisation of parts of society, it is termed (social) 'bias' and sought to be removed. In this opinion paper, we contextualize this discourse of bias in the ML community against the subjective choices in the development process. Through a mapping of how choices in data and model development construct subjectivity, or biases that are represented in a model, we argue that addressing and mitigating biases is near-impossible. This is because both data and ML models are objects for which meaning is made in each step of the development pipeline, from data selection over annotation to model training and analysis. Accordingly, we find the prevalent discourse of bias limiting in its ability to address social marginalisation. We recommend to be conscientious of this, and to accept that de-biasing methods only correct for a fraction of biases.

1 Introduction

Machine Learning is concerned with making decisions based on discernible patterns observed in data. Frequently, ML models and the bodies of data they act on are divorced from the context within which they are created, leading to an imposed 'objectivity' to these processes and their results. Given that supervised ML seeks to distinguish a set of given bodies of data from one another, and unsupervised ML aims to identify discernible bodies of data in the data provided;[1] both the underlying data and the model applied to it strongly influence what bodies are discovered, and what may be discovered within these bodies. ML models were initially hailed as objective, unimpeded by subjective human biases, and by extension by social marginalisation (O'Neil, 2016). However, more and more research suggests that social biases are common in ML models, and that such biases in the underlying data may be exacerbated by the ML models (Zhao et al., 2017). Accordingly, a number of research directions seek to identify (Shah et al., 2020; Bender and Friedman, 2018; Mitchell et al., 2019; Buolamwini and Gebru, 2018), reduce or remove social bias (Zhao et al., 2017; Agarwal et al., 2018) from ML models to protect against further marginalisation. However, previous work frequently assumes a positivist logic of social bias as an optimisation problem, i.e. that bias is a finite resource can be disentangled, isolated, and thus optimised for.

In this article, we revisit these assumptions. Drawing on a body of work from feminist Science and Technology Studies (STS) (Haraway, 1988) and examples from Natural Language Processing (NLP), we argue that (a) bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed; for which reason (b) an ongoing recognition and reflection on our own positions and the imaginary of objectivity found in subjective realities reflect political choices in the ML pipeline. By contextualising bias in these terms, we seek to shift the surrounding discourse away from bias and its elimination towards subjective positionality and its implications in the ML pipeline. This allows us to investigate how ML practitioners and researchers are complicit in processes of social exclusions in society.

2 Previous Work

Previous work on bias: (i) describes models and datasets with their intended uses and limitations;

[1] Bodies of data are amalgamated entities which exist by virtue of a strict separation from the material bodies they are derived from.

(ii) quantifies and analyzes disparities; or (iii) mitigates biases that are present in models and datasets.

Mapping Bender and Friedman (2018) propose 'data statements', a tool to describe and expose representational biases in the processes of developing datasets from collection to annotation. Analogously, Mitchell et al. (2019) propose 'model cards' to describe ML models and their behaviour across different populations that might be subject to a given model along with its intended use.

Quantification Shah et al. (2020) propose a mathematical framework quantifying biases in different steps in the NLP pipeline, basing their conceptualisation on work on ethical risks for NLP systems by Hovy and Spruit (2016). More practically, Buolamwini and Gebru (2018) identify how commercial facial recognition systems perform and fail for people with darker skin and women, and perform worst for women with dark skin.

Mitigation Two conceptualisations of bias can be found in the large body of work on addressing biases in models (e.g. Agarwal et al., 2018; Romanov et al., 2019; Kulynych et al., 2020): one in which bias is imagined as a finite quantity in the model that can be minimised by altering the model's representation (Agarwal et al., 2018; Romanov et al., 2019);[2] and one which, in a similar vein to our work, accepts the premise that ML, and more broadly optimisation systems, will contain social biases. Working with this assumption, they propose a class of systems that use optimisations logic to counteract the marginalisation a group experiences as the result of ML being applied to them.

3 The 'God Trick' of Objectivity

In her seminal STS work, Donna Haraway (1988) calls into question the notion of objectivity, arguing that the production of knowledge is an *active* process, in which we subjectively construct knowledge based on our very particular, subjective bodies. She argues that an 'objective' position like all other positions comes with its own limitations in what it obscures and highlights. In other words, an 'objective' position is no less subjective, insofar it privileges the point of view of a particular body marked by subjective social and political meanings and possibilities along lines of race, class, geography, gender etc. However, unlike other 'subjective' positions, an 'objective' position claims omniscience for itself by denying its particular embodiment, thereby obscuring its own subjective rootedness. This position can then be understood as a disembodied subjective position. By denying the subjectivity of its own body, the objective position elevates itself over other 'lesser subjective bodies', thus playing the 'God trick' (Haraway, 1988).

Through its disembodiment, the position of objectivity claims to be 'universal' and free from embodied socio-political meaning and is therefore applicable in all contexts and can thus be imposed upon all other subjective positions (Mohanty, 1984). Consequently, embodied positions are mired in a particular (as opposed to 'universal') context and their particularised experiences of embodied positions can safely be rejected, as accepting them would threaten the omniscient claim of objective study. However, as Haraway (1988) argues, subjectively embodied positions allow for things to be made visible, that are otherwise invisible from the disembodied position. For instance, in the context of *n-word* usage, an exclusive focus on its derogatory use would imply understanding the word through a disembodied and universalised position, which is a position often (but not always) occupied by the white human body in our world. It is only through an engagement with the particularised experiences of black bodies that the rich cultural meaning crafted in African-American communities reveal themselves. (Rahman, 2012).

4 Embodiment in the ML Pipeline

Haraway's (1988) critique of objectivity makes it possible to understand subjectivity or bias in ML in a way that recognises its potential to create social marginalisation, without inherently reducing it to a problem which can be optimised. We argue that in ML, the disembodied or objective position exists: (i) in the person designing the experiment and pipeline by developing methods to apply to a dataset of *others*; (ii) in the data which is often disembodied and removed from context, and potentially given adjudication by externalised others that may not be aware of the final use of their work; and (iii) in the model trained on the embodied data subjects.[3] We note that once data are ready to be

[2] This line of work has the dual aims of minimising discrimination, while maximising performance for a given metric.

[3] We highlight here the inherent self-contradiction in ML taking the position of objectivity while tacitly accepting that it is subject to disembodied data as evidenced by the fields of domain adaptation and transfer-learning.

processed by the model, we can consider the model to embody the data, as it is limited to the bodies of knowledge it is presented with. Thus, all other positions, i.e. those not represented in the training data, become disembodied. This can help explain why ML practitioners frequently call for 'more' and 'more diverse' data (Holstein et al., 2019) to address models that are unjust. However, simply adding more data without addressing whom the datasets embody and how is unlikely to yield the desired result of more just and equitable models.

Embodiment of the designer A lack of diversity in ML teams is often attributed as a source of socially biased technologies with corresponding calls for increasing embodying diverse experiences (West et al., 2019). The embodied designers, through data and modeling choices, project an embodiment of self into the technologies they develop. Cnsidering Haraway (1988), it is only through the recognition of different embodiments and promoting them that certain perspectives, understandings, and uses can be achieved.

4.1 Embodiment in Data

Datasets, following Haraway (1988), can be understood as a form of knowledge that does not simply exist but is produced (Gitelman, 2013) through embodied experiences. Subjectivity in datasets can come from a variety of sources, including but not limited to the source of the data (Gitelman and Jackson, 2013), the data sampling method (Shah et al., 2020), the annotation guidelines (Sap et al., 2019), and the selection of annotators (Waseem, 2016; Derczynski et al., 2016). Here, we ground our discussion of how subjectivity manifests itself in ML models through various processes of meaning-making, modeling choices, and data idiosyncracies, by drawing on examples from NLP. A common denominator that we seek to highlight, is the subjective and embodied nature of data and subsequent classifications; that by taking a position of objectivity, we cannot do justice to the needs and wants of individual or in fact discernible communities.

High-level tasks A range of NLP tasks are highly sensitive to subjective values encoded in the data. This includes high-level tasks that require semantic and pragmatic understanding, e.g. machine translation (MT), dialogue systems, metaphor detection, and sarcasm detection among others. In MT, research has identified a range of issues, including stylistic (Hovy et al., 2020) and gender bias (Vanmassenhove et al., 2018). Issues that pertain to the reinforcement of sexist stereotypes have been the object of academic and public scrutiny. A classic example is the stereotypical translation of English *doctor* (unmarked for gender) to German *Arzt* (marked for masculine), while *nurse* (unmarked) is translated to *Krankenschwester* (feminine). Here, the 'objective' position is truly a patriarchal one, which delegates more prestige to men and less to women. The translations above may be correct in certain contexts, but they are not necessarily the only correct ones. This exemplifies the overarching problem that, in MT, there is rarely one single 'gold' translation for a given sentence, yet most training and evaluation algorithms assume just that.

The issue of highly subjective 'truths' in data extends to several other tasks. In the area of text simplification, for instance, numerous datasets postulate that some words, sentences or texts are difficult, while others are simple. These labels are typically provided by a group of human annotators, and while there might be clear majorities for the labeling of certain items, the disembodied position and generalisational power of the annotations will never do justice to the subjective embodiments of text difficulty both across user groups (language learners of different L1 backgrounds, dyslexics, etc.) and just as much within these groups.[4]

For abusive language detection, the causes and effects of embodiment in different stages have been considered in a dataset for offensive language use (Davidson et al., 2017). Waseem et al. (2018) argue that a consequence of embodying a white perspective of respectability is that almost all instances of the *n-word* are tagged in the positive classes. Sap et al. (2019) show that by indicating the likely race[5] to the annotators, annotators sought to align their embodiment of 'offensive' with the dialect of the author. Further, Davidson et al. (2019) argue that the initially sampled data may itself contain social biases due to a disembodied perspective on slurs.

Core NLP tasks However, the issues outlined above are far from limited to high-level NLP tasks. Even core NLP tasks such as part-of-speech (POS)

[4]There is some merit in the meta-information on certain task-relevant demographic variables of individual annotators provided in the datasets for the Complex Word Identification 2018 Shared Task. Further, recent work recognises that text simplification systems must build on personalised models (Yimam and Biemann, 2018; Lee and Yeung, 2018; Bingel et al., 2018).

[5]As assumed through the prediction of dialect.

tagging are sensitive to the subjective nature of choices in the ML pipeline. Consider the Penn Treebank tagset (Marcus et al., 1993), the *de-facto* standard for describing English word classes in NLP. Behind this collectively accepted 'objective' truth is a linguistic theory that licenses a certain set of POS tags while not recognising others. The linguistic theory, in turn, is subjective in nature, and typically informed by observations on specific kinds of language. The tagset is thus better suited to describe the kind of English its underlying theory was built on rather than other varieties, sociolects or slang. This becomes more drastically apparent when a tagset developed for English is, for better or worse, forced upon some other languages.

4.2 Embodiment in Modeling

While datasets are a large source of how a model may be embodied, ML models also encode which positions, or embodiments, are highlighted. In our exploration of how models exacerbate disembodied positions, we primarily focus on supervised methods, as unsupervised methods are highly volatile to the subjective choices of the researcher, e.g. how data is represented and which parameters the model is subject to. Model behaviour can be seen as being on a spectrum ranging from globally acting models, i.e. models that compound multiple senses of word usage with little regard to its local context; and locally acting models, which seek to embody the datum in the context it is created in, e.g. context-aware models (Garcia et al., 2019; Devlin et al., 2019).

By virtue of the subjective nature of grounding datum in context, there is a large variation in how locally acting models may be developed. Through transfer learning, knowledge produced outside of the target task training set can alter what the model embodies. For instance, should a dataset embody the language production within multiple sociolects, a pre-trained language model (Devlin et al., 2019)[6] or mixed-member language models (Blodgett et al., 2016) may provide deeper information about the sociolects in question through using the context of a sentence to identify which context to situate the representation of a document.[7] Beyond language models, multi-task learning models can encode information about the creator of the datum (Benton et al., 2017; Garcia et al., 2019) thus allowing models to take into account the embodiment of the datum. Transfer learning can similarly be applied such to push the model to embody the context the datum is derived from. For instance, Romanov et al. (2019) encode the demographic of a document's author, deterring models from learning stereotypified representations of marginalized, and in the case of NLP often under-represented, communities.

5 Discussion

If subjective choices or biases masquerading as disembodied 'objective' positions permeate throughout the ML pipeline – and we argue here that they do – the quest for objectivity or bias-free ML becomes redundant. Rather, such a quest for objectivity or a universal 'truth' may even further harm already marginalised social groups by obscuring the dominance of certain bodies over others. Any effort to obscure only deepens the power of dominant groups and hurts marginalised communities further by justifying the imposition of experiences of dominant bodies upon marginalised bodies under the guise of 'objective' or 'bias-free'.

By recognising the positionality of the designers of ML models, one can account for what (and whom) one's own position, and the models derived from it, allow and penalise, and the political consequences of these. As data permeate the ML pipeline, a consideration of how data is embodied can allow for answering specific questions embodied in context; that the contexts which create data are present in every step of the dataset creation pipeline; and that as contexts change, so does the applicability of data. Further, models themselves privilege some views over others, and while transfer learning provides some avenues for embodying data in the model, what positions are given space remains a political question.

Thus, rather than asking how to eliminate bias and subjective experiences from ML, shifting to consider embodiments would ask us to reflect on the subjective experiences that are given voice. And, crucially, such a shift would require us to ask and reflect upon which subjective experiences, or rather, the subjective experiences of which bodies we need to account for to give voice to socially marginalised groups.

[6] As sociolects and dialects may not be equally distributed in training data for contextual models (Dunn, 2020), similar issues of which bodies are embodied plague such models (Tan and Celis, 2019)

[7] While 'context' here is limited to what is in the sentence, language production is situated within a larger socio-political context of society.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden. PMLR.

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 for Computational Linguistics*, 6:587–604.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*.

Noa Garcia, Benjamin Renoust, and Yuta Nakashima. 2019. Context-aware embeddings for automatic art analysis. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 25–33, New York, NY, USA. Association for Computing Machinery.

Lisa Gitelman, editor. 2013. *"Raw data" is an oxymoron*. Infrastructures series. The MIT Press, Cambridge, Massachusetts.

Lisa Gitelman and Virginia Jackson. 2013. Introduction. In Lisa Gitelman, editor, *"Raw Data" Is an Oxymoron*, pages 1–14. MIT Press, Cambridge, Massachusetts.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3).

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. Can you translate that into man? commercial machine translation systems include stylistic biases. acl. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda F. Gürses. 2020. Pots: Protective optimization technologies. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 177–188.

John Lee and Chak Yan Yeung. 2018. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330.

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 Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Chandra Talpade Mohanty. 1984. Under western eyes: Feminist scholarship and colonial discourses. *boundary 2*, 12/13:333–358.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Jacquelyn Rahman. 2012. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In Jennifer Golbeck, editor, *Online Harassment*, pages 29–55. Springer International Publishing, Cham.

Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems: Gender, race and power in ai. Retrieved from https://ainowinstitute.org/discriminatingsystems.html.

Seid Muhie Yimam and Chris Biemann. 2018. Par4sim–adaptive paraphrasing for text simplification. *arXiv preprint arXiv:1806.08309*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.