Disembodied Machine Learning: On the Illusion of Objectivity in NLP

Zeerak Waseem
University of Sheffield

Smarika Lulz
Humboldt University, Berlin

Joachim Bingel
Hero I/S

Isabelle Augenstein
University of Copenhagen

Machine Learning 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 enables marginalisation of parts of society, it is termed (social) 'bias' and sought to be removed. In this paper, we contextualise this discourse of bias in the ML community against the subjective choices in the development process. Through a consideration 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 (ML) 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, Schwartz, and Hovy 2020; Bender and Friedman 2018; Mitchell et al. 2019; Buolamwini and Gebru 2018), reduce or remove social biases (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.

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

We revisit these assumptions and question solutionist approaches that dominate the ML literature. Drawing on 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 are inescapable and thus cannot simply be removed; therefore (b) requires an ongoing reflection on the positions and the imaginary objectivity that ML researchers and practitioners find in subjective realities reflect political choices in the ML pipeline. By contextualising bias in these terms, we seek to shift the discourse away from bias and its elimination towards subjective positionality.

2. Previous Work

Previous work on bias in ML: (i) maps models and datasets with their intended uses and limitations; (ii) quantifies and analyses disparities; or (iii) mitigates biases 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. Similarly drawing on Haraway (1988), Rettberg (2020) identifies how data is situated and viewed through disembodied positions in the aggregation and display of personal data in mobile applications.

Quantification. Shah, Schwartz, and Hovy (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. Turning to language, Gonen and Goldberg (2019) highlight that methods for debiasing word-embeddings leave traces that allow for reconstructing gendered spaces in 'debiased' word embeddings.

Mitigation. Two conceptualisations of bias can be found in the large body of work on addressing model biases (e.g. Agarwal et al. 2018; Romanov et al. 2019; Kulynych et al. 2020; Bolukbasi et al. 2016): 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, similar to our work, accepts the premise that ML, and more broadly optimisation systems, contain social biases (Kulynych et al. 2020). Working with the latter assumption, Kulynych et al. (2020) 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

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

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 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, Whittaker, and Crawford 2019). The embodied designers, through data and modelling choices, project an embodiment of self into the technologies they develop. Considering Haraway (1988), it is only through the recognition of different embodiments and promoting them that certain perspectives, understandings, and uses can be achieved. Thus diverse representation in designers in a team may aid in highlighting discriminatory outcomes of machine learning systems, it does not foster questions of subjective positioning giving this explicit attention.

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.

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 can stem from various sources, including the data source (Gitelman and Jackson 2013), the sampling method (Shah, Schwartz, and Hovy 2020), the annotation guidelines (Sap et al. 2019), and the annotator selection process (Waseem 2016; Derczynski, Bontcheva, and Roberts 2016).

We ground our discussion of how subjectivity manifests itself in ML through processes of meaning-making, modelling choices, and data idiosyncrasies. A common denominator we seek to highlight is the subjective and embodied nature of data and subsequent classifications; that by taking a position of objectivity, one cannot do justice to the needs of individual or 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. In MT, research has identified a range of issues, including stylistic (Hovy, Bianchi, and Fornaciari 2020) and gender bias (Vanmassenhove, Hardmeier, and Way 2018).

Issues pertaining 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 a patriarchal one, which delegates more prestige to men and less to women. The translations above may be correct in certain, but not all contexts.This exemplifies the overarching problem that there is rarely one single 'gold' label for a given document (Reiter 2018), yet most training and evaluation algorithms assume just that.

In text simplification, numerous datasets postulate that some words, sentences or texts are difficult, while others are simple. These labels are typically provided by human annotators, and while there might be clear majorities for the labelling 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, Thorne, and Bingel (2018) argue that a consequence of embodying a white perspective of respectability is that almost all instances of the *n-word* are tagged as the positive classes. Sap et al. (2019) show that by indicating the likely race[5] to the annotators, they seek to align their embodiment of 'offensive' with the author's dialect. Further, Davidson, Bhattacharya, and Weber (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 tasks such as part-of-speech (POS) tagging are sensitive to the subjective nature of choices in the ML pipeline. Consider the Penn Treebank tagset (Marcus, Marcinkiewicz, and Santorini 1993), the *de-facto* standard for describing English word classes. Behind this

4 There is some merit in the meta-information on task-relevant demographic variables of individual annotators 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, Paetzold, and Søgaard 2018).

5 As assumed through the prediction of dialect.

collectively accepted 'objective' truth is a linguistic theory that licenses a certain set of POS tags while not recognising others. The 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 (Tommasel, Rodriguez, and Godoy 2018).

4.2 Embodiment in Modelling

While datasets are a large source of how a model may be embodied, ML models also encode which positions, or embodiments, are highlighted. 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, Renoust, and Nakashima 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. Transfer learning can provide one possible avenue for locally acting models. Through transfer learning, knowledge produced outside of the target task training set can alter what a model embodies. For instance, should a dataset embody the language production in multiple sociolects, a pre-trained language model (Devlin et al. 2019)[6] or mixed-member language models (Blodgett, Green, and O'Connor 2016) may provide deeper information about the sociolects in question by examining the sentential context.[7] It is important to note that the large-scale datasets for language models rely on disembodying the data from the bodies creating them to identify collective embodiments. Similarly, multi-task learning models can offer a path to embodying the creator of the datum through author attribute prediction as auxiliary task(s) (Benton, Mitchell, and Hovy 2017; Garcia, Renoust, and Nakashima 2019), thus allowing models to take into account the embodiment of the datum.

5. Discussion

If subjective choices or biases masquerading as disembodied 'objective' positions permeate through the ML pipeline – and we argue that they do – the quest for objectivity or bias-free ML becomes redundant. Rather, such a quest for objectivity or a universal 'truth' may 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'.

Designers of ML models and pipelines become complicit in how these marginalise when they fail to recognise their own positionality. Through a recognition of one's embodiment, designers can account for what (and whom) their position and models derived from it, allow and penalise, and the political consequences thereof. 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

6 Similar issues affect contextual models (Tan and Celis 2019) as sociolects and dialects may not be well represented in their training data (Dunn 2020).

7 While 'context' here refers to sentential context, language production is situated within a larger socio-political context.

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.

The discourse on bias in ML does look to account for these political consequences. However, it pins the problem down to the presence of subjective, embodied or "biased" positions in ML models and seeks their eradication. Thus, we propose to let go of fairness as a matter of bias elimination in a solutionist endeavour without regard for subjective experiences. Shifting to consider embodiments would instead require one to reflect on the subjective experiences that are given voice, as well as which bodies one needs to account for to give voice to socially marginalised groups.

References

Agarwal, Alekh, 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, PMLR, Stockholmsmässan, Stockholm Sweden.

Bender, Emily M. 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.

Benton, Adrian, 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, Association for Computational Linguistics, Valencia, Spain.

Bingel, Joachim, 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.

Blodgett, Su Lin, 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, Association for Computational Linguistics, Austin, Texas.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., pages 4349–4357.

Buolamwini, Joy 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, PMLR, New York, NY, USA.

Davidson, Thomas, 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, Association for Computational Linguistics, Florence, Italy.

Davidson, Thomas, 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.

Derczynski, Leon, 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, The COLING 2016 Organizing Committee, Osaka, Japan.

Devlin, Jacob, 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, Association for Computational Linguistics, Minneapolis, Minnesota.

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

Garcia, Noa, 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 âĂŹ19, page 25âĂŞ33, Association for Computing Machinery, New York, NY, USA.

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

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

Gonen, Hila and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Association for Computational Linguistics, Minneapolis, Minnesota.

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

Holstein, Kenneth, 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 âĂŹ19, page 1âĂŞ16, Association for Computing Machinery, New York, NY, USA.

Hovy, Dirk, 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*, Association for Computational Linguistics, Seattle, Washington.

Hovy, Dirk 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, Association for Computational Linguistics, Berlin, Germany.

Kulynych, Bogdan, 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.

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

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313âĂŞ330.

Mitchell, Margaret, 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* âĂŹ19, page 220âĂŞ229, Association for Computing Machinery, New York, NY, USA.

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

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

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

Reiter, Ehud. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Rettberg, Jill Walker. 2020. Situated data analysis: a new method for analysing encoded power relationships in social media platforms and apps. *Humanities and Social Sciences Communications*, 7(5).

Romanov, Alexey, 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, Association for Computational Linguistics, Minneapolis, Minnesota.

Sap, Maarten, 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, Association for Computational Linguistics, Florence, Italy.

Shah, Deven, 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*, Association for Computational Linguistics, Seattle, Washington.

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

Tommasel, Antonela, Juan Manuel Rodriguez, and Daniela Godoy. 2018. Textual aggression detection through deep learning. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 177–187, Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Vanmassenhove, Eva, 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, Association for Computational Linguistics, Brussels, Belgium.

Waseem, Zeerak. 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, Association for Computational Linguistics, Austin, Texas.

Waseem, Zeerak, 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*. Springer International Publishing, Cham, pages 29–55.

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

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

Zhao, Jieyu, 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, Association for Computational Linguistics, Copenhagen, Denmark.