Women are Surgeons: Natural Language Processing Models and Biases Mitigation

Jiashu Xu Emily Artiano
University of Southern California
Los Angeles, California, USA
{jiashuxu, artiano}@usc.edu

Disclaimer The preprint follows the format of machine learning conferences like ACL, CVPR and EMNLP. This particular format follows NeurIPS template[1]. Due to machine learning being currently under a rapid development, conferences are preferred by researchers over journals due to a quicker cycle of reviews and publication. Papers submitted to machine learning conferences should be structured by sections like introduction, related works *etc*. Note that the introduction section is mainly serving as motivation of the paper; while the related work section is giving the definition of terms used in this work. Those conference do not particularly look for rhetoric devices but generally more targeted toward abstractness (*e.g*. formulate a task/algorithm in mathematical formulation and agnostic to environment) and mathematical preciseness/rigorousness (*e.g*. proof of a theorem must be sound). However this work is not intended to scrutinize the algorithmic theoretical foundation, rather a critic for effectiveness of algorithms. Therefore, we will target towards analysis of advantages/disadvantages and implications. Also, the following paper will mention "evaluation of effectiveness", which means using the algorithm in various tasks/environments. It is sometimes difficult to find papers using the exact same algorithm, since simply running the algorithm on a new task is not considered too novel and will not likely to be accepted in conferences. Researchers generally use a modified variant of the algorithm, and by evaluation I mean the evaluation of those kinds of algorithm variants. Another note: hyperlinks/references are clickable.

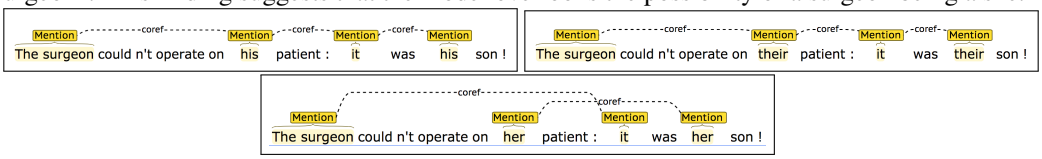
> A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, "I can't operate on this boy, he's my son!" **How can this be?**
>
> A famous riddle

Abstract

Natural Language Processing (NLP) models are widely deployed in millions of devices and services ranging from smartphone assistants to Discord chatbots, yet they are found to propagate and amplify biases regarding social identity including nationality, gender, and race. Given that natural language models are increasingly deployed in sensitive areas such as job hiring and loans, careless use of models will in turn discriminate against end-users by allocating or withholding opportunities or resources along the lines of specific social identity. Therefore it is vital for researchers and practitioners to investigate and combat biases. In this paper, we primarily focus on two approaches to mitigate biases and provide a comprehensive evaluation of each algorithmic design choice of those mitigation approaches under a wide range of English NLP tasks.

[1]Please refer to `https://neurips.cc/Conferences/2022/PaperInformation/StyleFiles`.

Figure 1: Rudinger et al. [28] found that Stanford CoreNLP coreference system resolves a male and neutral pronoun as coreferent with "The surgeon", yet unable to relate female pronoun with "The surgeon". This finding suggests that the model overlooks the possibility of a surgeon being a she.



1 Introduction

Since the introduction of AlexNet [19] in 2012, machine learning becomes ubiquitous. Machine learning models have been widely deployed in modern society and have obtained greater capacity every day. Astonishing achievements perpetuate our daily life: smartphone cameras that recognize human faces accurately [31], Instagram filters that create artistic styles [25], digital smart assistants that answer every question we might have [3] *etc*.

However, as researchers are exploring the boundary of Artificial Intelligence[2], danger lurks in the cover of the current machine learning hype. The very fact that machine learning models have significantly improved over the past few years makes them increasingly used in sensitive areas such as job hiring [8, 24] and loans [22], in which decisions will irreversibly and directly impact thousands of people. Given the importance, it is not unnatural to demand a guarantee that such algorithms will be fair. Yet such appeal seems to be futile [15, 5, 9]. Human biases often propagate [20, 9], sometimes being amplified [15, 34], to machine learning models, which are built from human-generated data. It is therefore urgent for researchers and practitioners to inspect and act proactively concerning understanding and combating the biases so that models, and services backed by those models, will not discriminate against any end-user group.

Natural Language Processing (NLP) is an emerging sub-field of machine learning in which the models deal with textual data (language). NLP models are deeply integrated into daily human routine, ranging from smartphone assistants to Discord chatbots [18, 3]. No one can argue against the lofty intention of those applications and how much users will be benefited from those services. However, language is complicated, as we humans use it as a conduit for exchanging profound social and contextual information. It is a proxy for social biases (and thus discrimination): existing research has found that much social information such as age, gender, or race can be correctly or incorrectly inferred from only a few lines of sentences, sometimes even one word [4]. Consequently, given that language is the most prevalent data surrounding our daily life, they inevitably suffer from conscious or unconscious biases and gender/racial disparity. NLP models trained based on those data are more susceptible to biases. Learning directly on that corpus without meticulous anatomy of data might lead to a misbehaved model, as shown in the example of a Microsoft chatbot becoming a racist after feeding twitter data [2]. Similarly, considering the classic riddle in the epigraph, many people are reportedly unable to solve the riddle [33], an epitome of underlying implicit gender bias in the natural language text. Most people overlook the possibility that *a surgeon can be a she*. Rudinger et al. [28] observe that the contemporary NLP model is also unable to relate the role of "surgeon" and "mother" as the same entity, shown in Fig. 1. The above examples demonstrate that existing unfair NLP models embed historical patterns of linguistic discrimination and reproduce systematic stereotypes. We emphasize that deploying those unfair NLP models in services prevent certain "non-standard" end-user group from fully taking advantage of services and by and large accrue uneven benefits to end-user groups, which is only harmful to social justice.

The contributions of this paper are two-fold: (1) we give an overview of NLP and bias, and show that NLP algorithms do have biases, and (2) we analyze two types of NLP biases mitigation methods and evaluate their effectiveness under a wide range of English NLP tasks.

[2]In this work, we do not distinguish between machine learning, machine learning algorithm, machine learning model, and artificial intelligence.

2 Related Work

In this section, we provide relevant context to readers.

Natural Language Processing The paradigm of training a machine learning model is to collect datasets with annotation, input those data to the model, and optimize the model with the goal of making the model understand the input[3]. In contrast to other machine learning sub-fields like computer vision, whose domain is images consisting of continuous-valued pixels [19], natural language processing focuses on textual data, and the domain is discrete text strings. Therefore, it is uneasy to optimize NLP machine learning algorithms until a decade ago [18] when Mikolov et al. [21] purposed a sea change. Instead of viewing "apple" as a five-character word, this method regards it as a dense vector that encodes semantic meaning. A dense vector, or word vector [21], is more suitable in that computer then does not need to cope with sparse and discrete words but instead a fully differentiable continuous representation [7, 5]. However, one implication of operating on high dimensional vectors instead of words and sentences is that machine learning models become hard to interpret. As the model becomes larger and deeper, it is increasingly difficult to explain why a specific decision or prediction is made [20, 7], leaving challenges to mitigate biases.

Machine Learning Bias A machine learning algorithm is biased if there exists a performance disparity among the user groups [15]. A biased algorithm perform poorly in one specific user group, lower than the average overall performance. The definition of the user group is quite loose. It can refer to the actual end-users who access the machine learning-backed service. For example, consider the scenario of predicting which of the two candidates is suitable for a job posting [8, 24]. If the two candidates have the same resume with the only difference of gender, then the prediction should be equal-likely, thus an algorithm that prefers one candidate is marked biased. User group can also refer to some specific components of the input. Consider the word vectors [21]. Garg et al. [13] have found that when the user sends a sentence to a model that contains an Islam adjective, the model will relate this sentence, whatever it is, to terrorists. The algorithm is biased in that the word vector component of the input misleads the model to correlate with terrorists. In summary, machine learning bias is *unfair*. It is the unjust and prejudicial treatment of people of specific social attributes, including race, sexual orientation, education, or income.

3 Natural Language Processing Algorithms have Biases

We first demonstrate that NLP algorithms do have biases, and such biases can be harmful to end-users.

There is no universal way to quantify biases. Most of the NLP algorithms used nowadays are black boxes. Given the complexity of computing predictions [27], it is hard to quantify biases and pinpoint which sub-part of the algorithm is the culprit, let alone mitigating biases. One practical workaround concocted by researchers is to design an indirect indicator to quantify the biases.

One of the most widely-deployed approaches is WEAT [6]. This work curates a list of concepts, each concept containing a series of idiomatic words commonly used in this concept, and proposes to use the vector distance of word vectors [21] in different pairs of concepts as an indicator of biases. For example, a pair of concept for gender bias can be `science` and `art`. Given a word "woman", we measure the distance of this word to every word in concept `science`, say "surgeon" and the distance of every word in concept `art`, say, "artist". If "woman" is closer to "artist" compared to "surgeon", models that use these word vectors are going to more frequently connect females with artists, but less likely to connect females with surgeons. In this case, we conclude that gender bias in word vectors does exist. Every model that uses such word vectors is susceptible to gender bias as well, and will neglect that a surgeon can be a she (shown in Fig. 1). Every concept idiomatic word series is collected based on psychological studies, thus the metric is considered convincing by many NLP researchers [20].

[3]Thus the name machine *learning*. Yet another layer of complexity is that the precise definition of understanding depends on the task. For computer vision dog-vs-cat image classification, the understanding means based on the image alone the model is able to recognize whether the image contains a dog or a cat; while for natural language processing toxicity classification, the understanding means the model is able to correctly tag toxic sentences.

Equipped with this tool, several studies were conducted to investigate various variants of word vectors that are used in virtually every NLP downstream application. Bolukbasi et al. [5] found that gender bias is embedded in the word vectors related to occupations. "Housekeeper" is around 50% closer to "woman" compared to "man", while "engineer" and "mechanic" is around 50% farther away from "woman" compared to "man". Additionally, Garg et al. [13] discovered that racial bias is also embedded: a word like "Asian" is closer to words such as "outsider", while a word like "Islam" is closely related to the "terrorist". Furthermore, Goldfarb-Tarrant et al. [14] found that such biases do not merely embed in English corpus, but in Spanish corpus as well, indicating that word vectors suffer from biases regardless of language.

Services built on top of such biased algorithms will lead to chaos. As correctly noted by The Guardian, "although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice" [1]. No one will feel gratitude if they are harassed by the police because a serious discussion of Islam theology triggered the machine learning terrorist prevention algorithm [13], African Americans will find it more difficult to communicate African-American Vernacular English (AAVE) since a majority of input methods on market do not support smart auto-completion or audio recognition for AAVE [9, 15], and women will be less empowered as most of the machine learning algorithms are male-orientated [34, 5, 28, 16, 11]. In other words, those minorities who are the most marginalized and in the deepest desire for such technology are ironically the ones who are most likely to be systematically excluded from the technology.

4 Biases Mitigation

Hope is not all lost, and in recent years we have witnessed several works that focus specifically on mitigating those biases and building a fair NLP model. Previous works that mitigate biases can be largely grouped into two categories: fair data collection [9] and optimization calibration [15, 34]. The two directions are not mutually exclusive and it's possible to integrate two methods to yield a more unbiased model (refer to §4.2 for discussion). In this work, we investigate a total of three algorithms from these two directions and layout a detailed comparison of these three algorithms in Tab. 1. We analyze the algorithms in four dimensions. Cost-efficiency means whether the method is practical and efficient in terms of implementation; resilience stands for the resilience to bias amplification problem (see §4.2); generalization evaluates whether the method is transferable in a variety of NLP tasks; and extensibility indicates whether the method is able to extend to other models or configurations.

Table 1: Comparison of three algorithms investigated in this paper.

	Fair Data Collection (§4.1)	**Optimization Calibration (§4.2)**	
	Dixon et al. [9]	**Hashimoto et al. [15]**	**Zhao et al. [34]**
method	data re-sampling	worst group performance	worst group performance + overall performance
cost-efficiency	✓	✗	✓
resilience	✗	✓	✓
generalization	✗	✓	✓
extensibility	✗	✗	✓

4.1 Fair Data Collection

One direction focuses on the data collection. Machine learning after all is recognizing patterns from large-scale data. To this end, data collection is by no means indispensable. However, such taxing jobs as collecting and annotating data have been shown to be expensive and error-prone [23]. On the one hand, human annotators might have biases and such biases will be reflected in the annotated dataset. On another hand, annotation by itself leads to prototypical biases. Annotation requires categorization of objects, yet, according to Rosch [26], "there may be some central, prototypical notions of items that arise from stored typical properties for an object category", implying that annotation, *i.e*. categorization process, subjects to prototypical proportions by nature. As a result, many of the published datasets are not inherently fair. For example, men are over-represented in

web-based news articles [16] or Twitter [11], both are large-scale corpus that is used to train the latest NLP models. What's more, to satisfy machine learning's ever-growing demand for more data, datasets are collected without fair scrutinization. Therefore, it is reasonable to presuppose a fair data collection leads to a fair model after training.

In practice, other than hoping future datasets to be fairer, it is more cost-efficient to modify the existing biased datasets. Dixon et al. [9] take a critical step toward utilizing such techniques to mitigate biases in the toxicity classification tasks, in which the algorithm tags whether the input sentence is toxic or not. To be concrete, Dixon et al. [9] re-sample the original (possibly biased) dataset to be more equal in quantity with respect to some attributes, say, sexual orientation. For example, if sentences that contain gay-related words are 5% less compared to other sentences, they will duplicate those sentences by 5%. Naive as it sounds, this method is found to be effective on the false positive rate, the percentage of a seemly innocent sentence like "I am a gay man" will be classified as toxic [9, Table 5, Figure 5]. Considering sentences that contain gay-related words, due to the rarity in the original biased dataset, there is a 10% false-positive rate, yet after training with this approach, the false positive rate drops to 2.5%, indicating that the model is able to learn that gay-related words by themselves are not red flags for toxic language through capturing better toxicity semantics from the sentence thanks to a more balanced and fair dataset.

Although effective in this specific task of toxicity classification, Dixon et al. [9] don't seem to generalize well to other tasks. We have compared with other works that use this algorithm on other NLP tasks, including Goldfarb-Tarrant et al. [14] on coreference resolution task, Gardner et al. [12] on question answering, and Zhao et al. [34] on visual semantic role labeling. We found that generally Dixon et al. [9] don't yield a much fairer model. One limitation is that re-sampling will always duplicate a small portion of the data (*i.e.* sampled multiple times), thus lacking diversity. The trained model might be able to memorize such duplicated data but failed to *learn* from it. As a consequence, even if "I am a gay man" is classified correctly by the model, a small perturbation of the sentence like "They are gay men" would still not be understood by the model and be tagged as toxic. Therefore for unseen data in other tasks, the model performs poorly, *i.e.* poor generalization. Moreover, it is difficult for such a naive method to extend to multiple attributes by merely sampling because re-sampling [9] is only capable of accounting for one single attribute at a time. But good-performing and fair models require the multi-facet property of the input dataset to develop a deep understanding of the input. As an example, instead of merely accounting for the sexual orientation of the subject in the sentence to predict whether the input sentence is toxic or not, the model might need to refer to the surrounding context, similar to what humans do. With only one control variable, it is hard to synthesize the dataset for the model to learn comprehensively.

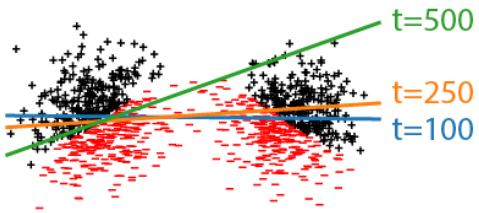


Figure 2: An illustrative example of bias amplification. Image is taken from Hashimoto et al. [15]. After iteratively training, the model becomes increasingly unfair.

4.2 Optimization Calibration

Another direction to building a fairer model is optimization calibration, namely calibrating the process of model optimization. In this direction, instead of relying on assumption that training data is fair, researchers inject various kinds of optimization components into the model training process such that the model is made aware of the potential biases in the dataset and accordingly counteract biases. This approach has received more discussions compared to the direction described in §4.1 due to a finer level of granularity researchers have control over [20].

Moreover, it is more resilient to the biases amplification problem [34, 15]: if the model learns completely from the training dataset, then the prediction should be roughly proportional to the biases in the dataset; however in reality, Zhao et al. [34] and Hashimoto et al. [15] both observe that after the model trains on the biased dataset, the biases of the dataset become amplified by the model. Consider a motivating example shown in Fig. 2. The two clusters are drawn from two independent Gaussian distributions and will be the input of the model. They can be viewed as two user groups. In the beginning, the model is predefined to be almost fair since it gives a similar performance for both

groups. But after iteratively training, the model will prefer one group (the right cluster) over another, as shown in the green line `t = 500` in which the left cluster is sacrificed (having a poor performance) to obtain higher performance on the right cluster. Dixon et al. [9] can generate a fair dataset with respect to one single attribute but have no means to prevent the model from becoming increasingly unfair over time.

In light of the benefits provided by optimization calibration, two promising approaches in this direction emerge.

Distributionally Robust Optimization Hashimoto et al. [15] take advantage of distributionally robust optimization theory [10]. The intuition of this approach is not to train a model to obtain the best overall performance, as it leads the model to sacrifice one user group to "cheat" for higher performance, but to enforce the model to optimize over the *worst user group performance*. Consider again the example in Fig. 2, if it turns out that during training the left cluster is hardly learned by the model, the performance of that cluster will then be relatively low, leading the model to focus more on the left cluster and counteract the poor performance on the left cluster. Since the worst user group might change during the training, such a method is also adaptive to the scenarios of multiple worst user groups. Indeed, Hashimoto et al. [15] demonstrate that for keyboard auto-completion system task, it can largely improve the minority group, that is, African-American Vernacular English, leading to a 15% African-American user retention improvement and 0.3 better user satisfaction[4]. Sagawa et al. [29] further generalize this method to other tasks including natural language understanding task and female face recognition, demonstrating a good generalization for this approach.

However, one of the major shortcomings of such an algorithm is that, due to the complexity and high non-linearity of deep machine learning models, it is difficult to train a deep and complex model using this schema. Hashimoto et al. [15] use a convex model with limited capacity, and to the best of our knowledge, there is no work that successfully applies distributionally robust optimization theory on an overparameterized models (complicated but with high capacity), hurting the extensibility of this algorithm. This leaves a dilemma: to produce a machine learning model to enhance downstream user experience, we need a fair and overparameterized model with great capacity, yet such a model can't be produced fairly using Hashimoto et al. [15] unless it is a model with limited capacity. Moreover, since the optimization objective becomes harder, training directly will lead to quadratic computational overhead, and is generally considered difficult to train smoothly [29], making it inefficient with respect to cost.

Constraints Injection Zhao et al. [34] take another route to combat biases, namely injecting constraints for overall performance. In the context of distributionally robust optimization theory [15], what Zhao et al. [34] purpose can be viewed as jointly optimizing worst group performance *and* overall performance. Such design makes the model focus on both metrics since both are crucial for a fair and useful model. To overcome optimizing difficulty (*i.e.* cost-inefficiency) and poor extensibility, they discard the direct optimizing schema but reformulate the optimization objective into training constraints. Those constraints are injected into model training and indirect optimization techniques such as Lagrangian Relaxation is used to solve the optimization numerically. Since this method is agnostic to the chosen model, researchers are free to choose an overparameterized model. Overall in the task of visual semantic role labeling task, they have decreased the amplification bias by 40.5%. Moreover, such an approach is transferable to other datasets or tasks as well. Jia et al. [17] applied a variant of such strategy on imSitu dataset, and Savoldi et al. [30] used the same strategy on the machine translation task. This strategy is also applicable even outside of the NLP domain. For example, Wang et al. [32] used this approach to mitigate bias in computer vision object classification tasks. They all observe that such an approach is effective in helping mitigate biases.

Despite empirically-tested effectiveness in all four dimensions, we have found that all previous works that utilize this algorithm mostly reduce the biases amplified by the model, leaving biases stemmed from the training dataset biases untouched. This is partially due to the constraints injected in the optimization process do not reflect the original biases in the training dataset, as a result leading to little effort in mitigating the original biases. One further potential improvement of Zhao et al. [34] might be combing fair data collection technique (§4.1), *i.e.* forcing the training dataset to be fair through re-sampling [9] and then injecting constraints in optimization. Caveats, however, exist in that injected constraints must comply with the re-sampled dataset, so a compatible combination strategy

[4]In this work, user satisfaction is qualified as a scalar ranging from 0 to 5.

must be devised to integrate those two directions seamlessly. We leave this combination direction to future works.

5 Conclusion

In an era in which machines are capable of beating the best go players and achieving super-human performance in several understanding and reasoning benchmarks, we iteratively strengthen machines in the hope that they will obtain better and better cognition. We have proceeded too rapidly and paid too little attention to biases implicitly or explicitly demonstrated by machines, completely neglecting that as humans have biases, machines that imitate human behaviors will learn to be biased. As we are cheering for another human baseline being beaten by the machines, we should not forget the intention of building such powerful machines, that is, benefiting every user regardless of social identity. Fairness is always an indispensable property of such machines; and machines must be regulated to achieve fairness. In light of this, we have investigated the NLP biases and two approaches to mitigate biases in the English NLP domain. Albeit preliminary, we hope that this work can provide insights to the community of NLP practitioners and researchers, and be helpful in devising a fair model in production.

References

[1] The guardian view on machine learning: people must decide | editorial | the guardian. `https://www.theguardian.com/commentisfree/2016/oct/23/the-guardian-view-on-machine-learning-people-must-decide`. (Accessed on 06/06/2022).

[2] Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day - the verge. `https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist`. (Accessed on 06/01/2022).

[3] E. Adamopoulou and L. Moussiades. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer, 2020.

[4] J. Baugh. Racial identification by speech. *American Speech*, 75(4):362–364, 2000.

[5] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

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

[7] J. Camacho-Collados and M. T. Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, 2018.

[8] L. Cohen, Z. C. Lipton, and Y. Mansour. Efficient candidate screening under multiple tests and implications for fairness. *arXiv preprint arXiv:1905.11361*, 2019.

[9] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[10] J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[11] D. Garcia, I. Weber, and V. R. K. Garimella. Gender asymmetries in reality and fiction: The bechdel test of social media. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[12] M. Gardner, Y. Artzi, V. Basmova, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[13] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644, 2018.

[14] S. Goldfarb-Tarrant, R. Marchant, R. M. Sánchez, M. Pandya, and A. Lopez. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*, 2020.

[15] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[16] S. Jia, T. Lansdall-Welfare, and N. Cristianini. Measuring gender bias in news images. In *Proceedings of the 24th International Conference on World Wide Web*, pages 893–898, 2015.

[17] S. Jia, T. Meng, J. Zhao, and K.-W. Chang. Mitigating gender bias amplification in distribution by posterior regularization. *arXiv preprint arXiv:2005.06251*, 2020.

[18] W. Khan, A. Daud, J. A. Nasir, and T. Amjad. A survey on the state-of-the-art machine learning models in the context of nlp. *Kuwait journal of Science*, 43(4), 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[20] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[22] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research*, 9(5):583–597, 2002.

[23] J. Podesta, P. Pritzker, E. J. Moniz, J. Holdren, and J. Zients. Big data: Seizing opportunities and preserving values. *Executive Office of the President*, 2014.

[24] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[25] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[26] E. Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

[27] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[28] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL `https://aclanthology.org/N18-2002`.

[29] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[30] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021.

[31] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[32] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

[33] M. Wapman and D. Belle. Undergraduate thesis. `https://mikaelawpman.com/category/gender-schemas`. (Accessed on 06/06/2022).

[34] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 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, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.