Ethically Sourced Modeling: A Framework for Mitigating Bias in AI Projects within the US Government

Melanie Laffin
Booz Allen Hamilton, Georgia Institute of Technology
laffin_melanie@bah.com

Abstract

The increasingly widespread use of Natural Language Processing (NLP) in AI applications must be continually monitored for biases and false associations, especially those surrounding protected or disadvantaged classes of people. We discuss methods and algorithms used to mitigate such biases and their weak points, using real world examples in civilian agencies of the US government.

1 Introduction

The notion of bias mitigation in AI is motivated by the interdisciplinary nature of AI. While statistical bias within machine learning (ML) algorithms has been well studied, there also has been a significant interest in studying bias propagation and fairness from a humanistic standpoint. In particular, motivation can be drawn from an interest in bias mitigation with regard to protected classes, such as sex, gender, citizenship, genetic information, and race within regulated domains such as education, housing, and health. Not only is there an ethical obligation to reduce bias in such domains, protecting underprivileged members of protected classes in regulated domains is legally mandated through congressional acts such as the Civil Rights Acts of 1964 and 1991, the Genetic Information Nondiscrimination Act, and the Immigration Reform and Control Act. As a result, we suggest all aspects of an AI solution be analyzed for fairness and bias analysis and mitigation must be performed.

We discuss an application of bias detection and mitigation in Artificial Intelligence. Our application will be in the field of financial and health services, specifically regarding how to mitigate biases found in using ML and NLP applications.

1.1 Background

In the age of analytics, it is imperative that architects and developers analyze their solutions for any potential bias toward a specific group. While many solutions look for statistically significant bias, this analysis may not take into consideration underrepresented groups who are members of a federally protected class [Sun *et al.*, 2019]. For example, any modeling done on a sample of a population entering a hospital may not account for less underprivileged groups in race or gender due to such groups being less represented in the population of data. Such biases must be recognized and immediately mitigated, especially for projects that are in pursuit of providing services or studies by the US Government.

While there are algorithms developed to identify and mitigate bias, it's important to note that

1. Any algorithm created will introduce its own biases, which must also be identified and mitigated
2. Data may be too sensitive to upload to some tools. As a result, careful consideration while identifying and mitigating bias and other markers for inequality requires detailed analysis to inform decision making regarding the mitigation. We recommend using a customized methodology that amalgamates several methods for an end to end bias mitigation tool, which are run locally and therefore does not expose private data.

2 Survey of Major Frameworks

There are several major frameworks openly available that can be applied to identifying and mitigating bias in AI solutions,.

2.1 AI Fairness 360 Tool

The AI Fairness 360 Tool is made widely available by IBM. The package consists of a set of fairness metrics for datasets and models, explanations for these metrics, and algorithms to mitigate bias in datasets and models, which includes an interactive Web application. The tool has been engineered to conform to standards in data science, improving usability for practitioners. The architectural design also enables users to easily extend the toolkit with new algorithms and metrics. [Bellamy *et al.*, 2018]

2.2 What-If Tool

The What-If Tool is available to the public as a Google product. The What-If tool is an open-source application that allows minimal coding to probe, visualize, and analyze ML systems. The What-If Tool lets practitioners test performance in hypothetical situations, analyze the importance of different data features, and visualize model behavior across multiple models and subsets of input data. It also lets practitioners measure systems according to multiple ML fairness metrics. [Wexler *et al.*, 2020]

2.3 The need for customized tools

While both the AI Fairness 360 Tool and What-If tool have many great capabilities, there is concern about using their platform and API within the US Federal Government due to privacy issues. Therefore, a customized methodology as described later in this paper that can be run privately and without exposing any data outside of government systems is preferred.

3 Application to Government Agencies

We present two use cases from real-world examples found in a financial agency of the government, and one that specializes in human health.

The financial example involves analyzing complaints of non-compliant or illegal behavior and triaging them. These complaints are funneled through an email box which is open to the public, and thus receives a mass volume of email – most of which results in an unproductive complaint. Here, we use NLP and ML to predict or suggest a risk rating which performs triaging into cases that are more likely to be urgent and productive.

The second use case takes applications for funding and other artifacts surrounding already funded contracts and using NLP to categorize them into specific contract types for archival purposes. While this use case is less risky since the decision for funding has been made, we must still be mindful of any bias or "false relationships" that are embedded in either the data or solution.

In both cases, we developed the following customized methodology to attempt to mitigate any bias that may have crept into either the data or the ML model. As a result of using and analyzing existing tools and discussing ethical frameworks, we have developed a methodology for these agencies to follow when performing future AI-related tasks and to check for any latent bias in these solutions that may affect underrepresented or legally protected classes.

Since this data being owned by US Government agencies, special consideration is given to the repeatability of these results as the data is not publicly available and is sensitive toward financial and sensitive research projects related to human health data. Moreover, any results must be especially scrutinized as they immediately affect the US population at large.

4 Methodology Overview

Regardless of which use case we considered, the high-level approach included injecting checking for bias in every step of the solution development lifecycle. For our purposes, we consider the solution development lifecycle to be preprocessing (prior to training the model), in-processing (training the model), and post-processing (analyzing the output of the model). We do not include post-deployment monitoring as part of the methodology, but arguably many of the methods mentioned may be applied during the operational maintenance of a ML model. We use each of the following methods to create an end-to-end bias mitigation tool. The use of all methods requires careful analysis post use of that specific method to see where there are potential biases.

4.1 Pre-processing techniques

Reducing bias in a prepared data set prior to training the ML model can be achieved through several preprocessing techniques, which may depend on the type of ML model, in particular if it is a classification or prediction model.

Optimized Pre-Processing for Discrimination Prevention

Data preprocessing is controlled with special consideration for controlling discrimination, limiting distortion in individual data samples, and preserving utility. The algorithm takes a randomized mapping transforming the original dataset into a new dataset where discrimination is minimized. The objectives of this method include limiting the dependence of the transformed outcome on protected variables, restricting the mapping to minimize or avoid certain large changes in outcome (e.g. mapping a low credit score to a much higher one), and that the statistical distribution of the transformed dataset is similar to that of the original one. [Kamishima *et al.*, 2012]

Certifying and Removing Disparate Impact

As defined here, disparate impact is a form of indirect and unintentional discrimination in which certain hiring, promotion or employment decisions disproportionately affect members of legally protected groups. In this method of bias mitigation, the goal is to make the transformed dataset have a disparate impact value of above 80%, which again assumes to only have two classes – privileged and underprivileged. This method defines a number of metrics to consider on the original dataset, and then uses it to certify Disparate Impact within a dataset. [Kamiran *et al.*, 2012]

4.2 In-processing techniques

There are several techniques that are also targeted toward removing bias and discrimination while the ML model is being trained.

Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees

The meta-algorithm for classification takes a general class of fairness constraints as input. Classification tasks are often formulated as constrained optimization problems that maximize accuracy. A classifier with fairness constraints uses tailored algorithms that are also constrained to common fairness metrics including statistical parity and equalized odds. This algorithm can handle non-convex "linear fractional" constraints (which includes fairness constraints such as predictive parity) for which no prior algorithm was known. Empirically, it was observed that this meta-algorithm is fast, can achieve near-perfect fairness with respect to various fairness metrics, and the loss in accuracy due to the imposed fairness constraints is often small. [Celis *et al.*, 2019].

Fairness-Aware Classifier with Prejudice Remover Regularizer

This method is focused on classification algorithms and utilizes regularization called a prejudice remover. The prejudice remover actually utilizes two regularizers: the first is a standard regularizer that removes overfitting of the model, whereas the second computes the minimum of the prejudice index defined as a function of the overall data population, the underprivileged class (or sensitive class), and the privileged class (nonsensitive class). [Feldman *et al.*, 2015]

Adversarial debiasing

Adversarial debiasing relies on adversarial training in order to remove bias from representations learned by the model. If the original model produced a representation that primarily encodes information about an attribute (e.g. race or sex), an adversarial model could recover and predict that attribute without that representation. Equivalently, if the adversary fails to recover any information about the attribute, then there must be a successfully learned representation of the input that is not substantially dependent on the attribute. [Zhang *et al.*, 2018]

Variational ”fair” autoencoders

In contrast to the more standard uses of neural networks as regressors or classifiers, Variational Fair Autoencoders (VFAEs) are powerful generative models, which store latent attributes as probability distributions. As a result, VFAEs allow for easy random sampling and interpolation to aid in testing for bias. Moreover, VFAEs characterize fairness as a representation that is invariant with some respect to the known aspect of the dataset. Because this method uses a semi-supervised approach, it can especially be useful for unlabeled data. VFAEs may be combined with adversarial neural nets (VFAE-GANs) in order to combine the power of both in order to detect bias. By combining a VFAE with an adversarial network, learned feature representations in the adversarial neural net discriminator serve as a basis for the VAE reconstruction objective. Thereby, element-wise errors are replaced with feature-wise errors to better capture the data distribution which allows for additional methods to capture potential biases. [Larsen *et al.*, 2015]

4.3 Post-Processing Techniques

The final aspect of the ML pipeline also may benefit from bias mitigation techniques specific to post-processing.

Fairness, calibration & equality of opportunity

This method focuses on analyzing and changing model predictions to satisfy specific fairness definitions. Specifically, it considers equalized odds, equal opportunity, and oblivious measures to analyze. Once this analysis is done, the method takes a derived predictor and derives a scoring function to score the model and provide guidance on any changes that should be made during the pre-processing or in-processing parts of the ML pipeline. [Hardt *et al.*, 2016]

Decision theory for discrimination-aware classification

Assuming that the most discrimination occurs close to the decision boundary, and thus exploits the low confidence region of a classifier for discrimination reduction. This method exploits, respectively, the reject option of probabilistic classifier(s) and the disagreement region of general classifier ensembles to reduce discrimination. Notably, this method does not require data modification nor classifier tweaking. [Kamiran *et al.*, 2012]

4.4 Solution Framework

Currently, there are no specific frameworks or methodologies that are widely used to identify or mitigate bias for AI solutions within agencies at the US Government. This solution overviewed in the methodology section builds such a framework to be repeated and reused by amalgamating different aspects of various attempts at identifying and mitigating bias in ML and NLP. It our suggestion to use each of the methods described in the methodology section to find and mitigate potential bias in the ML model.

We note that anything created by human beings will reflect the person who made it and carry their inherent and possibly unconscious bias. The framework and methodologies discussed above will contain our own biases as direct corollary. Moreover, a lack of data regarding protected classes with the data may introduce biases that should not exist, for example, if an email came in from a person with a typical name of a certain ethnicity. As a result, we are always looking for new studies or different methodologies/algorithms/frameworks to reduce bias and make a more equitable AI ecosystem.

- Sharing AI Ethical Frameworks and customized algorithms and methodologies with our clients to engage in the discussion of any potential bias within the data provided as a result of their business process [Floridi and Cowls, 2019]
- Balancing lack of statistical significance with underrepresentation in the dataset while cleaning the data and preparing for use in the model. For example, not removing all people of a protected class such as age, gender, sexuality, etc. because they are statistically less represented or anomalous
- Notating any known bias issues with the word vector, dataset or model used (e.g. GloVe is well known to have issues with gender bias [Vera, 2019])
- Running the data through both the original model, and then through bias detection and mitigation algorithms listed in the methodology to use all tools available to find and share reports of bias for an equitable dataset to perform ML tasks
- Discussing the results with the business process owners (or data owners) and determining what, if any, bias must be mitigated or called out in order to maintain fairness and equality for legally protected classes such as race or age

Every step is iterative and meant to improve the overall solution for a more equitable result.

5 Conclusion

By utilizing this customized and repeatable framework for bias and fairness detection and mitigation, we provide US Government Agencies the ability to easily and comprehensively understand bias in both their datasets and any ML/NLP models that they utilize to make important decisions that affect the broader US population.

References

[Bellamy *et al.*, 2018] Rachel Bellamy, Kuntal Dey, Michael Hind, Samuel Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Ramazon Kush, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 10 2018.

[Celis *et al.*, 2019] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 319–328, New York, NY, USA, 2019. Association for Computing Machinery.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[Floridi and Cowls, 2019] Luciano Floridi and Josh Cowls. A unified framework of five principles for ai in society. *Harvard Data Science Review*, 06 2019.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[Kamiran *et al.*, 2012] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *2013 IEEE 13th International Conference on Data Mining*, pages 924–929, Los Alamitos, CA, USA, dec 2012. IEEE Computer Society.

[Kamishima *et al.*, 2012] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[Larsen *et al.*, 2015] Anders Larsen, Søren Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. 12 2015.

[Sun *et al.*, 2019] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics.

[Vera, 2019] M L Vera. Exploring and mitigating gender bias in glove word embeddings. 2019.

[Wexler *et al.*, 2020] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.