Tools for Identifying AI Biases for Machine Learning Models

Austin Crow
George Mason University
acrow2@gmu.edu

Patcharaporn Adchariyavivit
George Mason University
padchari@gmu.edu

Victoria Golden
George Mason University
vgolden@gmu.edu

***Abstract*—There is no doubt about the usefulness of machine learning in today's data environment. Machine learning algorithms are used across many domains and across a variety of problems. These algorithms are seen in e-commerce such as Amazon's recommendations , financing through loan applications, image processing, autonomous vehicles, and speech recognition among many others. Acknowledging the market penetration of machine learning and its relevance to many big data related challenges, it is important to address the effect of bias in machine learning.**

***Index Terms*—AI, Fairness,Algorithm, Context**

I. Introduction

Bias in the context of machine learning refers to "any basis for choosing one generalization over another, other than strict consistency with the instances" . Bias comes from a variety of sources including humans, machines, and nature. Due to it being present in virtually every dataset, machine learning methodologies base their results on a foundation which can lead to possible unfairness in the affected groups. According to Mehrabi et al., fairness is "the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [3]. Due to the omnipresence of machine learning in today's society, developers should strive to identify the underlying biases that exist in their data in order to ensure no entity is treated unfairly.

Machine learning bias is not something that is as complicated as it may sound, but rather it is an unconscious or conscious skew in data. A great description of bias is "... omissions and deliberate choices of inclusion may show a particular bias" [4]. There are many different types of bias and not all of which are bad. But a rule of thumb is to always clean the data and if something appears to be an outlier it is likely worth looking into. The most common examples of data bias are sample, exclusion, measurement, recall, observer, racial, and association bias [5]. If you are cleaning data that you are not familiar with it can be challenging because it is hard to determine whether or not the data is skewed in an unintentional way. When cleaning data, it is best to exercise caution as the data may be valuable without appearing so during this phase of the data analytics lifecycle.

Bias affects all types of data sets from any time frame, this bias is often harmful when coming up with an algorithm for machine learning. The machine learning algorithm can only work off the data that you use to teach it. If you leave this skewed data in, it will perpetuate the bias to further generations of data. This perpetual bias can create unfair work environments, whether it be based on gender, ethnicity or anything else that should not have an impact on the work that is being performed. An example by Stolzfus indicating the effects of underlying bias is "If it perceives that men hold the vast majority of executive jobs, and the machine learning process involves filtering through the raw data set and returning corresponding results, it's going to return results that show a male bias." [4]. There is no proven reason why men should hold an executive position over women, so a machine learning process like this will only serve to hurt the company using it or show bias data that pushes false reasonings.

Bias is introduced into a dataset through a variety of means. Examples include population bias where only survey recipients who responded offer input, instrument-dependent bias [6] where the sampling instrument measures data in a way that alters the input received, among many other types of bias. Bias has many definitions often suited to the project, domain, or objective. One definition of machine learning bias can be defined as "any basis for choosing one generalization over another, other than strict consistency with the instances" [2]. Bias in the context of statistics can be defined as "a model or statistic is unrepresentative of the population" [7]. In either case, bias has the opportunity to drive a machine learning algorithm's results in favor of one subgroup over another causing unfairness. According to Caliskan et al., "machine learning can acquire stereotyped biases from textual data" [8]. Another example from Datta et al. describes personalized ads from Google showed discrimination by suggesting ads that promised large salaries more frequently for males as compared to females "simulated male ads from a certain career coaching agency that promised large salaries more frequently than the simulated females." [9]. Another example by Thelwall shows a core component of some algorithms is the ability to deduce the meaning of words by associating them with other words that tend to occur in the same document. Using this approach can lead to conservative implications, such as that homemaker is part of the "meaning" of the word woman and that programmer is part of the meaning of the term man [10].

Bias in big data shows that people often make conscious or unconscious decisions that can have major effects on people's lives. Big Data gets used to create machine learning algorithms, basing the algorithm off of biased data will lead to the

same biases down the line. Some of the mitigation strategies are to be diligent when cleaning the data to make sure that there are not obvious biases being shown. Also, diversifying the data making sure it comes from different sources, as well as having a robust amount of data sets. As with all data, including big data, bias is always a concern that developers and analysts need to be cognizant of. Nowadays, many developers try to integrate bias mitigation steps by developing private tools that generally have the same goal. There are 3 different steps during the process to incorporate these methods which include pre-processing, in-processing, and post-processing. The pre-processing starts at the beginning of searching through the data. If a dataset is found to have bias, steps to mitigate can be addressed at this point. Some data sets may contain unwanted biases, an example could be selection bias where ”t is usually associated with research where the selection of participants isn't random” [11]. Collecting multiple sources of data is an easy way to prevent selection bias using diverse samples to represent the population. The pre-processing method allows developers to catch unbalanced and unfair data before entering the in- process stage. For in-process, meta-algorithms (machine learning algorithms that learn from other machine learning algorithms) collect fairness metrics as input then returns new classifiers that are optimized in favor of the fairness metric. The last category is post-processing. Because the data was trained already, adjustments based on bias will be trained classifiers. This method spends less time than others because it uses trained data so there is no need to look back to the original dataset. However, for this method the accuracy needs to be validated. [12]

II. Method

The proposed methodology for this assessment will begin with the identification of publicly available tools to identify biases within a dataset. Each tool identified will then be compared to the other available tools to understand their capabilities. Tools being considered for comparison IBM AI Fairness 360, Google Tensorflow Fairness Indicators, Pymetrics Audit-AI, Datascience.com Labs Skater

After understanding how each tool functions, including similarities and dissimilarities, the authors will devise a ranking mechanism to score performance of each tool to ultimately identify the best available option for general use.

In order to find a dataset with a known bias, the authors will perform extensive literary review of research articles. The research articles studied will be related to machine learning projects. The identified dataset must have one or more of the following characteristics : Protected attributes, Significant volume, Contain individual or group bias

The dataset may require pre-processing in order to be utilized by each of the tools evaluated. The pre- processing that occurs will be done in accordance with best practices documented by the tools being used. The pre-processing that occurs will not affect any intrinsic bias as to keep biases consistent across processed data sets and tools [1]–[8].

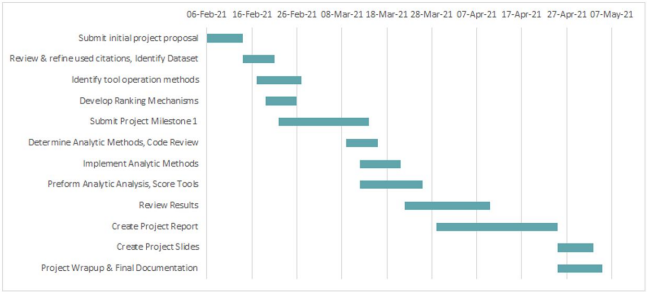

Fig. 1. Number of cases/deaths across time

Fairness and AI Applications in social media can be detected based on different tools in Social media [9] [10], [11] [12], [4] . [10], and [11] [13]. [14]–[30]

Each tool will be used with the identified biased dataset. The tools will then be evaluated utilizing the previously identified ranking mechanism. The authors then will select the ”best” tool according to their ranking system, compare and analyze the implications of utilization by each tool, and recommend the results to the reader. By understanding the capabilities of each tool, the authors hope to recommend a relevant publicly available tool to help improve machine learning projects through identification of biases in data.

A. Ranking Mechanisms Method and Evaluation Criteria

The authors decided to use the Analytic Hierarchy Process (AHP) as the means for scoring each tool against each other. This process was developed by Dr. Thomas L. Saaty in the 1970s. The process ”allows the decision makers to visually structure a complex problem in the form of a hierarchy having at least two levels: objectives (criteria for evaluation) and activities (productions, courses of action, etc.).” [18]. The AHP excels in decision making processes when it comes to ranking or priority setting for projects [19]. The authors chose to evaluate the tools based on the following objectives outlined in the next sections.

The applicability of this objective refers to scenarios such as availability of documentation, pre-processing requirements, etc. Tools will be scored based on presence or absence and extensiveness.

Applicability refers to scenarios such as relevance to machine learning algorithms, frequency of updates, language support diversity, and use throughout a project. Relevance to machine learning algorithms refers to the support a tool provides across issues being tackled by any given machine learning algorithm. Use throughout a project refers to a tool's ability to support identifying bias in either pre-processing, in-processing, or post-processing stages.

The applicability of this objective refers to a tool's ability to work across the many development environments available in today's IT landscape, support requirements within a language, etc. Each tool will be evaluated on its ability to support varying sizes of datasets as well as its speed to process such varying

sizes of datasets. Accuracy. Each tool will be evaluated on its ability to identify bias from a known biased dataset.

III. Essential Research Development

The timeline below will provide an approximate schedule for execution of project goals, starting with the initial project proposal creation and submission. Each section of the projected timeline may be extended or contracted as necessary to maintain an accurate project road map.

References

[1] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *IEEE 2021 World AI IoT Congress,AIIoT2021*, 2021.

[2] A. Adekunle, M. Meehan, D. Rojas-Alvarez, J. Trauer, and E. McBryde, "Delaying the COVID-19 epidemic in australia: evaluating the effectiveness of international travel bans," *Australian and New Zealand Journal of Public Health*, vol. 44, pp. 257–259, July 2020.

[3] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, "Ontology creation model based on attention mechanism for a specific business domain," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[4] M. Bielecki, D. Patel, J. Hinkelbein, M. Komorowski, J. Kester, S. Ebrahim, A. J. Rodriguez-Morales, Z. A. Memish, and P. Schlagenhauf, "Air travel and COVID-19 prevention in the pandemic and peripandemic period: A narrative review," *Travel Medicine and Infectious Disease*, vol. 39, p. 101915, Jan. 2021.

[5] M. Heidari and J. H. Jones, "Using bert to extract topic-independent sentiment features for social media bot detection," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0542–0547, 2020.

[6] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani, "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak," *Science*, vol. 368, pp. 395–400, Mar. 2020.

[7] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in *IEEE 2021 World AI IoT Congress,AIIoT2021*, 2021.

[8] M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network to recommend airline tickets by using online reviews," in *IEEE 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2020*, 2020.

[9] M. Heidari, J. H. J. Jones, and O. Uzuner, "Offensive behaviour detection on social media platforms by using natural language processing models," 2021.

[10] S. H. Bae, H. Shin, H.-Y. Koo, S. W. Lee, J. M. Yang, and D. K. Yon, "Asymptomatic transmission of SARS-CoV-2 on evacuation flight," *Emerging Infectious Diseases*, vol. 26, pp. 2705–2708, Nov. 2020.

[11] E. M. Choi, D. K. Chu, P. K. Cheng, D. N. Tsang, M. Peiris, D. G. Bausch, L. L. Poon, and D. Watson-Jones, "In-flight transmission of SARS-CoV-2," *Emerging Infectious Diseases*, vol. 26, pp. 2713–2716, Nov. 2020.

[12] N. C. Khanh, P. Q. Thai, H.-L. Quach, N.-A. H. Thi, P. C. Dinh, T. N. Duong, L. T. Q. Mai, N. D. Nghia, T. A. Tu, L. N. Quang, T. D. Quang, T.-T. Nguyen, F. Vogt, and D. D. Anh, "Transmission of SARS-CoV 2 during long-haul flight," *Emerging Infectious Diseases*, vol. 26, pp. 2617–2624, Nov. 2020.

[13] T. W. Russell, J. T. Wu, S. Clifford, W. J. Edmunds, A. J. Kucharski, and M. Jit, "Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study," *The Lancet Public Health*, vol. 6, pp. e12–e20, Jan. 2021.

[14] S. Chen, S. Owusu, and L. Zhou, "Social network based recommendation systems: A short survey," in *2013 International Conference on Social Computing*, pp. 882–885, 2013.

[15] S. Lin, C. Liu, and Z.-K. Zhang, "Multi-tasking link prediction on coupled networks via the factor graph model," in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 5570–5574, 2017.

[16] M. Heidari, J. H. J. Jones, and O. Uzuner, "Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter," in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[17] Y. Chu, F. Huang, H. Wang, G. Li, and X. Song, "Short-term recommendation with recurrent neural networks," in *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 927–932, 2017.

[18] C. Yang, X. Chen, T. Song, B. Jiang, and Q. Liu, "A hybrid recommendation algorithm based on heuristic similarity and trust measure," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 1413–1418, 2018.

[19] S. Ji and J. Liu, "Interpersonal ties and the social link recommendation problem," in *2019 6th International Conference on Systems and Informatics (ICSAI)*, pp. 456–462, 2019.

[20] M. Heidari and S. Rafatirad, "Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment," in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[21] J. Wang, H. Song, and X. Zhou, "A collaborative filtering recommendation algorithm based on biclustering," in *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 803–807, 2015.

[22] S. Chen, S. Owusu, and L. Zhou, "Social network based recommendation systems: A short survey," in *2013 International Conference on Social Computing*, pp. 882–885, 2013.

[23] M. Heidari and S. Rafatirad, "Semantic convolutional neural network model for safe business investment by using bert," in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[24] A. Gatzioura, J. Vinagre, A. M. Jorge, and M. Sànchez-Marrè, "A hybrid recommender system for improving automatic playlist continuation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1819–1830, 2021.

[25] Z. Liao, Y. Song, Y. Huang, L.-w. He, and Q. He, "Task trail: An effective segmentation of user search behavior," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 3090–3102, 2014.

[26] M. Heidari, J. H. J. Jones, and O. Uzuner, "An empirical study of machine learning algorithms for social media bot detection," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[27] C.-Y. Chi, Y.-S. Wu, W.-r. Chu, D. C. Wu, J. Y.-j. Hsu, and R. T.-H. Tsai, "The power of words: Enhancing music mood estimation with textual input of lyrics," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6, 2009.

[28] A. Gatzioura, J. Vinagre, A. M. Jorge, and M. Sànchez-Marrè, "A hybrid recommender system for improving automatic playlist continuation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1819–1830, 2021.

[29] H. Yang, C. He, H. Zhu, and W. Song, "Prediction of slant path rain attenuation based on artificial neural network," in *2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 1, pp. 152–155 vol.1, 2000.

[30] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.