Exploring Gender Bias and Toxic Comments Using Artificial Intelligence: Trends and Implications

1st Marina Adriana Mercioni
Politehnica University Timisoara
Timisoara, Romania
marina.mercioni@upt.ro

2nd Stefan Holban
Politehnica University Timisoara
Timisoara, Romania
stefan.holban@cs.upt.ro

***Abstract*— Nowadays, most systems use artificial intelligence algorithms to automate tasks and reduce the time required for execution. Moreover, it must estimate the bias risks that can be introduced within the system. Based on these considerations, quantitative measures and prioritization strategies can be established for those inadequate situations, choosing an appropriate method to overcome gender bias. In this study, the impact of gender bias on an annual salary risk score due to gender bias was analyzed to identify and reduce it as much as possible in machine learning algorithms and on text data provided to a virtual assistant. The study finds that gender bias can influence our decisions by illustrating hypotheses on how algorithms affect prioritization decisions and strengthen stereotypes by favoring men against women. Recommendations to lower gender bias can include training programs for poor people that face substantial barriers to accessing education; training programs for people with a low level of education or no access; access to all kinds of jobs for women; assurance of diversity and inclusiveness; and algorithms that are fair and trained with the definite goal of reducing gender bias.**

Keywords*— *algorithm, artificial intelligence, bias, deep learning, detection, ethical, gender, machine learning, mitigation, risk, salary, workplace

I. Introduction

Artificial intelligence (AI) [1] technologies have been around for a while, but due to a lack of computer power, they became increasingly popular when graphics processing unit (GPU) [2] development increased computing capacity. In recent years, tremendous improvements in AI have been made thanks to the availability of large amounts of data and innovative algorithms.

AI is defined as a computer that replicates cognitive functions, such as learning and problem solving, that humans identify with the human mind. AI has the ability to improve efficiency, accuracy, precision, and performance across a wide range of fields as a result of various applications developed that can execute jobs that were previously done manually by people [3]. Advanced web search engines (Google) [4], recommendation systems that trail YouTube, Amazon, and Netflix [5-7], understanding human speech such as Siri [8] or Alexa [9], self-driving cars [10], automated decision-making, and competing at the highest level in strategic game systems succeeding to exceed human capability such as chess playing (IBM's Deep Blue succeeds to beat world chess champion Garry Kasparov) are just a few examples of AI-powered applications [11]. However, in addition to the numerous benefits of AI, it is critical that everyone understands the possible hazards and ethical problems [12] that AI raises, a worry that applies to both tech and non-tech users. For AI solutions to be deployed efficiently, they must adhere to the values and ethical data best practices proposed through the development of federal standards to ensure the building blocks for dependable, robust, and trustworthy AI systems [13]. In offering robustness solutions, AI can pose compliance and ethical concerns in areas like data protection, transparency, fairness, explainability, and inconsistency. Moving ahead, the focus of this research is on the concern that arises as a result of bias because it has the potential to do severe harm. Bias is a systematic error in an AI system; in other words, it is a "*disproportionate weight in favour of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. Biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief*" [14]. There are many types of bias discussed in the literature, including algorithmic bias, human bias, cognitive bias, statistical bias, prejudice bias (ageism, racism, sexism, and so on), and societal bias, but the current paper is just interested in gender bias.

Furthermore, the primary goal will be to comprehend gender bias in the workplace and its consequences for individuals. Only by understanding where gender bias reductions can be used and how they interact making continual progress in avoiding and resolving negative consequences.

This work attempts to understand the ways in which gender prejudice affects day-to-day activities in order to develop successful and cost-effective algorithms to counteract it.

The following is a breakdown of the paper's structure. The previous work in terms of gender prejudice is described in Section 2. The gender bias modeling section of this research is introduced in Section 3. The flow of eliminating gender bias is depicted in Section 4 of the results. Finally, Section 5 summarizes the findings and suggests future study directions.

II. Prior Work

According to a recent study, technology companies profit from the spectrum of white femininity by programming it into

979-8-3503-4929-0/24/$31.00 ©2024 IEEE

artificially intelligent virtual assistants (AI VAs), causing harm to people of color [15-16].

Sex or racial bias in AI-based cardiac magnetic resonance (CMR) segmentation has been investigated [17].

Given the high rate of maternal mortality in the United States each year due to pregnancy or its consequences and the large racial and ethnic disparities in pregnancy-related mortality, this is a major concern [18]. There have also been some inconsistencies in facial recognition, such as racial recognition due to AI prejudice, which has resulted in many black women being misidentified as men [19]. According to another study, popular applications that have already been built exhibit clear discrimination based on skin color. Incorrect, incomplete, or unvarying data sets on which the application is being trained are one cause of unfair and biased results [20]. It was also studied when and how harm might be introduced throughout the machine learning life cycle, presenting high-risk infections at different stages (Fig. 1) within an end-to-end solution, including both data generation and model building and implementation, to identify, anticipate, prevent, and mitigate undesirable consequences [21].

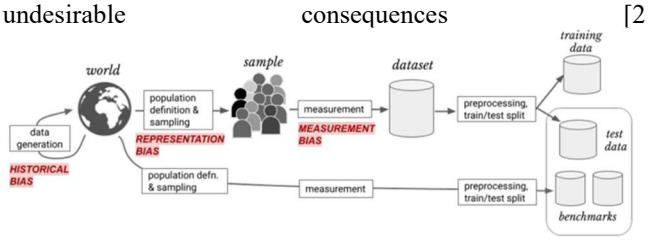

Fig. 1. Data generation [21]

Virtual assistants are a widely used product in today's society, appearing in cell phones, speakers, business apps, driving cars, and other places. The majority of them come with a female name and voice by default [22].

Virtual assistants are now in charge of over a billion actions every month, ranging from the most basic—such as checking the time—to the most crucial—such as contacting emergency services. To put it another way, it must pay close attention in order to decrease bias and improve performance. Gender-neutral virtual assistants, which convey that intelligent technologies do not need to have an assigned gender, are one method to avoid these issues. It is often overlooked that it collects gender binary data to create a viable virtual assistant.

III. GENDER BIAS

Gender bias is easy to introduce in many aspects of the AI development process, from data collection to solution implementation. To see that, the study was divided into two parts: one is an analysis of gender bias in a common workplace based on structured data, and the other is a text check for existing bias in it so that we are aware when we include it in our virtual assistants.

A. Annual salary earnings Dataset

To examine gender prejudice, the Disparate Impact Remover [23] algorithm was used, which was already integrated by IBM Research within AI Fairness 360 [24], to detect and mitigate gender bias during the preprocessing phase. The data utilized may be found on Kaggle [25], and it covers the individual's annual income as a result of numerous factors. It is influenced, as expected, by the individual's educational level, age, gender, occupation, and other dataset factors. The income column (dependent variable) in the dataset (Fig. 2) has two classes: *<=50K* and *>50K*. The remaining independent variables (features) are linked to demography and other personal characteristics.

workclass	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hour_per_week	country	ann_salary
State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50k
Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50k
Private	HS-grad	9	Divorced	Handlers-cleaners	Husband	Black	Male	0	0	0	United-States	<=50k

Fig. 2. Annual salary earnings data

B. Toxic comment data on Wikipedia

The dataset contains many Wikipedia comments [26–27] that have been rated for harmful behavior by humans (Fig. 3). The following are the most common types of toxicity: toxic, severe toxic, obscene, threat, insult, and identity hate.

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
6	0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
12	0005c987bdfc9d4b	Hey... what is it..\n@ I talk .\nWhat is it.....	1	0	0	0	0	0
16	0007e25b2121310b	Bye! \n\nDon't look, come or think of comming ...	1	0	0	0	0	0
42	001810bf8c45bf5f	You are gay or antisemmitian? \n\nArchangel WH...	1	0	1	0	1	1
43	00190820581d90ce	FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!	1	0	1	0	1	0
...	...	...	...	...	...	...	...	...
312450	ff84f0367ea58abb	am sorry for being a dickhead! I cannae help i...	1	0	1	0	1	0
312479	ff91c3d8a3e34398	NIGEL IS A CRAZY IDIOT!!!	1	0	0	0	1	0
312649	ffdf6854b41d9102	==Fourth Baldrick possibly being cleverer than...	1	0	0	0	0	0
312690	ffebe90c8d5acaba	" \n\n == IRAN == \n That's right, Iran. It wa...	1	0	1	0	0	0
312726	fffac2a094c8e0e2	MEL GIBSON IS A NAZI BITCH WHO MAKES SHITTY MO...	1	0	1	0	1	0

21384 rows × 8 columns

Fig. 3. Toxic comment data on Wikipedia

The wiki-news-300d-1M.vec file was used, which contains "*1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus, and statmt.org news dataset (16B tokens)*" in this experiment [28]. The procedure for unbiasing gender data to provide a correct outcome is shown in Fig. 4.

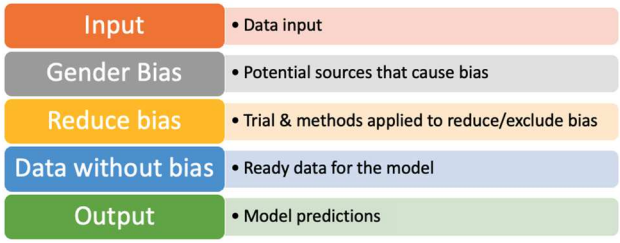

Fig. 4. The unbiasing data

A type of long-short-term memory (LSTM) [29] was used as the architecture. In addition, the FastText open-source library was used to train text representations and classifiers.

IV. Results

A. Annual salary earnings Dataset

An unbalanced dataset was found (Fig. 5) (binary class, earning less than 50k and greater than 50k), and in Fig. 6, the gender distribution of females and males can be shown, with the males' class being twice as big as the females' class as follows:

Fig. 5. Earnings data

Fig. 6. Gender data unbalance

The linear regression technique with standardization (Standard Scaler) was applied on 32561 data. At the split, 20% of the data was kept for testing phase.

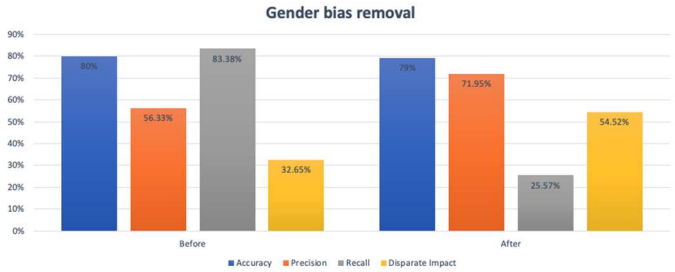

Fig. 7. Gender data debiased

As shown in Fig. 7, the disparate impact increased from 32.65% to 54.52% after using the Disparate Impact Remover method (where bigger corresponds to better, the ideal value should be 100 percent).

B. Toxic comment data on Wikipedia

This experiment tackled a natural language processing (NLP) problem that is frequently encountered in virtual assistant solutions with the goal of better understanding the risk of online abuse and harassment, which means many people stop expressing themselves and give up on seeking different perspectives, resulting in a lack of effectiveness in facilitating conversations, leading many communities to limit, or completely shut down user comments. An architecture was defined by an embedding layer, a LSTM layer (25 units), a GlobalMaxPool1D, two dropout layers (dropout rate of 0.01), two dense layers, and an Adam optimizer (learning rate of 3e-5).

	(2,)	(0,)	(1,)	(4,)	(5,)	(3,)
data	12140	21384	1962	11304	2117	689
train	8741	15396	1413	8139	1511	496
test	2428	4277	392	2261	438	138
validation	971	1711	157	904	168	55

Fig. 8. Data split shape

The split for each label can be seen in the figure above: training data 80%, test data 20% (Fig. 8), using the following encoding: 0: toxic; 1-severe toxic; 2-obscene; 3-threat; 4-insult; 5-identity hate. In Table I are presented different performance metrics (precision, recall, and F1-Score) computed as in Fig. 9 [30].

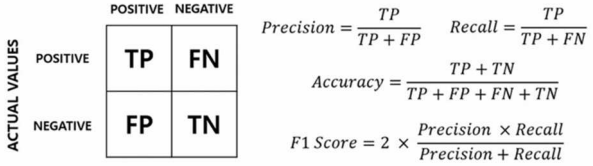

Fig. 9. Performance metrics

These measures are critical for assessing how well classifiers function, particularly in cases where there is an imbalance in the classes. A high precision indicates a cautious labelling of a positive sample by the classifier. A high recall indicates that the majority of the positive samples were properly labelled by the classifier. The balance between these two measures is offered by the F1-score.

The following table shows the results for each case:

I. CLASSIFICATION REPORT

Architecture	*Label*	*Precision*	*Recall*	*F1-Score*
LSTM	toxic	0.33	0.35	0.34
	severe toxic	0	0	0
	obscene	0.34	0.34	0.34
	threat	0	0	0
	insult	0.35	0.30	0.32
	identity hate	0.26	0.15	0.19

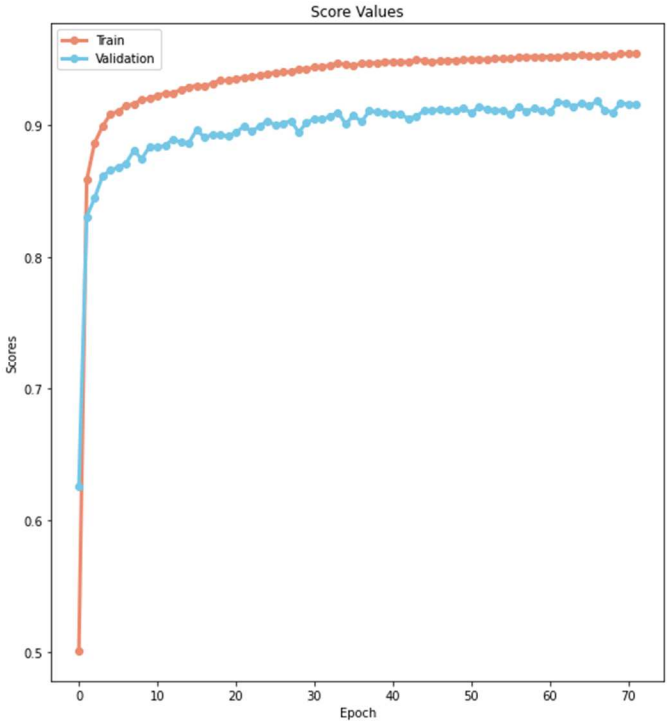

Fig. 10. Validation Area Under the Curve

As shown in Figure 10, the validation Area Under the Curve (AUC) did not improve from 91.80% (run the model with batch size = 64, max epochs = 100, patience = 10, eval metric = auc_score) monitoring 'val_auc' and a loss function of 0.0877. Table I shows that the worst results were obtained by threat and severe toxic, with insufficient sample data for training available for both labels as a factor.

V. CONCLUSIONS AND FUTURE WORK

The current study sought to identify prejudice and devise strategies to reduce it as much as feasible, as you can see. Men often benefit more than women, as was seen in the first case when the yearly income was assumed. It is crucial that we comprehend the possibility of bias in systems that employ text data, as evidenced by the second case, where bias was found to be introduced even on an NLP work. Thus, in order to solve the challenges facing the world, it is imperative that concerns be

raised and that people understand the biases that such institutions uphold as well as how to change them.

It trained a million 300-dimensional word vectors using transfer learning on Wikipedia and the statmt.org news dataset, however the classification report's low precision and recall were caused by insufficient training data for those labels. The fact that they had not come across instances when these edge corners appeared throughout the context as opposed to just embracing a few related words because of word embeddings might also have contributed to that circumstance. An embedding layer that learns task-specific embeddings is also a part of the design. Using previously learned word embeddings that have been trained on a bigger corpus is another avenue for research that aims to improve accuracy and recall. Although biases can be addressed by AI, they can also be inherited or reinforced. Establishing a workplace that is inclusive, varied, safe, and balanced is essential to eradicating gender prejudice.

In order to remove any biases that could be present in a system and produce a more robust solution, it plans to carry out in-depth evaluations of alternative designs at different depths as a future research topic.

REFERENCES

1. Russell, Stuart J.; Norvig, Peter (2009), "*Artificial Intelligence: A Modern Approach*" (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall. ISBN 978-0-13-604259-4.
2. *"NVIDIA Launches the World's First Graphics Processing Unit: GeForce 256*", Nvidia. 31 August 1999. Archived from the original on 12 April 2016. Retrieved 28 March 2016.
3. Jordan, M. I.; Mitchell, T. M. (16 July 2015), "*Machine learning: Trends, perspectives, and prospects*", Science. 349 (6245): 255–260.
4. Stefan Buettcher, Charles L. A. Clarke, Gordon V. Cormack, "*Information Retrieval: Implementing and Evaluating Search Engines*", MIT Press, Cambridge, MA, 2010.
5. Paul Covington, Jay Adams, Emre Sargin, "*Deep Neural Networks for YouTube Recommendations*", Google, RecSys '16 September 15-19, 2016, Boston, MA, USA, DOI: http://dx.doi.org/10.1145/2959100.2959190.
6. Greg Linden, Brent Smith, and Jeremy York, "*Recommendations Item-to-Item Collaborative Filtering*", 1089-7801/03/$17.00©2003 IEEE IEEE INTERNET COMPUTING.
7. Carlos A. Gomez-Uribe and Neil Hunt. 2016, "*The Netflix Recommender System: Algorithms, Business Value, and Innovation*," ACM Trans. Manage. Inf. Syst. 6, 4, Article 13 (January 2016), 19 pages. DOI:https://doi.org/10.1145/2843948.
8. Sheetal Reehal, "*Siri –The Intelligent Personal Assistant*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 5, Issue 6, June 2016.
9. Xinyu Lei et al., "*The Insecurity of Home Digital Voice Assistants – Amazon Alexa as a Case Study*", 2019.
10. Sorin Grigorescu et al., "*A Survey of Deep Learning Techniques for Autonomous Driving*", 2020.
11. Newborn, Monty (1997), "*Kasparov versus Deep Blue: Computer Chess Comes of Age*", (1st ed.). Springer. ISBN 978-0-387-94820-1.
12. Stahl B.C. (2021), "*Ethical Issues of AI*", In: Artificial Intelligence for a Better Future. SpringerBriefs in Research and Innovation Governance. Springer, Cham. https://doi.org/10.1007/978-3-030-69978-9_4.
13. "*Legislation Related to Artificial Intelligence*", https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx, last accessed 2021/10/19.
14. Steinbock, Bonnie (1978). "*Speciesism and the Idea of Equality*", Philosophy. 53 (204): 247–256. doi:10.1017/S0031819100016582.
15. Taylor C. Moran, "*Racial technological bias and the white, feminine voice of AI VAs*", Communication and Critical/Cultural Studies, Volume 18, 2021 - Issue 1, Pages 19-36.
16. Ravi B. Parikh et al., "*Addressing Bias in Artificial Intelligence in Health Care*", JAMA. 2019;322(24):2377-2378. doi:10.1001/jama.2019.18058, 2019.
17. Esther Puyol-Antón et al., "*Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-based Segmentation*", doi: https://doi.org/10.1101/2021.07.19.21260749, 2021.
18. Petersen EE, Davis NL, Goodman D, et al., "*Vital signs: pregnancy-related deaths*", United States, 2011–2015, and strategies for prevention, 13 states, 2013–2017. MMWR Morb Mortal Wkly Rep2019;68:423–9. 10.15585/mmwr.mm6818e1.
19. Daniel C. Wood, "*Facial Recognition, Racial Recognition, and the Clear and Present Issues With AI Bias, Robotics, Artificial Intelligence & Law*", May–June 2021, Vol. 4, No. 3, pp. 219–221.
20. Buolamwini, J., Gebru, T.: "*Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*", Proceedings of Machine Learning Research 81:1 — 15, 2018, Conference on Fairness, Accountability, and Transparency.
21. Harini Suresh and John Guttag, "*A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*", 2021.
22. Christiane B. Oliveira et. al., "*An analysis of the reproduction of gender bias in the speech of Alexa virtual assistant*", Proceedings XIII Congress of Latin American Women in Computing 2021, October 25–29, 2021, San José, Costa Rica.
23. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "*Certifying and removing disparate impact*", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.
24. *AI Fairness 360 IBM Research*, https://aif360.mybluemix.net/.
25. *Adult income dataset*, https://www.kaggle.com/wenruliu/adult-income-dataset.
26. Yanran Li et al., "*DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*", 2017.
27. *Toxic comment classification challenge*, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.
28. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, "*Advances in Pre-Training Distributed Word Representations*", https://fasttext.cc/docs/en/english-vectors.html, 2017.
29. Sepp Hochreiter et. al., "*Long Short-Term Memory*", Neural Computation (1997) 9 (8): 1735–1780.
30. Seol, D.H.; Choi, J.E.; Kim, C.Y.; Hong, S.J. Alleviating Class-Imbalance Data of Semiconductor Equipment Anomaly Detection Study. *Electronics* 2023, *12*, 585. https://doi.org/10.3390/electronics12030585.