IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 4, Ser. I (Jul. –Aug. 2022), PP 81-86
www.iosrjournals.org

Addressing Bias in Natural Language Processing (NLP) Models: Techniques and Implications

Godavari Basavaraj
Research scholar Department of Computer Science OPJS University Churu (Raj)

Dr.Vijay Pal Singh
Associate Professor Department Of Computer Science OPJS university Churu. (Raj)

ABSTRACT
The reduction of bias in training data and the improvement of algorithmic fairness may be accomplished by the use of procedures such as data preparation, data augmentation, and the oversampling of groups that are underrepresented. Another essential component is the creation of algorithms that take into consideration the principle of fairness. There have been a lot of different approaches that researchers have developed in order to guarantee that the decisions that machine learning algorithms make are objective and fair. It is necessary to include fairness constraints into the training approach in order to ensure that the model achieves the highest possible level of accuracy and justice. Researchers and practitioners have proposed a wide variety of mitigation strategies, some of which include improving the quality of the data and developing algorithms that are expressly fair. The use of artificial intelligence (AI) has the potential to revolutionize a broad variety of industries and dramatically improve the quality of life for many individuals. Despite this, bias continues to be a substantial barrier to the development and use of artificial intelligence systems. The term "bias" refers to the phenomenon that occurs when there are systematic errors in the decision-making processes that lead to unfair outcomes. It is possible for bias in artificial intelligence to arise from a variety of sources, including human interpretation, algorithm design, and data collection.

Keywords: *Bias , Natural Language Processing , Techniques ,Implications*

I. INTRODUCTION

Multiple strategies are required in order to counteract prejudice based on algorithmic processes. The ability to recognize and eliminate biases in training data is a key component that must be addressed. To do this, it is necessary to conduct a comprehensive analysis of the data that is used in the training of machine learning models. This analysis is necessary in order to detect any potential biases and to take the necessary steps to reduce or eradicate them. The reduction of bias in training data and the improvement of algorithmic fairness may be accomplished by the use of procedures such as data preparation, data augmentation, and the oversampling of groups that are underrepresented. Another essential component is the creation of algorithms that take into consideration the principle of fairness.

There have been a lot of different approaches that researchers have developed in order to guarantee that the decisions that machine learning algorithms make are objective and fair. It is necessary to include fairness constraints into the training approach in order to ensure that the model achieves the highest possible level of accuracy and justice. Researchers and practitioners have proposed a wide variety of mitigation strategies, some of which include improving the quality of the data and developing algorithms that are expressly fair. This study provides a comprehensive analysis of the sources and consequences of prejudice in artificial intelligence by analyzing data biases, algorithmic biases, and user biases, as well as the ethical implications of each of these types of biases.

This article takes a look at the current state of the art in mitigation strategy research, discussing its inadequacies, problems, and the need of interdisciplinary collaboration. The academic community, politicians, and researchers are all in agreement that fairness and bias are very important in artificial intelligence. The purpose of this survey study is to investigate the complex and multifaceted issues that are associated with prejudice and fairness in artificial intelligence. These issues include proposed solutions, the sources of bias, and its impacts. In order to help ongoing attempts to develop artificial intelligence systems that are more ethical and responsible, the project's overarching objective is to provide knowledge on the causes, impacts, and mitigation approaches of fairness and prejudice in artificial intelligence.

The use of artificial intelligence (AI) has the potential to revolutionize a broad variety of industries and dramatically improve the quality of life for many individuals. Despite this, bias continues to be a substantial barrier to the development and use of artificial intelligence systems. The term "bias" refers to the phenomenon that occurs when there are systematic errors in the decision-making processes that lead to unfair outcomes. It is possible for bias in artificial intelligence to arise from a variety of sources, including human interpretation, algorithm design, and data collection. Machine learning models are a kind of artificial intelligence (AI) system that has the ability to detect and replicate bias patterns that are seen in the training data. This may result in outputs that are unfair or discriminatory. The different types of bias that may be found in artificial intelligence, such as algorithmic, user, and data bias, will be discussed in this section, along with examples of how these biases manifest themselves in the real world.

AI-Related Bias, Including Algorithmic, User, and Data Bias

The process of machine learning may be broken down into many stages, including data collection, algorithm design, and user interactions, all of which have the potential to yield bias in artificial intelligence. This section examines the several factors that contribute to bias in artificial intelligence (AI), as well as specific examples of each kind of bias, including algorithmic, user, and data bias.

Skewed outputs are the consequence of data bias, which occurs when machine learning models are trained on data that is either insufficient or inadequately representative. This may potentially place when information is obtained from sources that are biased, when there is a lack of information or information that is insufficient, or when there are errors present. Nevertheless, algorithmic bias is the term used to describe the situation in which the algorithms used in machine learning models have inherent biases that manifest themselves in the results they produce. This might happen if the decision-making criteria that algorithms utilize are biased, or if the algorithms use biased premises as their foundation. User bias is the term used to describe the phenomenon that occurs when users of artificial intelligence systems mistakenly or actively insert their own prejudices or assumptions into the system. The occurrence of this phenomenon might take place when users interact with the system in a way that is influenced by their own personal experiences or when they supply training data that is biased.

There have been a number of different approaches proposed in order to lessen the impact of various sources of bias. Some of these approaches include dataset augmentation, bias-aware algorithms, and user feedback techniques. The process of adding more diverse data to training datasets is known as dataset augmentation. This is done with the goal of improving representativeness and reducing bias. Building bias-aware algorithms refers to the process of developing algorithms that take into consideration different types of prejudice and use strategies to lessen the impact that these biases have on the results produced by the system. The strategies of user feedback involve soliciting the views of individuals in order to identify and rectify biases that are introduced by the system.

Equitable Distributions with Strategic Partners

There is a lot of discussion around the problem of splitting up public resources or things among a large number of people who desire them. It is the choices or aspirations of these strategic players that take precedence over the resources that are accessible to them. The agents may be able to modify their preferences in order to obtain the intended effects, which is especially true with AI-based agents. In the field of game theory, there are a number of desirable characteristics that the allocations need to fulfill. For the purposes of the disciplines of economics and game theory, mechanism design is the process of developing economic mechanisms or incentives for the purpose of achieving objectives in tactical contexts in which individuals behave rationally. The next discussion will elaborate on the fact that it is very pertinent in a number of complex situations.

- **Auctions** are used to a significant degree by government bodies as well as in online markets. They provide an atmosphere that is an appropriate setting for the development of processes that include money. During auctions, many agents compete with one another for commodities as diverse as real estate, frequency spectrum, airport slots, and other such resources. It is the responsibility of the social planner to ascertain the appropriate distribution and payment schedule for certain objectives. Both maximizing profits and promoting social welfare might be considered to be the aim. There has been a considerable rise in the number of web-based auctions, which has resulted in the development of mechanisms that can accommodate vast and complex environments.
- **Internet advertisement** now constitutes a multimillion-dollar business due to its rapid growth. When a user enters a phrase into one of the several search engines (such as Google or Bing), a large number of links that are pertinent to the query are shown. In addition, a number of sponsored links that correspond to the advertisements featured by certain sponsors are shown as well. When a person clicks on such a link, they are sent to the website that is pertinent to their inquiry. The advertiser is entitled to a set amount in exchange for sending the visitor to the page in question. These search engines often make use of processes that were developed with the help of multi-armed bases (MABs). The advertisements that are shown are determined by these systems based on the amount of clicks that are received, and an auction is held to determine the price that will be charged for the slots.

Expert sourcing refers to a method that is widely used by requesters in order to collect information or opinions from a big community. It makes it possible for a huge number of individuals with a diverse variety of experiences to take part in any activity that is outsourced. In addition to the jobs, there are also business application testing, online product ratings, and data collecting from sources that are based in the real world. A significant number of these tasks do not have any responses, which means that we are unable to verify the correctness of the player contributions or ascertain the reality of the situation. Consequently, this indicates that strategic individuals may be rewarded to influence the system by providing random data, rather than fulfilling the tasks that are currently being performed. For the purpose of discouraging such unethical strategies, academics propose a variety of incentive strategies that are based on specific reward systems.

OBJECTIVES OF THE STUDY

1. To study on AI-Related Bias, Including Algorithmic, User, and Data Bias
2. To study on Bias's Impact on Decision-Making and Social Repercussions

Fair Decisions for Groups in Part B

Recent years have seen a rise in the use of machine learning models as prediction models. This is because these models may help improve the decision-making process. making judgments in significant areas like as employment, the criminal justice system, and even medicine. The list of different use cases may be seen below.

- **Recruitment**. In the process of hiring new employees, organizations often make use of models that are formed using machine learning. Machine learning algorithms provide assistance with each and every stage of the process, including the assessment of applicants and the prediction of the prerequisites for successful recruitment. The prior data is also used in order to establish the manner in which job advertisements are strategically shown. These companies often have the goal of increasing their profits in the near term. On the basis of a number of factors, including confidential information, they look for the most qualified candidates who have shown the highest level of performance. These models are often trained on biased data, which manifests itself in the projections and eventually has an effect on the judgments that are made.
- **Criminal Justice.** The areas of forensics, law, and criminology have all begun to use prediction models that are powered by artificial intelligence. Computer programs are used to make decisions on risk assessments, parole and punishment, as well as bail. Due to the fact that incorrect decisions may have an immediate effect on a huge number of other people, these models are of utmost significance in the lives of the incarcerated individuals being considered.
- **Healthcare**. The healthcare sector is characterized by a persistent level of rivalry for resources that are both costly and limited. There is a growing trend in the healthcare business to use machine learning for the purpose of enhancing diagnostic tools, predicting the adverse effects of medications, scheduling visits, and even ranking patients according to their preferred treatment options. Continuous training is performed on these models in order to ensure that they provide the greatest results at the lowest feasible cost. Nevertheless, we cannot help but worry if the projections are biased in favor of certain demographic groupings, so leading to an even greater division in society.
- **Recommender Systems.** There has been a rise in the significance of recommender systems in our lives as a result of the growth of services such as Netflix, Amazon, and Youtube. These e-commerce systems provide product recommendations to clients based on the interests that they have expressed. When it comes to advertising on the internet, they provide suggestions for relevant information to individuals based on the interests they have expressed. By analyzing the profiles of the users, these systems make an effort to recommend items that are relevant to the individual customers. It is common for the efficiency of a recommender system to be the primary focus. On the other hand, we exclude the bias that it brings about via its unfair implications. In addition, the exploitation of sensitive data by these algorithms may give rise to problems about information privacy.

Comprehensive Global Learning Algorithms:

Bias is defined as the systematic error that occurs throughout the decision-making process and leads to the production of unfair consequences. It is possible for bias in artificial intelligence to arise from a variety of sources, including human interpretation, algorithm design, and data collection. A machine learning model is an example of an artificial intelligence system. This model has the potential to recognize and recreate bias patterns in the training data, which might result in outputs that are unfair or prejudiced by the system. The identification and elimination of bias in artificial intelligence is of the utmost importance in order to ensure that these systems are fair and equitable for each and every user.

Types of bias in machine learning:

1. Data bias:

The term "data bias" refers to the situation in which machine learning algorithms are developed using training data that is either not representative of the population or biased. This leads to the model producing predictions that are not correct. This bias might be the result of a variety of factors, such as errors in the sampling process, the underrepresentation of certain demographic groups, or past injustices that are reflected in the statistics.

2. Algorithmic bias:

A prejudice that is established in algorithms and produces skewed outcomes is referred to as "algorithmic bias," and the word "algorithmic bias" defines. During the design phase of an algorithm, judgments may be made about the selection of features, the architecture of the model, and the optimization techniques. These biases may be the outcome of these decisions. There is a possibility that some groups may be treated unjustly as a consequence of algorithmic prejudice, which has the potential to perpetuate or intensify societal prejudices that already exist.

3. Evaluation bias:

A biassed evaluation is the result of using performance metrics that are either biased or inadequate when attempting to assess the effectiveness of machine learning systems. It is possible, for instance, that discrepancies in prediction accuracy across different demographic groups will escape unreported if accuracy is the sole measure that is employed. There is a possibility that assessment bias may disguise algorithmic prejudice, which will also make it more difficult to address fairness issues in machine learning models.

Bias's Impact on Decision-Making and Social Repercussions

Due to the existence of bias in machine learning algorithms, there is a possibility that social dynamics and decision-making processes may be greatly impacted. Additionally, it has the capacity to sustain existing disparities while also having the potential to exacerbate existing gaps in access to opportunities, resources, and rights. Additionally, biased algorithmic judgments have the potential to undermine the legitimacy of algorithmic control across a variety of industries and to reduce the level of trust that users have in automated systems.

In addition, the presence of bias in machine learning algorithms has the potential to exacerbate systemic disparities, prejudice, and stereotypes, so contributing to the dissatisfaction of society and the intensification of social divisions. As a result, eliminating bias in machine learning algorithms is essential for preserving ethical norms, creating social cohesion, and ensuring justice and equality in a society that is becoming more computerized.

Examples of AI Bias in the Real World

There have been several reports of instances of discrimination in artificial intelligence systems in a variety of fields, including the criminal justice system and the healthcare industry. A well-known example is the COMPAS system, which is used in the criminal justice system of the United States of America. This method anticipates the likelihood of a defendant committing another crime. The findings of an investigation conducted by ProPublica indicate that African-American defendants were disproportionately classified as high-risk even when they did not have any prior convictions, which is indicative of the existence of institutional discrimination against them. Through another research, it was revealed that a related system that was used in the state of Wisconsin had similar biases.

- **In healthcare,** Someone made the discovery that an artificial intelligence system that was used to predict the mortality rates of patients had a bias against patients who were African-American. African-American patients had a larger possibility of obtaining higher-risk ratings from the system, according to a study that was conducted by Obermeyer and colleagues. This was the case even in situations when other characteristics, such as age and health status, were the same. As a result of this discrimination, African-American patients may not be able to access medical treatment or could receive services that are not up to their standards.

However, the employment of facial recognition technology by law enforcement agencies is just another example of bias in artificial intelligence systems. According to the findings of a study conducted by the National Institute of Standards and Technology (NIST), facial recognition software generated a higher proportion of false positives when it came to those with darker skin tones. This was due to the fact that the program was much less accurate for these individuals. If this bias is allowed to continue, it might lead to serious consequences such as wrongful arrests or convictions.

In the end, if generative artificial intelligence (GenAI) systems grow more widespread, there is a larger possibility that they may suffer from harmful biases. Stable Diffusion, OpenAI's DALL-E, and Midjourney were among the text-to-image models that were shown to have racial and stereotype biases in their outputs. This was discovered via an investigation. One of the more shocking examples of prejudice in GenAI was this.

When asked to develop images of CEOs, these models demonstrated gender bias since they favored to generate photos of males; this phenomenon was seen. This bias is a reflection of the current reality, which is that women are underrepresented in CEO posts in the real world. When asked to develop images of criminals or terrorists, the models created a far higher number of photographs of people of color than any other kind of person.

- **COMPAS Algorithm Bias Detection and Mitigation:** A criticism was leveled against the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm that is used in the United States criminal justice system. The program was accused of possessing racial discrimination. Researchers were able to discover bias by analyzing the predictions made by the algorithm, which demonstrated that particular ethnic groups had higher rates of making mistakes than others. For the purpose of reducing bias and providing more fair findings for predicting recidivism, researchers advised the use of recalibrated risk ratings, which were less impacted by demographic variables.
- **Google Photos Image Recognition Bias:** The controversy that ensued was triggered by the fact that black persons were incorrectly labeled as gorillas in photographs that were shot using the image recognition feature of Google Photos. This occurrence brought to light the fact that there is racial discrimination included in the algorithms that are used for machine learning. In response, Google modified its photo recognition algorithms and removed the gorilla label in an attempt to avoid biases of this sort from occurring in the future. They also included bias detection tools in order to identify and rectify any biases that may exist in the algorithms that are used to categorize pictures.
- **Success Story:** A toolkit called the Fairness Toolkit was developed by LinkedIn in order to eliminate biases that may be present in the artificial intelligence algorithms that are used to promote job openings. LinkedIn revealed biases in job recommendations based on area, gender, and race via the study of historical data and model performance monitoring. These biases were uncovered by LinkedIn for the first time. The Fairness Toolkit enables LinkedIn to minimize biases and improve the fairness of its job recommendation algorithms, which will result in more equitable results for job seekers. This will be beneficial to job seekers since they will benefit from more equitable results.
- **Challenges:** Facial Recognition Systems: These systems have come under fire for being biased, especially when it comes to incorrectly recognizing members of ethnic minorities who are underrepresented in the population. These prejudices may provide unfair results in fields like surveillance and law enforcement. In order to successfully eliminate biases, careful examination and monitoring are necessary. Other challenges in maintaining fairness include the dearth of varied training data and inherent biases in face recognition algorithms.

Implications of Bias in AI on People and Society, Such As Perpetuation And Discrimination of Existing Inequalities

Bias in artificial intelligence has the potential to have major negative repercussions on both individuals and society as a whole., biased artificial intelligence systems have the capacity to both perpetuate and exaggerate already-existing inequities, giving rise to a significant reason for concern about discrimination. For instance, the use of biased algorithms within the criminal justice system may lead to the unfair treatment of some groups, particularly people of color, who are more likely to be erroneously convicted or given harsher sentences. This may result in the criminal justice system's tendency to treat certain groups unfairly. It is possible that prejudice in artificial intelligence might have a detrimental impact on a person's ability to get essential services such as banking and healthcare. It has been suggested that biased algorithms may lead to the underrepresentation of certain groups in credit scoring systems. These groups include people of color and those who come from lower socioeconomic backgrounds. As a consequence, it may be more challenging for these persons to get loans or mortgages.

Moreover, bias in artificial intelligence may contribute to the perpetuation of prejudice and gender stereotypes. For instance, when facial recognition algorithms are trained on data that is mostly made of male faces and fail miserably to recognize female faces, this contributes to the perpetuation of gender bias in security systems., when generative artificial intelligence (GenAI) models are asked to produce CEO visuals, they often depict CEOs as men, which only serves to propagate preconceived notions. There is a possibility that bias in artificial intelligence might lead to the development of new forms of discrimination, such as those based on skin color, ethnicity, or even physical appearance, in addition to maintaining that which currently exists.

II. CONCLUSION

The development of approaches that are intrinsically stochastic was something that we looked into. Learning the underlying stochastic parameters is accomplished via the use of MAB algorithms. The automation of the process of developing a payment rule for a mechanism that is based on Thompson sampling has been our primary objective over the last several years. At this point in time, to the best of our knowledge, this is the very first occasion that a data-driven strategy has been used to simulate an MAB mechanism. A neural network model was used in the process of designing the payment rule. Using our network as the payment rule, we are able to ensure the WP-DSIC and EPIR mechanism while also assuring that the utilities of the agents are not significantly different from one another. This technique, in contrast to the other strategies, ensures that payments are kept to a minimum and do not exceed the welfare value, which would be contrary to the auctioneer's rationale. In addition, this strategy assures that payments are maintained minimal. The aforementioned assertions have been validated via the process of comparing the Cost Index of the various approaches that were used in our experiments. In a nutshell, we show that neural networks are capable of learning effective redistribution mechanisms, provided that they are given the suitable initialization and a sufficiently defined ordering across valuation profiles.

REFERENCES

[1]. Smith, B. (2018). The ethical challenges of artificial intelligence. Cambridge Quarterly of Healthcare Ethics, 27(4), 599-609. doi:10.1017/S0963180118000038

[2]. Buolamwini, J., & Gebru T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1-15

[3]. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77-91.

[4]. Smith-Lewis A., & Srinivasan R., (2020). Responsible Artificial Intelligence: Building Trust Through Transparency & Explainability [Whitesection]. IBM Watson Health.

[5]. Russell, S., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3rd ed.). Pearson Education Limited.

[6]. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77-91.

[7]. Crawford K., Calo R., Barocas S., Beasley B., Friedler S., Kroll J. (2019).

[8]. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77-91.

[9]. Angwin, J., Larson, J., Mattu S., & Kirchner L. (2016). Machine Bias: There's Software Used Across The Country To Predict Future Criminals And It's Biased Against Blacks. ProPublica.

[10]. Angwin, J., Larson, J., Mattu S., & Kirchner L. (2016). Machine Bias: There's Software Used Across The Country To Predict Future Criminals And It's Biased Against Blacks. ProPublica.

[11]. California Legislative Information (2020). AB-2261 Privacy: facial recognition technology: pilot program.

[12]. Benneh Mensah, G. (2023). AI in the Legal System, Transparency, Interpretability and the Right to a Fair Trial: The Challenges and Implications for the Ghanaian Civil and Criminal Justice Systems. Preprint. 10.13140/RG.2.2.14854.96324.

[13]. Smith-Spark L., & Duffy H., (2019). Federal Court Rules Against Insurer That Relied Solely On AI Algorithm. CNN Business.