Understanding Bias in Machine Learning promulgated by Pre-trained Models

Group 5
Virginia Tech
CS 5024: Ethics and Professionalism in CS

Abstract

There have been various issues in bias and fairness in the field of AI. In this paper, we specifically focus on the current state of research in bias and fairness in one of the subsets of AI, Natural Language Processing(NLP). NLP models trained on crowd-sourced data reflect human biases. We start by providing a comprehensive survey on bias and fairness in natural language processing and then present the various forms that bias can take in applications. We illustrate the cases where bias becomes palpable in pre-trained language models. We discuss the development of fair algorithms and how they are being used in modern NLP applications as well as existing solutions for detecting or removing bias in most commonly used applications. After completing a comprehensive survey on bias and fairness as well as its solutions toward bias detection and removal in NLP, we further explore the challenges that we still face and future research directions. The aim of this comprehensive survey paper is to discuss the ethical problem of discrimination in NLP which is apparent through results and the state-of-the art solutions to address this bias problem.

Introduction

A bias is a certain pre-conceived notion that is held by a decision making entity. It, in itself is not a problem, but when a bias is wrongly held against a certain target group it can be pernicious. This held bias shapes itself as a stereotype, and is one that should not be promulgated or accepted through repeated reinforcement. Reinforcement is incidentally the way in which most modern machine learning models are trained. Since Machine Learning has become an integral part of most customer-facing software it is vital that it is kept in check and duly accounted for.

The use of traditional machine learning models provides us with controllable points of failure, such as the data or the model itself. On the other hand, with the advent of pre-trained language models we deal with a large unknown and work with a PTLM that can hold intractable biases.

Bias can be formalized in a machine learning context both theoretically and empirically. [Gajane et al.2018] define bias through notions of distributive justice. Fairness, a concept closely related to bias, is evaluated based on two central ideas,

i) Parity or Preference?

Should fairness in predictions mean achieving parity or is it expected to satisfy preferences?

ii) Treatment or impact?

Should fariness be sustained through treatment or impact?

These ideas instigate questions on whether a "fair" model is expected to treat all groups equally and ignore the disparity or take the disparity into account, provide preference to the marginalized and make predictions that create the most impact?
Parity-based ideas are fairness through unawareness, counterfactual measures,group and individual fairness and equal opportunity while preference-based ideas explore preferred treatment and impact.

Fairness through unawareness is a concept explored in the paper where one can claim that the model is being fair if it does not use protected-attributes in the prediction process. But through notions of distributive justice one can argue that this is just the model being "blind" even when prior knowledge is present in the training data.

This leads to the concept of counterfactual awareness, where the output of the predictor should remain the same even if its protect attribute is flipped. Group fairness, individual fairness and equality of opportunity all operate on the idea that every group should have an equal probability in the outcome, producing similar outputs with a similar true positive rate.
Preference-based fairness notions look to provide group benefit through its predictions. The paper then proposes two prospective notions of fairness through equality of resources and capability of opportunity.

Once fairness is formalized, it can be measured in machine learning systems by removing bias. The first step in dealing with such biases is to identify and quantify its presence and look for its source. With this information we can look for ways of debiasing or atleast accounting for the bias present.

Types of Bias

In this section, we cover many different types of bias that are prevalent in machine learning algorithms. [Suresh et al.2019] elaborates on seven types of bias that could potentially lead to danger for sensitive groups. The definition of each of the seven bias is as follows:

- Aggregation bias: It occurs when it is assumed that the trends found in aggregated data or groups also apply to individual data.
- Learning bias: It occurs when the learner uses a set of assumptions that degrades existing disparities to predict outputs of unencountered data inputs.
- Evaluation bias: It occurs when the dataset used to evaluate the trained model is not representative of the population, leading to a poor quality of model evaluation.
- Deployment bias: It occurs when the model trained is used or deployed in such a way that the developers didn't intend. It will highly likely to have poor model performance as the model was applied for unintended purposes.
- Measurement bias: It arises when there is a problem with data accuracy or how it was measured. For example, if one's photo was used to measure the quality of the work environment and employees knew they were being measured for how happy they are at the workplace, then the model may be biased.
- Representation bias: It mainly occurs when some groups or classes of the population are underrepresented or overrepresented in the dataset, leading to generalization problems where the model can't predict outcome values for previously unseen data.
- Historical bias: It occurs when some wrong prejudices and stereotypes lead to a bias in data itself.

[Mehrabi et al.2021] also introduces some additional types of bias as follows:

- Omitted variable bias: It occurs when one or multiple variables or attributes are missing or left out of the model.
- Sampling bias: This is similar to representation bias except the concept of sampling exists. It occurs mainly due to non-random sampling of subgroups.
- Linking bias: It occurs when attributes collected from users are actually different from those of real users.

These bias types may not be extensive, but they help to lead to a better understanding of existing bias in ML/DL systems and how it could result in poor quality when evaluated from an ethical and moral perspective.

Bias Probing

Taking the case of GPT-3, a language model used at internet-scale, it is important that the model is inclusive and multi-culturally sensitive. Keeping the purpose in mind, the model was probed for bias by [Brown et al.2020] during its inception. They probed for gender, racial and religious bias.

Gender bias manifested itself when investigated for associations between gender and occupation. They found that 83% of occupations they tested for, out of 388 occupations, were more likely to be followed by a male-identifier in GPT-3. These occupations usually were ones that required higher levels of education such as professor, banker or emeritus or ones that required hard labor such as mason or sheriff. They probed further by adding adjectives such as competent and found that occupations that require competency were associated with male identifiers while the opposite was associated with female identifiers.

The work performs co-occurrence tests for analyzing racial and religious bias. They discovered that Asians had a consistently high sentiment while Blacks had a consistently low sentiment. The model reflected the bias often present in the world when it came to religious contexts.
This important finding leads us to the reason behind bias in pre-trained language models. The models are trained on crowd-sourced data such as Wikipedia or Twitter. These websites capture human knowledge often in uninhibited form exhibiting the inherent bias present. Since this bias is now present in the training data, it is propagated to the model during training and then downstream.

Understanding Representational Embeddings

The seed for all bias in Natural Language Processing applications is the representation of the natural language, i.e. through word embeddings, which are vector representations of words. As explored by [Tolga Bolukbasi et al. 2016], we see a clear quantification of bias perpetrated by word embeddings. As words are mapped to a vector space using word embeddings, we can similarly map word embeddings to the “gender” subspace. Using this, we can determine the true leanings of neutral words’ embeddings with respect to gender. For example, we can establish direct bias if the word “programmer” leans more heavily towards “man”, as compared to the word “home-maker”. These words have no implicit gender reference and therefore should ideally lie at the midpoint of the words “man” and “woman” in our vector space. We can now map all word embeddings to this space to quantify the bias through a simple vector deviation from the “middle point” of the vector space, thus quantifying bias. This forms the crux of direct bias, which accounts for a majority of the usage of biased words and forms the essence of most bias-finding in different models.

A more nuanced setup to finding bias is the notion of indirect bias, which is subtler and more difficult to quantify. However, such cases of indirect bias are more human-like and hence ubiquitous in most crowd sourced data. An attempt at quantification is made by [Tolga Bolukbasi et al. 2016] by mapping the word embeddings of neutral words onto the direction of some known biased words. For example, the word embeddings of all occupations are mapped along the "softball-football" axis to see how these occupations are similar to these two sports. This is a more nuanced comparison but still explores the underlying bias and if a significant leaning is observed, it is indicative of bias in terms of the underlying "gender" category.

[Ryan Steed et al. 2021] borrows this idea of probing word embeddings and applies it to image embeddings which are generated by unsupervised image models. The practice of using proven sucessful NLP techniques for images is not uncommon in the ML/DL community. Transformers, a highly popular language model was applied to images successfully. This work executes that idea through the iGPT and SimCLRv2 models. Similar to the Word Embedding Association Test(WEAT) they introduce an Image Embedding Association Test(iEAT).They probe for bias through the next pixel prediction task. To establish a semantic relationship for an image, they provide visual stimuli to the model. They look to prove three hypotheses to probe for intersectional bias,

1. Intersectionality hypothesis: Testing for intersectional data between attributes would reveal more emergent biases than testing these attributes individually.
2. Race hypothesis: Biases present between racial groups would be more similar to biases amongst men rather than women. As an example, the differential bias between Black and White is more evident between a Black man and White man rather than a Black woman and White woman.
3. Gender hypothesis: biases between men and women will be most similar to biases between men and women belonging to a specific race.

They probed the model by providing a set of images - {A,B} and a set of targets - {X,Y}to check the association between A and X, B and Y. As an example, the baseline test was for {Flower,Insect} and {Sunset,Morgue} where Sunset and Morgue were accepted stimuli for pleasant and unpleasant. This is a widely-accepted bias and was considered the baseline for evaluation metric - Valence, which is the intrinsic pleasantness of things. If an image had positive valence then it was pleasant and negative valence indicated unpleasant. They find that skin color is a bias where the valence reduces as one goes from light to dark skin color. Another finding comes from the Gender-Career test where significant bias associating male to career-related attributes and women to family-related attributes. Intersectional valence tests such as White Female:Pleasant:: Black Male:Unpleasant are consistent with societal biases. The Gender-Science bias when applied to White Males and Black Females produce significant results. There are some cases where no bias is reported but there have been no findings that contradict the notion of human societal bias.

We see the effects of training data on word embedding and therefore model bias through the experimental results of [Xavier Ferrer et al. 2020]. By training word embeddings on a text corpus of reddit data, they were able to quantify the bias present in the data itself. A notable method of testing is that of co-occurrence, where a gender neutral word's context of usage is noted. Another interesting method of checking for bias is through clustering of similar words and noting any anomalous/biased groupings. The results of [Xavier Ferrer et al. 2020] indicate strong bias in the gender category. These results are indicative of the importance of training data in determining the bias of the word embeddings, the trained model and therefore the downstream application.

Solutions

A biased model is an imminent issue that should be handled at the source - data. But, it is expensive to re-annotate large-scale publicly-sourced datasets. It is possible that it is not correctable at the data level either as human bias is bound to be reflected in the annotations. The next strategy would be to correct the model.

[Tolga Bolukbasi et al. 2016] mention techniques to achieve a preliminary level of "debiasing" in these word embeddings. These can take form in a more rigorous hard debiasing or an information preserving soft debiasing. The debiasing achieved here takes advantage of the vector form of the word embeddings and the fact that the distance between two word vectors holds significance in the natural language domain. Therefore, a clear way to remove bias in the gender space is to make sure that all gender-neutral words are always equidistant from their respective gender poles. This would strip off any additional information specific to the chosen words outside of gender. For example, the word embedding for "grandfather" might encapsulate information with respect to babysitting outside of the gender subspace. To soften the blow in terms of model performance, we can now look to "tune" the model's bias correction through a softness/hardness scale.

Another measure proposed by [Alex Beutel et al. 2017] is the adversarial learning model for fair representation. Here, we modify our regular training procedure by not only minimizing the error (or loss) which the model predicts as compared to the ground truth but to also minimize the dependence of the results on a predefined sensitive attribute. They also define a tradeoff factor between the prediction accuracy and the dependence on the sensitive attribute (which could be race, gender, etc.). We must note that for a fixed data source, we currently need to make a tradeoff between model accuracy and model fairness. This would become another tunable parameter for models in the near future.

But, how about pre-trained models that have already been deployed? [Mitchell et al, 2019] proposes a framework for model transparency. They suggest that every model be released with a "model card" in order to help stakeholders standardize ethical practices and reporting. The

technical details of the model are covered under the model details, metrics, quantitative analyses, training data and evaluation data. The ethical aspects are detailed in the intended use, factors, ethical considerations and caveats and recommendation sections. This framework promotes accountability for the model developers. They are provided with an opportunity to contemplate their model from all ethical perspectives and declare any unexpected or biased behavior to the stakeholders. Expanding on the ethical considerations section,
i) Data: if the data has any sensitive information
ii) Human life: if the model makes any decision concerning human life such as safety or health
iii) Risk Mitigation and harms: what are the possible risks? How has it been mitigated and if anyone would be harmed by the model.
iv) Use-cases: Any unintended consequences that can be foreseen from a use-case.

Another section that touches upon the ethical aspects of a model is quantitative analyses. The developers are encouraged to perform intersectional and unitary tests to quantify fairness through parity on the different metrics. The idea promotes ethical contemplation of a model and attaches accountability to the developers.

[Shen et al, 2021] expand on this idea by proposing a Value Card that consists of a Model and a Persona Card. The purpose of this is largely educational. They hope to teach future practitioners about the social impacts of ML-based decision systems. Along with the Model cards they include Persona cards which outline the impact of this model on every stakeholder. Then finally the toolkit provides a Checklist Card where one lists the impacts, the degree, scale and direction (positive or negative) of this impact. This toolkit, although intended for future practitioners, can be useful for current as well, to examine the societal and ethical implications for any new model.

Discussion

A machine learning model is only as good as its data, and this is no different for a pre-trained model - be it a language model or a computer vision model. On reviewing all the papers highlighted, we see a common theme which ties the biases learnt to the pretraining data. The data used for pre training is usually crowd sourced, be it wikipedia data or news articles. The fact that our data sources are created by humans and represent human opinions or facts which are presented by humans allows for the seepage of human biases into the ML models.

As explained through the prototype theory [Daniel Aberra 2006] in [8], humans tend to implicitly add qualifying words when describing things which are out of the norm. For example, rarely do we refer to a banana as a “yellow banana” but we do hear ourselves describing unripe bananas as “green banana” or any other such qualifier. This might seem obvious to us in this context but we may be doing the same when it comes to other areas. For example, the usage “female doctor” or “male nurse” when fed into a language model starts treating the lone word “doctor” associated with a male usage and the word “nurse” with a female context respectively.

These kinds of usages could dominate news articles which try to make a reader quickly grasp knowledge even while running the risk of sounding biased. When fed into a model we see associations that have been highlighted by [Tolga Bolukbasi et al. 2016].

Another origin of bias through data as highlighted by [Ryan Steed et al. 2021] is the use of web-sourced data. Here they mention how web images have implicit representational bias which causes the model to develop bias. Web-sourced text data suffers from the same problems. The problem is that the embeddings infer contextual information which suffers from bias which takes the form of "ingroup favoritism, rather than out-group derogation". For example, if the model currently learns that weddings have a positive connotation and our data has instances of white grooms or more data points associating white people with weddings, the model implicitly learns a more positive connotation associated with the white race as compared to other underrepresented races in our pre-training data.

This dependence on data makes us rethink our demonization of pre-trained models. Is an effort in data engineering required more than the cutting edge ML effort we are focussing on today? Would it be easier to curb the bias at source rather than correcting it down the stream? Is it possible to find an intermediate measure where we account for bias at the data level and at the model level?

Conclusion

Through our paper, we review and collate the current literature concerning the bias problems seen in the Natural Language Processing domain of Machine Learning which is now promulgated by the use of Pre-Trained Language Models. We note that the problem of bias can manifest itself through every step of the ML pipeline starting from the data and ending with the inference from our downstream ML models. We also highlight the feasibility of action at various steps in the pipeline.

We do not limit ourselves to discussing the problems but also concern ourselves with the current solutions as well as the future directions of solving and accounting for ML bias. Bias accountability through model cards is a future step in this domain and could break the "black-box" image of current pre-trained language models. Generating and highlighting model cards could be made an essential part of ML pipelines and provide for greater transparency and traceability of bias, which provide for greater accountability.

Over the past decade or so we have seen great effort towards establishing and quantifying bias. As part of the future direction, we see a greater effort in taking accountability for bias and trying to correct for bias. In the near future, we expect to see bias percentages being reported as a standard model metric.

References:

1. Brown et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165. Retrieved from https://arxiv.org/abs/2005.14165.

2. Pratik Gajane, Mykola Pechenizkiy. 2018. On Formalizing fairness in prediction with Machine Learning. arXiv:1710.03184. Retrieved from https://arxiv.org/abs/1710.03184.

3. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. 2016. *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*. arXiv:1607.06520. Retrieved from https://arxiv.org/abs/1607.06520.

4. Xavier Ferrer, Tom van Nuenen, Jose M. Such,and Natalia Criado. 2020. Discovering and Categorising Language Biases in Reddit**.** Retrieved from **https://arxiv.org/abs/2008.02754**

5. MEHRABI, N et al., 2022. A Survey on Bias and Fairness in Machine Learning | ACM Computing Surveys. [online] ACM Computing Surveys. Available at: <https://dl.acm.org/doi/pdf/10.1145/3457607> [Accessed 6 May 2022].

6. Suresh, H. and Guttag, J., 2022. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.

7. Aberra, Daniel. (2006). Prototype Theory in Cognitive Linguistics.

8. Bias in the Vision and Language of Artificial Intelligence, Lecture URL : https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf

9. Ryan Steed and Aylin Caliskan. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3442188.3445932
10. Alex Beutel, Jilin Chen, Zhe Zhao, Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In Proceedings of 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning, Halifax, Canada, August 2017 (FAT/ML '17), 5 pages.

11. Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

12. Hong Shen, Wesley H. Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In Conference on Fairness, Accountability, and Transparency (FAccT ’21), March