Fortnightly Task 3

Student Name:- Mark Pereira

Student ID:- S3797413

Evaluating machine learning models for bias is becoming an increasingly common focus for different industries and researchers. In, this report I am going to research the Holstein et al. CHI'19 paper titled "Improving fairness in machine learning systems: What do industry practitioners need?" and address the following points. Investigation is done of various industry ML practitioners about the challenges and needs around fairness in Machine Learning. The authors interviewed to get a wide sense of challenges faced in defining fairness measures.

What is the state-of-the-art on evaluation of fairness in machine learning?

While Investigation many interviewees reported that their team only look for the training datasets and not for their ML models to improve their fairness in their products. Furthermore, survey respondents mentioned that they try to address fairness issues found in the products, the very commonly used strategy was "collecting more training data". Communication between models and data collectors is needed in order to improve the fairness, get more data for training but to get it is very rare since data collectors doesn't focus on particular data. In order to evaluate fairness in machine learning the test set should be well constructed and not biased. For e.g. Images of female doctors were often mislabelled as nurses which they ultimately traced to imbalances in their training data. People mostly think about attributes like ethnicity and gender but the biggest problem found is that it should be based on domain and problem. Interviewees also stated that they have practice of getting together as a team and try to imagine everything that could go wrong with their products, so that they can proactively monitor for those issues. The performance and progress is monitored by fairness metrics or key performance metrics and the automated tests are performed. Fairness auditing without access to demographics at an individual level. However, some interviewees reported that they could only use coarse-grained demographic information (e.g., region or organizational-level demographics) for fairness auditing. Prototyping conversational agents more rapidly, including methods for simulation conversational trajectories and then finding ways to automate the identification of risky conversation patterns that comes.
There should be domain-specific resources, such as ethical frameworks and case studies to guide the teams with their ongoing efforts around fairness. Changes to broader system design could be used to address the fairness in real world ML systems instead on the development of algorithmic methods. For example, images of female doctors mislabelled as nurses was solved by replacing the system outputs nurse and doctor with more generic health-care professional. It was suggested that a tool could be used to artificially generate essays which can paraphrase an essay in another subgroup's style, a different voice (or native speech) and without changing the linguistic content to know whether this could change the scores.

What are the main challenges on defining 'fairness' and fairness-aware measures in this context?

The challenge is with the automated essay scoring to score African American student fairly, the team needed the data of their highly scoring African students, the team need the data of their highly scores with the data collection team which is very rare. What is the right way to sample high scorers without having score all the essay? So there is some kind of different way to indicate from which schools to take the data from or to bother spending extra money to score. There is a challenge where specific user populations are less engaged with the product, to overcome such challenges the manager suggested to encourage product usage with targeted or specific populations. Also, to balance the data get the more data with less engaged users so there won't be any fairness. The other challenge is that which sub populations needed to be considered when developing specific kinds of ML applications to ensure that they collect sufficient data from these subpopulations or balance across then when curating datasets. Another challenge is potential issues after deployment through user complaints or by reading negative media coverage about their products. Several issues were known only after deployment of the system despite running user studies. Even with efforts of gathering training data to address the fairness issues were slowed down by the team's blind spots. Automated test were performed but it's really hard to fix things that you can't measure. Challenge was faced with respect to auditing, methods were abandoned by the teams citing limited time and resources to spend on building their own solutions. There were challenges in diagnosing whether specific issues (e.g. Complaints from customers) are symptomatic of broader, systematic problems or just "one-offs". There are challenges when making side effects to making changes to datasets or models to improve fairness in an unexpected ways that's harmed users experience. The challenge is that when targeting people based certain aspects of their person where they are confused how to do that in the most morally and ethically and even vaguely responsible way. Future research should explore ways to help Teams better understand biases that may be present in the humans embedded throughout the ML development pipeline, as well as ways to mitigate these biases.

Is it possible to measure fairness in the context of your WIL project? If so, how?

In our WIL project we are going to analyse the impact of covid on economies GDP of the various countries. We will compare the impact of covid cases with respect to the current situation of the cases in country and whether the number of cases impacted by covid are more or less and whether they are good enough to impact on the nation's economy .We have collected data from online source such as john Hopkins University, etc. Yes it is possible to measure fairness in our project. Fairness auditing without access to demographics at an individual level can be done as our project doesn't has demographic information and also it has country related information so this is currently under use.

REFERENCES:

[1]"Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?", *Arxiv.org*, 2019. [Online]. Available: https://arxiv.org/pdf/1812.05239.pdf.