



PAPER READING TASK

BIAS AND FAIRNESS IN NLP

MSS Sriharsha



STRENGTHS

- Built a **dataset** for evaluating fairness and bias in language models. This dataset is made, annotated and reviewed by human volunteers from different parts of India. This ensures that enough stereotypes are considered in the dataset.
- Analysis of common stereotypes in India and comparing them to the western context. Analysis refers to sentiment analysis of these stereotypes and corpus analysis. Corpus analysis refers to finding these common stereotypes in standard datasets like **Wikipedia** and **IndicCorp**. This helps in understanding whether a stereotype is positive or negative, and also test how biased a language model is by measuring the number of co occurrence of stereotypes.
- From the gender analysis of language models, the paper establishes that models perform differently for an Indian setting, thus making it more important to come up with newer and different fairness techniques which are being developed in the West.
-



WEAKNESSES

- The *dataset* constructed does not include all the stereotypes. It is based majorly on region and religion. It needs to span other social axes like caste, gender, literacy rates etc. It also needs to span multiple identities at the same time.
- Most of the stereotypes in the dataset are in English. This way the dataset falls short in obtaining stereotypes which are not in English. There are many stereotypes which are different in multiple languages. One needs to include them too.
- The paper does not specify a metric system for evaluating the bias in language models. Comparing bias between 2 models is hard since the underlying datasets may or may not be different.
-



IMPROVEMENTS

- A diverse set of human annotators can help cover the marginalized communities, while covering other societal parameters like literacy, access to healthcare and technology, financial status, etc. This also helps us to bridge the gap between different languages, regions, etc.
- Analyzing the underlying datasets can help us understand, why this bias arises in different language models. If the datasets, which are used to train these models, have a large amount of stereotypes, then one can conclude that the models trained on these datasets are likely to be biased and stereotypical.
- The above analysis can help us in removing the stereotypes in the pre-processing stage itself. This can help us train the model in a more fair manner. This is particularly useful in a legal setting, where AI can be used to evaluate legal cases, The presence of stereotypes can affect the decision making process of the model in such a setting.
- Intersectionality of the dataset constructed might improve and cover more number of stereotypes. This bridges the marginalized groups which are specific to certain regions.