Summary of paper

- The main point of this paper is to re-contextualizing NLP fairness for non-Western areas, especially for India. It is because of the diversity of our nation i.e., due to various axes of disparities such as Region, Caste, Gender, Religion, Ability, and Gender Identity and Sexual Orientation. This paper mainly focuses on the Region and Religion.
- This paper also curated dataset for the purpose of evaluating fairness of NLP models(LM) with respect to Indian context(only with Region and Religion).
- Demonstration of prediction biases and over-prevalence of social stereotypes in data and models. The above three points are the main contributions of the paper.
- Indian specific axes are Region and Caste, whereas gender, religion, ability, sexual orientation are global axes in India context.
- Bias evaluation in NLP relies on proxies of sub- groups in language, such as identity terms and personal names, to reveal the undesirable associations present in models and data used in training these models.
- In the Indian context, the paper mentions to have identified three major kinds of proxies: identity terms, personal names, and dialectal features.

- The paper has curated lists of India-specific identity terms along three different axes-
  - region-for example Punjabis for people of Punjab, Goans for people of Goa etc..
  - religion-popular religious groups like Hindu, Muslim, Christian, Sikh, Buddhist, Jain.
  - caste-like SC/ST, OBC, Brahmins, Shudras etc..

  The bias has been demonstrated using these terms, in the model DistilBERT-base-uncased.
- The bias is estimated by using "perturbation sensitivity analysis", that reveals biases by counterfactual replacement of terms of same semantic category in natural sentences.
- The result, as an analysis of the predictions by the model is, for caste, the model had significant negative association towards the terms OBC and Dalit, both of which represent historically marginalized groups; and for religion, we find negative association towards the terms Muslim and Hindu, while Jain and Christian have positive associations.
- The analysis was for gendered correlation in pretrained models using the DisCo metric is performed using MuRIL and mBERT with the task of predicting the masked token in a sentence. DisCo metric, which measures if the predictions of a language model have disproportionate association to a particular gender.
- For example, a template for the profession category of tokens is: “[it] are most likely to work as <MASK>." 16 For each tuple (i, t), we replace it in the template with identity term i and record if the token t, or its inflections occur in the top K (K=5) predictions of the model.

- Results for the gendered correlation in the NLP models is while using American names, it appears that MuRIL has a lesser amount of bias than mBERT. However, using Indian names reveals that while MuRIL learned to detect names better, it also learned more stereotypical associations around those names.
- It also finds that MuRIL shows consistently higher percentage of Stereotypical tuples in top 5 predictions suggesting that it has learned more stereotypes in the Indian context due to data sourced from India.
- Limitations of the dataset created, first we capture only two axes of disparities: region and religion, and in English. The approach towards filtering the set of tuples for annotation based on co-occurrence limit our data to only capture those stereotypes that are explicitly mentioned in text, but there might exist stereotypes in society that are not captured in corpora and hence will not be captured by our dataset.

Strengths of the paper

- This paper created a dataset (NLP-fairness-for-India) with stereotypical associations to identity terms to evaluate the bias encoded in the models and in the data as well.
- This paper lays a foundation to further research in this particular area, by mentioning various methods to quantify the bias in the NLP models such as DisCo metric to quantify the gender associations of a model and perturbation sensitivity analysis.
- The demonstration of the bias prevalent in the models (DistilBERT, MuRIL, mBERT) by the help of the NLP-fairness-for-India dataset created.

Weakness of the paper

- The paper only focuses on region and religion axes of social disparities. The reason for not able to cover the other axis of disparities as mentioned by the paper are, the access to diverse annotator pools who have familiarity and/or lived experiences of the marginalized groups especially as the public discourse around (dis)ability, gender identity and sexual orientation is relatively new and limited.
- The paper doesn't account for multilingual fairness research(especially for Indian languages like Hindi, Telugu etc.)
- Though the dataset creation was done by selecting people from various regions and religion, trying to ensure diversity, they may not accurately represent the societal biases.

Scope of improvement to the paper

- As the paper only focuses on region and religion axes of social disparities, we could create dataset with other axis and perform similar analysis to check the bias in the models. This would need crowdsourcing to be able to capture as much variance as possible,which would be representative about the India's population.
- As the paper doesn't account for multilingual fairness research(especially for Indian languages like Hindi, Telugu etc.),we could use Indian data and models trained from scratch on this data(for example - IndicBERT) to know the bias present in them.
- Propose methods, strategies to tackle the bias identified in the models as well as in the data.