PRECOG

TASK 4

KUVAM PAHUJA 2022101030

Summary

The research paper reported some highlights from extensive empirical analysis of social biases in NLP models.in the Indian context , The axes of disparities they considered include two India-specific axes: a) Caste,which is an inherited hierarchical social identity, that has been the basis of historical marginalization;and b) Region, or ethnicity associated with geographic regions of India, as well as four globally-salient axes that have unique manifestations in the Indian context: a) Gender, where there are different structural disparities in engagement of women in society as compared to the West; b) Religion,wherein the majority and minority religious groups differ compared to the west; c) (dis)Ability and 4)Gender Identity and Sexual Orientation, around which the social discourse and awareness in India is fairly recent. They analyzed various proxies in language data for these social groups such as identity terms, personal names, and dialectal features to study biases in NLP models.

On analyzing sentiment scores, we see that the model has learnt to associate higher negative sentiment towards marginalized sub-groups, such as ‘Dalit’ and ‘OBC’ (other backward castes) in caste, and ‘Muslim’ in religion. For state identities, the model has learnt to associate more negative sentiment with southern states like Andhra Pradesh and Telangana, and North-Eastern states like Mizoram and Manipur.

NLP models reflect societal biases around socio-demographic subgroups in the Indian context. To effectively address these issues they proposed a holistic research agenda for re-contextualizing fairness in NLP along three dimensions:

accounting for the societal context, bridging the technological gaps, and adapting to the local values and norms.

Societal Context

1. **Socially Situated Evaluation:** Paper highlights the importance of having diverse annotator pools with familiarity and lived experiences of marginalized groups to ensure fairness in evaluations. This is especially crucial in India where public discourse on issues like (dis)ability, gender identity, and sexual orientation is limited. Participatory approaches to co-create evaluation resources are suggested as a solution.

2. **Data Voids:** Entire communities might be excluded from language data due to disparities in literacy and internet access, which can lead to biases in language models. There's a risk of unintentionally excluding marginalized communities based on dialect or linguistic features while filtering data for quality. Participatory data curation, including collecting language data specifically from marginalized communities, is proposed to address these data voids.

3. **Intersectionality:** The paper emphasizes the exacerbation of biases due to the intersection of diverse axes in the Indian context. Differences in literacy, economic stability, technology access, and healthcare access across geographical, caste, religious, and gender divides contribute to disparate representation and access to language technologies. Bias evaluation and mitigation interventions need to consider these intersectional biases.

Bridging cross-lingual Technological gaps

1. **Performance Gaps Across Languages:** Despite India's linguistic diversity, there are significant discrepancies in NLP capabilities among languages and dialects. These disparities hinder equitable access to the internet, information, and representation in data and models. While progress has been made in narrowing this gap, more work is needed, especially for marginalized and endangered languages.
2. **Multilingual Fairness Research:** Current NLP fairness research largely relies on evaluation resources designed for Western languages. It's crucial to develop these resources for Indian

languages, such as Hindi, Bengali, and Telugu, as biases may vary across languages. Additionally, the impact of bias mitigation strategies in one language on others needs consideration, highlighting the need for a research agenda to address these complexities in the multilingual setting.

Aligning NLP Models with Indian Context

1. **Avoiding Value Imposition:** Fairness inquiries often rely on implicit assumptions rooted in Western values, which may risk imposing values on Indian contexts. While Western notions of fairness are typically based on egalitarianism and distributive justice, Indian philosophy emphasizes social restorative justice. Addressing value alignment challenges is crucial in deploying fairness interventions effectively.

2. **Accounting for Indian Justice Models:** India employs restorative justice measures, such as reservations, to address historical marginalization of communities like Dalits, other backward castes, Adivasis, and religious minorities. In NLP fairness research, it's essential to consider how interventions align with these established measures of justice, particularly in domains like educational institutes and government jobs.

MAJOR STRENGTHS OF PAPER

- Comprehensive coverage of various social axes pertinent to Indian society, ensuring a thorough examination of biases.
- The research agenda, aimed at re-contextualizing fairness in NLP within the Indian context, is robust and addresses a wide range of issues related to biases in NLP.
- Reliable metrics employed for bias calculation across different social axes have yielded significant and meaningful results, enhancing the paper's credibility and contribution to the field.

MAJOR WEAKNESSES OF PAPER

- Insufficient dataset robustness and quantity, particularly concerning religion and caste bias assessment.
- Lack of explicit disclosure regarding the implementation formulas for disco and normalization techniques.
- Limited comparison with models trained in the Indian context, hindering comprehensive understanding and interpretation of the results.

AREAS OF IMPROVEMENT

- **Enhance Dataset Robustness and Diversity**:

Address the weakness related to the dataset by sourcing a more comprehensive and diverse dataset, particularly for variables like religion and caste. This could involve collecting data from various sources to ensure representation across different demographics and socio-economic groups within the Indian context. Additionally, the dataset should be sufficiently large to provide robust insights into biases across these axes.

- **Provide Transparent Methodology**:

Address the lack of transparency in the methodology by explicitly detailing the formulas and procedures used for calculating metrics such as DISCO and normalization. This would enhance the reproducibility of the research and allow other researchers to validate and build upon the findings. Providing clear documentation of the methodology will also increase the credibility of the study.

- **Expand Model Comparison**:

To provide a more comprehensive analysis, expand the comparison of NLP models trained specifically on the Indian context. Including a wider range of models trained on Indian languages and datasets would offer a more nuanced understanding of biases and performance differences across various NLP frameworks. This comparative analysis would help readers gain insights into the strengths and weaknesses of different models and their suitability for addressing biases in the Indian context.