CS 626
Seminar Propasal

Anuj Attri (23M0808) Arnav Attri (23M0811)

February 1, 2024

Paper Selection: ”Re-contextualizing Fairness in NLP: The Case of India”

- **Code:** This paper lacks accompanying code, but we have taken the initiative to reproduce the results independently. Our implementation can be found on Github: `https://github.com/arnavcse/NLP-Project`. The repository includes our Google Colab notebook providing insights into the code implementation.

  *Please note that the **results of our findings** are included at the end of this document. Refer to the last pages for visual representations of our results.*

- **Paper ID:** Anthology ID: 2022.aacl-main.55
- **Conference:** Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)
- **Date:** November 2022

A Comprehensive Paper Review:

What:

- The research, titled ”Re-contextualizing Fairness in NLP: The Case of India,” meticulously investigates the fairness challenges in Natural Language Processing (NLP) models, particularly when applied to the diverse linguistic and cultural landscape of India.
- It aims to uncover biases and disparities within NLP models, emphasizing the need for contextually relevant and unbiased responses in India’s sociolinguistic context.

Why:

- **Motivation:**
  - The motivation arises from the necessity to align fairness considerations in NLP models with the intricate sociolinguistic tapestry of India, characterized by a myriad of languages, cultures, and social disparities.
  - Existing NLP models, often trained on global datasets, may lack contextual relevance and perpetuate biases when applied to the diverse Indian linguistic landscape.

How:

- **Methodology:**
  - **Perturbation Analysis:** The study employs perturbation sensitivity analysis to reveal biases in the default HuggingFace sentiment pipeline. For instance, the model's sensitivity to regional, caste, and religious identities is demonstrated through sentence perturbations.
  - **DisCo Metric:** The DisCo metric is utilized to analyze gender biases in pretrained models, showcasing the necessity of India-specific resources for evaluating biases in the Indian context.
  - **Stereotype Dataset Creation:** The authors build a stereotype dataset for the Indian context, highlighting the preferential encoding of stereotypical associations in both NLP data and models.
- **Examples:**
  - The perturbation analysis demonstrates significant shifts in sentiment scores based on regional identities. For instance, Mizoram and Telangana exhibit negative score shifts, emphasizing regional biases.
  - DisCo metric analysis reveals gender biases encoded in personal names, with differences in bias detection between MuRIL and mBERT, showcasing the importance of context-specific evaluations.
  - Stereotype dataset creation unveils the prevalence of societal stereotypes in NLP data and models, emphasizing the need for culture-specific considerations.

Key Terms and Key Points:

- **Social Disparities:** The research emphasizes the importance of accounting for social disparities, including region, religion, gender, and caste, in the evaluation of biases in NLP models.

- **Dialectal Features:** The study explores the sensitivity of sentiment models to dialectal features, showcasing concerns related to socio-economic class and regional identities.
- **Intersectionality:** Recognizing the interplay of diverse axes in the Indian context, the research highlights the importance of addressing intersectional biases in evaluating and mitigating biases in NLP models.

Conclusion:

- The research effectively contributes to re-contextualizing fairness in NLP models for the unique challenges presented by India's linguistic and cultural diversity.
- By employing innovative methodologies such as perturbation analysis, DisCo metric evaluation, and stereotype dataset creation, the study provides valuable insights into biases in NLP models and data.
- The proposed fairness framework, coupled with India-specific resources, lays a foundation for developing more inclusive and culturally sensitive NLP models tailored to India's nuanced context.
- The incorporation of examples, key terms, and key points from the research paper strengthens the understanding of the challenges and solutions proposed, making a compelling case for the necessity of context-specific fairness considerations in NLP.