Debiasing Contextual Embeddings

Project Description

Language models like BERT and RoBERTa capture powerful contextual word representations but also encode harmful social biases (e.g., gender, racial, and occupational stereotypes) from large-scale web data. These biases can lead to unfair outcomes in downstream NLP tasks such as sentiment analysis, coreference resolution, or hiring recommendation systems. This project aims to measure and reduce bias in contextual embeddings.

Datasets

We will use existing, publicly available datasets to measure bias and evaluate model performance:

- **StereoSet** (GitHub link) : A benchmark evaluating gender, race, religion, and profession bias via sentence-pair comparisons (stereotypical vs. anti-stereotypical). It provides quantitative bias scores for masked language models.
- **CrowS-Pairs** (GitHub link) : Contains ~1,500 sentence pairs for testing social biases (gender, race, age, etc.) by checking model preference for stereotypical continuations.
- **SST-2 (Stanford Sentiment Treebank)** (GLUE benchmark): A sentiment classification dataset that we will use to test downstream task performance of debiased models.

Tools

- **Programming Language:** Python
- **Libraries and Frameworks:**
  - **Transformers (Hugging Face) –** for pretrained BERT and RoBERTa models and debiasing implementations
  - **PyTorch –** for fine-tuning and embedding manipulation
  - **NumPy**, **Pandas –** for data preprocessing and computation
  - **Matplotlib –** for visualizing embedding bias and evaluation results
- **Environment:** Google Colab / Kaggle (GPU-enabled)

Models and Evaluation Approach

Our project will utilize BERT and RoBERTa as the primary contextual embedding models. We will use the pre-trained implementations available in the Hugging Face transformers library and fine tune them using counterfactual data augmentation to mitigate bias. We will build on pre-trained transformer architectures rather than implementing them from scratch, adapting and fine-tuning the models. The main hyperparameters to be tuned include learning rate, batch size and number of epochs to optimize model performance.

Evaluation will be done using SEAT/WEAT bias metrics to quantify bias levels and accuracy/F1 scores on downstream tasks such as sentiment analysis, ensuring a balance between fairness and utility.

Related work (not an exhaustive list)

- Caliskan et al., 2017 - Semantics derived automatically from language corpora contain human-like biases: Introduces WEAT, foundational intrinsic test that quantifies associations between target and attribute word sets. We use it as a core metric and to justify permutation-based significance testing.
- May et al., 2019 - On Measuring Social Biases in Sentence Encoders: Extends WEAT to sentence encoders like BERT/RoBERTa via templated texts.
- Zhao et al., 2018 - Gender Bias in Coreference Resolution: Evaluation and Debiasing by Data Augmentation: Canonical Counterfactual Data Augmentation recipe that improves fairness on downstream tasks.
- Bolukbasi et al., 2016 - Man is to Computer Programmer as Woman is to Homemaker: Although designed for word2vec/GloVe, it grounds our projection-based approach

Visualizations and Results to Produce

Intrinsic bias dashboards (per model & condition):

- WEAT/SEAT effect sizes with permutation p-values (tables and compact bar charts).
- Layer wise heatmaps of effect size accross layers and target-attribute sets (gender, race, occupation).

Utility–fairness trade-offs:

- Curves: x-axis = Bias scores(lower=better), y-axis = task metric (F1/Acc, higher = better) for CDA, Projection, CDA + Projection.

Reproducibility Tables:

- Summarized parameters, training time, and seed variance.

Timeline

- Week 1: Implement WEAT/SEAT + permutation tests; layerwise extraction; first plots.
- Week 2: Run CDA fine-tuning on BERT-base; Fine-tune downstream tasks with baseline vs CDA; collect metrics.
- Week 3: Projection & Ablations; Implement bias-subspace estimation + projection; Compare baseline vs projection vs CDA across layers (heatmaps).
- Week 4: Synthesis & Paper; Counterfactual eval on tasks (swap inputs, measure robustness); Write paper draft; create figures; complete Ethics & Limitations.

Responsibilities

- **Sai Charan Reddy Avula:** Evaluation & Bias Metrics Lead.
- **Kirubhaharan Joseph Abraham:** Debiasing Methods Lead.
- **Sujith Peddireddy:** Downstream Tasks & Trade-off Analysis Lead.