Project 1: Detecting Bias in Language Models - Report

Megha Chakraborty

Key Idea

This project explores the concept of word embedding models and their encoding of cultural biases. Using the Word Embedding Association Test (WEAT), we uncover biases in language models trained on diverse datasets and discuss the ethical implications of these biases in real-world applications.

---

1. Objectives

- Understand how language models encode cultural biases.
- Use WEAT to detect biases in models trained on datasets like Twitter, Wikipedia, and Common Crawl.
- Analyze and mitigate the impact of biases in downstream applications.

---

2. Data and Methods

Data Sources

- **Twitter:** Reflects societal stereotypes due to its less curated nature.
- **Wikipedia:** Relatively balanced and curated dataset, showing fewer biases.
- **Common Crawl Corpus:** Includes a mix of curated and societal data, showing intermediate bias levels.

Methods

- **Word Embedding Association Test (WEAT):** A statistical tool for measuring bias in word embeddings.
- **Evaluation Criteria:** Association scores between names (e.g., European vs. African) and attributes (pleasant vs. unpleasant), career vs. family, and religion-related terms.

---

3. Key Findings

Race/Ethnicity Bias

- European names showed stronger associations with "pleasant" attributes compared to African names.
- This highlights biases in the training data, potentially perpetuating stereotypes in downstream tasks.

Gender Bias

- Male names were more associated with "career" attributes, while female names showed stronger ties to "family" attributes.
- Twitter and Common Crawl datasets amplified these biases due to societal stereotypes, while Wikipedia showed relatively balanced results.

Religion Bias

- Christianity was more positively associated with "pleasant" attributes, whereas Islam had a closer association with "unpleasant" attributes.
- These findings reflect Western cultural biases embedded in the training corpus.

---

4. Evaluation

Visualizations

- **Box Plots for Bias Analysis:**
    - Race/Ethnicity: Clear separation in association scores for European vs. African names.
    - Gender: Stronger career-family bias in less curated datasets.
    - Religion: Western cultural biases favor Christianity over Islam.

Ethical Implications

- Models trained on societal data risk amplifying biases present in their corpus.
- Results emphasize the importance of dataset inspection and fairness audits.

---

5. Insights and Learning

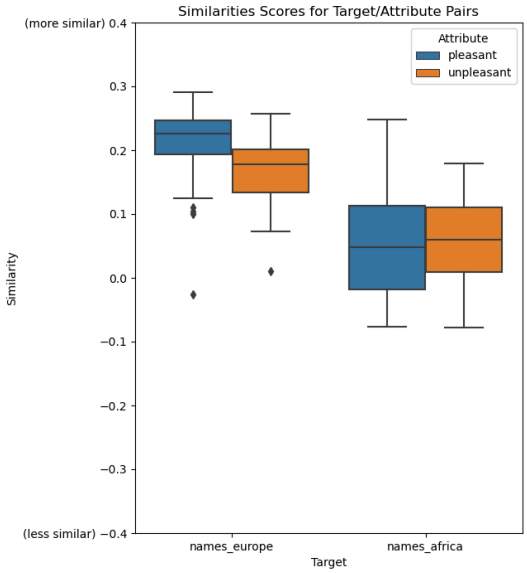
Experience

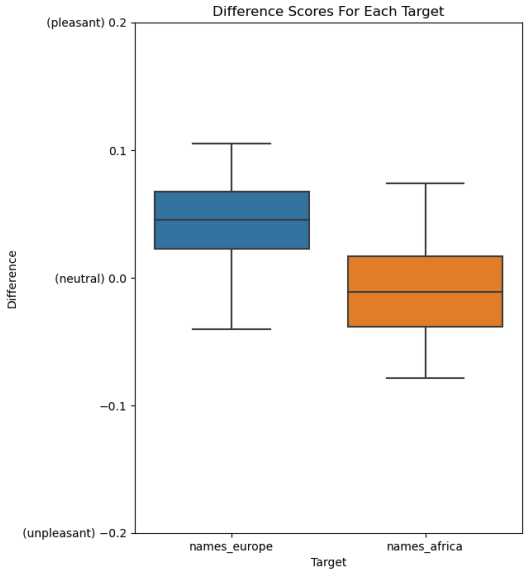
- Learned how societal structures and imbalanced datasets impact word embeddings.
- Highlighted the importance of using tools like WEAT for systematic bias analysis.

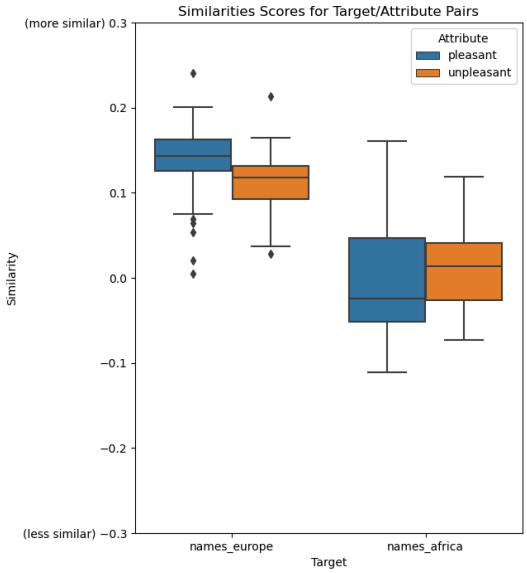
---

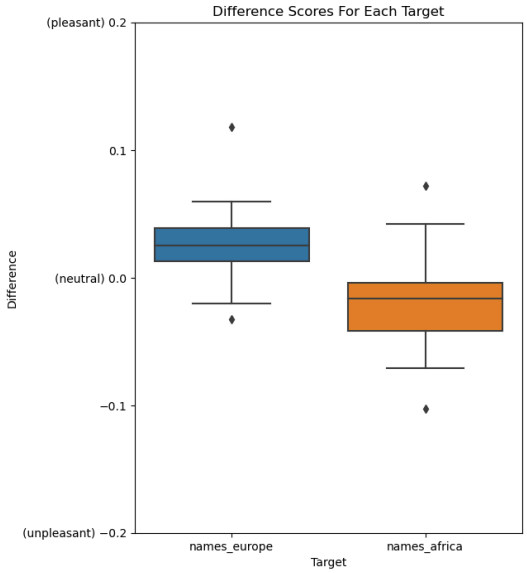
6. Screenshots

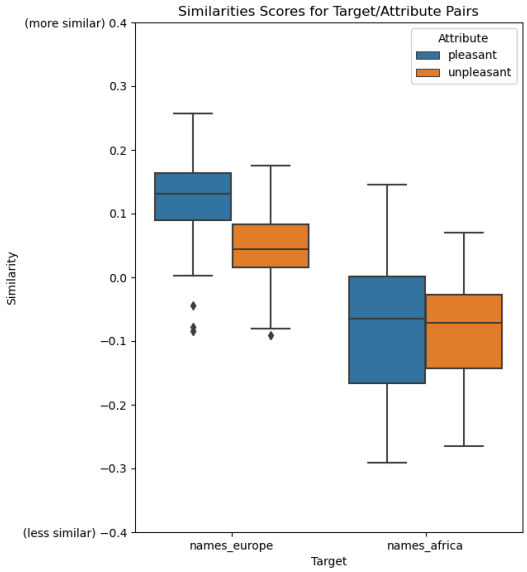
- Screenshots of box plots and WEAT results showcasing biases in datasets and models.

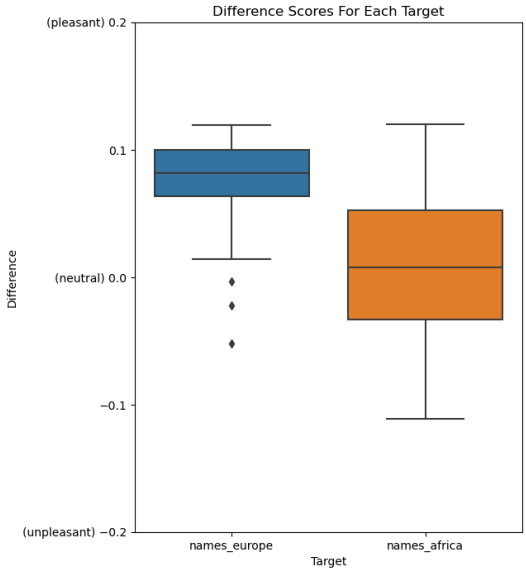












---

7. Implications and Future Directions

- The results stress the critical role of balanced and representative data in reducing biases.
- Future work should focus on:
    - Developing advanced debiasing methods.
    - Incorporating diverse cultural perspectives in training data.
    - Conducting fairness audits before deploying models in sensitive applications.