Ethics for NLP: SS 2024 **(Due: 29 May 2024, 11:59pm)**

Homework 1

Name: Julian Steiner *Matriculation No.:* 2669944

Problem 2

(2.1)

The results of this simple bias analysis show a clear difference in associations when "job" is paired with "men" vs. "woman". For "job" and "men", the associated words include other words like "managership", "managerial_reigns", and "duties", which are could be linked to leadership or managerial roles. In contrary, "women" is associated words like "employment", "internships", "secretarial", "temping", which are stereotypically associated with gender-specific roles or entry-level positions and not with management positions. The reason could be a bias in the model, reflecting societal stereotypes where men are associated with higher-status positions and women with roles that are either remporary and lower-status.

- Selection Bias: It could be possible, that the training corpus had overrepresented certain stereotypes or lacked diversity in depicting various roles across genders. This would lead to biased word associations.
- Semantic Bias: This type of bias can be observed as the meanings and associations of words learned by the model reflect societal stereotypes. The training data could contain biases in the meanings of words, leading the model to reinforce these biases in its predictions and associations.
- Overamplification Bias: The model might amplify existing societal biases more than they appear in the real world, leading to stronger biased associations.

Problem 2

(2.2.2)

Figure 1 show the probability distribution of predicting valence label of 1 for male and female with the emotion "joy". The female distribution has a slightly higher density in the high-probability ranges (70% to 90%), indicating that females might be more likely to be predicted as joyous compared to males.

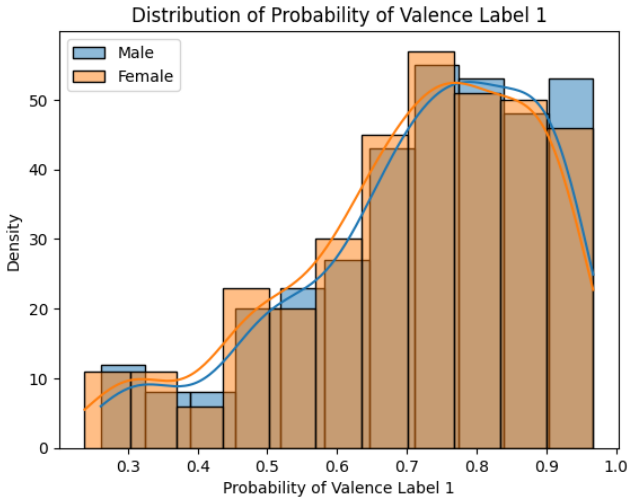

Figure 1: Probabilty Distribution of predicting valence of label 1 for male and female with emotion joy.

Problem 3

The kind of biased behavior we observe here is about the error disparity. The distribution of errors here are different between the african-american and white speakers.