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 most probability ranges, indicating that females might be more likely to be predicted as joyous compared to males.

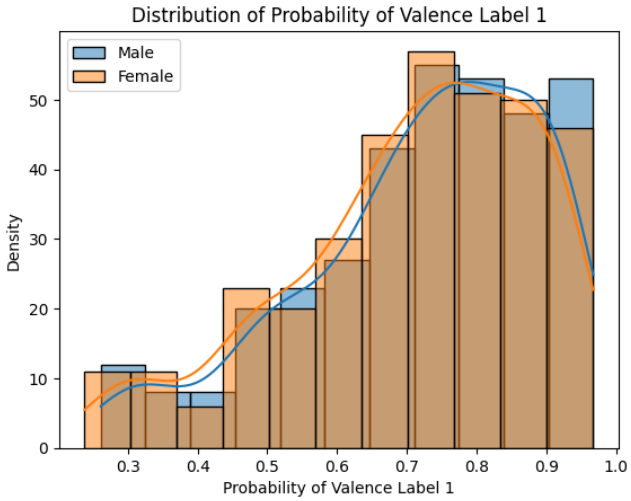

Figure 1: Probabilty Distribution of predicting valence of label 1 for male and female with emotion joy.

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. The Word Error Rate (WER) is consistently higher for African-American speakers compared to white speakers.

The result for this could be of multiple bias sources:

- Label Bias: The labels used for training could be biased, for example, due to the annotators of the data. If the majority of the annotators were white speakers, it could be possible that this had led to a bias in the training data.
- Selection Bias: It could be possible that the training data does not adequately represent the diversity of the population using the system. The dataset could contain significantly more audio samples from white speakers than from African-American speakers. The result would be that the model will be better at recognizing patterns it has seen more frequently.
- Overamplification Bias: It could be possible that there is a overamplification bias. The model could overamphasize linguistic features that are more prevalent in white speech.
- Semantic Bias: It could be possible that the training data may not include sufficient representation of the linguistic styles common among African-American speakers. The consequence is that the model could have biases in how the interpret and process language from different groups.
- Design Bias: I don't think that the system should only work for white speaker and not for African-American speaker. It should work for both groups. It could be possible that more resource of white speakers were available to train the models. The consequence we can see in the results. The system perform suboptimally for the African-American speaker.