HW3 Report (template)

JohnSmith, 2020345678, Computer Science and Engineering

**Question 1.** **(10pt)** After executing the EDA/Modeling/Inference code on the Hotel Review dataset, analyze and discuss your observations. **Do you think the training has been appropriately processed? Provide your reasoning based on the plot result above.**We highly encourage you to attach specific examples and what observations you made about them.

|  |
| --- |
|  |

**Question 2.** **(10pt)** Have you observed anything on the relationship between the model performance and bias measure, even if it is not significant in our assignment? If so, share your opinions on why such a relationship could exist.

|  |
| --- |
|  |

**Question 3.** **(10pt)** We explored one debiasing techniques in pre-processing. Consider whether there are other debiasing techniques that could be applied using the same training dataset. **Are there additional debiasing methods beyond the two covered in this assignment?**  If so, identify and describe them; if not, explain why.

|  |
| --- |
|  |

**Question 4.** **(20pt)** Debiasing methods can be applied not only during pre-processing and post-processing but also within the model training phase. While integrating debiasing techniques into the training process can be challenging due to resource constraints or project limitations, it is feasible and can yield benefits sometimes (e.g., see [Zhang et al., "Mitigating Unwanted Biases with Adversarial Learning," 2018]).

Imagine you are using a Large Language Model (LLM) API in your application, and you notice that the model exhibits a specific type of bias—for instance, it generates gender-stereotypical responses when asked about certain professions or provides culturally biased answers in specific contexts.

***Given that you cannot modify the underlying model directly, discuss how you might interact with the LLM to mitigate this bias from a Human-AI Interaction and Interface perspective.* *Propose at least one method to overcome the bias interactively without altering the model itself.*** Consider strategies such as prompt engineering, designing user interfaces that promote fairness, or implementing feedback mechanisms. Explain how your proposed method addresses the identified bias and enhances the overall user experience. Support your discussion with specific scenarios or examples.

|  |
| --- |
|  |