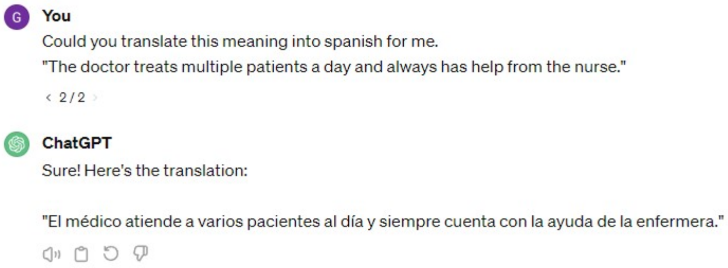
Theory 1 D7047E

May 1, 2024

1 Theoretical Tasks

1.1 Task 1.1

In figure 1 chatGPT (version 3.5) was asked to translate a English sentence which contained the words ”doctor” and ”nurse” which in English are not connected to a gender to Spanish where these words translate to ”médico/médica” and ”enfermero/enfermera” depending on gender. Here the modeled translated doctor to the masculine version ”médico” and nurse to the feminine version ”enfermera”. This is a clear example of bias in the LLM. We tested the exact prompt as in 1 a couple of times and in all of our attempts ”doctor” was translated to the masculine version and ”nurse” to the feminine version. Our responses was either as in figure 1 or as in 2.

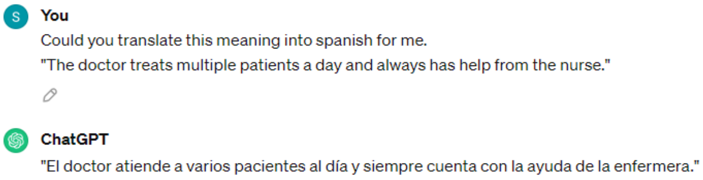


You
Could you translate this meaning into spanish for me.
"The doctor treats multiple patients a day and always has help from the nurse."

ChatGPT
Sure! Here's the translation:

"El médico atiende a varios pacientes al día y siempre cuenta con la ayuda de la enfermera."

Figure 1: Example of bias within chatGPT 3.5



You
Could you translate this meaning into spanish for me.
"The doctor treats multiple patients a day and always has help from the nurse."

ChatGPT
"El doctor atiende a varios pacientes al día y siempre cuenta con la ayuda de la enfermera."

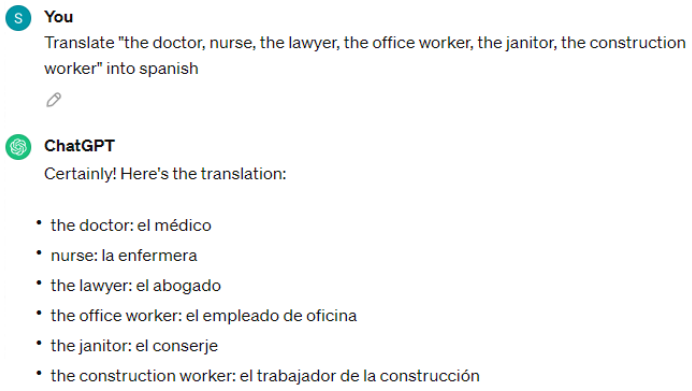
Figure 2: Another response to the same prompt as in 1

We also tested translating all the examples given and got results according to 3 and here only ”nurse” was translated to the feminine version of the word.

2 Metrics Tasks

2.1 Fill out all the missing values, and put an explanation of why accuracy may not be the best metric

Given the confusion matrix



You

Translate "the doctor, nurse, the lawyer, the office worker, the janitor, the construction worker" into spanish

ChatGPT

Certainly! Here's the translation:

- the doctor: el médico
- nurse: la enfermera
- the lawyer: el abogado
- the office worker: el empleado de oficina
- the janitor: el conserje
- the construction worker: el trabajador de la construcción

Figure 3: Caption

	Predicted	
Actual	**Negative**	**Positive**
Negative	**990**	**10**
Positive	**20**	**30**

The classes are imbalanced. Accuracy may therfore not be the best metric to evaluate the performance of the classifier. It doesn't provide a clear picture of how well the classifier is performing, especially in predicting the minority class (Positive cases). In this case, accuracy could be misleadingly high because it doesn't penalize misclassifications in the minority class as much as it should. The values for the confusion matrix are:

- True Negative (TN): 990.
- False Positive (FP): 10.
- False Negative (FN): 20.
- True Positive (TP): 30.

Precision for this is given by

$$P = \frac{TP}{TP + FP} = \frac{30}{30 + 10} = 0.75. \tag{1}$$

Recall for this is given by

$$R = \frac{TP}{TP + FN} = \frac{30}{30 + 20} = 0.6. \tag{2}$$

Accuracy is given by

$$A = \frac{TP + TN}{TP + TN + FP + FN} = \frac{990 + 30}{990 + 30 + 20 + 10} = \frac{34}{35} \approx 0.9714. \tag{3}$$

F1-score is given by

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{60}{60 + 10 + 20} = \frac{2}{3} \approx 0.6667. \tag{4}$$

Given the confusion matrix

	Predicted	
Actual	**Negative**	**Positive**
Negative	**9000**	**50**
Positive	**100**	**850**

the values are:

- True Negative (TN): 9000.
- False Positive (FP): 50.
- False Negative (FN): 100.
- True Positive (TP): 850.

Precision for this is given by

$$P = \frac{TP}{TP + FP} = \frac{850}{850 + 50} = \frac{17}{18} \approx 0.9444. \tag{5}$$

Recall for this is given by

$$R = \frac{TP}{TP + FN} = \frac{850}{850 + 100} = \frac{17}{19} \approx 0.8947. \tag{6}$$

Accuracy is given by

$$A = \frac{TP + TN}{TP + TN + FP + FN} = \frac{9000 + 850}{9000 + 850 + 50 + 100} = \frac{34}{35} \approx 0.9714. \tag{7}$$

F1-score is given by

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{1700}{1700 + 100 + 50} = \frac{34}{37} \approx 0.919. \tag{8}$$