Programming Task

While I devoted significant effort to understanding the paper, the time commitment unfortunately left me unable to begin work on the second task. Although I made headway on the first task, I ultimately couldn't complete the code.

Approach for Task 1:

The objective of the task is to answer the following questions:

- Is a particular NLP model biased?
- Is a model stereotypical towards a specific social axis(religion, region, etc)?
- Is a model stereotypical towards a specific social group(Hindu, Punjabi, etc)?
- How would you quantify bias through this method?

I have considered the region_annotations.tsv file as my input file to train the BERT model. In the input file, the columns stereotypical, non stereotypical , not sure and total represent bias present in the society(widely held opinion).

BIAS in the model (irrespective of societal stereotypes)

Firstly, we consider only the first three columns of the dataset that is, the column token id, identity_term and token.
In the new dataset, if there is a token that is common to all the identity_terms then we remove it, because it does not contribute to any bias for the model.
Now we use Masked prompting on the final dataset using the handcrafted templates to analyse how the structure of the prompt is influencing the answer.
For example: Bihari mostly work as [MASK].

If we use the same approach to find the bias for religion_annotations.tsv, we could find which social axis region or religion in this case is influencing the masked token output given by the model.
For example: Hindus are [MASK].

We can’t find the intersectional bias because we do not have the data relating the region and religion in this case.

To find if the model is biased towards a certain social group, we can use the reverse/converse of the prompts used above .
For example: Architects are mostly [MASK].
This will give the bias towards a particular social group.

Srivarshitha Medarametla