Advanced NLP Exercise 2

Submission: Until June 2 2025, 23:59 PM

You should submit all the answers in a single pdf file through Moodle.

1 Open Questions (40 points)

1. We discussed gender bias as one prominent societal bias. Look in the ACL anthology[1] for one additional source of societal bias (e.g., race or religion), and describe at high-level either a dataset (e.g., how many samples? from what domain?) or a debiasing technique (e.g., did it debias the embeddings? balance the data? etc.).

2. In class, we discussed four categories of efficiency techniques: training, inference, modeling, and data methods. Select **three inference methods**, each representing a different category.

   - For each method, provide the following:
     (a) A brief description of the method.
     (b) The type of resource(s) it aims to save (e.g., memory, runtime, compute).
     (c) Whether the resource savings come at the expense of other resources. For instance, in distillation, while we save memory and compute during inference, additional training time and compute are required to train the student model.
   - For every pair among the three methods you've chosen, discuss whether the methods can be combined or if they are mutually exclusive.

[1] https://aclanthology.org

2 Practical Exercise (60 points)

In this exercise you will try to identify the causes of failure of models, characterize the failure and look for the reasons for such failures.

Choose a model First, choose a model from the **Models to Use** section below. For all models, the path appended to the name is the path in the HuggingFace hub (some have an online API; if you choose a model without an API, simply download it using the Transformers package).

Break the model Find a specific type of input on which the model tends to be wrong. You can take inspiration from cases we have seen in the lecture (gender bias, tokens that are correlated with contradiction in NLI, etc.). The input should use a zero-shot style prompts. Give **at least five** examples of such inputs in the format of Table 1. Describe in a few words the type of input the model struggles with, as exemplified by the input you provided.

Index	Input	Model output
1	I have never had so much fun in a movie: I played Angry Birds from the 10th minute onwards	Positive (very confident)
2	The movie was great, the acting was phenomenal, and the storyline was fascinating. That is, for the first two minutes; After that I fell asleep.	Positive (very confident)
3	This movie was as fun as getting punched in your nose	Positive (very confident)
4	The vacation was horrible: hanging in the pool all day, eating great food, and sleeping a lot. A total disaster	Negative (very confident)
...	...	...

Table 1: Example of a table showing examples where a sentiment analysis model predicts a wrong sentiment, for sentences with sarcasm.

Why is the model wrong? Make a hypothesis (or hypotheses) as to why the model is wrong for this type of input. Test the hypothesis by giving the model inputs that are similar to the problematic input type but have some variation so that you can map the extent of the problem.

For example: If you claim that a sentiment analysis model predicts a wrong sentiment for sarcastic sentences (e.g., "That was the best meal I have ever eaten. At least since breakfast"), you might suggest that this is because the model is fixating on the words with strong sentiment at the beginning of the sentence ("the best meal I have ever eaten"). To check this, you could try other sentences with positive words but negative sentiment (or vice versa), such as sentences with negation words ("The movie was not great and the acting was not phenomenal") or sentences comparing the subject to another experience ("I always enjoy Chuck Norris movies. This time was different"). Document the inputs and outputs for the model in a table similar to Table 1.

Try another model Try entering the inputs you tried with the current model into another model for the same task from the model list. Does this model break on these examples as well? Document the inputs and outputs for the model in a table similar to Table 1.

Propose a solution Propose a solution to the problem. The solution should not be "annotate data of this type and train the model". It can be

- An automated method of collecting self-supervised data that will mitigate this problem.
- Changes in the architecture of the model.
- Any other thing you can think of.

Continuing the example of sarcasm in sentiment analysis models from before (e.g., "That was the best meal I have ever eaten. At least since breakfast"), you might note that this type of sentences are composed of two parts: the positive part ("That was the best meal I have ever eaten") and the part that makes us flip the

sentiment ("At least since breakfast"). An architectural change could be to first build a constituency tree of the sentence, apply the model to each constituent, and look for cases where there are two main constituents with a different sentiment.

Models to use

- meta-llama/Meta-Llama-3-8B-Instruct
- Qwen/Qwen2.5-7B-Instruct
- deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
- mistralai/Mistral-7B-Instruct-v0.3