Machine Learning Engineer NLP Task

Part 1:

You will find attached a small subset of a news dataset. This dataset is annotated with Named Entity Recognition (NER) tags. The tags can be one of three categories: Person, Organization, and Location.

Your task is to develop a python code that does the following:

- Read the dataset and clean it from all the HTML tags to produce raw unannotated text
- Extract NER tags in the text using two approaches of your choice. One has to be statistical, and the other transformer based
- Evaluate the produced tags from the two approaches and compare them with the ground truth

Deliverables:

- A Github repository containing your solution
- A ReadMe detailing the steps you took to solve the problem, your approach, your data handling techniques, any problems you faced, and the results of your tests.

Notes:

- Use proper data handling and data loading
- Explore multiple solutions and showcase why you chose the final one
- Be clear on your choice of metrics and why you chose them
- Make sure to write clean, modular, and well documented code.
- Make sure to keep your approach optimum as much as possible for deployment, minimal model size and number of computations.

You are not expected to reach optimal results, but it is very critical for you to showcase how you approach the problem and to outline your thought process.

Part 2:

Since their introduction in the famous “Attention is All You Need” paper, transformers have been taking the world by a storm. Given that, a lot of papers are published yearly on the topic making it hard to keep up to date and discern what’s a good paper to consider.

In that regard, you are tasked with evaluating the following paper “Multimodal Few Shot Learning with Frozen Language Models” available here. You should deliver, from your perspective, an analysis of the paper detailing your feedback: What do you like about it? What do you dislike? What do you think are potential areas of improvement and what are the main contributions?

You are an ML Engineer at a company, and you are tasked with building a classifier for clothing articles.

Deliverables:

- A Github repository containing your solution
- A ReadMe detailing the steps you took to solve the problem, your approach, your data handling techniques, any problems you faced, and the results of your tests.

- Report the overall receptive field of your model and discuss how it can be increased or decreased with examples.
- Report the estimated/calculated number of FLOPS and MACCs per layer (mainly convolutional and fully connected layers, ignore the rest), and discuss how can we decrease it with examples, also highlight the most computationally expensive layers.

Best of luck!