NLP Task

OCEAN scores analysis: Suppose you're given the task to build a tool that takes a user's tweet data (all the tweets they've ever made), and predicts their OCEAN scores to an 80% accuracy. Let's assume we can get OCEAN scores against 100k twitter users. This is a lofty goal to achieve even with this size of data, and we must plan and tackle the problem smartly.

Question 1:

The first logical step is to find research papers that have done something similar to set a benchmark. How will you find papers against this problem? List the most promising papers you find.

Answer 1:

This problem belongs to multi-label classification, which generalizes the traditional (single label) classification setting by allowing multiple labels to simultaneously belong to an instance but here we're going to predict the probability score of each label to predict the OCEAN score percentages for each label(Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) therefore I will start looking for papers related to multi-label classification for NLP type problems.

Following are the most promising papers I found
1)https://arxiv.org/pdf/2106.03103.pdf
2)https://arxiv.org/pdf/2105.05614.pdf
3)https://link.springer.com/chapter/10.1007/978-3-030-73197-7_1
4)https://www.aclweb.org/anthology/D19-1044.pdf
5)https://www.archives-ouvertes.fr/hal-01179430/document

Question 2:

How will you plan the experimentation towards building a robust model? What techniques can you employ based on your initial research?

Answer 2:

If the dataset is provided as specified in the task manual then the solution is straightforward either BERT (language model) or RNN with LSTM architecture can be used for this type of problem as far as the experimentation towards building a robust model is concerned, learning curves might help moreover ensembling methods can be used to get a more robust model.

Question 3:

How will you document your efforts and report progress?

Answer3:

List all the papers used, discuss every step of the ML pipeline, and finally use a classification report to shed light on the progress