Q) Write a Jupyter Notebook to conduct a small task with a transformer and explain what you are trying to solve.

I tried to solve a binary task of detecting whether a text is sexist or not.
Link to the notebook:
https://github.com/debashish05/Explainable_Detection_of_Online_Sexism/blob/main/Experiment%206%20DPT/multi-mlp-concat-base-dpta.ipynb
This problem was a part of SemEval task, and I have presented a systems paper in ACL SemEval Workshop which is currently at review.

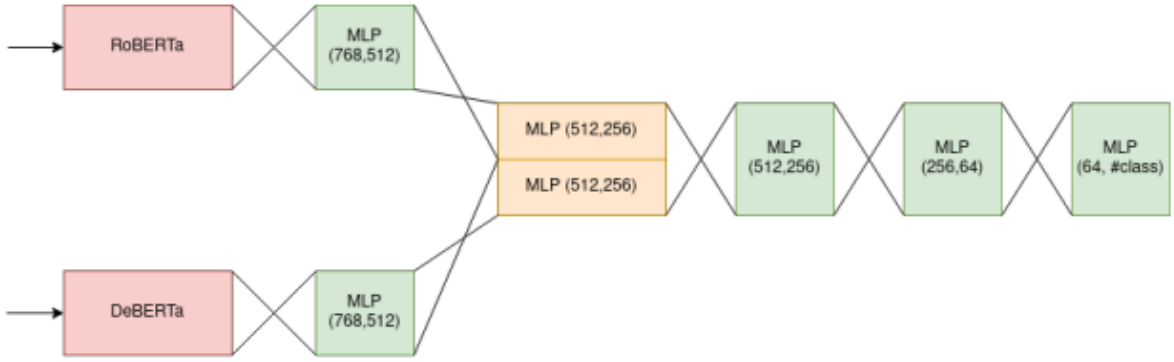
The Steps are as follows:

- **Data Processing:**
  - The dataset used is https://github.com/debashish05/Explainable_Detection_of_Online_Sexism/blob/main/Experiment%206%20DPT/edos_labelled_aggregated.csv which is provided by Semeval 2023 Task 10: Explainable detection of Online Sexism.
  - The data contains the following fields rewire_id, text, label_sexist, label_category, label_vector, split.
  - Rewire id is unique to the text, we need to detect if the text is sexist or not. Here in this task we have used only the label_sexist feature only indicating the text is sexist or not. Split specifies whether it is a part of train, validation or test split.
  - Data is imbalanced here. Oversampling should have been tried but for simplicity we left the split as it is.
  - To determine the number of sequence lengths used in the transformer we plotted the length of all text and took 64 as the max length. If the text is less than this we will pad or else if more than truncate.

- **Data Cleaning:**
  - We tried some data cleaning like emojis to text, converting text to lower converting link to tags, removing punctuations and removing words with numbers.
  - Do lemmatization.
  - Then later we split on the base of spaces.
  - But by experimenting this doesn't give best results. The reason for this is we removed some data that may be useful. And the most important thing is when we pass a text to the tokenizer in the transformer, it does subword tokenization, so even if some word is not in vocabulary, it will try to get approximate embeddings for that.

- **Loading Model:**
  - Here we have merged the last layer of RoBERTa and DeBERTa transformers followed by the MLP layers.

- We have taken the pretrained version of these two and trained masked language modeling on text from online platforms, so that it will adapt to the domain of the problem. The models are hosted in https://huggingface.co/debashish-roy



Architecture of the system

- **Tokenization and Converting the data to appropriate form**
  - Use the tokenizer for the respective to tokenize the text. Get the input ids and attention mask
  - Converted the data into tensors.
  - Later made a data loader for the data.

- **Model and Hyperparameter:**
  - Define the model as per the architecture.
  - Used cross entropy loss.
  - Used AdamW optimizer with learning rate of 1e-5
  - Run for 20 epochs.
  - Based on 20 epochs, choose the model which gives the best validation dataset.

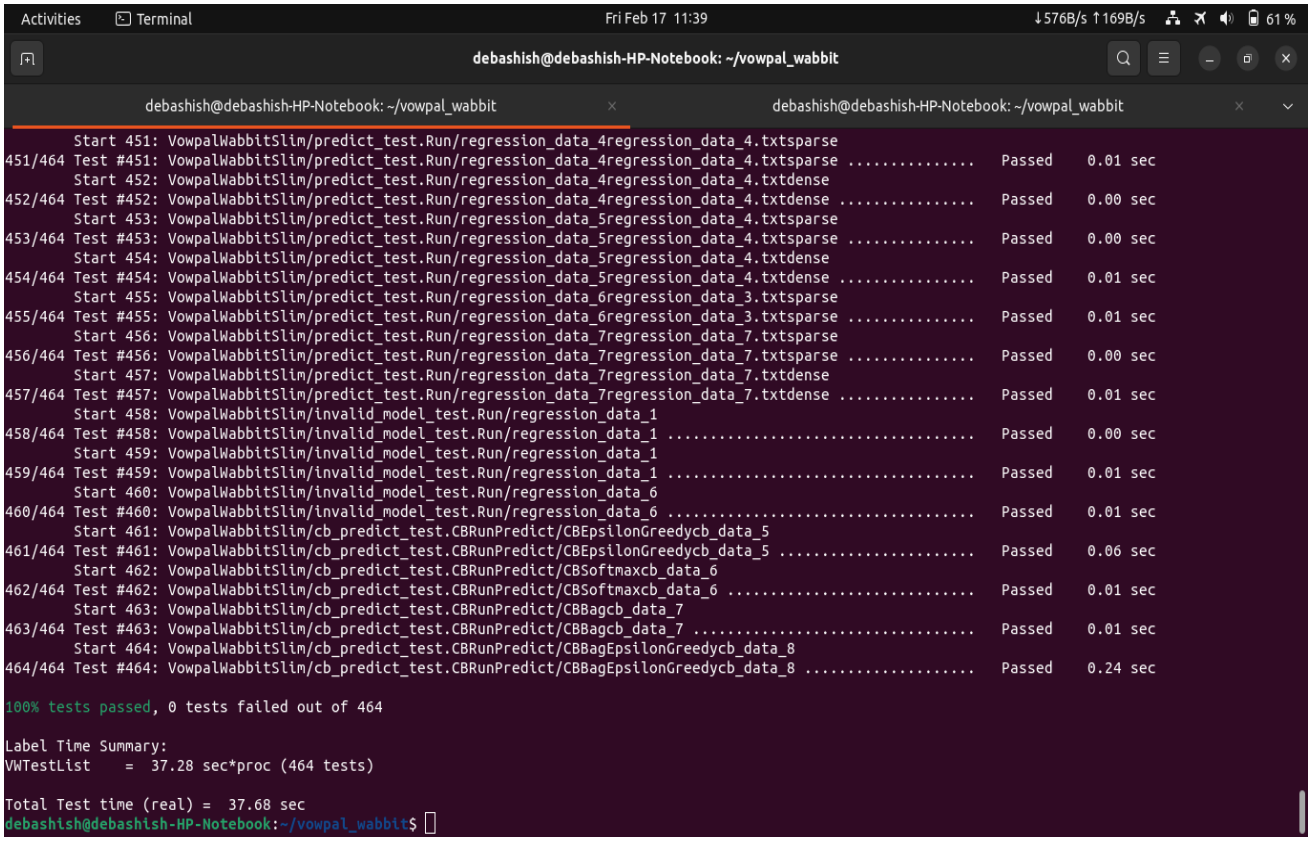
- **Inference**
  - Based on the input text, we will tokenize the text and use the freezed model to the output.
  - After finding the optimal parameter, we have used all the data and deployed the final model.

- **Multiple Experiments for the same problem**
  - Multiple experiments are performed and are present in https://github.com/debashish05/Explainable_Detection_of_Online_Sexism
  - Results for all the experiment are as follows:

S.No	Experiment	Task A Macro F1	
		Val	Test
1	RoBERTa last layer + MLP	82.90	81.47
2	DeBERTa last layer + MLP	83.24	81.99
3	RoBERTa avg of all layer + MLP	79.85	78.65
4	DeBERTa avg of all layer + MLP	80.40	78.62
5	RoBERTa+DeBERTa+ (Embeddings of these two are concatenated, before concatenating these are passed through MLP) + MLP	82.96	82.26
6	Experiment 5 + Domain Adaptive Pre Training with unlabelled text	**84.27**	82.66
7	Joint Learning for task B using task (A and B's data), last layer is of 5 neurons with labels from task B and one non-sexist text	NA	NA
8	RoBERTa+DeBERTa+ (Embedding of RoBERTa and DeBERTa are concatenated and passed through MLP) + Domain Adaptive Pretraining	84.13	**83.9**

Q) Build Vowpal Wabbit from the source in Linux.

Results of Unit Tests



Results of Integration Tests