Baseline Models & Results Generation:

Sub-Task 1: Span Identification

1. Create a folder called NLP in google drive
2. Upload the folder project_5_data (provided as part of the project or included in the code folder) in the NLP folder. (the sub-folders ‘datasets’ and ‘propaganda-techniques-scorer’ are required.)
3. The required folder path in drive would look like /NLP/project_5_data/datasets, /NLP/project_5_data/propaganda-techniques-scorer
4. Navigate to codebase/baselines/notebooks
5. Run the ***NLP_Task_SI_baseline.ipynb*** notebook in Google Colab.
6. The score generation script is also included as part of the code.
7. To run the scoring script manually, use the following command

python3 propaganda-techniques-scorer/task-SI_scorer.py -s ***datasets/roberta_base_SI_output.txt*** -r datasets/dev-labels-task-si/

8. The above params are provided assuming the terminal is in the code folder.
9. Only the ***roberta_base_SI_output.txt*** is generated file. This is generated and stored in google drive as part of the notebook run. Rest of the files are provided as part of the project set-up/start-up code.

Sub-Task 2: Technique Classification

- Running the model and generating the output:
  1. The code requires Glove embeddings (~1GB) which is not included in the codebase folder. Download the Glove embeddings and place the **glove.6B.300d.txt** file in the **/codebase/baselines** folder
  2. Go to the codebase/baselines folder in the submitted code folder
  3. Run the python file ***baseline-task-TC_LogisticRegression.py***
  4. This will generate an output text file with the name ***baseline-output-TC_dev.txt***

- Generating the scores:
  1. Copy the above generated file into the codebase/propaganda-techniques-scorer/data folder
  2. Navigate to codebase/propaganda-techniques-scorer and run the ***task-TC_scorer.py*** with the following parameters

python3 codebase/propaganda-techniques-scorer/task-TC_scorer.py -s ***data/baseline-output-TC_dev.txt*** -r ***data/dev-task-flc-tc.labels*** -p ***data/propaganda-techniques-names-semeval2020task11.txt***

3. The above params are given assuming the terminal is in codebase/propaganda-techniques-scorer folder. If not, please update the paths accordingly to the local system paths.
4. Only the ***baseline-output-TC_dev.txt*** is generated file. Rest of the files are provided as part of the project set-up/start-up code.

Final System Models & Results Generation:

Sentence Classification:

1. Create a folder called NLP in google drive.
2. Upload the folder project_5_data (provided as part of the project or included in the code folder) in the NLP folder. (the sub-folders 'datasets' and 'propaganda-techniques-scorer' are required.)
3. ***Dataset Generation:*** Go to the folder codebase/final_system. Run the notebook ***NLP_Sent_Classifier_Data_creation.ipynb*** in Google Colab. This will generate 2 files - train_sentence_classification.txt and dev_sentence_classification.txt in the following folder: /content/drive/MyDrive/NLP/
4. Training and Scoring the Model: Run the notebook ***NLP_Sent_Classifier_final.ipynb*** in Google Colab. This will need the two files generated in previous step.

Sub-Task 1: Span Identification:

1. Create a folder called NLP in google drive
2. Upload the folder project_5_data (provided as part of the project or included in the code folder) in the NLP folder. (the sub-folders 'datasets' and 'propaganda-techniques-scorer' are required.)
3. Go to the folder codebase/final_system and upload the notebooks ***NLP_Task_SI_final_berta.ipynb*** and ***NLP_Task_SI_final_roberta.ipynb*** to Google Colab.
4. Run the above notebooks to generate the models and results.

Sub-Task 2: Technique Classification

1. Create a folder called NLP in google drive
2. Upload the folder project_5_data (provided as part of the project or included in the code folder) in the NLP folder. (the sub-folders ‘datasets’ and ‘propaganda-techniques-scorer’ are required.)
3. Go to the folder codebase/final_system and upload the notebooks ***NLP_Task_TC_final_bert.ipynb*** and ***NLP_Task_TC_final_roberta.ipynb*** to Google Colab.
4. Run the above notebooks to generate the output files as done in the baselines.
5. Generating the scores:
    1. Copy the above generated files into the local system ***codebase/propaganda-techniques-scorer/data*** folder
    2. Navigate to *codebase/propaganda-techniques-scorer* and run the ***task-TC_scorer.py*** with the following parameters

python3 codebase/propaganda-techniques-scorer/task-TC_scorer.py -s ***data/file_name_from_output.txt*** -r ***data/dev-task-flc-tc.labels*** -p ***data/propaganda-techniques-names-semeval2020task11.txt***

    3. The above params are given assuming the terminal is in codebase/propaganda-techniques-scorer folder. If not, please update the paths accordingly to the local system paths.
    4. Replace the ***file_name_from_output.txt*** with the name of the files that were generated in the notebook.

Deploying the Propaganda Detection Application:

1. Go to the codebase/Flask-App folder.
2. Each of the above task viz, Sentence Classification, Span Identification, Technique Classification would have generated and saved a torch model in the drive. Download the models and place them in the ***codebase/Flask-App/models*** folder.
3. Ensure that the model names are as follows ***RoBERTa_Task_SentClassify.pt (for Sentence Classification), RoBERTa_Task_SI.pt (Span Identification), RoBERTa_Task_TC.pt (Technique Classification).***