Milestone II – (Max 40 points)

A. How to deliver your code (10 points):
1. Divide the project into three files.
    a. Training.py: This must include all code related to loading training data, pre-processing and training model. This also includes the code related to validation data, and model evaluation during training.
    b. Model file: The final trained model must be saved as a pickle file. Please find instructions on how to save a trained model in Pytorch and how to save it in an sklearn .
    c. Test.py: This file must include all testing code for performing prediction/inference on test set.
2. If your preprocessing technique involves using other pre-trained models, you should include your data features as a train_features.npy file and test_features.npy. The extracted features (for both train and test set) must be dumped as *.npy files. Do not submit the pre-trained model. Please reach out to us, if you are facing issues in this part.
3. If your data file is larger than 5 MB, please do not submit your data. Instead, include the links and the sources where the data can be downloaded from. Your training code should include functions to read train_features.npy file. It must also include the code to convert the original data to train/test feature numpy arrays.
4. This milestone has 2 deliverables
    a. the code + data + model zip file (as indicated above),
    b. a report in a **pdf** format containing all the above details in Times New Roman, Font Size 12, Maximum 2 pages.
5. Any files with greater than 2 pages or with different font sizes, WILL NOT be evaluated.
6. **IMPORTANT**: Please merge these 2 pages with the previous 2 pages of your milestone 1. So you will submit 4 pages of the report in this milestone (max 2 pages from the previous milestone + max 2 pages from this milestone).
7. Please name the files as <member1_id>_<member2_id>_< member3_id>.pdf and <member1_id>_<member2_id>_< member3_id>.zip.

Submit all the above files as a zipped folder on avenue.

B. How to organize your report (30 points):

1. **Team** members roles and responsibilities (**1 point**): This should be same as Milestone I.

2. **Context (1 point):** This should be same as submitted in Milestone I with only slight modifications permitted. It must answer the following questions.
    a. What is the problem you have identified?

      b. Why is it a challenging problem?

      c. How can it be solved through machine learning?

      d. What aspects of the problem are you going to solve?

      e. Why is it relevant to the world? How does it help?

3. **Dataset (2.5 points):** This should include all the details pertaining to the dataset. If you included these details in milestone I, then this can be same as Milestone I. If your Milestone I did not have these details, please add them in Milestone II.
    a. Provide details like
        i. size,
        ii. number of samples,
        iii. number of features,
        iv. where is the data coming from and,
        v. the link to source from where you have downloaded the data.
    b. Slight modifications to dataset are permitted at this stage, like change in size, features, and number of samples.

4. **Preprocessing (8 points):** Here you need to provide more specific details about preprocessing than milestone I.

    a. Include the details of all techniques used for pre-processing the data.
    b. Feel free to use diagrams and tables to provide clear specifics in short space.
    c. This must include details like:
        i. sampling ratio (test-train split),
        ii. feature selection strategies applied,
        iii. removing any correlated features,
        iv. any data augmentation techniques used,
        v. dimensionality reduction, or one-hot encoding.
        vi. any other techniques used.
    d. If you are using any embedding models (for text), please include the embedding model specifications and link.
    e. If you are using any pretrained models for feature extraction for images, please include the model specifications of the image model and link.
    f. No need to submit the pretrained models used for feature extraction (only extracted features, see A2).

5. **Model Specifications (10 points):** Include all the details of your **proposed** model.
    a. Include details like
        i. hyper-parameters selected,
        ii. model configuration, (layers, computational units in layers, etc.)
        iii. type (SVM, neural network),
        iv. family (supervised, unsupervised).
    b. If you are using more than one model, you may use a table to include above details to save space.

     c. Include all training parameters (e.g. epochs, batch size, learning rate, etc.)
     d. Also include any regularization techniques applied, like early stopping, dropout. Briefly explain the details of the technique. For instance, if you used early stopping, include the number of epochs after which you calculate the validation loss, and the stopping threshold value.
     e. Include the training loss, or any other model progress tracking strategy you incorporate. You may use charts to show these details. For instance, if you are computing training loss after every 10 epochs, please mention this.

6. **Evaluation (2.5 points)**: In this section, share your evaluation strategy. This section can be the same as Milestone I. However, if you are making any changes to this strategy, please include them.
     a. This includes evaluation strategy during training time- cross validation/ training loss, test-train split)
     b. This also include evaluation during test time (precision, accuracy, recall, benchmarking, etc.)
     c. Include specifics of the evaluation. Feel free to use bullet points.

7. **Preliminary Results (2.5 points):** Share any preliminary results like
     a. accuracy, precision, recall.
     b. Validation Loss / Training loss charts
     c. Test set performance on benchmarks.
     d. This part does not have to be very exhaustive for this milestone.
     e. No need to perform a very comprehensive evaluation at this point. Some basic numbers are fine.

8. **Limitations (2.5 points)**: At this point, you may find out that your approach may or may not work. Please keep a note of what works and what does not work. This is a short section for you to reflect on the approach based on the preliminary results you are reporting. Maximum 3-4 lines is acceptable. No detailed reflection necessary.