Criteria for evaluating your project:

1. Quality of your newly added data
2. Quality of the code that you wrote to complete the project
3. Quality of the report that you will submit as follows. Quality of the report will be judged based on the following metrics
   a. Writing quality (no grammatical error, no typo, no ambiguous sentences)
   b. Clarity of your writeup in the report
   c. Reproducibility of your claims between your code and your paper
   d. Quality of your produced solution
   e. Any improvements (e.g., model performance) that you could make over the original paper

Report Template:

Follow the ACM word template to write the report:

https://www.acm.org/binaries/content/assets/publications/taps/acm_submission_template.docx

Report Writeup Guidelines

The report can be composed of any number of pages.

Page 1 should have declaration of team and individual contribution as follows

1. Title of the project
2. Team Member
3. Summary of Contributions: It is expected that each member of a team contributes equally. Explain the contribution of each team member in data collection + coding + writeup of the report as follows:
   a. For data collection, explicitly mention which part of the data was collected and/or labeled by who.
   b. For coding, explicitly link blocks in Databricks notebook of the coding of each individual member.
   c. For writeup, explicitly note each section of the report (what was written by who).
4. Link to Databricks notebook where your project was coded

Starting from Page 2, you will write the report as follows

The report consists of five sections as follows.

1. **Abstract.**
   Summarize the overall report along the following dimensions: (1) context (2) objective (3) method (4) results (5) conclusions. Please see https://www.sciencedirect.com/science/article/abs/pii/S0950584919301387 as an example.

2. **Introduction.**
    a. Motivate the problem
    b. Provide background
    c. You will answer the following questions in this draft:
        i. How was the new data labeled/collected?
        ii. How does the newly added data compare with the original data?
        iii. How was the data preprocessed?
        iv. How do the models perform on the original data vs the new + original data?
        v. How does the performance of the models change based on the choice of hyper parameters?
        vi. How are the misclassifications of the best performing model distributed?
3. **Results**

    Each question has a subsection. Each answer to a question has two sub-subsections: approach and results. Please follow the guidelines clearly while writing. It is expected that each team member has equally contributed to produce the results and to write the results. A straightforward way to show your contributions would be to divide the questions equally among the team members.

    a. How was the new data labeled/collected? (Q1)
        i. Approach. There are two ways depending on how the additional data was collected. Option 1. You manually labeled the new data. Option 2. You used some pre-defined approach to collect the new data (e.g. Dr. AI project).
            1. Approach for Option 1. Explain how additional data is collected (e.g., using GitHub API).
            2. Approach for Option 2. Explain how additional data is collected
        ii. Results.
            1. Results for Option 1. Explain the agreement analysis with examples. Explain how disagreements are resolved with each other
            2. Results for Option 2. Explain how you ensured the quality of the new data. Provide statistics backing up your claim.

    b. How does the newly added data compare with the original data? (Q2)
        i. Approach. Use a set of metrics to compare the new data with the original data. The choice of metrics can differ depending on the type of project.
        ii. Results. (use tables with results and then explain the tables to highlight the key findings)
            1. Introduce the original data
            2. Provide summary statistics of the original data (use tables and then explain it in paper. See the original paper that y)
            3. Provide summary statistics of the new data
            4. Show any difference between the original and new data and explain why it happened.

    c. How was the data preprocessed? (Q3)
        i. Approach. Explain how the data was preprocessed

   ii. Results. (use tables with results and then explain the tables to highlight the key findings) Show summary statistics of the data after the preprocessing.

d. ***How do the models perform on the original data vs the new + original data? (Q4)***

   i. Approach. Use the following performance metrics: Precision, Recall, F1-score. You are allowed to use other performance metrics. Precision = TP / (TP + FP), Recall = TP / (TP + FN), F1-score = 2*Precision*Recall/(Precision + Recall)

   ii. Results. (use tables with results and then explain the tables to highlight the key findings) Compare the model performance on the new, new+old, old data. Please show TP, FP, TN, FN for each setting for each metric.

e. ***How does the performance of the models change based on the choice of hyper parameters? (Q5)***

   i. Approach. Explain the different parameters in the model that are available.

   ii. Results. (use tables with results and then explain the tables to highlight the key findings) Show how the choice of the different parameter value can improve/degrade the model performance

f. ***How are the misclassifications of the best performing model distributed? (Q6)***

   i. Approach. Pick the best performing model on new+old dataset from Q5. Analyze the cases where the best performing model was wrong.

   ii. Results. (use tables with results and then explain the tables to highlight the key findings) Report the misclassification as follows.

      1. Randomly pick 200 misclassified records.
      2. Manually label the reason of misclassification
      3. Explain each reason with example

4. **Discussions**

   Discuss the implications of your developed models, e.g., think of real-world scenarios where your models could be useful. Each team member should come up with at least one scenario. Clearly mention who came up with which.

5. **Conclusions**

   Conclude and summarize the report

   References

   Reference of existing papers. See https://giasuddin.files.wordpress.com/2020/11/opinerusagedoctosem2020-2.pdf how references are provided and how the references are used in the paper.