IIT Kharagpur Data Science Hackathon 2025 Round 2 submission

Team Name :- AI Devilops

Report by - Deepak Kaura (Team Leader)

Problem Statement -

The manual evaluation of research papers for conference submission is laborintensive, time-consuming, and demands significant expertise. This hackathon challenges participants to develop an AI-driven system using the Pathway Framework to streamline the process of conference selection and research paper evaluation. The system will harness advanced language models, comparative analysis techniques, and streaming data frameworks to automate and optimize these tasks.

Participants will have access to a dataset of high-quality, benchmark research papers from conferences or conferences. The objective is to evaluate new submissions, compare them with these benchmark papers, and recommend the most suitable conferences or conference with formal justification.

Task-1: Research Paper Publishability Assessment

In the academic and research domains, the quality, and publishability of research papers play a critical role in advancing knowledge and fostering innovation. However, the process of determining whether a paper meets the standards for publication can be both time-consuming and subjective, often requiring expert review. With the increasing volume of research outputs, there is a growing need for automated systems to assist in evaluating the quality and suitability of papers for publication. This challenge not only offers an opportunity to innovate but also holds the potential to streamline the publication process and enhance its objectivity.

The task involves developing a framework that can classify research papers as either "Publishable" or "Non-Publishable" based on the evaluation of their content. The goal is to create a robust system capable of identifying critical issues, such as inappropriate methodologies, incoherent arguments, or unsubstantiated claims, that affect the suitability of a paper for publication. For instance, a research paper that applies methodologies or techniques that are not well-suited to the problem being addressed, without adequate justification or adaptation to the context, would be considered unsuitable. Similarly, a paper that presents arguments that are unclear, disorganized, or lack logical coherence, or one that claims results that appear unusually high or unrealistic without sufficient evidence or proper validation, would also fall into the "Non-Publishable" category

A dataset of 150 research papers is provided for classification, with 15 labeled papers available for reference to guide the development of the framework. The framework should be designed to accurately classify papers into the appropriate category, ensuring that it can handle a wide range of research topics and maintain consistency across different types of content. The proposed framework must be capable of systematically analyzing these and other aspects of research papers to ensure a reliable and objective evaluation process. The solution should demonstrate high accuracy in detecting such issues, ensuring its applicability across a range of research domains and scalability for future use with larger datasets.

Classifying and Organizing Extracted PDF Files with Subcategories -

```
     Index  pdf_name           Labels  Subcategory
0        1   P001.pdf        Published         CVPR
1        2   P002.pdf    Not Published         None
2        3   P003.pdf        Published        EMNLP
3        4   P004.pdf    Not Published         None
4        5   P005.pdf        Published          KDD
..     ...        ...              ...          ...
130    131   P131.pdf        Published         CVPR
131    132   P132.pdf    Not Published         None
132    133   P133.pdf        Published        EMNLP
133    134   P134.pdf    Not Published         None
134    135   P135.pdf        Published          KDD

[135 rows x 4 columns]
```

Data Visualization
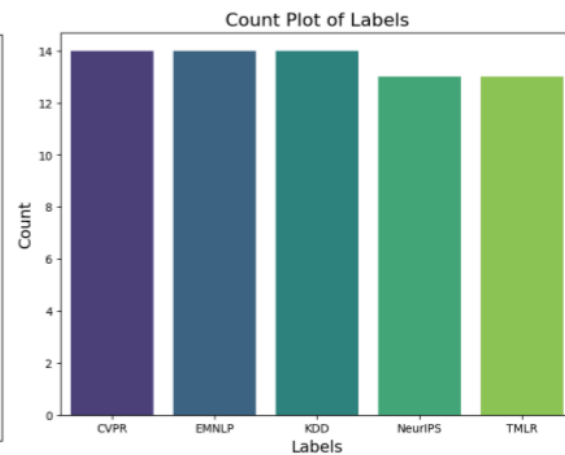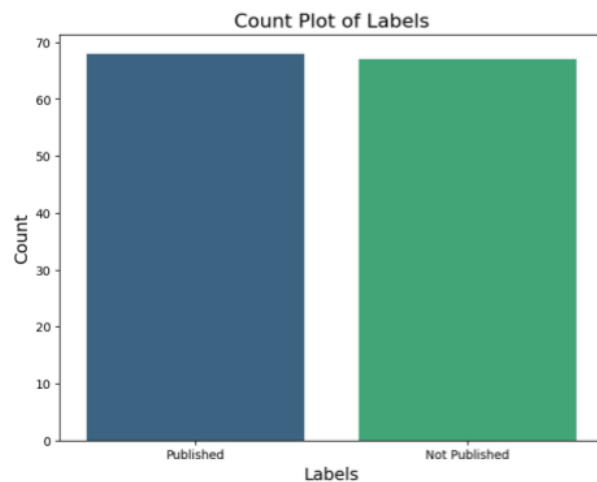
Data Visualization (continue..)

Count Plot of Labels with Subcategories
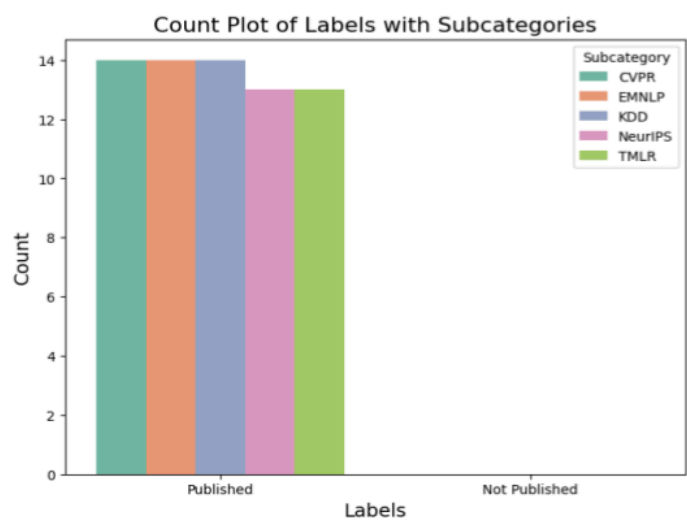
Encoding Combined Labels and Subcategories for PDF Classification (Glimpse) :-

```
:]  Class Mapping: {'Published_CVPR': 0, 'Published_EMNLP': 1, 'Published_KDD': 2, 'Published_NeurIPS': 3, 'Published_TMLR': 4, nan: 5}
        pdf_name        Labels Subcategory Combined_Label_Subcategory  \
    0    P001.pdf     Published        CVPR              Published_CVPR
    1    P002.pdf  Not Published        None                        NaN
    2    P003.pdf     Published       EMNLP             Published_EMNLP
    3    P004.pdf  Not Published        None                        NaN
    4    P005.pdf     Published         KDD               Published_KDD
    ..        ...           ...         ...                        ...
    130  P131.pdf     Published        CVPR              Published_CVPR
    131  P132.pdf  Not Published        None                        NaN
    132  P133.pdf     Published       EMNLP             Published_EMNLP
    133  P134.pdf  Not Published        None                        NaN
    134  P135.pdf     Published         KDD               Published_KDD

         Labels_encoded
    0                 0
    1                 5
    2                 1
    3                 5
    4                 2
    ..              ...
    130               0
    131               5
    132               1
```
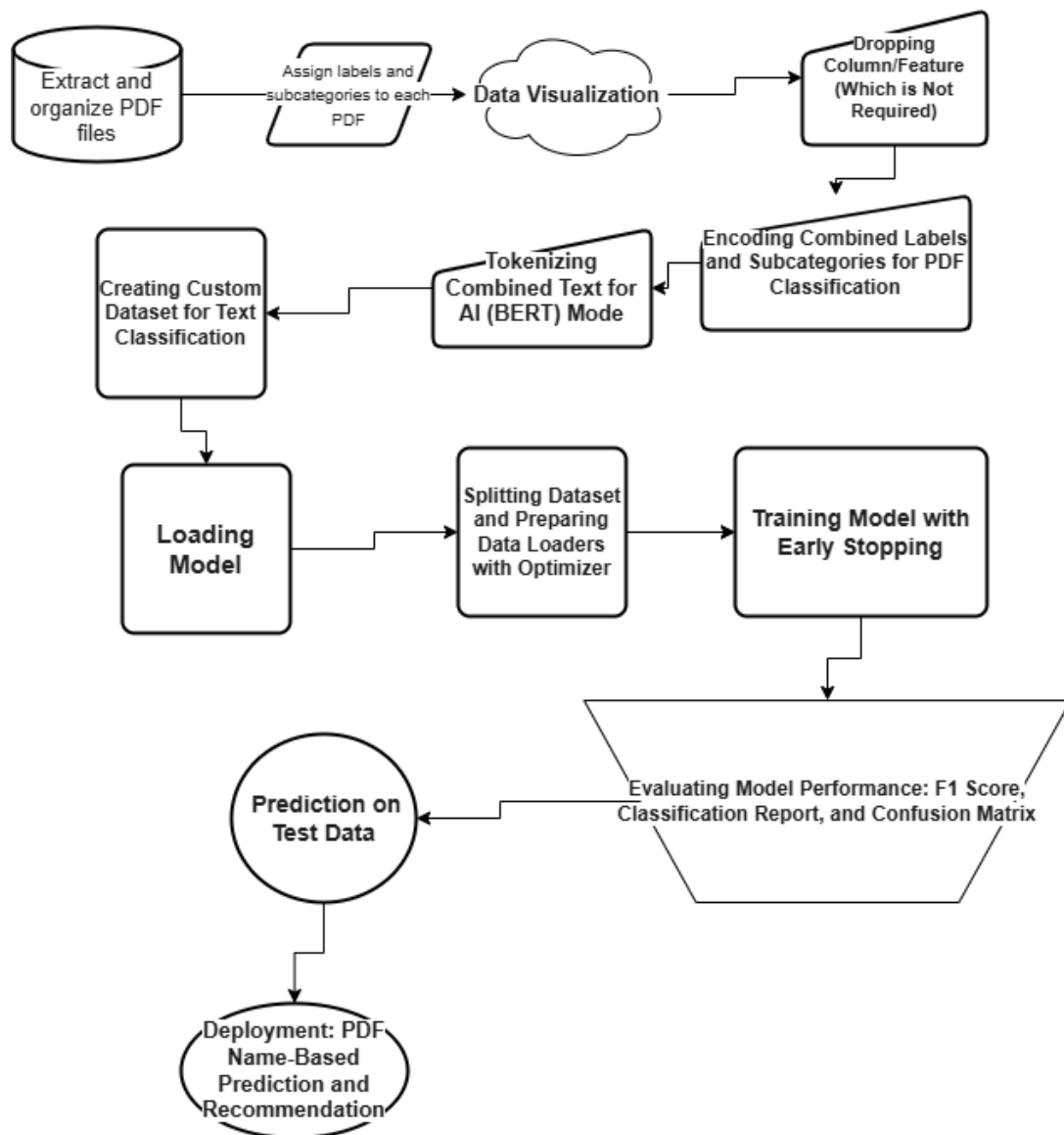
✓ 0s   completed at 8:56 PM

Model Work Flowchart -

Extract and organize PDF files → Assign labels and subcategories to each PDF → Data Visualization → Dropping Column/Feature (Which is Not Required)

Encoding Combined Labels and Subcategories for PDF Classification → Tokenizing Combined Text for AI (BERT) Mode → Creating Custom Dataset for Text Classification → Loading Model → Splitting Dataset and Preparing Data Loaders with Optimizer → Training Model with Early Stopping

Evaluating Model Performance: F1 Score, Classification Report, and Confusion Matrix → Prediction on Test Data → Deployment: PDF Name-Based Prediction and Recommendation

Model Training (Glimpse) -

```
item = {key: torch.tensor(val[idx]) for key, val in self.encodings.items()}
] Epoch 0: 100%|████████| 7/7 [00:31<00:00,  4.50s/it, loss=1]
  Epoch 0: Train Loss = 1.3784, Validation Loss = 1.2556
  Validation loss improved. Saving model.
  Epoch 1: 100%|████████| 7/7 [00:22<00:00,  3.24s/it, loss=0.918]
  Epoch 1: Train Loss = 0.9365, Validation Loss = 1.0869
  Validation loss improved. Saving model.
  Epoch 2: 100%|████████| 7/7 [00:22<00:00,  3.22s/it, loss=0.802]
  Epoch 2: Train Loss = 0.7160, Validation Loss = 0.8465
  Validation loss improved. Saving model.
  Epoch 3: 100%|████████| 7/7 [00:20<00:00,  2.95s/it, loss=0.743]
  Epoch 3: Train Loss = 0.5667, Validation Loss = 0.5439
  Validation loss improved. Saving model.
  Epoch 4: 100%|████████| 7/7 [00:23<00:00,  3.41s/it, loss=0.261]
  Epoch 4: Train Loss = 0.3767, Validation Loss = 0.3540
  Validation loss improved. Saving model.
  Epoch 5: 100%|████████| 7/7 [00:22<00:00,  3.23s/it, loss=0.256]
  Epoch 5: Train Loss = 0.2585, Validation Loss = 0.2365
  Validation loss improved. Saving model.
  Epoch 6: 100%|████████| 7/7 [00:22<00:00,  3.16s/it, loss=0.0512]
  Epoch 6: Train Loss = 0.1681, Validation Loss = 0.1440
  Validation loss improved. Saving model.
  Epoch 7: 100%|████████| 7/7 [00:22<00:00,  3.26s/it, loss=0.111]
  Epoch 7: Train Loss = 0.1144, Validation Loss = 0.1006
  Validation loss improved. Saving model.
```
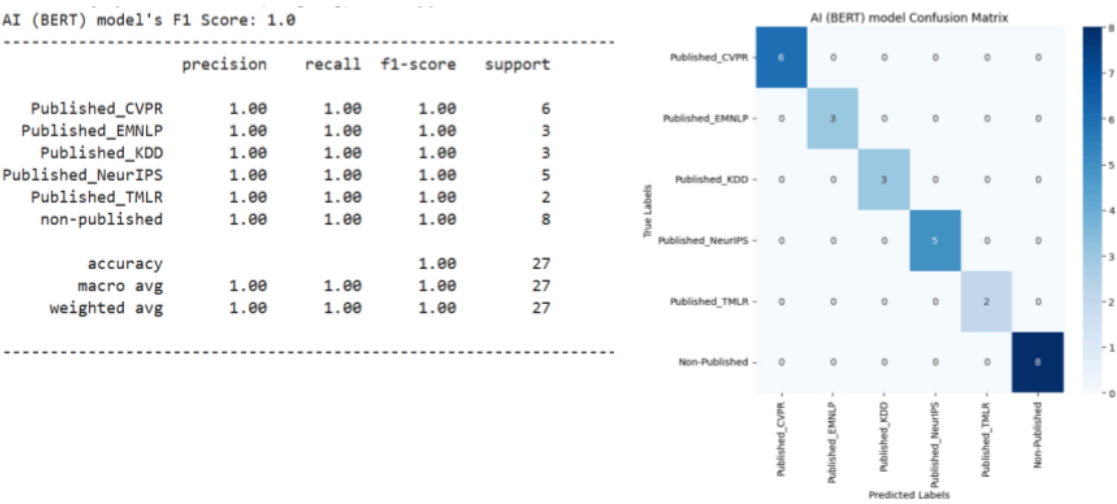
0s   completed at 8:56 PM

Epoch 12: 100%|████████| 7/7 [00:21<00:00, 3.14s/it, loss=0.0363]
Epoch 12: Train Loss = 0.0451, Validation Loss = 0.0407
Validation loss improved. Saving model.
Epoch 13: 100%|████████| 7/7 [00:22<00:00, 3.20s/it, loss=0.0406]
Epoch 13: Train Loss = 0.0415, Validation Loss = 0.0377
Validation loss improved. Saving model.
Epoch 14: 100%|████████| 7/7 [00:20<00:00, 2.96s/it, loss=0.0372]
Epoch 14: Train Loss = 0.0399, Validation Loss = 0.0357
Validation loss improved. Saving model.
Epoch 15: 100%|████████| 7/7 [00:22<00:00, 3.15s/it, loss=0.0377]
Epoch 15: Train Loss = 0.0363, Validation Loss = 0.0340
Validation loss improved. Saving model.
Epoch 16: 100%|████████| 7/7 [00:22<00:00, 3.21s/it, loss=0.0489]
Epoch 16: Train Loss = 0.0360, Validation Loss = 0.0326
Validation loss improved. Saving model.
Epoch 17: 100%|████████| 7/7 [00:21<00:00, 3.00s/it, loss=0.0313]
Epoch 17: Train Loss = 0.0335, Validation Loss = 0.0317
Validation loss improved. Saving model.
Epoch 18: 100%|████████| 7/7 [00:21<00:00, 3.13s/it, loss=0.0342]
Epoch 18: Train Loss = 0.0346, Validation Loss = 0.0311
Validation loss improved. Saving model.
Epoch 19: 100%|████████| 7/7 [00:22<00:00, 3.22s/it, loss=0.0312]
Epoch 19: Train Loss = 0.0325, Validation Loss = 0.0309
Validation loss improved. Saving model.

Evaluating Model Performance: F1 Score, Classification Report, and Confusion Matrix -

```
AI (BERT) model's F1 Score: 1.0
--------------------------------------------------------------
                    precision    recall  f1-score   support

    Published_CVPR       1.00      1.00      1.00         6
   Published_EMNLP       1.00      1.00      1.00         3
     Published_KDD       1.00      1.00      1.00         3
  Published_NeurIPS      1.00      1.00      1.00         5
    Published_TMLR       1.00      1.00      1.00         2
     non-published       1.00      1.00      1.00         8

          accuracy                           1.00        27
         macro avg       1.00      1.00      1.00        27
      weighted avg       1.00      1.00      1.00        27

--------------------------------------------------------------
```


AI (BERT) model Confusion Matrix

Prediction on Test Data -

```
     pdf_name        Labels        predictions
98   P099.pdf      Published    Published_TMLR
67   P068.pdf  Not Published    Non-Published
105  P106.pdf  Not Published    Non-Published
19   P020.pdf  Not Published    Non-Published
42   P043.pdf      Published   Published_EMNLP
62   P063.pdf      Published   Published_EMNLP
12   P013.pdf      Published   Published_EMNLP
110  P111.pdf      Published    Published_CVPR
125  P126.pdf  Not Published    Non-Published
128  P129.pdf      Published    Published_TMLR
44   P045.pdf      Published     Published_KDD
40   P041.pdf      Published    Published_CVPR
27   P028.pdf  Not Published    Non-Published
100  P101.pdf      Published    Published_CVPR
119  P120.pdf  Not Published    Non-Published
31   P032.pdf  Not Published    Non-Published
51   P052.pdf  Not Published    Non-Published
4    P005.pdf      Published     Published_KDD
84   P085.pdf      Published     Published_KDD
81   P082.pdf  Not Published    Non-Published
85   P086.pdf  Not Published    Non-Published
26   P027.pdf      Published  Published_NeurIPS
16   P017.pdf      Published  Published_NeurIPS
```

Semi-Deployment: PDF Name-Based Prediction and Recommendation

```
Enter the PDF name: P111.pdf
Recommendation for the Label and Subcategory of the Specified PDF:
     pdf_name     Labels       predictions
110  P111.pdf  Published  Published_CVPR
```