# Hackathon Submission Report

## Project Title

Predicting Research Paper Publishability and Conference Classification Using SciBERT

## Team Details

* Team Name: Team Lebu
* Team Members: Chitrak Saha

Antareep Ghosh

Anit Ketan Suin

Debarupa Sinha

## Abstract

This study addresses the automation of research paper classification into publishable and non-publishable categories (Task 1) and the subsequent prediction of suitable conferences (Task 2) for publishable papers. Employing SciBERT, a domain-specific pre-trained transformer model, we attained robust performance with accuracy metrics of 0.75 for Task 1 and 0.920 for Task 2. These results underscore the efficacy of transformer-based approaches in enhancing academic workflows.

## Problem Statement

The process of determining whether a research paper is publishable and identifying the most appropriate conference for submission is labor-intensive and subjective. This project proposes a machine learning solution to:

1. Classify papers as publishable or non-publishable (Task 1).
2. Predict the optimal conference for publishable papers (Task 2).

## Methodology

### Data Preparation

1. **Dataset**: The dataset comprised labeled research papers. Task 1 involved binary labels (publishable/non-publishable), while Task 2 included multi-class conference labels (e.g., CVPR, NeurIPS, EMNLP).
2. **Preprocessing**:
   * Extracted and cleaned text data from research papers.
   * Removed noise, including stopwords and special characters.
   * Tokenized and vectorized the text using the SciBERT tokenizer.
3. **Data Split**: The dataset was divided into training and testing subsets to ensure reliable evaluation.

### Model Architecture

We utilized SciBERT (allenai/scibert\_scivocab\_uncased) for both tasks. SciBERT’s pretraining on scientific text renders it particularly suitable for this domain.

1. **Task 1: Pushability Classification**
   * Fine-tuned SciBERT for binary classification.
   * Output Layer: Single neuron with sigmoid activation to classify papers as publishable or non-publishable.
2. **Task 2: Conference Prediction**
   * Fine-tuned SciBERT for multi-class classification.
   * Output Layer: Softmax activation for predicting specific conference labels.

### Training

* **Optimizer**: AdamW
* **Learning Rate**: 5e-6
* **Batch Size**: 5
* **Epochs**: 5

## Results

### Task 1: Pushability Classification

* **Accuracy**: 0.75
* **Precision, Recall, F1-Score**: Detailed metrics are included in the appendix.

### Task 2: Conference Classification

* **Accuracy**: 0.920
* **Confusion Matrix and Metrics**: Detailed results are included in the appendix.

## Challenges and Solutions

1. **Data Imbalance**: The dataset exhibited a significant imbalance, with fewer publishable papers compared to non-publishable ones. Instead of oversampling, we relied on SciBERT’s capability to generalize from contextual embeddings.
2. **Sequence Length Constraints**: SciBERT’s maximum token length of 512 necessitated chunking lengthy texts and aggregating predictions.
3. **Resource Limitations**: The computational expense of training transformer models was mitigated by efficient batching and gradient accumulation techniques.

## Conclusion

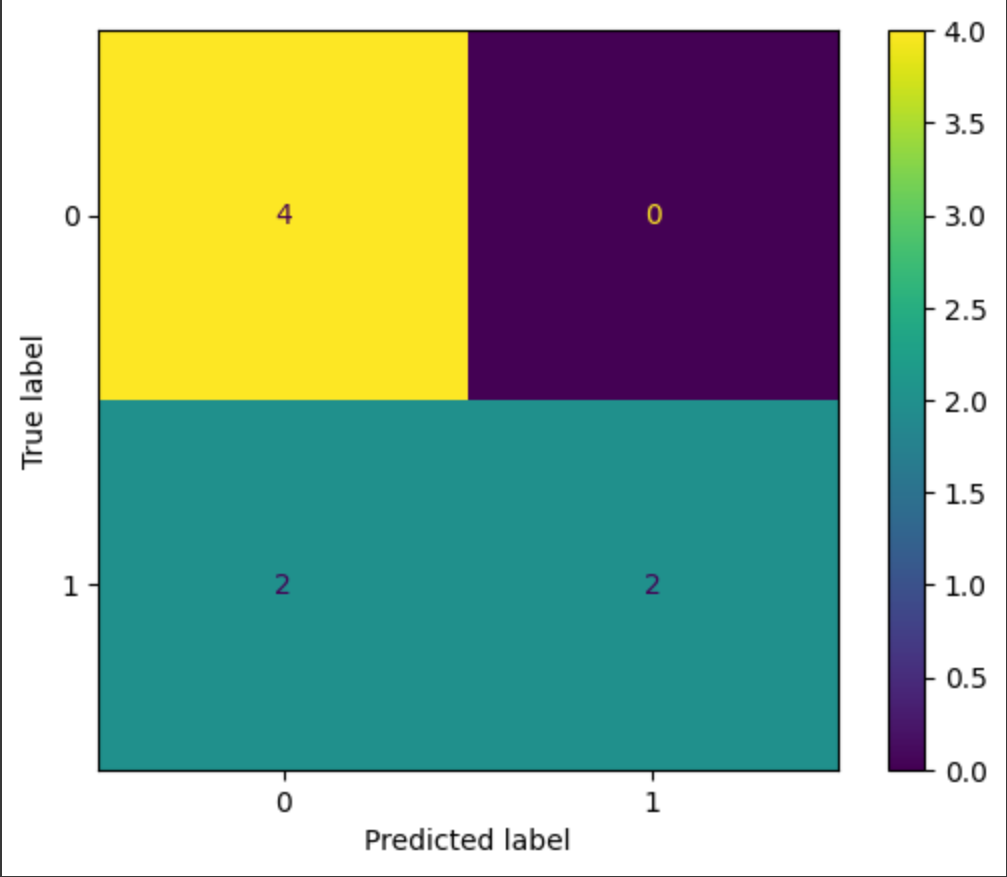
This project effectively demonstrates the application of SciBERT for academic workflow automation. With accuracy scores of 0.75 for Task 1 and 0.920 for Task 2, the model exhibits substantial potential for deployment in real-world scenarios, significantly reducing the time and effort required for manual classification tasks.

## Future Work

1. **Enhancing Task 1 Accuracy**: Exploring advanced data augmentation techniques to improve publishability classification.
2. **Real-Time Integration**: Developing a user-friendly web interface for seamless adoption by researchers.
3. **Pathway Integration**: Incorporating Pathway connectors to facilitate streamlined data ingestion and automated result dissemination.

## Appendix

### Task 1: Classification Report



### Task 2: Classification Report

