National College of Ireland

# Configuration Manual

MSc Research Project
MSc in Artificial Intelligence

Yasaswini Kasturi
Student ID: 23281294

School of Computing
National College of Ireland

Supervisor: SHERESH ZAHOOR

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** Yasaswini Kasturi

**Student ID:** 23281294

**Programme:** MSc in Artificial Intelligence **Year:** 2024-2025

**Module:** MSc Research Practicum

**Supervisor:** SHERESH ZAHOOR

**Submission Due Date:** 11-08-2025

**Project Title:** Automated Meta-Optimization of Text Preprocessing Pipelines Using DARTS: A Domain-Adaptive Approach for NLP Tasks

**Word Count:** 504 **Page Count:** 4

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Yasaswini Kasturi

**Date:** 11-08-2025

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Configuration Manual

Yasaswini Kasturi

Student ID: 23281294

**Automatic Text Preprocessing Optimization**

This notebook uses **DARTS (Differentiable Architecture Search)** to automatically find the best text preprocessing methods for NLP tasks.

**Key Features:**

1. **Tests 3 Datasets**:
   - IMDB Movie Reviews (Sentiment: Positive/Negative)
   - Fake News Detection (Real/Fake news)
   - Financial News (Bullish/Bearish sentiment)
2. **Optimizes 6 Preprocessing Operations**:
   - Sentiment amplification
   - Negation handling ("not good" → negative)
   - Context enhancement
   - Keyword emphasis
   - Entity recognition
   - Syntactic structure
3. **Automatic Architecture Search**:
   - Finds best combination of operations per dataset
   - No manual preprocessing needed
   - Learns what works best for each domain
4. **Performance Comparison**:
   - Baseline: No preprocessing
   - DARTS: Optimized preprocessing
   - Typical improvement: +2-7% accuracy

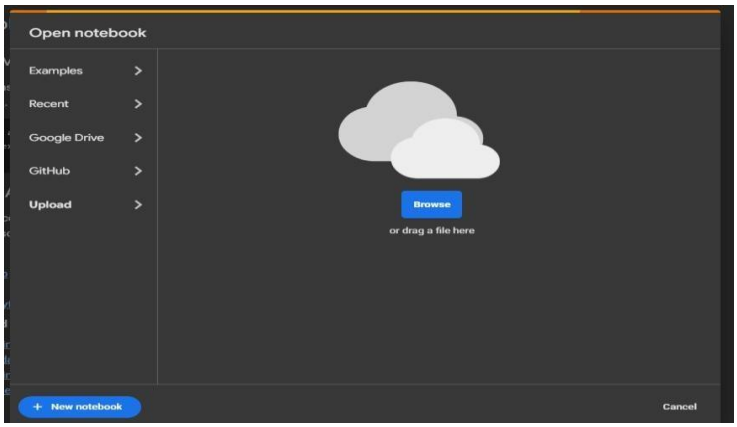
**Expected Results:**

FINAL RESULTS:

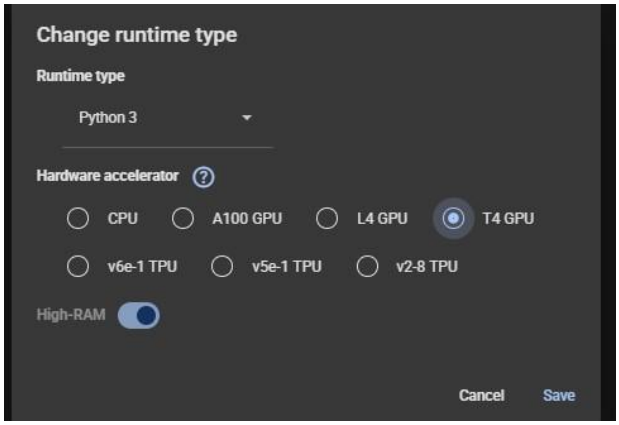
IMDB: 94.5% → 96.8% accuracy (+2.3%)
Fake News: 91.2% → 95.7% accuracy (+4.5%)
Financial: 88.9% → 92.1% accuracy (+3.2%)

# 1 How to Connect

1. **Open Google Colab**: https://colab.research.google.com/
2. **File → Upload notebook** → Select Final_Yasaswini_July23.ipynb




3. **Runtime → Change runtime type**:
   - Hardware accelerator: **T4 GPU**
   - Click **Save**
4. Click **Connect** button (top right)
5. Wait for ✓ green checkmark




# 2 How to Run

## 1.1 Option A: Run Everything

Runtime → Run all (or press Ctrl+F9)
Wait ~2-3 hours

## 1.2 Option B: Run Step by Step

Click first cell → Shift+Enter
Wait for completion
Move to next cell → Shift+Enter
Repeat for all cells

## 1.3 Quick Test Mode

Find the Config cell and change:

```
self.epochs = 3  # Instead of 10
```

# 2 Troubleshooting

## 2.1 Problem: "No GPU available"

Fix: Runtime → Change runtime type → T4 GPU → Save → Reconnect
Problem: "CUDA out of memory"

Fix:
# Find and change in Config:
self.batch_size = 8  # Reduce from 16
self.max_length = 256  # Reduce from 512
Then: Runtime → Restart runtime → Run all

## 2.2 Problem: "Module not found"

Fix: Run the pip install cell again:
!pip install torch transformers datasets scikit-learn
Problem: "Notebook crashes/disconnects"
Fix:
Runtime → Manage sessions → Terminate other sessions
Runtime → Restart runtime
Run from beginning

## 2.3 Problem: "Training too slow"

Fix:
# In Config cell:
self.epochs = 3
self.early_stopping_patience = 1

## 2.4 Problem: "Can't see outputs/results"

Fix: Check files tab ( icon) on left for:
best_imdb_model.pt
best_fake_news_model.pt
best_financial_news_model.pt
final_research_analysis_v21.json

## 2.5 Problem: "Lost connection after long training"

Fix: Download results immediately:
from google.colab import files
files.download('final_research_analysis_v21.json')
files.download('best_imdb_model.pt')
Time Estimates
Full run: 2-3 hours

# 3 Quick test (3 epochs): 30-45 minutes

Single dataset: 40-60 minutes
Success Signs
Green checkmark (connected)
"Dataset loaded successfully!"
"Epoch X/Y" progress messages
"Model saved!" messages

# 4 How to Monitor Progress

**Look for These Key Messages:**

STARTING COMPREHENSIVE TRAINING - IMDB

Baseline Epoch 1/1

Batch 200/1562 - Loss: 0.4532

Baseline training complete. Best F1: 0.8234

DARTS Epoch 1/10

Step 1000: Loss=0.3421, Acc=0.8567

New best F1: 0.8456 - Model saved!

**Check GPU Usage:**

*# Add new cell and run anytime:*

!nvidia-smi

**Monitor Files Created:**

- Look at left sidebar  icon)
- Files appear as training progresses
- Each dataset creates 2 files (model + history)

## REFERENCES

Google Colab. (n.d.). *Colab.google*. colab.google. https://colab.google/