National College of Ireland

Configuration Manual

MSc Research Project
MSc in Data Analytics

RAMAKRISHNA REDDY VENREDDY
Student ID: x23318503

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ramakrishna Reddy Venreddy
Student ID:	x23318503
Programme:	MSc in Data Analytics **Year:** 2024-2025
Module:	Research Project
Lecturer:	Jorge Basilio
Submission Due Date:	15th Sep 2025
Project Title:	MedReview:A Graph-Enhanced Framework for Comparing Domain Specific and General Language Models in Consumer Drug Review Analysis
Word Count:	827 **Page Count:** 07

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: Ramakrishna Reddy Venreddy

Date: 15th Sep 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

Ramakrishna Reddy Venreddy

Student ID: x23318503

1. Project Overview

- **Description**: An advanced medical review analysis system that uses Graph Neural Networks (PyTorch Geometric) to analyze drug reviews, extracting medical entities and building knowledge graphs for enhanced sentiment analysis and effectiveness prediction.
- **Main Objectives**: Compare 12 transformer models (BERT, RoBERTa, ClinicalBERT, BioBERT, etc.) with and without graph enhancement, achieving 10-15% performance improvement through semantic bridging.
- **Expected Outputs**: Sentiment classifications (3-class), drug effectiveness scores (0-1), multi-aspect ratings (6-class), and comprehensive performance comparison tables.

2. Google Colab Setup Guide

Step 1: Initial Setup

1. **Open Google Colab**
    - Navigate to colab.research.google.com
    - Click "New notebook" or File → Upload notebook to upload enhanced_medreview_v2.1.1.ipynb
2. **Connect to T4 GPU**
    - Runtime → Change runtime type
    - Hardware accelerator: GPU
    - GPU type: T4 (recommended) or A100
    - Click "Save"
    - Connect to runtime (top-right button)
3. **Mount Google Drive** (for saving results)
    - The code will auto-mount, or manually run:

```
from google.colab import drive
drive.mount('/content/drive')
```

Step 2: Dependencies & Environment

Required Installations (auto-installed by the script):

!pip install torch-geometric

!pip install transformers torch datasets networkx requests scikit-learn tabulate

!pip install torch_geometric *# For Graph Neural Networks*

API Keys Configuration (Optional but recommended):

 # Line 358-359: Update with your API keys (or use defaults)

UMLS_API_KEY: str = "your-umls-api-key-here" *# <- CHANGE THIS (optional)*

BIOPORTAL_API_KEY: str = "your-bioportal-key" *# <- CHANGE THIS (optional)*

Figure 1 API Keys

Data Source Configuration:

 # Line 362: HuggingFace dataset (auto-downloads)

DATASET_NAME: str = "forwins/Drug-Review-Dataset" *# No change needed*

Step 3: Critical Configuration Points

A. Sample Size Configuration

 # Line 365-367: Adjust dataset size

REVIEWS_PER_RATING: int = 5000 *# <- CHANGE THIS (reduce to 1000 for faster testing)*

MAX_TOTAL_REVIEWS: int = 50000 *# <- CHANGE THIS (total samples to process)*

```
123    BIOPORTAL_BASE_URL: str = "https://data.bioontology.org"
124
125    # Dataset Configuration - INCREASED SAMPLES
126    DATASET_NAME: str = "forwins/Drug-Review-Dataset"
127    REVIEWS_PER_RATING: int = 5000
128    RANDOM_SEED: int = 1999
129    MAX_TOTAL_REVIEWS: int = 50000
130
131    # ENHANCED ENTITY EXTRACTION
132    MAX_API_ENTITIES: int = 300
133    ENTITY_EXTRACTION_BATCH_SIZE: int = 20
134    NER_PROCESSING_SAMPLES: int = 3000
135    MIN_ENTITY_CONFIDENCE: float = 0.3
136    MEDICAL_ENTITY_BOOST: float = 2.0
137    API_VALIDATION_BOOST: float = 3.0
138
```

B. Training Parameters

 # Line 387-390: Adjust training settings

BATCH_SIZE: int = 16 *# <- REDUCE to 8 if GPU memory error*

BASELINE_EPOCHS: int = 2 *# <- INCREASE to 3-4 for better baseline*

GRAPH_EPOCHS: int = 8 *# <- REDUCE to 4-6 for faster training*

```
149
150    # Training Configuration - DIFFERENTIAL EPOCHS
151    MAX_LENGTH: int = 256
152    BATCH_SIZE: int = 16
153    LEARNING_RATE: float = 2e-5  # Slightly higher for graph models
154    BASELINE_EPOCHS: int = 2    # REDUCED for baselines
155    GRAPH_EPOCHS: int = 8       # INCREASED for graph models
156    WARMUP_STEPS: int = 500
157    GRADIENT_ACCUMULATION_STEPS: int = 4
```

C. Google Drive Paths (for saving results)

 # Line 376-378: Update save locations

DRIVE_MOUNT_PATH: str = "/content/drive"

RESULTS_DRIVE_PATH: str = "/content/drive/MyDrive/MedReview_Results" *# <- CHANGE THIS*

MODELS_DRIVE_PATH: str = "/content/drive/MyDrive/MedReview_Models" *# <- CHANGE THIS*

D. Model Selection (choose which models to compare)

 # Line 447-524: Comment out unwanted models

self.MODELS_TO_COMPARE = {

'bert_baseline': {...}, *# Keep or comment*

'bert_graph_enhanced': {...}, *# Keep or comment*

 # Add '#' at line start to disable any model

}

```
# 12 Model Comparison
self.MODELS_TO_COMPARE = {
    'bert_baseline': {
        'model_name': 'bert-base-uncased',
        'is_medical': False,
        'use_graph': False,
        'description': 'BERT Base Baseline',
        'category': 'General'
    },
    'bert_graph_enhanced': {
        'model_name': 'bert-base-uncased',
        'is_medical': False,
        'use_graph': True,
        'description': 'BERT Base + Enhanced PyTorch Geometric GNN',
        'category': 'General'
    },
    'roberta_baseline': {
        'model_name': 'roberta-base',
        'is_medical': False,
        'use_graph': False,
        'description': 'RoBERTa Base Baseline',
        'category': 'General'
    },
    'roberta_graph_enhanced': {
        'model_name': 'roberta-base',
```

E. Graph Configuration

MAX_GRAPH_NODES: int = 500 *# <- REDUCE to 200 for faster processing*

GRAPH_LAYERS: int = 3 *# <- REDUCE to 2 for simpler graphs*

GRAPH_ATTENTION_HEADS: int = 8 *# <- REDUCE to 4 for less memory*

3. Execution Instructions

Quick Run (Default Settings)

1. Upload notebook to Colab
2. Connect to T4 GPU
3. Run all cells: Runtime → Run all
4. Wait ~2-4 hours for completion

Fast Test Run (Reduced Settings)

 # Modify these values before running:

REVIEWS_PER_RATING = 500 *# Line 365*

BASELINE_EPOCHS = 1 *# Line 389*

GRAPH_EPOCHS = 2 *# Line 390*

 # Then run all cells

Custom Dataset

 # Line 362: Point to your dataset

DATASET_NAME: str = "your-username/your-dataset" *# <- CHANGE THIS*

 # Or load local CSV:

 # df = pd.read_csv('/content/your_data.csv') # Add after line 2890

4. Output Files & Locations

Results are saved to Google Drive:

- /MyDrive/MedReview_Results/
  - enhanced_v2.1.1_results_[timestamp].json - Final results
  - enhanced_v2.1.1_checkpoint_[model].json - Training checkpoints
- /MyDrive/MedReview_Models/
  - [model_name]_enhanced_v2.1.1_best.pt - Trained model weights

Console Output includes:

- Performance comparison tables
- Graph impact analysis
- Research insights and conclusions

5. Troubleshooting

Issue	**Solution**
GPU Memory Error	Reduce BATCH_SIZE to 8 (line 387)
Training Too Slow	Reduce REVIEWS_PER_RATING to 1000 (line 365)
API Connection Failed	APIs are optional; system uses fallback
PyTorch Geometric Error	Re-run installation cell or restart runtime
Drive Mount Failed	Manually mount using code in Step 1.3

6. Expected Runtime

Configuration	Estimated Time
Full Run (5000 samples/rating)	3-4 hours
Medium Run (2000 samples/rating)	1.5-2 hours
Test Run (500 samples/rating)	30-45 minutes
Quick Test (100 samples/rating)	10-15 minutes

7. Key Parameters Summary

 # Essential parameters to customize:

config = EnhancedSystemConfig(

REVIEWS_PER_RATING = 5000, *# Dataset size per rating*

BASELINE_EPOCHS = 2, *# Training epochs for baseline*

GRAPH_EPOCHS = 8, *# Training epochs for graph models*

BATCH_SIZE = 16, *# Training batch size*

MAX_GRAPH_NODES = 500, *# Graph complexity*

RESULTS_DRIVE_PATH = "/content/drive/MyDrive/MedReview_Results"

)

8. Quick Validation

After setup, verify installation:

import torch

import torch_geometric

print(f"PyTorch: {torch.__version__}")

print(f"PyTorch Geometric: {torch_geometric.__version__}")

print(f"GPU Available: {torch.cuda.is_available()}")

print(f"GPU Name: {torch.cuda.get_device_name(0) if torch.cuda.is_available() else 'None'}")

Expected output:

- PyTorch: 2.x.x

- PyTorch Geometric: 2.x.x
- GPU Available: True
- GPU Name: Tesla T4

Ready to Run! Follow steps 1-3, adjust parameters as needed, and execute. Results will be automatically saved to your Google Drive.

References:

K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting readmission," arXiv (Cornell University), Jan. 2019, doi: 10.48550/arxiv.1904.05342. Available: https://arxiv.org/abs/1904.05342

He, P., Liu, X., Gao, J., & Chen, W. (2020b, June 5). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv.org. https://arxiv.org/abs/2006.03654