Automating Course Articulation: A Deep Metric Learning Framework Using Public Data

Thesis Outline & Terminology Guide

Mark S. Kim

May 2025

Thesis Outline

Abstract

- **Problem:** The manual process of determining course equivalency is a major obstacle for student mobility and educational equity, causing credit loss and graduation delays that disproportionately affect underrepresented students.
- **Limitations of Previous Work:** Past automated methods have been hampered by their dependence on sensitive student data or the impracticality and opacity of using large language models (LLMs) for direct classification.
- **Proposed Solution:** This thesis introduces a novel framework that uses only public course catalog data, separating semantic representation from classification.
- **Methodology:** The approach uses deep metric learning to fine-tune embedding models on course text. These specialized embeddings are used to create a composite distance vector, which then serves as a feature set for traditional machine learning classifiers.
- **Results:** The framework achieves state-of-the-art accuracy, with F1-scores over 0.99.
- **Conclusion:** The result is a computationally efficient, scalable, and privacy-preserving tool to automate course articulation, reduce administrative burden, and promote a more equitable educational system.

Chapter 1: Introduction

1.1 The California Context

- California's three-tiered public higher education system (UC, CSU, CCC) is the largest in the U.S.
- The course articulation process, facilitated by the ASSIST repository, is a manual, inefficient, and intractable task.

1.2 The National Problem of Student Transfer

- Transferring institutions is a common part of the modern student journey.
- Transfer enrollment is resurgent post-pandemic, with growth driven by diverse and returning student populations.

1.3 Consequences for Students: Inefficiency and Inequity

- **Credit Loss:** Transfer students lose a significant percentage of their credits.
- **Financial and Academic Setbacks:** Credit loss increases time-to-degree and cost, while negatively impacting student persistence.
- **Educational Equity:** Low-income and underrepresented minority students are disproportionately harmed by these systemic barriers.

1.4 Thesis Contribution and Roadmap

- **Contribution:** This thesis develops a computational framework to automate course articulation using deep metric learning on public data.
- **Roadmap:** The thesis is organized into chapters covering background, methodology, results, and discussion.

Chapter 2: Background and Related Work

- **2.1 Keyword and Statistical Methods:** Early attempts (e.g., TF-IDF) lacked true semantic understanding.
- **2.2 Static Semantic Representations:** Models like Word2Vec and GloVe were an improvement but are context-insensitive.
- **2.3 Contextual Semantic Representations:** Transformer models like BERT generate richer, contextual embeddings.
- **2.4 Direct LLM Classification:** Promising accuracy but limited by cost, opacity, and prompt sensitivity.
- **2.5 Enrollment-Based Approaches:** Methods like course2vec are powerful but rely on private data and lack generalizability.
- **2.6 Research Gap:** A need exists for a hybrid framework that decouples semantic representation from classification.

Chapter 3: Methodology

- **3.1 Phase 1: Direct LLM Classification:** An exploratory phase using Google's Gemini Pro to establish a baseline.
- **3.2 Phase 2: The Decoupled Pipeline Framework:**
  - **PPM Corpus:** A large dataset from the Program Pathways Mapper (PPM) with 2,157 courses.

  - **Feature Engineering:** A novel **Composite Distance Vector (Δ_c)** combines element-wise difference and cosine similarity.

- **3.3 Model Architecture and Training:**
  - **Embedding Models:** Selection of BGE, GIST, NVE, and SFR for analysis.
  - **Fine-Tuning:** Deep metric learning with triplet loss functions to create a domain-specific model.
  - **Downstream Classifiers:** Evaluation of KNN, SVM, Random Forest (RF), and XGBoost.
- **3.4 Evaluation Framework:** A multi-stage evaluation using F1-score for efficacy and timing for efficiency.
- **3.5 Misclassification Analysis:** A qualitative diagnosis to find root causes of errors, distinguishing between model-specific and data-inherent issues.

Chapter 4: Experimental Setup and Results

- **4.2 Baseline Performance:** Direct LLM classification achieved 90.5% accuracy but had practical limitations.
- **4.3 Core Component Validation:** An ablation study confirmed the superiority of the Composite Distance Vector.
- **4.4 Domain-Specific Fine-Tuning:** The fine-tuned **BGE-ft** model was statistically superior to all off-the-shelf models.
- **4.5 Downstream Classifier Performance:** SVM was the most accurate, while RF and XGBoost were far more efficient.
- **4.6 Qualitative Diagnosis:** The majority of remaining errors were systematic and stemmed from data quality issues (e.g., semantic divergence, vague descriptions, labeling errors).

Chapter 5: Discussion, Future Work, and Conclusion

- **5.1 Discussion:** The framework is vindicated, the impact of fine-tuning is critical, and the bottleneck has shifted from model-centric to data-centric problems.
- **5.2 Limitations:** Performance is capped by data quality; the fine-tuned model's generalizability needs testing; the scope of "equivalency" is limited to text; complex articulations are not natively handled.
- **5.3 Future Work:** Focus on data-centric strategies (human-in-the-loop), expand the framework into a recommendation engine, and explore more advanced modeling techniques.
- **5.4 Conclusion:** The thesis delivers a practical, scalable, and privacy-preserving tool that contributes to a more equitable and efficient educational ecosystem.

Standardized Naming Convention Reference Guide

This guide categorizes and defines the standard abbreviations for all technical entities mentioned in the thesis.

Embedding Models

These are the core deep learning models used for generating semantic representations of course text.

Abbreviated Name	Full Name	Entity Type
BGE	BAAI/bge-small-en-v1.5	Embedding Model
GIST	avsolatorio/GIST-Embedding-v0	Embedding Model
NVE	nvidia/NV-Embed-v2	Embedding Model
SFR	Salesforce/SFR-Embedding-2_R	Embedding Model
BGE-ft	Fine-tuned BGE Model	Model Variation

Machine Learning (ML) Classifiers

These are the traditional machine learning algorithms used in the downstream classification task.

Abbreviated Name	Full Name	Entity Type
KNN	K-Nearest Neighbors	ML Classifier
SVM	Support Vector Machine	ML Classifier
RF	Random Forest	ML Classifier
XGB	XGBoost	ML Classifier
LR	Logistic Regression	ML Classifier
RIDGE	Ridge Classifier	ML Classifier
LASSO	Lasso Classifier	ML Classifier
LDA	Linear Discriminant Analysis	ML Classifier
QDA	Quadratic Discriminant Analysis	ML Classifier

Established Models & Architectures

These are foundational models and architectures with widely recognized names that will retain their original capitalization for consistency with existing literature.

Framework Components

These are other key technical components of the methodology. They are generally referred to by their full name to maintain clarity.

Abbreviated Name	Full Name	Entity Type
Word2Vec	Word2Vec	Static Embedding Model
GloVe	Global Vectors for Word Representation	Static Embedding Model
course2vec	course2vec	Enrollment-Based Model
TF-IDF	Term Frequency-Inverse Document Frequency	Statistical Method
BERT	Bidirectional Encoder Representations from Transformers	Model Architecture
SBERT	Sentence-BERT	Model Architecture
PaLM2	Pathways Language Model 2	Large Language Model
Gemini	Gemini Pro v1.0	Large Language Model
GPT-4	Generative Pre-trained Transformer 4	Large Language Model

Abbreviated Name	Full Name	Entity Type
(N/A)	BatchAllTripletLoss	Loss Function
(N/A)	BatchSemiHardTripletLoss	Loss Function
(N/A)	BatchHardTripletLoss	Loss Function
(N/A)	BatchHardSoftMarginTripletLoss	Loss Function
AdamW	AdamW Optimizer	Optimizer
(N/A)	CosineAnnealingWarmRestarts	Learning Rate Scheduler
(N/A)	GroupByLabelBatchSampler	Data Sampler
PCA	Principal Component Analysis	Dimensionality Reduction
t-SNE	t-Distributed Stochastic Neighbor Embedding	Dimensionality Reduction
PaCMAP	Pairwise Controlled Manifold Approximation	Dimensionality Reduction
UMAP	Uniform Manifold Approximation and Projection	Dimensionality Reduction