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

Unified Outline & Terminology Guide

Mark S. Kim

May 2025

Thesis Outline: Expanded Version

Abstract

- **Problem:** The manual process of course equivalency determination acts as a significant barrier to student mobility and educational equity. This leads to substantial credit loss and delayed graduation, which disproportionately harms underrepresented students.
- **Limitations of Previous Work:** Past automated methods have been limited by a reliance on sensitive student enrollment records or the operational intractability and opacity of using large language models (LLMs) for direct classification.
- **Proposed Solution:** The thesis introduces and validates a novel framework that decouples semantic representation from classification. It uses only publicly available course catalog data, making it a privacy-preserving solution.
- **Methodology:** The framework leverages deep metric learning to fine-tune contextual embedding models on public course text. These specialized embeddings are then used to construct a novel composite distance vector, which serves as a rich feature set for training traditional machine learning classifiers.
- **Results:** When evaluated on a real-world dataset, the proposed framework achieves state-of-the-art accuracy, with F_1-scores exceeding 0.99.
- **Conclusion:** The final result is a computationally efficient, scalable, and privacy-preserving tool for institutions to automate course articulation, reduce administrative burden, and foster a more equitable educational ecosystem.

Chapter 1: Introduction

- **1.1 The California Context: A Critical Case Study**
  - California's public higher education system is the largest in the U.S., comprising 149 colleges and universities serving nearly 2.9 million students.

- A foundational principle is student mobility, particularly the pathway from a two-year community college to a four-year university.
- The Articulation System Stimulating Interinstitutional Student Transfer (ASSIST) is the state's official repository for articulation agreements, but it is fundamentally a display for agreements that are negotiated manually.
- This manual process is inefficient, slow, and intractable due to the sheer number of institutions.

- **1.2 The National Problem of Student Transfer**
  - Transferring between institutions is a normative part of the modern student's academic journey.
  - In the fall of 2023, transfer enrollment constituted 13.2% of all continuing and returning undergraduates.
  - Transfer enrollment has seen a post-pandemic resurgence, growing by 5.3% from Fall 2022 to Fall 2023.
  - Over half of all returning learners re-enroll at a new institution, highlighting the critical role of the transfer system.

- **1.3 Consequences for Students: Inefficiency and Inequity**
  - **Credit Loss:** A 2017 U.S. Government Accountability Office (GAO) report estimated that transfer students lost an average of 43% of their credits. More than half of all transfer students lose some credits.
  - **Financial and Academic Setbacks:** Credit loss increases the time-to-degree, delays entry into the workforce, and increases the total tuition burden.
  - **Educational Equity:** Low-income students and students from historically underrepresented racial and ethnic groups are more likely to rely on transfer pathways. Recent growth in transfer enrollment has been driven disproportionately by Black and Hispanic students. The barriers imposed by an inefficient system disproportionately harm these student populations.

- **1.4 Thesis Contribution and Roadmap**
  - This thesis confronts the problem by developing a novel computational framework to automate course articulation.
  - It leverages deep metric learning on publicly available course catalog text, introducing a privacy-preserving and scalable method.
  - Primary contributions include the development of a highly accurate framework, an innovative feature engineering technique, and a computationally efficient approach that avoids the issues of previous methods.

Chapter 2: Background and Related Work

- **2.1 Keyword and Statistical Methods (e.g., TF-IDF)**
  - These methods are computationally simple but lack semantic depth.

  - Their fundamental limitation is a complete lack of semantic understanding; they cannot grasp that "calculus" and "differentiation" are related concepts.

- **2.2 Static Semantic Representations (e.g., Word2Vec, GloVe)**
  - These models represent a leap toward semantic understanding by representing words as dense vectors.
  - However, they produce a single, fixed vector for each word regardless of context, failing to account for polysemy (e.g., "bank" in "river bank" vs. "bank account").

- **2.3 Contextual Semantic Representations (e.g., BERT, SBERT)**
  - The transformer architecture, particularly models like Bidirectional Encoder Representations from Transformers (BERT), revolutionized NLP by enabling contextual embeddings.
  - Architectures like Sentence-BERT (SBERT) were developed to produce semantically meaningful embeddings for entire sentences or paragraphs that can be efficiently compared.

- **2.4 Direct LLM Classification**
  - Preliminary work for this thesis explored using LLMs like GPT-4 and Gemini directly for classification.
  - This approach was found to be computationally expensive, highly sensitive to prompt phrasing, and opaque ("black box"), providing no quantifiable similarity score.

- **2.5 Enrollment-Based Approaches (e.g., course2vec)**
  - These models predict course similarity by analyzing student co-enrollment patterns. The model 'course2vec' learns embeddings from the patterns of courses students take together.
  - This approach is constrained by its reliance on large-scale, proprietary institutional data, which raises significant privacy concerns and limits its generalizability. It cannot be used to compare courses between two institutions with no prior history of student transfer.

- **2.6 Research Gap**
  - The limitations of existing methods point toward a need for an intelligent, hybrid framework that decouples deep semantic representation from final classification. This conceptual gap forms the central motivation for the proposed methodology.

Chapter 3: Methodology

- **3.1 Phase 1: Direct LLM Classification Approach**
  - An exploratory phase to establish a performance baseline using a small, manually curated dataset.
  - The dataset was constructed from 5 required lower-division Computer Science courses at San Francisco State University (SFSU) and their articulated equivalents across 63 different California public colleges.

- From this, a final stratified random sample of 400 pairs (200 equivalent, 200 non-equivalent) was created for the evaluation set.
- Google's PaLM2 and its successor, Gemini Pro v1.0, were selected for this phase.

- **3.2 Phase 2: The Decoupled Pipeline Framework**
  - **PPM Corpus and Data Preparation:** A larger dataset was procured from the Program Pathways Mapper (PPM), initially containing 2,217 courses. After filtering to ensure robust partitioning, the final corpus contained 2,157 courses across 157 distinct C-ID classes. The dataset was partitioned into a stratified 50/50 train/test split.
  - **Input Document Normalization:** A consistent input document was created by concatenating four fields for each course: department name, department course number, course title, and the full course description.
  - **Feature Engineering: Composite Distance Vector (Δ_c):** A novel composite feature vector, Δ_c, was designed to represent the relationship between two courses. It is constructed by concatenating the element-wise difference of two course embedding vectors (A and B) with their cosine similarity:

$$\Delta_c = (a_1 - b_1, ..., a_k - b_k, \frac{A \cdot B}{\|A\|\|B\|})$$

  - This design provides the classifier with both granular, dimension-specific (local) disparities and a single, normalized measure of overall (global) alignment.

- **3.3 Model Architecture and Training**
  - **Embedding Model Selection:** Based on a preliminary analysis, three models were selected for the primary analysis: BAAI/bge-small-en-v1.5 (BGE), avsolatorio/GIST-Embedding-v0 (GIST), and nvidia/NV-Embed-v2 (NVE). Salesforce/SFR-Embedding-2_R (SFR) was added at a later stage.
  - **Metric Learning with Triplet Loss:** The BGE model was fine-tuned using a metric learning approach. The goal of Triplet Loss is to train the embedding function $f(x)$ such that the distance between an anchor (A) and a positive (P) is smaller than the distance to a negative (N) by a margin, α.

$$L(A, P, N) = \max(d(f(A), f(P)) - d(f(A), f(N)) + \alpha, 0)$$

  - Four primary batch-based triplet loss implementations were evaluated: BatchAllTripletLoss, BatchSemiHardTripletLoss, BatchHardTripletLoss, and BatchHardSoftMarginTripletLoss.
  - **Optimization Protocol:** The training used the AdamW optimizer paired with the CosineAnnealingWarmRestarts learning rate schedule.
  - **Downstream Classifiers:** Four models were selected for the final, in-depth analysis due to their strong performance: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and XGBoost.

- **3.4 Evaluation Framework and Metrics**
  - A multi-stage evaluation funnel was designed, moving from a broad review to a focused analysis on the PPM Corpus.

- During fine-tuning, the `BinaryClassificationEvaluator` from the `sentence-transformers` library was used to monitor embedding quality.
- The F_1-Score, the harmonic mean of Precision and Recall, is used as the primary metric for reporting final classification performance.

- **3.5 Misclassification Analysis Methodology**
  - A qualitative diagnosis was designed to move beyond aggregate scores and understand the root causes of model failures.
  - This involved a two-pronged methodology:
    1. Analyzing shared misclassifications across all models to identify systematic, data-inherent errors.
    2. A granular, qualitative case-based review of specific false positive and false negative examples.

Chapter 4: Experimental Setup and Results

- **4.2 Baseline Performance: Direct LLM Classification**
  - Experiments on the Initial Dataset using Google's Gemini Pro v1.0 achieved a peak accuracy of 90.5% when using full, unprocessed raw text as input.
  - Analysis of confusion matrices revealed a consistent conservative bias, where the model was far more likely to produce a false negative than a false positive.

- **4.3 Core Component Validation: The Composite Distance Vector**
  - An ablation study showed that the performance of linear models, like Logistic Regression, improved dramatically with the inclusion of the cosine similarity term (the F_1-score surged from 0.686 to 0.901).
  - More complex, non-linear models (KNN, SVM, RF) were not significantly affected, suggesting they are capable of inferring the global relationship from the local difference vector alone.

- **4.4 Domain-Specific Adaptation: Embedding Model Fine-Tuning**
  - The single best-performing fine-tuning configuration was the combination of `BatchSemiHardTripletLos` with the `cr_v1` cosine annealing schedule, which achieved the highest peak validation F_1-score of 0.8104.
  - **Statistical Validation:** The fine-tuned model (`bge-ft`) was proven to be statistically superior to its competitors. A one-way Analysis of Variance (ANOVA) confirmed a significant difference between model groups ($F = 6.2856$, $p < 0.001$). A subsequent Games-Howell post-hoc test showed that `bge-ft` demonstrated a statistically significant improvement over its base model, `bge`, with a mean F_1-score increase of 0.0109 ($p < .001$).

- **4.5 Final Stage Evaluation: Downstream Classifier Performance**
  - **Efficacy vs. Efficiency:** While all four finalist models achieved high accuracy, Random Forest and XGBoost were superior in efficiency. Their inference times are clustered under 0.1 seconds, an order of magnitude faster and more predictable than KNN and SVM.

- **Test Data Performance:** On the held-out test data, the Support Vector Machine (SVM) model achieved both the highest mean F_1-score ($\mu = 0.975973$) and the lowest standard deviation ($\sigma = 0.009120$), making it the most accurate and consistent performer.
- **Statistical Analysis:** An ANOVA confirmed a significant difference among the classifiers ($F(3, 316) = 3.1293$, $p = 0.02596$). A Games-Howell post-hoc test revealed that the SVM classifier performed statistically significantly better than both Random Forest and XGBoost ($p = 0.0229$).

- **4.6 Qualitative Diagnosis: Misclassification Analysis**
  - **Shared Misclassifications:** A high degree of overlap in errors was found across models, providing strong evidence of systematic, data-inherent issues. A total of 211 distinct course pairs were misclassified by every single embedding model evaluated.
  - **False Positives (FPs):** A primary cause is high topical overlap without true equivalence. An example is the pair of sequential physics courses PHYS-4D and PHYS-4A from Foothill College, which cover modern and classical mechanics, respectively.
  - **False Negatives (FNs):** A primary cause is semantic divergence in descriptions, where officially equivalent courses are described with vastly different terminology. An example is a pair of CDEV-100 courses, one described with language from developmental psychology ("milestones") and the other with language from social justice pedagogy ("anti-bias curriculum"). In some cases, errors were traced to data quality issues, such as a course labeled SOCI-125 (statistics) having a description for an LGBTQ+ studies class.

Chapter 5: Discussion, Future Work, and Conclusion

- **5.1 Discussion of Results**
  - The proposed decoupled pipeline successfully addresses the privacy, scalability, and interpretability challenges of prior approaches.
  - The statistical superiority of the fine-tuned `bge-ft` model, even over much larger models, provides powerful evidence that for specialized domains, targeted adaptation is more effective than sheer scale.
  - The primary bottleneck for achieving near-perfect automation has now shifted from being model-centric to data-centric.

- **5.2 Limitations of the Current Study**
  - The framework's performance is fundamentally capped by the quality and content of the public course descriptions.
  - The fine-tuned `bge-ft` model is specialized for California's public college system and may not generalize to other systems without re-tuning.
  - The framework simplifies articulation into a binary classification and does not natively handle complex one-to-many or many-to-many articulation rules.

- **5.3 Future Work**
  - The most critical future work involves data-centric strategies, such as developing an interactive, human-in-the-loop system for expert review.

- Active development is underway to evolve the framework into a full-scale course recommendation engine with a conversational interface.
- Future research could explore using graph neural networks to identify one-to-many and many-to-many relationships.

- **5.4 Conclusion**
  - This thesis confronted the challenge of manual course articulation by designing, developing, and validating a novel computational framework using only publicly available data.
  - The work's primary contribution is a privacy-preserving, scalable, and computationally efficient pipeline achieved through two key innovations: the application of deep metric learning to create a bespoke embedding model, and the design of a novel composite distance vector.
  - The result is a practical tool that can help institutions reduce administrative workload and foster a more transparent and equitable 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.

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

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
(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