

Chapter 1

Introduction

California's public higher education system is a titan of American academia, a complex, three-tiered structure comprising the University of California (UC), California State University (CSU), and the California Community Colleges (CCC) [11]. Collectively, these 149 colleges and universities serve nearly 2.9 million students, forming the largest public higher education system in the United States [11, 22, 7, 6]. A foundational principle of this system is the promise of student mobility, particularly the pathway from a two-year community college to a four-year university. However, the mechanism designed to facilitate this movement, the process of determining course equivalency, or articulation, is a formidable, largely manual process that creates significant barriers for students.

At the heart of this process is the Articulation System Stimulating Interinstitutional Student Transfer (ASSIST), the state's official public repository for articulation agreements [4]. While ASSIST provides a centralized platform for students and advisors to view established

equivalencies, it is fundamentally a display for agreements that are negotiated and updated manually by articulation officers at each individual campus [3]. Given the sheer number of institutions, the process of defining and maintaining these agreements is a task of bleak combinatorics, rendering it inefficient, slow, and inherently intractable [17]. This manual paradigm places a considerable burden on academic advisors and administrative staff, who must meticulously review course descriptions and syllabi to compare content, rigor, and learning outcomes [17]. The result is a system that struggles to keep pace with the needs of a vast and mobile student body, making California a critical case study for a problem that extends far beyond its borders.

The challenges exemplified by California's system are a microcosm of a systemic crisis in American higher education. The act of transferring between institutions has become a normative component of the modern student's academic journey. Data from the National Student Clearinghouse Research Center reveals that in the fall of 2023, transfer enrollment constituted 13.2% of all continuing and returning undergraduates [8]. This trend is not static; it represents a post-pandemic resurgence in student mobility, with transfer enrollment growing by 5.3% from Fall 2022 to Fall 2023 and an additional 4.4% in Fall 2024 [8, 9]. This mobile population is increasingly diverse, comprising not only students following traditional two-year to four-year pathways but also a substantial number of returning learners who have previously paused their education. Over half of these returning students opt to re-enroll at a new institution, underscoring the critical role of the transfer system in providing flexible pathways to degree completion [10].

The consequences of this systemic inefficiency are borne almost entirely by the students, manifesting in significant academic and financial setbacks. The most direct and damaging outcome is the loss of earned academic credit. A comprehensive 2017 report by the U.S. Government Accountability Office (GAO) estimated that students who transferred between 2004 and 2009 lost an average of 43% of their credits in the process [27]. This finding is echoed across numerous studies, with reports indicating that more than half of all transfer students lose at least some credits, and approximately one-fifth are forced to repeat courses for which they have already received a passing grade at a previous institution [1].

This loss of credit creates a cascade of negative consequences. It invariably leads to an increased time-to-degree, delaying graduation and entry into the workforce. Each repeated course also carries a financial cost, increasing the total tuition burden and potentially exhausting a student's eligibility for federal financial aid programs like Pell Grants and Direct Loans [27]. A process that is often undertaken to save money (for example, by starting at a less expensive community college) can paradoxically result in a greater overall financial commitment, trapping students in a cycle of additional coursework and debt [5].

The friction and frustration inherent in the transfer process also have a measurable impact on student persistence and graduation. Studies have shown that transfer students, as a group, tend to have lower retention and graduation rates than their peers who begin and end their studies at the same institution [21]. This issue transcends mere administrative inefficiency and becomes a critical matter of educational equity. Low-income students and students from historically underrepresented racial and ethnic groups are more likely to

begin their postsecondary journey at community colleges and rely on transfer pathways to attain a bachelor's degree [26]. The recent growth in transfer enrollment has been driven disproportionately by Black and Hispanic students [8]. Therefore, the barriers imposed by an inefficient articulation system such as credit loss, increased cost, and delayed graduation, disproportionately harm the very student populations that institutions are striving to support.

A clear and troubling feedback loop emerges from this analysis. The fundamentally manual and inefficient nature of course articulation is a direct cause of credit loss. This credit loss imposes a tangible academic and financial burden on students which falls most heavily on underrepresented and low-income students, who are a large and growing segment of the transfer population. This disproportionate impact, in turn, undermines institutional goals of improving student retention and closing persistent equity gaps in degree attainment. Thus, the seemingly low-level administrative task of determining course equivalency is revealed to be a significant driver of systemic inequity in higher education. Addressing this challenge through robust automation is not merely an operational optimization; it is a necessary intervention to foster a more equitable, efficient, and supportive educational ecosystem for all students.

1.1 Contributions

This research makes several key contributions to the fields of educational data mining and natural language processing, offering a practical and powerful solution to the long-standing challenge of course articulation.

Primary Contributions

1. **A High-Accuracy, Automated Framework:** This thesis develops and validates a novel framework for determining course equivalency that achieves state-of-the-art accuracy, with F_1 -scores exceeding 0.99 on a challenging real-world dataset. Crucially, it accomplishes this using only publicly available course catalog text, making it broadly applicable.
2. **An Innovative Feature Engineering Technique:** It introduces a composite distance vector, Δ_c , that uniquely combines element-wise embedding differences with cosine similarity. This technique provides a richer input signal for classification and is shown to demonstrably improve the performance of downstream machine learning models, particularly linear classifiers.
3. **A Computationally Efficient and Scalable Approach:** The research demonstrates that by decoupling semantic representation from classification, it is possible to harness the power of deep contextual embeddings without the high computational costs, API dependencies, and opaque nature of direct LLM-based classification. This

makes the proposed solution more efficient, scalable, and practical for institutional deployment.

4. **A Privacy-Preserving Methodology:** By relying exclusively on public course descriptions, the proposed method circumvents the significant privacy, security, and data access challenges associated with techniques that require sensitive student enrollment records. This makes the framework more ethically sound and generalizable across any pair of institutions, regardless of their data-sharing agreements.

Thesis Roadmap

The remainder of this thesis is structured to provide a comprehensive account of this research.

1. **Chapter 2: Background and Related Work** will provide a detailed in-depth survey of the landscape of student transfer automation and the evolution of technological interventions.
2. **Chapter 3: Methodology** will offer a deep dive into the data collection and preparation processes, the specific embedding models evaluated, the construction of the feature vectors, and the theoretical underpinnings of the machine learning classifiers employed.
3. **Chapter 4: Experimental Setup and Results** will detail the experimental design, the datasets used for training and validation, and a comprehensive analysis of the

classification performance, including ablation studies and model comparisons.

4. **Chapter 5: Discussion and Future Work** will interpret the results in a broader context, discuss the limitations of the current study, and outline promising avenues for future research, including the development of a full-scale course recommendation system and the exploration of fine-tuning techniques.
5. **Chapter 6: Conclusion** will summarize the key findings of the thesis and reiterate the significance of its contributions to both academic research and the practical administration of higher education.

Chapter 2

Related Works

The manifest inefficiencies and inequities of manual course articulation, exemplified by the challenges within California’s vast system, have prompted a range of research efforts aimed at automating the process [15, 20, 16, 13, 28]. These technological interventions have evolved in sophistication, mirroring the broader advancements in natural language processing (NLP) and machine learning. A critical review of this literature reveals a clear trajectory from simple statistical methods to complex deep learning models, with each stage introducing new capabilities while also exposing new limitations. This evolution illuminates the path toward a more robust and scalable solution.

Foundational Approaches: Keyword and Statistical Methods

The earliest attempts at automating course comparison relied on foundational text analysis techniques that, while computationally simple, lack semantic depth. The most basic

systems are essentially search engines or databases that depend on exact keyword matching or pre-populated tables of known equivalencies [14]. These systems are inherently brittle; they cannot recognize semantic variations (e.g., equating “Introduction to Programming” with “Fundamentals of Computer Science I”) and require continuous manual updates to remain relevant [24].

A more advanced statistical method, Term Frequency-Inverse Document Frequency (TF-IDF), improves upon keyword matching by vectorizing documents and weighting terms based on their importance. A term’s frequency within a single document (TF) is balanced against its rarity across a collection of documents, or corpus (IDF) [2]. This allows the model to assign higher importance to distinctive terms (e.g., “calculus”) and lower importance to common words (e.g., “the,” “a,” “is”) [2]. TF-IDF has been a workhorse for information retrieval and has been applied to course similarity tasks [2]. However, the fundamental limitation of TF-IDF and other bag-of-words models is their complete lack of semantic understanding. They treat words as discrete, unrelated tokens and cannot grasp that “calculus” and “differentiation” are related concepts, nor can they distinguish between different meanings of the same word.

Static Semantic Representations

The development of word embeddings represented the first major leap toward a true semantic understanding of course content. Models like Word2Vec and GloVe, trained on

vast text corpora, learn to represent words as dense vectors in a high-dimensional space, where words with similar meanings are positioned closer to one another. For example, the vectors for “car” and “automobile” would be near each other, while being distant from the vector for “planet.” This innovation enabled a more nuanced comparison of texts than was possible with TF-IDF. In the context of educational data mining, these techniques have been applied to content-based course recommendation systems, typically by creating a single vector representation for a course description by averaging the vectors of all its constituent words [18].

Despite this advancement, these models produce static embeddings. Each word is assigned a single, fixed vector regardless of its context. This is a significant drawback, as it fails to account for polysemy: words with multiple meanings. For instance, the word “bank” would have the same vector in the phrases “river bank” and “bank account,” despite their disparate meanings. Furthermore, the common practice of averaging all word vectors to create a document-level representation is a crude heuristic that can dilute or lose critical semantic information, especially in complex or lengthy descriptions.

Contextual Semantic Representations

The introduction of the transformer architecture, and specifically models like Bidirectional Encoder Representations from Transformers (BERT), revolutionized NLP by enabling the generation of contextual embeddings. In these models, the vector representation of a

word is dynamically influenced by the words surrounding it in a sentence [12]. This allows the model to disambiguate word meanings and capture a much richer, more accurate semantic representation of the text. Architectures such as Sentence-BERT (SBERT) were subsequently developed to fine-tune these models specifically for the task of producing semantically meaningful embeddings for entire sentences or short paragraphs, which can then be efficiently compared using metrics like cosine similarity [23].

The most recent evolution in this domain involves the direct application of large-scale generative models, or Large Language Models (LLMs) like GPT-4 and Gemini, for classification tasks. Through sophisticated prompt engineering and in-context learning, these models can be instructed to perform pairwise comparisons of course descriptions and render a judgment on their equivalency. The preliminary work for this thesis explored this very approach, using Google’s Gemini Pro to classify course pairs. While these experiments yielded promising accuracy, they also uncovered significant practical and theoretical limitations. The direct use of LLMs for this task is computationally expensive and inefficient, as it requires repeatedly sending full text descriptions to a model API for every comparison. Performance is acutely sensitive to the exact phrasing of the prompt, necessitating a costly and time-consuming iterative tuning process. Most critically, the decision-making process of an LLM is a “black box,” providing a categorical output (e.g. “equivalent”) without a quantifiable similarity score or confidence level. This opacity makes it difficult to rank potential matches, set decision thresholds, or provide transparent justifications for the model’s conclusions. This approach is also ill-suited for handling more complex articulation scenarios,

such as one-to-many or many-to-many course mappings.

Enrollment-Based Approaches

It is essential to situate this research within the context of parallel efforts that leverage different data sources. A notable body of work has demonstrated that course similarity and prerequisite relationships can be predicted with high accuracy by analyzing student enrollment data. Models such as *course2vec* learn course embeddings not from their textual descriptions, but from the patterns of which courses students tend to take together [19]. The underlying principle is that courses frequently taken in the same semester or in sequence likely share a functional or topical relationship [19].

While this behavioral approach is powerful, its reliance on large-scale, proprietary institutional datasets of student records presents two major obstacles. First, it raises significant data privacy and security concerns, as it involves the analysis of sensitive student information [25]. Second, it limits the scalability and generalizability of the solution. The model is only applicable at institutions that can provide access to such data, and it cannot be used to compare courses between two institutions that have no history of student transfer between them. This approach is therefore not a universal solution for the broader course articulation problem.

The evolution of these varied approaches reveals a fundamental trade-off: as models gain greater semantic power, they tend to become more computationally intensive, less inter-

pretable, and more demanding of specialized or private data. The limitations of direct LLM classification (cost, opacity) and enrollment-based methods (data privacy, limited access) point toward the need for a new paradigm. An ideal solution should harness the semantic power of large pre-trained models without inheriting their operational burdens. This suggests that the next logical step is not simply a larger, more complex end-to-end model, but rather a more intelligent, hybrid framework. Such a framework would decouple the task of deep semantic representation from the task of final classification, allowing each component to be optimized for what it does best. This conceptual shift forms the central motivation for the methodology proposed in this thesis. Table 2.1 summarizes the primary methods for determining course transferability.

Table 2.1: Comparative Taxonomy of Course Equivalency Determination Methods

Approach	Key Characteristics	Data Source(s)	Semantic Capability	Strengths	Limitations
Manual Review	Human experts (advisors, faculty) compare syllabi descriptions.	Course Catalogs, Syllabi	High (Human-level)	Nuanced, context-aware, trusted by faculty.	Extremely slow, not scalable, subjective, prone to inconsistency.
Keyword/TF-IDF	Bag-of-words representation, statistical term weighting.	Course Catalogs	None to Low	Simple, computationally cheap, easy to implement.	Fails to capture synonyms, context, or true semantic meaning
Static Embeddings (Word2Vec/GloVe)	Pre-trained word vectors, often averaged for document representation.	Course Catalogs	Medium	Captures word-level semantics, better than TF-IDF.	Context-insensitive, averaging vectors loses information.
Enrollment-Based (e.g., course2vec)	Embeddings learned from student co-enrollment patterns.	Proprietary Student Records	High (Behavioral)	Captures functional relationships between courses, highly predictive.	Requires access to sensitive private data, not generalizable, privacy concerns.
Direct LLM Classification	End-to-end classification using prompt engineering.	Course Catalogs	Very High	High accuracy potential, understands complex language.	Computationally expensive, “black box” opacity, prompt sensitive, no quantifiable similarity score, risk of hallucinations.
Proposed Method (Embeddings + ML)	Deep contextual embeddings as features for traditional classifiers.	Course Catalogs	Very High	State-of-the-art accuracy, computationally efficient, quantifiable, uses public data only.	Relies on the quality of the pre-trained embedding model.

Bibliography

Bibliography

- [1] Public Agenda. *Beyond Transfer: Insights from a Survey of American Adults*. <https://publicagenda.org/resource/beyond-transfer/> (visited on 06/30/2025).
- [2] Akiko Aizawa. “An information-theoretic perspective of tf-idf measures”. In: *Information Processing & Management* 39.1 (2003), pp. 45–65. ISSN: 0306-4573. DOI: [https://doi.org/10.1016/S0306-4573\(02\)00021-3](https://doi.org/10.1016/S0306-4573(02)00021-3). <https://www.sciencedirect.com/science/article/pii/S0306457302000213>.
- [3] ASSIST. *Frequently Asked Questions*. <https://resource.assist.org/FAQ> (visited on 06/30/2025).
- [4] ASSIST. *General Information*. <https://resource.assist.org/About/General-Information> (visited on 06/30/2025).
- [5] Leticia Tomas Bustillos et al. *The Transfer Maze: The High Cost to Students and the State of California*. The Campaign for College Opportunity, Sept. 17, 2017.

- [6] California Community Colleges Chancellor's Office. *Management Information Systems Data Mart*. 2024. https://datamart.cccco.edu/Students/Student%5C_Headcount%5C_Term%5C_Annual.aspx (visited on 09/13/2024).
- [7] California State University Office of the Chancellor. *Enrollment*. 2024. <https://www.calstate.edu/csu-system/about-the-csu/facts-about-the-csu/enrollment> (visited on 09/13/2024).
- [8] National Student Clearinghouse. *College Transfer Enrollment Grew by 5.3% in the Fall of 2023*. <https://www.studentclearinghouse.org/news/college-transfer-enrollment-grew-by-5-3-in-the-fall-of-2023/> (visited on 06/30/2025).
- [9] National Student Clearinghouse. *College Transfer Enrollment Grew for Third Straight Year*. <https://www.studentclearinghouse.org/news/college-transfer-enrollment-grew-for-third-straight-year/> (visited on 06/30/2025).
- [10] National Student Clearinghouse. *DATA DIVE: Returning Learners Lead Transfer Population*. <https://www.studentclearinghouse.org/nscblog/data-dive-returning-learners-lead-transfer-pop/> (visited on 06/30/2025).
- [11] Kevin Cook. *California's Higher Education System*. 2024. <https://www.ppic.org/publication/californias-higher-education-system/> (visited on 09/06/2024).

- [12] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. <https://arxiv.org/abs/1810.04805>.
- [13] Weijie Jiang and Zachary A Pardos. “Evaluating Sources of Course Information and Models of Representation on a Variety of Institutional Prediction Tasks.” In: *International Educational Data Mining Society* (2020).
- [14] Shamrock Solutions LLC. *Transfer Credit Automation: How Universities Are Simplifying Course Equivalency*. Overland Park, KS, USA. <https://www.shamrock-solutions-llc.com/post/transfer-credit-automation-universities> (visited on 06/30/2025).
- [15] H Ma et al. “Course recommendation based on semantic similarity analysis”. In: *2017 3rd IEEE International Conference on Control Science and Systems Engineering*. 2017, pp. 638–641.
- [16] Z. A Pardos, H Chau, and H Zhao. “Data-assistive course-to-course articulation using machine translation”. In: *Proceedings of the Sixth Conference on Learning@ Scale*. 2019, pp. 1–10.
- [17] Zachary Pardos, Hung Chau, and Haocheng Zhao. “Data-Assistive Course-to-Course Articulation Using Machine Translation”. In: (July 2019). DOI: 10.1145/3330430.3333622.

- [18] Zachary A. Pardos, Hung Chau, and Haocheng Zhao. “Data-Assistive Course-to-Course Articulation Using Machine Translation”. In: *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*. L@S ’19. Chicago, IL, USA: Association for Computing Machinery, 2019. ISBN: 9781450368049. DOI: 10.1145/3330430.3333622. <https://doi.org/10.1145/3330430.3333622>.
- [19] Zachary A. Pardos, Zihao Fan, and Weijie Jiang. *Connectionist Recommendation in the Wild: On the utility and scrutability of neural networks for personalized course guidance*. 2018. arXiv: 1803.09535 [cs.AI]. <https://arxiv.org/abs/1803.09535>.
- [20] Zachary A. Pardos, Zihao Fan, and Weijie Jiang. “Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance”. In: *User Modeling and User-Adapted Interaction* 29.2 (Apr. 2019), pp. 487–525. ISSN: 0924-1868. DOI: 10.1007/s11257-019-09218-7. <https://doi.org/10.1007/s11257-019-09218-7>.
- [21] Stephen Porter. “Assessing Transfer and Native Student Performance at Four-Year Institutions”. In: *39th Annual Forum of the Association for Institutional Research*. June 1999.
- [22] Regents of the University of California, The. *Fall enrollment at a glance*. 2024. <https://www.universityofcalifornia.edu/about-us/information-center/fall-enrollment-glance> (visited on 09/13/2024).

- [23] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: 1908.10084 [cs.CL]. <https://arxiv.org/abs/1908.10084>.
- [24] Natenaile Asmamaw Shiferaw et al. *BERT-Based Approach for Automating Course Articulation Matrix Construction with Explainable AI*. 2024. arXiv: 2411.14254 [cs.LG]. <https://arxiv.org/abs/2411.14254>.
- [25] Sharon Slade and Paul Prinsloo. “Learning Analytics: Ethical Issues and Dilemmas”. In: *American Behavioral Scientist* 57.10 (2013), pp. 1510–1529. DOI: 10.1177/0002764213479366. eprint: <https://doi.org/10.1177/0002764213479366>. <https://doi.org/10.1177/0002764213479366>.
- [26] The National Task Force on the Transfer and Award of Credit. *Reimagining Transfer for Student Success*. Report to Congressional Requesters. American Council on Education, Mar. 2020. <https://www.gao.gov/products/gao-17-574>.
- [27] United States Government Accountability Office. *Higher Education: Students Need More Information to Help Reduce Challenges in Transferring College Credits*. Report to Congressional Requesters GAO-17-574. United States Government Accountability Office, Aug. 14, 2017. <https://www.gao.gov/products/gao-17-574>.
- [28] Yinuo Xu and Zach A. Pardos. “Extracting Course Similarity Signal using Subword Embeddings”. In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. LAK ’24. Kyoto, Japan: Association for Computing Machinery, 2024, pp. 857–

863. ISBN: 9798400716188. DOI: 10.1145/3636555.3636903. <https://doi.org/10.1145/3636555.3636903>.