Clustering Beyond the Classroom: NLP-Driven Insights into Curriculum Standards K-12

Angeline Dorvil
CAI-5307 Natural Language Processing
University of South Florida

Abstract

Curriculum standards serve as foundational guidelines for K-12 education, yet their complexity and variability across states and school districts pose significant challenges for educators and policymakers. This study investigates the application of transformer-based language models, DistilBERT, RoBERTa, and GPT-2, to classify and cluster curriculum standards into meaningful grade-level groupings. Multiple clustering schemes, ranging from fine-grained single-grade distinctions to broader, content-based categories are explored. The results reveal that while granular classifications are difficult and yield low accuracies, more conceptually aligned or aggregated clusters facilitate near-perfect classification performance. Moreover, all three models exhibit remarkably similar results under coarse-grained schemes, suggesting that the inherent structure of curriculum content outweighs architectural differences between models. These findings have implications for how curriculum standards are developed, highlighting the potential benefits of organizing standards into more coherent groups that reflect natural learning progressions. This project contributes to the growing body of literature in educational NLP and offers a computational perspective on curriculum design, potentially guiding future reforms and enhancements in educational policy and practice.

1 Introduction

Curriculum standards serve as foundational K-12 educational goals, typically determined at the state level, to guide school districts and educators in designing effective lesson plans and semester long curricula. Although these standards outline what students should achieve by the end of certain grade clusters (e.g., K-5 before entering middle school), local educational bodies and individual teachers retain considerable flexibility in how these standards are met. Further complicating matters, states employ different methods to assess student progress. For instance, Florida, which formerly relied on the Florida Standards Assessments (FSA), has transitioned to the Florida Assessment of Student Thinking (FAST) aligned with the B.E.S.T. standards, an example of the evolving, dynamic landscape of curriculum evaluation.

This diversity and complexity create significant challenges in ensuring consistent, appropriate grade level alignment of curriculum standards. In many cases, standards overlap in content and difficulty across adjacent grades, making single-grade classification less reliable. As a result, grouping multiple grades into broader clusters (e.g., Elementary, Middle School, High School) or organizing them by developmental proximity offers a more practical approach. Applying Natural Language Processing (NLP) techniques can help educators and policymakers systematically classify and align standards across various states and organizations. By analyzing large, heterogeneous datasets, NLP based models can help identify patterns, streamline curriculum planning, and ensure that students acquire the required skills at the appropriate educational stage.

While recent educational NLP research has emphasized tasks such as reading comprehension, question generation, grammatical error correction, and essay scoring fields highlighted in recent ACL

SIGEDU workshops, direct classification of curriculum standards by grade level or cluster remains underexplored. To address this gap, this project compares the performance of several transformer-based language models (DistilBERT, RoBERTa, and GPT-2) on classifying English Language Arts (ELA) curriculum standards at varying levels of granularity. Specifically, it systematically examines multiple clustering schemes, from broad educational phases to more narrowly defined groups and individual grades. Through this exploration, the project aims to identify the most effective models and approaches for curriculum classification, ultimately laying a foundation for future research in educational NLP and more informed curriculum development.

2 Related Previous Research

While prior studies have applied NLP methods to educational settings, most efforts focus on areas like lesson planning, question generation, or content adaptation rather than directly classifying curriculum standards by grade level. For example, some recent research explores how middle school teachers can co-design and co-teach curricula that integrate NLP generative AI technologies in STEM classrooms (Kwon & Kim, 2024; Katuka et al., 2024). Although these approaches demonstrate the usefulness of NLP in guiding instructional strategies, they do not directly address the challenge of aligning educational standards across multiple grades or states.

More broadly, research on curriculum design has examined approaches at various educational levels, but they have often not focused on the unique complexities of K-12 environments. Many existing studies center on higher education or professional training curricula, with the assumption that these methodologies might generalize to K-12 settings. However, the lack of direct K-12 curriculum classification research means current methods remain insufficient for tackling the heterogeneity of standards that vary across states and grades.

This gap in the literature reinforces the need for an approach that applies NLP models to K–12 curricula. By systematically classifying educational standards and examining different clustering schemes, this project aims to address the current limitations, ultimately improving the alignment and coherence of instructional guidance within and across diverse educational contexts.

3 Methodology

3.1 Data Collection and Preprocessing

This study utilizes English Language Arts (ELA) curriculum standards sourced as JSON files from OpenSALT, an open-source platform designed for managing and hosting competency frameworks and academic standards. OpenSALT provides aligned, standards-based data that can be readily integrated into NLP pipelines. Although various states and organizations often publish curriculum standards in diverse formats, including PDF and proprietary structures, this project focused exclusively on ELA standards available in JSON format. Approximately 25 state-level ELA curriculum datasets and an additional comprehensive JSON file (covering all 50 states from a previously constructed curriculum database) were incorporated, resulting in a combined dataset size of 37,005 entries.

From this initial dataset, standards without associated grade information were removed, yielding a cleaned dataset of 36,918 samples. Each curriculum standard text was stripped of extraneous whitespace and confirmed to be free of problematic characters. Because the JSON files were already relatively clean, no additional extensive text normalization was required. To ensure consistent labeling, grades such as “KG,” “01,” and “02” were mapped to numeric codes (e.g., Kindergarten → 0, Grade 1 → 1, Grade 2 → 2) and further aggregated into clusters where necessary. For instance, multiple lower grades might be grouped into a single developmental cluster, allowing for both fine-grained (individual grade) and coarse-grained (cluster-based) classification tasks. After label standardization, the dataset maintained a relatively balanced distribution across the primary labels, with counts of ~12,000 entries for each developmental cluster (K-4, 5-8, 9-12), ensuring no severe imbalance issues at this stage.

All text was tokenized using Hugging Face tokenizers, and sequence lengths were verified to remain within the maximum token limits for the chosen models (512 tokens for DistilBERT and RoBERTa, and 1024 for GPT-2), making chunking unnecessary. The final datasets were split into training, validation, and test sets to support model fine-tuning and evaluation.

3.2 Clustering Schemes

To capture the inherent complexity and overlap in K-12 curriculum standards, multiple clustering schemes were explored. The **content-based** clustering reflects a grouping of grades that are substantially similar in their curricular content, recognizing that certain skills and competencies develop over multiple adjacent grade levels (e.g., K-4, 5-8, 9-12). In contrast, the **standard school division** clustering follows common U.S. educational structures, grouping grades into standard school divisions (e.g., K-5, 6-8, 9-12) that are frequently observed in public school systems. Beyond these conventional groupings, two additional schemes were introduced to examine real-world constraints: **two-grade clusters** and **three-grade clusters**. These scenarios simulate classrooms where teachers must address multiple grades simultaneously, either due to overlapping material or practical constraints such as staffing shortages, resulting in instruction that spans two or three consecutive grades. Finally, a **single-grade** classification scheme was tested to gauge the feasibility of distinguishing each grade's standards in isolation. While this fine-grained approach is theoretically appealing, it often yielded poor accuracy due to the substantial overlap in curricular content across neighboring grades. Collectively, these clustering schemes allowed for a more nuanced understanding of how curricular standards are structured and how they might be more effectively aligned or redesigned.

3.3 Model Architecture

Three transformer-based models, **DistilBERT**, **RoBERTa**, and **GPT-2** were employed to investigate the effectiveness of different architectures for curriculum standard classification. DistilBERT, a lightweight and distilled version of BERT, served as an accessible baseline that required relatively fewer computational resources. RoBERTa, known for its robust optimization of BERT's pretraining, was incorporated to test a more advanced model architecture with improved masking strategies and typically stronger performance. Meanwhile, GPT-2, originally designed as a generative model, offered a contrasting approach to classification, allowing exploration of whether a generative transformer could match or surpass the performance of encoder-based models when adapted for a classification task.

All models were fine-tuned using Hugging Face's sequence classification frameworks, which added a linear classification head atop the pretrained Transformer layers. This classification layer, initialized with random weights, mapped the model's contextualized representations to the target classes (e.g., three-grade clusters). For GPT-2, which lacks a native padding token, the *pad_token* was explicitly set to the *eos_token* to facilitate batch processing. Additionally, GPT-2's larger maximum sequence length (1024 tokens) was leveraged to accommodate longer curriculum statements without the need for chunking.

Default parameters for dropout, initialization, and optimization were largely retained, while a few hyperparameters were adjusted based on common fine-tuning practices. Learning rates differed slightly between models (e.g., 2e-5 for DistilBERT and RoBERTa, 5e-5 for GPT-2), and a consistent batch size of 16 was used to balance training stability with available GPU memory. Each model was fine-tuned for up to 12 epochs using the Hugging Face Trainer API, with evaluation after every epoch and best-model checkpointing enabled. This setup ensured a standardized training environment that allowed for fair comparisons across architectures.

3.4 Training Procedure

The dataset was split into training, validation, and test sets to ensure robust model evaluation and reduce the risk of overfitting. First, 20% of the data was reserved for testing, while the remaining 80% was designated as a provisional training set. From this training portion, an additional 10% was split off to form the validation set. This process, implemented using the Hugging Face Dataset library and a fixed random seed (42), ensured reproducibility and resulted in an approximate 72%/8%/20% split for training, validation, and testing, respectively.

Prior to model training, the text data was tokenized using a model-specific tokenizer (e.g., RoBERTa's tokenizer), with padding and truncation employed to align sequences to a uniform length (up to 512 tokens in this example). Because curriculum statements were generally within the maximum token limit, chunking was not necessary. The labeled datasets were then converted into PyTorch tensors to be compatible with the training pipeline.

Model fine-tuning was conducted using the Hugging Face Trainer API, which simplified the training loop, evaluation scheduling, logging, and checkpoint management. Each model, DistilBERT, RoBERTa, or GPT-2 was fine-tuned for up to 12 epochs, with a fixed learning rate (as stated in Model architecture section) and a batch size of 16. Evaluation was performed at the end of each epoch to monitor validation metrics, and the best-performing model checkpoint, determined by highest validation accuracy, was automatically saved. Performance was primarily assessed using accuracy, with additional metrics like precision, recall, and F1-score computed for a more comprehensive evaluation. This systematic, reproducible approach ensured that model comparisons were fair and grounded in consistent experimental conditions.

3.5 Evaluation Metrics

A range of metrics were employed to thoroughly evaluate model performance at various stages of the training and testing process. During fine-tuning, both training and validation loss were tracked to monitor convergence and detect potential overfitting, while validation accuracy offered an immediate measure of how well each model performed on unseen data at the end of each epoch. After completing fine-tuning, a held-out test set provided a final, unbiased assessment of generalization capabilities. On this test set, accuracy served as a straightforward indicator of overall predictive correctness, while precision, recall, and F1-score, computed using the classification_report function from scikit-learn offered a more detailed understanding of model strengths and weaknesses across individual classes. In some cases, confusion matrices were generated to visualize misclassifications and identify patterns in the errors. Additionally, raw class probabilities (obtained via softmax outputs) were examined to gauge model confidence. This multi-faceted evaluation strategy ensured that both the global performance and class-level nuances were captured, ultimately guiding the interpretation of results and informing subsequent refinements in modeling approaches.

4 Results

4.1 Overview of Metrics

Figure 1 summarizes the performance of all three models, DistilBERT, RoBERTa, and GPT-2, across the five clustering schemes: single-grade, two-grade, three-grade, standard-based, and content-based. Notably, each model produced the same metrics within a given clustering configuration, rendering differences between architectures negligible for this dataset.

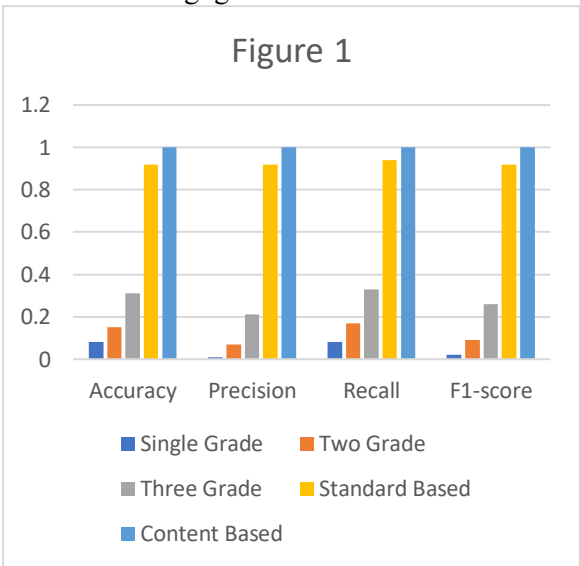

Figure 1: Performance metrics across clustering schemes. Each bar represents the average model performance (DistilBERT, RoBERTa, GPT-2) for a given clustering scheme. Single-grade, two-grade, and three-grade schemes struggle to achieve high performance, while the standard-based scheme and content-based scheme yield significantly better accuracy, precision, recall, and F1-score.

4.2 Overview of Performance

The results highlight a clear relationship between the granularity of the clustering scheme and classification accuracy. Fine-grained classifications (e.g., single-grade) were extremely challenging, with accuracy barely surpassing 8% and F1-scores around 2%. Performance improved incrementally when grades were combined into two-grade (15% accuracy) and three-grade (31% accuracy) clusters, suggesting that some coarse grouping aids classification. More substantial gains were observed in the standard-based scheme, where accuracy reached 92% with a corresponding F1-score of 92%. Finally, the content-based clustering achieved a perfect 100% accuracy and F1-score across all three models, demonstrating that broader, thematically aligned clusters are substantially easier to predict.

4.3 Preliminary Insights

These results underscore the complexity and overlap inherent in K-12 curriculum standards. The near-perfect performance at coarser levels of clustering suggests that curriculum standards might be more naturally organized into broad categories rather than strictly delineated grade boundaries. The uniformity of results across three distinct transformer architectures implies that model choice, under these conditions, is less critical than how the clusters are defined. Overall, this finding raises questions about the suitability of fine-grained, single-grade standards for practical classification tasks and hints that curriculum design or evaluation methods may benefit from more aggregated, conceptually aligned groupings.

Limitations

The results of this study shed light on the interplay between curriculum structure, model architectures, and classification performance. Notably, all three models, DistilBERT, RoBERTa, and GPT-2, exhibited nearly identical metrics when tasked with broader, more coherent clustering schemes. This consistency suggests that the inherent structure of the curriculum data and the models' capacity to capture general linguistic patterns overshadowed potential differences in model architectures. In other words, when the curriculum standards were grouped into content-based or standard-based clusters, even lightweight or generative-oriented models performed on par with more robust, encoder-based counterparts. This finding implies that the complexity of the model itself is less critical when the curriculum content is naturally more predictable.

However, the difficulty of fine-grained classification at the single-grade level highlights an ongoing challenge. Here, performance dropped sharply, indicating that the subtle distinctions between grades may not be explicitly encoded or easily inferred from the text of the standards. This limitation points to potential gaps in the standards' representation: if computational models struggle to differentiate closely related grades, it may reflect a lack of clarity in the standards themselves rather than a shortcoming of the models. These findings invite a deeper examination of how standards are authored and whether more explicit grade-level distinctions could make them more amenable to automated analysis.

The implications extend beyond algorithmic performance. If broader, conceptually aligned clustering schemes yield more reliable classification results, educators and policymakers might consider rethinking how standards are structured. Perhaps a more tiered approach, aligning closely with naturally occurring developmental or conceptual progressions, would facilitate both human interpretation and machine-driven analysis. Such changes could enhance the consistency and clarity of curricular materials, ultimately supporting more data-driven educational planning.

Despite these insights, this project has limitations. The models were fine-tuned without additional domain-specific pretraining, and the chosen dataset, though diverse, may not represent all curriculum structures or niche educational contexts. Moreover, the predefined clustering schemes used here are just one way to interpret these standards; exploring alternative groupings could yield further insights.

Future research can address these limitations and open new avenues for improvement. Incorporating domain-specific pretraining on educational corpora might enhance model sensitivity to subtle curricular nuances. Examining larger or more advanced models, such as GPT-4, could determine whether scaling model size or complexity helps resolve fine-grained classification challenges. Another important direction involves analyzing misclassifications in greater detail to identify patterns of ambiguity or inconsistency in the standards, potentially guiding revisions in curriculum design. By refining both the computational approaches and the structure of the curriculum standards themselves, it can create a more cohesive, data-driven understanding of educational frameworks.

Ethics Statement

This project utilized publicly available educational curriculum standards from state and district sources, ensuring that no personally identifiable information (PII) or sensitive data were included in the dataset. All data were obtained and used in accordance with the respective organizations' terms of use. The models employed (DistilBERT, RoBERTa, and GPT-2) are pretrained on large, publicly available corpora that may contain inherent biases. The findings presented

here should be interpreted in the context of improving curriculum design and alignment with computational models, rather than as critiques of specific frameworks or normative judgments about educational quality. Computations were conducted using Google Colab's GPU resources, with care taken to minimize environmental impact through efficient training practices, limited training epochs, and the careful selection of model configurations.

References

Marks, J., Ward, D. and Nadeau, G. (2024) *Competency frameworks*, *OpenSALT*. Available at: https://opensalt.net/cfdoc/ (Accessed: 23 November 2024).

Katuka, Gloria & Chakraburty, Srijita & Lee, Hyejeong & Dhama, Sunny & Earle-Randell, Toni & Celepkolu, Mehmet & Boyer, Kristy & Glazewski, Krista & Hmelo-Silver, Cindy & Mcklin, Tom. (2024). *Integrating Natural Language Processing in Middle School Science Classrooms: An Experience Report*. 639-645. 10.1145/3626252.3630881.

Kwon, K. & Kim, K. (2024). Exploring Middle School Teachers' Experience in Co-design and Co-teach an NLP-Generative AI focused curriculum in STEM classroom. In J. Cohen & G. Solano (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 2520-2522). Las Vegas, Nevada, United States: Association for the Advancement of Computing in Education (AACE). Retrieved December 13, 2024 from https://www.learntechlib.org/primary/p/224335/.

Messinger, S. (2012) *SIRFIZX/standards-data: JSON representations of the Common Core Standards*, *GitHub*. Available at: https://github.com/SirFizX/standards-data.git (Accessed: 28 November 2024).