DLSU Research Congress 2025
De La Salle University, Manila, Philippines
June 26 to 28, 2025

Automated Classification of Course Outcome–Program Outcome Alignment Using Transformer-Based Language Models

Gilfred Allen M. Madrigal[1,2] and Melvin K. Cabatuan[1]
[1] *De La Salle University, Manila*
[2] *Technological University of the Philippines, Manila*
**Corresponding Author: gilfred_madrigal@dlsu.edu.ph*

Abstract: The accurate alignment of Course Outcomes (COs) with Program Outcomes (POs) plays a critical role in ensuring curriculum quality, enabling continuous improvement, and fulfilling accreditation requirements in outcome-based education frameworks. Traditionally, this mapping process has been labor-intensive, subjective, and vulnerable to inconsistencies due to its reliance on expert judgment. To address these limitations, this study explores the automation of CO-PO alignment classification by leveraging the capabilities of transformer-based language models. A multi-class classification pipeline was developed and evaluated using BERT (Macro F1: 0.79), DistilBERT (Macro F1: 0.86), RoBERTa (Macro F1: 0.83), and Sentence-BERT with Logistic Regression (SBERT + LR, Macro F1: 0.81). The dataset consisted of carefully labeled CO-PO pairs derived from Commission on Higher Education (CHED) Memorandum Orders for Electrical, Electronics, and Mechanical Engineering programs in the Philippines. These textual CO-PO pairs were transformed into rich contextual embeddings using the selected transformer models. A dedicated classification layer was then fine-tuned to predict the alignment level, categorized as Introductory, Enabling, or Demonstrative. Model performance was rigorously evaluated using metrics such as accuracy, macro precision, recall, and macro F1-score. Among the models, DistilBERT performed the best across all alignment categories (Demonstrative F1: 0.90, Enabling F1: 0.84, Introductory F1: 0.85) and achieved a Macro Average AUC of 0.94. This paper emphasizes how transformer models may simplify the process of aligning CO-POs, reduce subjective judgment and human mistakes, and increase both consistency and efficiency, through which human error could be avoided. These developments could change curriculum planning and provide a road towards more consistent and high-quality teaching results in higher education establishments.

Key Words: Course Outcome–Program Outcome Alignment; Transformer-Based Language Models; Natural Language Processing (NLP); Automated Curriculum Mapping; Outcome-Based Education (OBE)

1. INTRODUCTION

Effective Course Outcome (CO) to Program Outcome (PO) mapping relies on maintaining curriculum continuity, enhancing quality, and aligning with the standards, including the Accreditation Board for Engineering and Technology (ABET). Recognizing the importance of streamlining this process, exploring efficient alternatives to traditional methods warrants attention from academic institutions. The conventional manual approach, however, is often labor-intensive, susceptible to subjective interpretation, and prone to errors (Chor et al., 2024). Regardless of its significance, the use of technology in CO-PO mapping is still somewhat understudied.

Recent advancements in Natural Language Processing (NLP), particularly through transformer-based models, offer promising opportunities for automating this process. Models such as Bidirectional Encoder Representations from Transformers (BERT), Distilled BERT (DistilBERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), and

Sentence-BERT (SBERT) can evaluate semantic similarity between outcome statements by generating deep contextual embeddings. These models use self-attention mechanisms, allowing them to understand language more effectively than traditional techniques like Term Frequency–Inverse Document Frequency (TF-IDF) and Support Vector Machines (SVM), as demonstrated in various NLP applications (Acheampong et al., 2020; Teng & Varathan, 2023; Özkurt, 2024).

BERT, introduced by Google in 2018, revolutionized NLP by using bidirectional training of transformer encoders, enabling it to consider both left and right context simultaneously. This is achieved through self-attention, which weighs the relevance of surrounding words to deeply understand meaning. Pre-trained on large corpora, BERT provides robust contextual embeddings that significantly improve performance in downstream tasks like classification and similarity analysis (Devlin et al., 2019; Özkurt, 2024).

Developed by Hugging Face, DistilBERT is a streamlined and faster version of BERT. Its main aim is to preserve much of BERT's language understanding prowess in a significantly smaller and more efficient model. This is accomplished using "knowledge distillation," a technique where a smaller "student" model (DistilBERT) learns from a larger, pre-trained "teacher" model (BERT). According to Sanh et al. (2019), DistilBERT's reduced number of layers and parameters makes it advantageous for applications with real-time processing needs or limited computational resources. Despite its smaller size, DistilBERT reportedly retains over 95% of BERT's language understanding abilities, establishing it as a compelling and practical option.

RoBERTa, a notable evolution from the original BERT, was developed by Facebook AI with a significant emphasis on refining the pre-training methodology, though it maintains BERT's fundamental Transformer architecture. This enhanced model incorporates several key improvements, including pre-training on a considerably larger and more varied dataset, a longer training period, the exclusion of the Next Sentence Prediction (NSP) task used in BERT, and the implementation of dynamic masking during the pre-training phase. These carefully considered modifications appear to contribute to more resilient and generally enhanced language representations when contrasted with the original BERT, frequently resulting in improved performance across a wider spectrum of downstream natural language processing tasks, as evidenced by the findings of (Liu et al., 2019).

SBERT enhances the original BERT model using a siamese and triplet network structure to compare better and understand sentence pairs. It encodes sentences into fixed-dimensional vectors, enabling efficient similarity computations using metrics like cosine similarity. Consequently, SBERT is advantageous for applications assessing semantic textual similarity, including clustering and ranking text. This efficiency makes it particularly useful for tasks like CO-PO mapping, where rapid and accurate comparison of numerous sentence pairs is essential (Reimers & Gurevych, 2019).

These transformer-based models have not been extensively investigated in terms of their potential to automate CO-PO mapping, despite their apparent potential. Conventional approaches have primarily relied on expert-driven heuristics or traditional machine learning algorithms, which do not possess the profound semantic comprehension required for optimal outcomes (Acheampong et al., 2021). This study aims to address that gap by developing an NLP pipeline that employs BERT, DistilBERT, RoBERTa, and SBERT to autonomously categorize CO-PO associations. The models were assessed using accuracy, precision, recall, and F1-score, while fine-tuning and embedding procedures were examined to determine the optimal method. This comprehensive assessment ensured a balanced evaluation of model performance across various aspects of predictive capability.

This research utilizes transformer models to advance educational outcome mapping, offering scalable and efficient methods for curriculum assessment and quality assurance in higher education.

2. METHODOLOGY

2.1 Research Design

This study used a supervised multi-class classification approach to automate CO-PO mapping with transformer-based language models. Each CO-PO pair was classified into one of three alignment levels: Introductory (I), Enabling (E), or Demonstrative (D), representing increasing levels of

competence and application.

To support this task, the dataset, summarized in Table 1, was directly derived from the Commission on Higher Education (CHED) Memorandum Orders (CMOs) that define the curricular standards for three engineering programs in the Philippines: Electronics Engineering (ECE), Electrical Engineering (EE), and Mechanical Engineering (ME). These official documents served as authoritative sources of CO and PO statements, providing a standardized structure aligned with national qualifications and accreditation frameworks.

Table 1. Dataset Composition and Partitioning

Program	CO-PO Pairs	Avg CO Length (Words)	Avg PO Length (Words)	Avg CO Length (Chars)	Avg. PO Length (Chars)	Training Samples (80%)	Testing Samples (20%)
ME	462	11.22	10.09	78.92	78.11	367	95
ECE	316	9.85	12.38	71.00	89.26	254	62
EE	233	11.74	11.76	82.94	90.60	187	46
Total	**1,011**	**≈ 10.91**	**≈ 11.19**	**≈ 77.37**	**≈ 84.47**	**808**	**203**

As shown in Table 1, a total of 1,011 CO-PO pairs were extracted, with a distribution of 462 pairs from ME, 316 from ECE, and 233 from EE. Each CO-PO pair was manually annotated with one of the three alignment labels—Introductory (I), Enabling (E), or Demonstrative (D)—based on information within the CMOs. The average length of the CO and PO texts varied slightly across the programs, with ME exhibiting average CO and PO text lengths of approximately 11.22 and 10.09 words, respectively, while ECE had averages of 9.85 and 12.38 words, and EE showed averages of 11.74 and 11.76 words. This dataset was subsequently partitioned into training (80%) and testing (20%) sets, resulting in 808 samples for training and 203 samples for evaluation. The stratified split ensured a proportional representation of the alignment labels across both subsets within each program (ME: 367 training, 95 testing; ECE: 254 training, 62 testing; EE: 187 training, 46 testing), providing a robust foundation for training and evaluating the automated alignment models.

Annotation was carried out following the CMO framework to ensure consistent labeling across samples. Preprocessing involved text normalization and the structured formatting of CO-PO pairs, separated by a [SEP] token to prepare the data for subsequent analysis. All procedures were implemented using Python in Google Colaboratory (Colab), an online cloud-based platform that supports efficient and scalable text processing through the integration of relevant libraries.

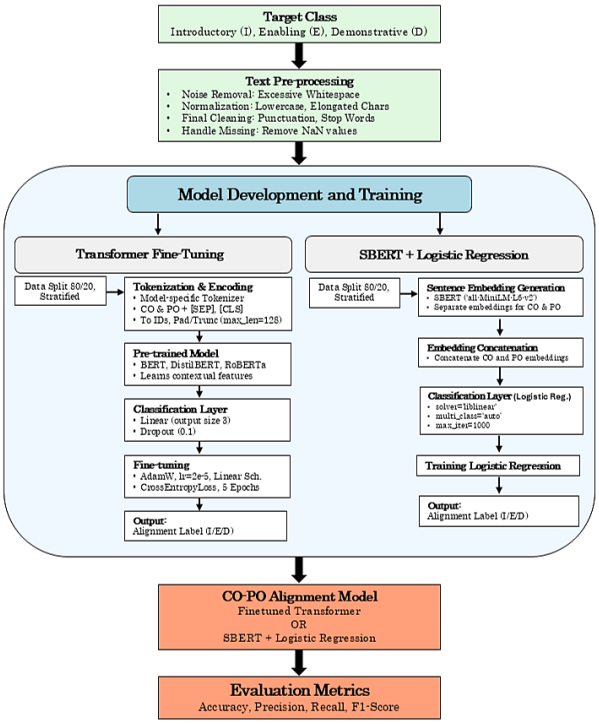

Fig. 1. Research Workflow for Automated CO-PO Alignment Classification

The research methodology, visually depicted in the block diagram (Fig. 1) presented above, details the essential phases of automated CO-PO alignment classification. Initially, the target alignment categories (I, E, or D) are defined, and a labeled CO-PO dataset undergoes Text Pre-processing. Subsequently, the workflow evaluates two distinct modeling strategies: (1) fine-tuning transformer architectures (BERT, DistilBERT, RoBERTa) with a classification layer, and (2) employing SBERT embeddings with a Logistic Regression classifier, both aimed at predicting the alignment category. The performance of these models is then assessed using standard classification metrics. The presented block diagram effectively encapsulates the key stages of our research design for automated CO-PO alignment classification.

2.2 Model Architecture

This study automated CO-PO alignment using four transformer-based models: BERT, DistilBERT, RoBERTa, and SBERT. These models were selected for their strong contextual understanding, which is essential for accurately classifying CO-PO pairs based on alignment levels.

Through bidirectional context, BERT yielded deep semantic insights. DistilBERT presented a faster, lighter substitute with like performance. RoBERTa enhanced generalization with longer pretraining. SBERT excelled in sentence-pair tasks, including CO-PO mapping with its Siamese network design.

CO-PO pairs were tokenized into subword units, then processed to generate contextual embeddings. These embeddings, capturing semantic relationships, were used by a fine-tuned classification head to predict the alignment level: I, E, or D.

2.3 Experimental Setup

Table 2. Sample CO-PO Pairs and Manual Mapping Labels

Program	Code	Course	CO Text	PO #	PO Text	Map
ECE	PHYSENG-M	Physics for Engineers	apply the Newton's Laws of motion	POa_ ECE	apply knowledge of mathematics and science to solve engineering problems	I
ME	BES7-ME-M	Mechanics of Deformable Bodies	apply the concepts of stress and strain	POa_ ME	apply knowledge of mathematics and science to solve complex mechanical engineering problems	E
EE	PEE7-M	EE Law, Codes, and Professional Ethics	discuss the existing laws, codes, and guidelines in the practice of the electrical engineering profession	POf_ EE	recognize ethical and professional responsibilities in engineering practice	D
ME	PME5-M	Industrial Plant Engineering	apply basic design concepts to industrial plants systems and equipment.	POa_ ME	apply knowledge of mathematics and science to solve complex mechanical engineering problems	D

Experiments ran on an Intel Core i7 12th Gen CPU and NVIDIA RTX 3050 Ti GPU using TensorFlow, PyTorch, and Hugging Face Transformers in Python. Preprocessing occurred in Google Colab, with an 80/20 train-test split.

Table 2 provides representative examples of CO to PO mappings, using three relationship levels: I, E, and D. These levels reflect the depth and maturity of outcome alignment across the curriculum.

In Physics for Engineers (ECE), with course code PHYSENG-M, the CO "apply the Newton's Laws of motion" is mapped as "I" to the PO "apply knowledge of mathematics and science to solve engineering problems," indicating that the course introduces essential concepts early in the program and aligns with fundamental scientific principles. Mechanics of Deformable Bodies (ME), with course code BES7-ME-M, which includes the outcome "apply the concepts of stress and strain," is categorized as "E". This suggests that the course builds theoretical and analytical skills in mechanics to support more advanced mechanical engineering program objectives. Courses such as EE Law, Codes, and Professional Ethics (EE), code PEE7-M, and Industrial Plant Engineering (ME), code PME5-M, are classified as "D". In EE Law, the ability to "discuss the existing laws, codes, and guidelines in the practice of the electrical engineering profession" directly assesses the PO to "recognize ethical and professional responsibilities in engineering practice." Similarly, in Industrial Plant Engineering, the CO to "apply basic design concepts to industrial plants systems and equipment" demonstrates a direct application of knowledge to solve complex mechanical engineering problems. These "D" level courses require students to apply their learning in more advanced or professional contexts, directly assessing PO achievement.

2.4 Ethical Considerations and Limitations

This study followed ethical guidelines by utilizing publicly available CHED CMOs for CO and PO data, avoiding privacy concerns. Expert-validated data was handled responsibly to prevent misuse.

However, limitations exist. The dataset focused on three engineering programs (BSECE, BSEE, BSME), which may affect generalizability across disciplines. While expert annotations added value, some subjectivity may have influenced labeling despite standardization efforts. The significant

computational requirements for training transformer models (BERT, DistilBERT, RoBERTa, SBERT) may restrict accessibility for resource-constrained institutions. Furthermore, performance may differ when utilized in domain-specific languages.

Automated CO-PO mapping presents a continuous challenge, necessitating persistent model refinement and careful monitoring for data bias. This study offers valuable insights into the application of machine learning for enhancing educational quality assurance and optimizing outcome alignment.

3. RESULTS AND DISCUSSION

3.1 Model Performance Evaluation

To evaluate the performance of the transformer-based models—BERT, DistilBERT, RoBERTa, and SBERT—this study utilized four widely recognized evaluation metrics: accuracy, precision, recall, and F1-score. These metrics assess different dimensions of classification effectiveness, offering a comprehensive view of each model's ability to predict the alignment between COs and POs. The definitions and formulas are outlined below:

Accuracy: Accuracy measures how often the model makes correct predictions overall. It accounts for both correctly identified positive and negative cases, giving a general sense of model effectiveness. It is mathematically expressed as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(Eq. 1)}$$

Here:

- TP (True Positives): Correctly predicted positive instances
- TN (True Negatives): Correctly predicted negative instances
- FP (False Positives): Incorrectly predicted positive instances
- FN (False Negatives): Incorrectly predicted negative instances

Precision: Precision evaluates a model's capacity to minimize false positives by quantifying the ratio of accurately predicted positive instances to the total number of cases classified as positive. Mathematically, it is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(Eq. 2)}$$

Recall: Recall quantifies the model's proficiency in detecting all truly positive instances. This metric is particularly important in scenarios where failing to identify a positive case (a false negative) carries significant consequences. Mathematically, it is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(Eq. 3)}$$

F1 Score: The F1 Score provides a harmonic mean of precision and recall, balancing both metrics to offer a single measure of a model's effectiveness, especially in the presence of class imbalance. It is expressed as:

$$\text{F1 Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad \text{(Eq. 4)}$$

These evaluation metrics were applied to the classification outputs to assess each model's capability in mapping COs to POs, ensuring a reliable analysis framework for handling complexity and data imbalance in educational assessments.

Table 3. Comparative performance of transformer-based models on CO-PO alignment classification.

Model	Acc.	Prec.	Rec.	F1
BERT	0.78	0.78	0.79	0.79
DistilBERT	**0.86**	**0.86**	**0.86**	**0.86**
RoBERTa	0.83	0.84	0.83	0.83
SBERT + LR	0.80	0.82	0.80	0.81

Acc. = Accuracy, Prec. = Macro Precision, Rec. = Macro Recall, F1 = Macro F1-score

Table 3 demonstrates that DistilBERT attained the highest overall performance, achieving an accuracy of 86% and a macro-averaged F1-score of 0.86. This reflects a robust balance between precision and recall across all alignment categories. RoBERTa subsequently exhibited similar efficacy with an accuracy of 83% and a macro-averaged F1-score of 0.83. BERT (accuracy: 78%, F1-score: 0.79) and SBERT (accuracy: 80%, F1-score: 0.81) demonstrated marginally reduced overall performance. DistilBERT demonstrates superior performance, underscoring its suitability for automating CO-PO alignment tasks.

3.2 Performance Analysis

Table 4 presents the precision, recall, and F1-score for each alignment level (I, E, D) across all four models, revealing varying classification patterns where DistilBERT showed robust performance with high F1-scores peaking at 0.90 (D) and 0.84 (E), BERT had strong Demonstrative precision (0.82) but weaker Enabling performance (F1: 0.74), RoBERTa displayed balanced performance with top Enabling recall (0.86) and notable F1-scores for Introductory (0.82) and Demonstrative (0.85), and SBERT + LR consistently had lower F1-scores, highlighting the different effectiveness of transformer architectures in discerning CO-PO relationships.

Table 4. Performance of transformer-based models on CO-PO alignment.

Model	AL	Prec.	Rec.	F1
BERT	D	0.82	0.84	**0.83**
	E	0.74	0.73	0.74
	I	0.79	0.79	0.79
DistilBERT	D	**0.89**	**0.91**	**0.90**
	E	**0.83**	**0.86**	**0.84**
	I	**0.88**	0.83	**0.85**
RoBERTa	D	0.86	0.84	**0.85**
	E	0.79	**0.86**	0.82
	I	**0.86**	0.79	0.82
SBERT + LR	D	0.89	0.79	0.84
	E	0.78	0.80	0.79
	I	0.78	0.81	0.79

AL. = Alignment Level, Prec. = Precision, Rec. = Recall, F1 = F1-score

3.3 Confusion Matrix and Heatmap Analysis

The classification performance of the four models was visualized through their confusion matrices. DistilBERT's confusion matrix demonstrated a strong diagonal (Figure 2), indicative of high accuracy and minimal misclassifications across all alignment levels. In contrast, SBERT + Logistic Regression's confusion matrix (Figure 3) displayed more significant confusion, particularly between the E and I level. While the confusion matrices for BERT and RoBERTa (not explicitly shown in figures) revealed their error patterns, BERT exhibited some misclassification of the E level as I, and RoBERTa presented a well-defined diagonal with generally low misclassification rates.

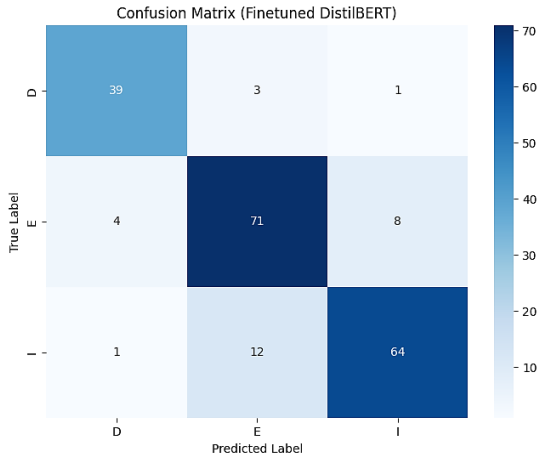

Fig. 2. Confusion Matrix for Finetuned DistilBERT

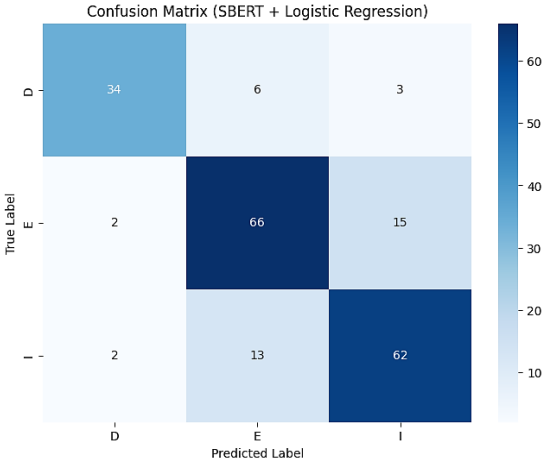

Fig. 3. Confusion Matrix for SBERT + Logistic Regression

3.4 ROC Curve and AUC Score

The discriminatory power of the models was evaluated using Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve (AUC) scores. Macro-average AUC scores (Table 5) indicated strong overall performance for all models in distinguishing between alignment levels, with DistilBERT and RoBERTa (both 0.94) highest, followed by BERT and SBERT + Logistic Regression

(both 0.92). The high AUC values suggest a good ability of the models to correctly rank instances.

Table 5. Macro-average AUC scores for the transformer-based models.

Model	Macro Avg AUC
BERT	0.92
DistilBERT	**0.94**
RoBERTa	**0.94**
SBERT + LR	0.92

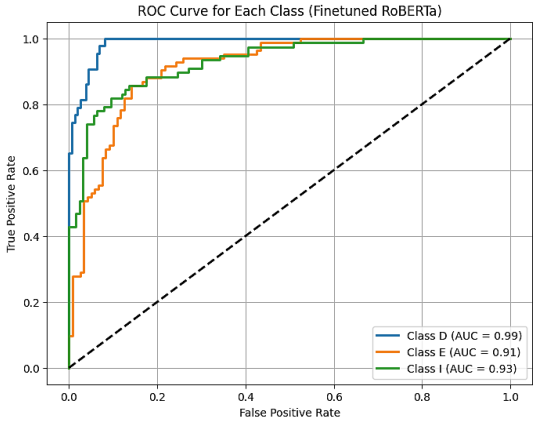

Fig. 4. ROC Curve for Each Class (Finetuned RoBERTa)

DistilBERT and RoBERTa achieved the highest Macro-average AUC score of 0.94, indicating strong discrimination between CO-PO alignment levels. DistilBERT's efficiency comes from BERT knowledge distillation, while RoBERTa benefits from its robust pre-training.

3.5 Precision-Recall Curve Analysis

Precision-Recall (PR) curves and Average Precision (AP) scores further illuminated model categorization, especially regarding class imbalance. DistilBERT's representative PR curves (Figure 5) consistently showed high precision across recall values, supported by strong AP scores (D: 0.94, E: 0.88, I: 0.90), indicating robust pertinent instance extraction. While RoBERTa (D: 0.96, E: 0.86, I: 0.91), BERT (D: 0.94, E: 0.81, I: 0.87), and SBERT + Logistic Regression (D: 0.91, E: 0.85, I: 0.86) performed reasonably, their AP scores were slightly lower than DistilBERT's in some classes, suggesting a less optimal precision-recall balance for those specific categories. Overall, PR curve analysis, via AP scores, highlighted DistilBERT's strong ability to maintain high precision across different recall levels.

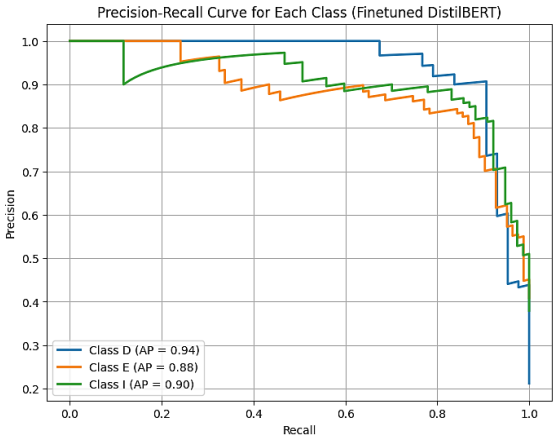

Fig. 5. Representative Precision-Recall Curves (e.g., DistilBERT)

3.6 Comparative Summary and Interpretation

DistilBERT demonstrated superior overall performance for CO-PO alignment, achieving the highest accuracy and F1-score, alongside top-tier AUC. Although RoBERTa performed exceptionally well with a comparable Precision-Recall profile, DistilBERT's consistent lead in crucial classification metrics renders it the most effective model overall. BERT performed competitively but showed limitations with the Enabling level, while SBERT + Logistic Regression consistently underperformed. This likely arises from DistilBERT's ability to efficiently distill knowledge from a larger model while maintaining a balanced architecture suitable for learning the semantic distinctions within CO-PO pairs. Ultimately, fine-tuned transformer architectures, with DistilBERT proving particularly

robust, are the most effective for this CO-PO alignment task.

4. CONCLUSIONS

Fine-tuned transformer models, especially DistilBERT and RoBERTa, outperformed SBERT + Logistic Regression in CO-PO alignment classification, with DistilBERT showing the best performance. These findings demonstrate the potential of transformer-based automation for improving curriculum mapping and quality assurance. Future research should explore broader datasets, advanced fine-tuning techniques, and enhance model transparency for educators, while also applying these methods to other educational alignment tasks.

5. REFERENCES

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. Artificial Intelligence Review, 54(8), 5789–5829. https://doi.org/10.1007/s10462-021-10026-2

Chor, W. T., Goh, K. M., Lim, L. L., Lum, K. Y., & Chiew, T. H. (2024). Towards a machine learning-based constructive alignment approach for improving outcomes and composition of engineering curriculum. Education and Information Technologies, 29(7), 8925–8959.

Commission on Higher Education (2017). CMO No. 88, Series of 2017: Policies, Standards and Guidelines for the Bachelor of Science in Electrical Engineering (BSEE) Program Effective Academic Year (AY) 2018-2019.

Commission on Higher Education (2017). CMO No. 97, Series of 2017: Policies, Standards and Guidelines for the Bachelor of Science in Mechanical Engineering (BSME) Program Effective Academic Year 2018 - 2019.

Commission on Higher Education (2017). CMO No. 101, Series of 2017: Policies, Standards and Guidelines for the Bachelor of Science in Electronics Engineering (BSECE) Program Effective Academic Year (AY) 2018-2019.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Özkurt, C. (2024). Comparative analysis of state-of-the-art Q&A models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 dataset. Chaos and Fractals, 1(1), 19–30.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper, and lighter. arXiv preprint arXiv:1910.01108.

Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. IEEE Access, 11, 55533–55560.