AI-Driven Textual Feedback Analysis in E-Training Using Enhanced RoBERTa

Rakan Saad Alotaibi, Fahad Mazyed Alotaibi, Sameer Abdullah Nooh, Abdulaziz A. Alsulami
Department of Information Systems-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—In corporate e-training environments, traditional metrics like course completion and quiz scores often fail to reflect actual job performance. Rich insights are embedded in unstructured textual feedback, yet they remain underutilized due to limitations in existing analytical models. This study proposes E-RoBERTa, an enhanced transformer-based model designed to predict employee job performance by analyzing open-ended feedback from digital training platforms. The model aims to improve accuracy, domain adaptability, and interpretability. E-RoBERTa integrates Domain-Adaptive Pretraining (DAPT) to fine-tune RoBERTa on corporate-specific language and introduces Dynamic Attention Scaling (DAS) to highlight semantically critical tokens. A real-world, GDPR-compliant dataset containing 16,000 feedback entries from 3,500 employees across multiple departments was used. Preprocessing included tokenization, sentiment tagging, and feature extraction. The model achieved superior performance with a macro F1-score of 0.875, outperforming standard RoBERTa, LSTM, and SVM baselines. Attention visualizations revealed alignment between influential tokens and human-interpretable performance indicators. E-RoBERTa provides a transparent and accurate framework for evaluating job performance through textual feedback. Its use of domain adaptation and dynamic attention mechanisms supports scalable, ethical, and explainable AI in corporate learning analytics, offering actionable insights for personalized interventions and strategic HR decision-making.**

Keywords*—*Job performance prediction; transformer models; enhanced RoBERTa; domain-adaptive pretraining (DAPT); dynamic attention scaling (DAS); natural language processing (NLP); explainable AI; textual feedback analysis; workforce analytics

I. Introduction

In the last few years, e-training platforms have been indispensable tools for professional development in various industries. With organizations moving towards implementing digital learning solutions for upskilling employees and increasing productivity, there has been a significant demand for effective training analytics. Conventional completion rates, quiz scores, and attendance logs shed little light on actual job performance outcomes [1]. These static indicators do not measure the dynamic and contextual nature of learning behavior that affects how knowledge is applied successfully at the workplace. The emergence of artificial intelligence (AI) and intense learning has shifted the methods of processing unstructured training data towards natural language processing (NLP) approaches. Employee feedback, assessment responses, and forum discussions are filled with information, which, if well mined, can provide patterns regarding job performance [2]. Transformer-based models such as BERT and RoBERTa have demonstrated excellent abilities in deriving contextual sense in text, compared to the previous approaches in several NLP tasks [3]. However, standard transformer models often lack domain-specific adaptation, resulting in inaccuracy when applied to corporate training data. Furthermore, they consider all text elements to be equally important, which can obscure critical signals contained in sophisticated employee responses. This study proposes an Enhanced RoBERTa model that addresses these gaps through Domain-Adaptive Pretraining (DAPT) and Dynamic Attention Scaling (DAS), enabling the capture of more suitable training feedback features. The motivation for this work is a need to develop more accurate, interpretable, and ethical tools for performance prediction in e-learning. Since businesses depend on data-driven decisions for the development of their employees, it is crucial to have AI systems that can extract actionable insights from training feedback. Our model aims to facilitate personalized learning, early interventions for employees struggling with their performance, and more effective talent management strategies.

Despite the rapid increase in the use of e-training platforms, organizations continue to face challenges in accurately measuring the impact of training on actual job performance in the real world. The existing evaluation methods, i.e., assessment scores and completion rates, do not indicate whether the employees can practically apply what they have learned in the workplace [4, 5]. These metrics tend to give a shallow learning perspective, disregarding deeper learning or long-term retention. Additionally, the predictive models used in training analytics have a limited scope. Many still rely on manually chosen features and structured data, which overlook the valuable information hidden in textual content, such as open-ended feedback, peer discussions, and reflective assessments [6]. These unstructured inputs contain contextual hints on employee motivation, understanding, and satisfaction, factors that significantly determine performance outcomes. Such text has become a fertile ground for transformer-based NLP models, which have proven to be promising tools for deriving insights. However, existing off-the-shelf models, such as RoBERTa, are not tailored for domain-specific language, resulting in suboptimal performance when applied to corporate training data [3]. Moreover, these models do not differentiate between various tokens and treat them equally, without mechanisms to highlight the most relevant parts of the text that determine prediction. Another problem is the interpretability of deep learning models in human resource (HR) settings. Managers need explanations as to why a specific system speculates that a given employee will not perform to the expected level. Black-

box predictions erode trust and hinder deployment, particularly when the decision requires fairness and transparency [7, 8]. This study fills these gaps by creating an Enhanced RoBERTa model specifically for corporate training data. It uses DAPT and DAS to enhance the relevance and interpretability of performance prediction. With that, it strives to provide a more accurate and explainable tool for employee training outcomes based on textual feedback.

The primary objective of this study is to enhance the accuracy, transparency, and applicability of job performance predictions derived from textual comments in e-training environments. Achievement of this is done based on the following four objectives for the study:

- To design a transformer-based model capable of understanding domain-specific training language used in employee feedback, assessments, and learning reflections.
- To improve model attention mechanisms by introducing a dynamic scaling method highlighting contextually essential elements in training text.
- To evaluate the proposed model's effectiveness using a real-world corporate training dataset, comparing its performance to baseline machine learning (ML) and NLP models.
- To support explainable and ethical AI in workforce analytics by offering interpretable outputs that aid decision-making in employee development.

Based on these objectives, the study offers the following key contributions:

- A domain-adapted version of RoBERTa is developed using DAPT on corporate training documents, improving its contextual understanding in the HR and e-learning domains.
- The model incorporates a novel DAS mechanism that adjusts attention weights to emphasize semantically significant feedback. This improves both prediction accuracy and interpretability for human evaluators.
- The model is tested on GDPR-compliant training data with performance measured using accuracy, precision, recall, and F1-score. Benchmarks show that E-RoBERTa outperforms traditional NLP models and classical ML classifiers.
- Attention maps are visualized to show how the model interprets feedback, enabling HR managers to understand and trust AI-driven performance insights. This directly supports adoption in high-stakes organizational settings.

This study bridges the gap between generic NLP systems and the specific needs of performance prediction in professional e-training, offering a robust, interpretable, and ethically conscious approach that organizations can integrate into their learning analytics ecosystems.

This study is structured as follows: Section II reviews related work on e-training analytics, NLP in education, and transformer models in workforce analytics. Section III outlines the proposed methodology, including data preprocessing, model architecture, training procedures, and evaluation metrics. Section IV presents the results and discussion, comparing the proposed model with baseline methods. Finally, Section V concludes the study and suggests directions for future work.

II. Related Work

A. E-Training Analytics and Machine Learning Approaches

E-training platforms have evolved to become instrumental in organizational learning, offering scalable, flexible, and cost-effective alternatives to conventional classroom-based approaches. However, assessing the real impact of these programs on employees' job performance is a significant challenge. Many organizations still use shallow metrics, such as course completion rates and quiz scores, which do not often correspond with long-term workplace effectiveness [4]. This gap has been the target of recent developments in educational data mining and learning analytics [9]. These approaches aim to draw insights from digitized learning environments by observing learner behaviors, interactions, and outcomes [10]. Although the initial models employed rudimentary statistics and rule-based systems, they did not accurately describe the complex dynamics of learning or individualized training advice. ML brought a paradigm shift in predictive analytics training. Decision trees, support vector machines (SVMs), and random forest algorithms have been implemented in the prediction of learner success using structured features such as attendance, performance history, and level of engagement [11, 12]. Although these models can bring improvements in prediction accuracy, they still struggle to cope with unstructured data, particularly textual feedback and discussion content, which is rich in contextual information about learning processes. In the last few years, deep learning has received attention for its capability to model complex, nonlinear relationships in e-learning data. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been shown to reveal temporal patterns in learning behavior [13]. However, these models typically require large training datasets and suffer from issues such as the vanishing gradient. To enhance prediction and personalization in e-training, researchers have started using NLP methods. Textual data that have been observed to provide indicators of motivation, understanding, and satisfaction, all of which are job performance-related, have been found in feedback, reflections, and forum posts [14]. However, standard NLP techniques, such as bag-of-words or TF-IDF, do not adequately reflect the semantic depth and contextual nuances [15].

B. Advances in NLP and Transformer Models

Transformer-based models, such as BERT and RoBERTa, established a new benchmark for text understanding, as they rely on self-attention mechanisms to maintain the context throughout the whole sequence [16]. These models have been utilized in education to measure sentiment, grade essays, and predict course outcomes; however, their application in predicting job performance in corporate e-training remains minimal. This study extends these bases by introducing a domain-adapted

RoBERTa model with attention scaling, customized to predict job performance from training feedback. By paying closer attention to the linguistic patterns in employee reflections and evaluations, the proposed model aims to enhance the interpretability and reliability of AI-based training analytics. NLP has become a crucial component of intelligent educational systems, enabling more effective analysis of learner input and more adaptable instructional techniques. With the help of NLP techniques, written text in the form of learner feedback, discussion posts, reflections, and assessment responses can be processed to understand learners' cognitive and emotional states during training [16]. Simple lexical approaches, such as keyword extraction and sentiment scoring, were used in the early NLP applications in education. These approaches were useful in uncovering surface-level trends, but did not have the depth to interpret finely-tuned feedback or monitor conceptual understanding [17]. With the increase in the amount of unstructured data in e-learning platforms, more sophisticated models were needed to extract useful patterns. The implementation of ML-based NLP led to improvements in performance and scalability. Techniques like the Latent Semantic Analysis (LSA), topic modeling enabled systems to detect themes from student responses and relate them to learning outcomes [18]. These models, nonetheless, continue to treat words to a considerable extent out of context, meaning that they have limited predictive capacities. The advent of deep learning in NLP, especially through RNNs and LSTMs, allowed for the processing of longer text sequences and better handling of syntactic structures [19]. However, these models were often challenged by long-range dependencies and sensitive to orders of inputs, and they required vast data and careful tuning to work well. Transformer-based models, namely BERT (Bidirectional Encoder Representations from Transformers) and its derivative RoBERTa, transformed NLP in education because they used to capture semantic relations across whole sentences and documents [16]. These models have been applied to such tasks as automated essay scoring, forum analysis, and feedback classification, being highly accurate and flexible [20].

C. Domain Adaptation and Explainable AI in Workforce Analytics

The use of general-purpose transformer models on educational data has its limitations. The pre-trained models are based on large, generic corpora (e.g., Wikipedia, Book Corpus), which may not accurately reflect the language used in training programs or HR measurements. Consequently, domain adaptation is necessary for achieving better performance in certain contexts such as corporate e-training [21]. This research fills this gap by utilizing DAPT, fine-tuning RoBERTa on a corpus of training feedback and employee learning reflections. Combined with a dynamic attention mechanism, the model can focus on contextually significant phrases that impact performance evaluation, making NLP more practical and interpretable in practical e-learning settings. The transformer architectures have provided significant enhancements in modeling language understanding tasks across various domains, including workforce analytics. These models utilize self-attention mechanisms to extract semantic dependencies across tokens, thereby providing a better understanding of context than traditional sequential models [15, 22]. Their ability to handle long-range relationships in text has made them ideal for applications such as resume screening, performance feedback analysis, and employee sentiment monitoring. Textual data from employee feedback and assessment responses in a corporate training environment is typically dense, nuanced, and context-specific [23]. Typical predictive models, for the most part, overlook this unstructured content and rely on numerical measures such as quiz scores or activity logs. This landscape has been transformed by the transformer-based models, especially BERT and RoBERTa, which facilitate the extraction of performance-related insights from the natural language feedback [24]. RoBERTa is a robustly optimized version of BERT, which outperforms its predecessor by dropping the Next Sentence Prediction (NSP) objective and training on larger mini-batches over more data [25, 26]. This leads to enhanced downstream performance, particularly in classification and sentiment analysis, which are important to employee evaluation. Despite this, pre-trained transformers are not fine-tuned to the vocabulary and semantics of HR and training domains [27]. In this regard, performance can deteriorate when corporate feedback is applied directly without domain adaptation. Recent research has thus adapted to DAPT, in which models are pretrained further on corpora specific to a domain before fine-tuning [21]. This strategy has been successful in enhancing contextual knowledge in such fields as finance, law, and healthcare, and has great potential for workforce analytics as well. Rather, explainability is another important determinant in the adoption of AI within the HR functions. Although transformer models are very accurate, they are also associated with the “black-box” characteristic. Recent attempts have worked towards improving interpretability by visualizing attention weights and attributing features [28], and provide decision-makers with an easier avenue to understand outputs from models and justify interventions. Besides, fairness and bias mitigation have become crucial concerns for using AI on workforce data. Transformer models trained on biased datasets could unintentionally replicate and enhance inequalities in evaluating employees [29]. Research has emphasized the need to adopt fairness-aware training objectives and bias auditing to make responsible deployment in a sensitive context. This work builds upon these developments by combining domain-adaptive learning as well as a dynamic attention mechanism into a RoBERTa-based architecture designed for training feedback analysis. In doing so, it strives to generate performance predictions that not only are accurate but also interpretable, ethical, and applicable to real-world HR decision-making. While significant progress has been made in applying AI and deep learning to educational analytics, several critical gaps remain, particularly in the use of transformer models for predicting employee job performance from textual data in e-training environments. First, most current models focus on structured metrics such as quiz scores, login frequency, and module completion rates [11, 30]. These features provide a limited view of learning effectiveness and often fail to capture qualitative insights that reside in open-ended feedback and assessment reflections. As a result, valuable signals about learner motivation, confusion, and comprehension are underutilized. Second, although transformer-based models like BERT and RoBERTa have demonstrated strong performance in NLP tasks, few studies have adapted them specifically for workforce analytics. Many existing works apply pre-trained models

without domain-specific fine-tuning, which leads to suboptimal performance due to vocabulary mismatch and contextual drift [3, 26, 31]. Third, interpretability remains a challenge. Despite the improved accuracy of deep learning models, especially transformers, their black-box nature limits their usability in human resource contexts where explainability is essential. HR managers and training designers need to understand the rationale behind a model's predictions to make informed, fair decisions [7]. Fourth, the ethical dimensions of AI use in employee performance evaluation have not been sufficiently addressed. Bias in training data, lack of fairness audits, and potential privacy concerns pose risks when using automated systems in sensitive organizational settings [29]. Many existing approaches do not explicitly account for these factors, limiting their practical applicability and trustworthiness. Finally, little research has explored the integration of dynamic attention mechanisms into transformer architectures to emphasize contextually important feedback features. Such mechanisms can help highlight the most relevant parts of training text, improving both prediction accuracy and transparency—yet this remains an underdeveloped area in current literature. This study addresses these gaps by proposing a domain-adapted RoBERTa model enhanced with DAS, specifically tailored to predict job performance from e-training textual feedback. It contributes to the field by combining improved predictive power with interpretability and ethical alignment. Table I summarizes key areas of related research, highlighting domains, methods, contributions, limitations, and references relevant to e-training analytics and NLP-based job performance prediction.

TABLE I. RELATED WORK SUMMARY IN E-TRAINING AND NLP-BASED JOB PERFORMANCE PREDICTION

Ref.	Domain	Study Focus	Methods/Mode	Key Contributions	Limitations Identified
[4, 10]	E-Training Evaluation	Measuring training effectiveness	Course completion, quiz scores	Simple, scalable training metrics	Poor correlation with real-world job performance
[11]	Educational Data Mining	Predictive learning analytics	Decision Trees, SVM, Random Forests	Structured data modeling (engagement, performance)	Neglects unstructured text; lacks semantic context
[13]	Deep Learning for E-Learning	Sequential learning behavior	RNN, LSTM	Captures temporal patterns in learning activities	Needs large datasets, suffers from vanishing gradients
[17, 18]	Early NLP Techniques	Feedback and forum analysis	Bag-of-Words, TF-IDF, LSA, Topic Modeling	Initial semantic interpretation from text	Out-of-context word treatment, low accuracy
[14, 16, 20].	Modern NLP in Education	Understanding learner emotions and cognition	Transformer models (BERT, RoBERTa)	High accuracy in sentiment analysis, grading, and reflection mining	Limited use in corporate/job performance settings
[21]	Domain Adaptation in NLP	Enhancing model relevance for training data	DAPT	Improves transformer model performance on specialized corpora	Underused in HR and e-learning systems
[3, 24]	Workforce Analytics	Job performance prediction from feedback	NLP, BERT, RoBERTa	Enables prediction from text-based performance indicators	Few applications tailored to domain-specific feedback
[7, 28]	Explainable AI in HR	Trust and transparency in predictions	Attention visualization, attribution tools	Supports decision justification via interpretable models	Limited adoption in sensitive HR environments
[29]	Ethical AI in Workforce Systems	Fairness, bias, and data privacy	Bias audits, ethical frameworks	Highlights the need for fairness-aware training and auditing practices	Lack of widespread implementation; risk of reinforcing historical bias
[7, 29] [3, 31]	Gaps in Current Approaches	Text-based job performance modeling	Generic transformers without fine-tuning	Identifies the need for context-aware, interpretable, and ethical model development	Misses attention on important feedback cues; poor generalization across domains

III. METHODOLOGY

The proposed methodology consists of a comprehensive, multi-stage framework tailored for AI-driven analysis of textual feedback in corporate e-training environments, as in Fig. 1. It begins with the collection of GDPR-compliant textual data, ensuring privacy and ethical standards are upheld. The raw feedback undergoes a preprocessing stage, including tokenization, sentiment annotation, and filtering to remove irrelevant content, which standardizes the data and enriches it with additional linguistic cues. Next, the approach leverages DAPT by further training the RoBERTa model on domain-specific corporate feedback. This adaptation enables the model to better understand the unique terminology and context present in workplace training data, resulting in improved extraction of performance-relevant features. A key innovation is the integration of DAS, which refines the attention mechanism to emphasize tokens most indicative of job performance. Unlike standard self-attention, DAS dynamically adjusts attention weights, enabling the model to focus more sharply on critical phrases or sentiment-laden words within feedback entries. Finally, the model is fine-tuned using supervised learning, with cross-entropy loss guiding the classification of employees into performance categories such as "Low", "Medium", or "High". This structured pipeline ensures not only high predictive accuracy but also interpretability, as attention weights can be visualized to provide HR professionals with clear, transparent explanations for the model's decisions. Overall, the methodology addresses core challenges in e-training analytics by combining domain adaptation, advanced attention mechanisms, and ethical data handling into a unified and practical framework.

A. Data Collection and Preprocessing

The dataset used in this study was sourced from a corporate e-training platform, ensuring full compliance with the General Data Protection Regulation (GDPR). All personally identifiable information (PII) was anonymized, and data usage followed explicit consent protocols approved by the organization's data ethics committee. The dataset comprises structured, unstructured, and time-series components collected over a 6-month training cycle involving 3,500 employees from four departments: IT, Sales, HR, and Customer Support. It contains over 220,000 individual records, summarized in Table II:

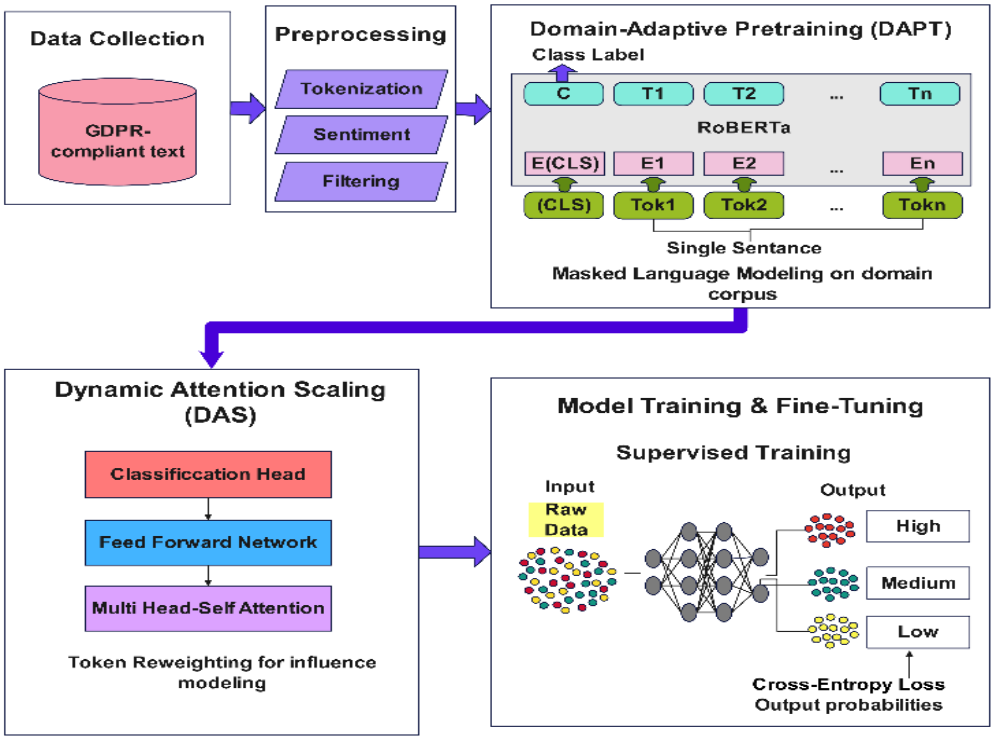

Fig. 1. Proposed methodology framework.

TABLE II. SUMMARY OF DATASET USED FOR MODEL TRAINING AND EVALUATION

Data Category	Type	Volume	Description
Employee Records	Structured	3,500	Anonymized user ID, department, experience
Assessment Scores	Structured	28,000+	Quiz and post-training test scores
Engagement Logs	Time-Series	140,000+ entries	Timestamps of login activity, module visits
Training Feedback	Unstructured (Text)	16,000 responses	Open-ended reflections and course reviews
Performance Ratings	Structured (Labels)	10,000 entries	HR-assigned job performance categories (Low–High)
Course Completion Flags	Structured (Boolean)	3,500 entries	Completion status of assigned training

Textual feedback was the primary source for training the model. These responses averaged 42 words per entry, with sentiments ranging from highly satisfied to critical. Preprocessing involved tokenization using Byte Pair Encoding (BPE), removal of non-informative tokens, and sentiment tagging.

Stopwords were removed, and the data was normalized to lowercase. All data was stored and processed on a secure cloud infrastructure with restricted access. Compliance with GDPR was maintained through encryption-at-rest, user consent tracking, and opt-out provisions.

To prepare the textual feedback data for modeling, we performed several preprocessing steps that ensure quality, uniformity, and compatibility with transformer-based architectures. Eq. (1) presents the raw textual dataset be denoted by:

$$D = [d_1, d_2, \ldots, d_N] \quad (1)$$

where, d_i represents an individual employee feedback entry, and $N = 16{,}000$ denotes the total number of textual responses.

Step 1: Normalization

Each feedback d_i is first normalized to lowercase and stripped of special characters, as presented in Eq. (2):

$$d_i' = normalize(d_i) = lowercase(remove\_punctuation(d_i)) \quad (2)$$

Step 2: Tokenization

We applied BPE tokenization using a vocabulary size of $V = 30{,}000$ shown in Eq. (3), which converts each document d_i' into a sequence of tokens:

$$T_i = BPE(d_i') = [t_{i1}, t_{i2}, \ldots, t_{iL_i}] \quad (3)$$

where, L_i is the token length of the i^{th} document, with an average length $\bar{L} \approx 64$ tokens.

Step 3: Stopword Removal

To reduce noise, a standard English stopword list S (e.g., "the", "is", "at") was used in Eq. (4):

$$T_i' = T_i \setminus S \quad (4)$$

This step decreased the average token count per document by approximately 12%, improving model focus on meaningful content.

Step 4: Padding and Truncation

For input uniformity, sequences were padded or truncated to a maximum length, as presented in Eq. (5):

$$T_i'' = \begin{cases} T_i'[:128], & if \ |T_i'| > 128 \\ T_i' \cup [PAD]^{128-|T_i'|}, & if \ |T_i'| < 128 \end{cases} \quad (5)$$

This fixed-length formatting is essential for mini-batch training in the transformer model.

Step 5: Sentiment Annotation (Optional Feature)

Each document was optionally annotated with a sentiment score $s_i \in [-1,0,+1]$, derived using a pre-trained sentiment classifier, where:

- -1: negative sentiment
- 0: neutral
- $+1$: positive

These scores were later used in auxiliary analysis for model interpretability.

As a result, the cleaned dataset was transformed into a matrix $X \in \mathbb{R}^{N\times 128}$ representing tokenized feedback, ready for embedding and input to the E-RoBERTa model. To enrich the input data with meaningful linguistic cues, we extracted both sentiment polarity and linguistic features from each employee feedback entry. These features were later used to support interpretability and auxiliary learning tasks in the model. Each preprocessed feedback sample d_i was analyzed using a pretrained sentiment classifier based on a fine-tuned BERT model. The classifier assigned a sentiment label $s_i \in [-1,0,+1]$, representing negative, neutral, or positive sentiment, respectively. The sentiment score was computed in Eq. (6):

$$s_i = arg\,max(Softmax(f(d_i))) \quad (6)$$

where, $f(\cdot)$ is the classifier output vector (logits) for the three sentiment classes. Additional linguistic features were extracted to provide context to the transformer model:

- Text Length ℓi: Total number of tokens in d_i
- Positive/Negative Word Count: Based on a sentiment lexicon L
- Named Entity Count: Using spaCy NER pipeline
- TF-IDF Scores: Used selectively for attention visualization

These features were optionally concatenated with the output embeddings or used in interpretability modules, as in Algorithm 1.

Algorithm 1: Sentiment and Feature Extraction

Input: Preprocessed feedback corpus $D = [d_1, d_2, \ldots, d_n]$

Output: Sentiment labels $S = [s_1, s_2, \ldots, s_n]$, Feature set F

1. For each feedback d_i in D do
2. Tokenize d_i using BERT tokenizer
3. Predict sentiment using pretrained BERT classifier → s_i
4. Compute text length $\ell_i = len(d_i)$
5. Count sentiment words from lexicon → pos_i, neg_i
6. Apply NER model to get entity $count \rightarrow e_i$
7. Calculate TF-IDF vector (optional) → $tfidf_i$
8. Store $F_i = [\ell_i, pos_i, neg_i, e_i, tfidf_i]$
9. End For

Return S, F

These enriched representations improve model contextualization and serve as auxiliary signals during training and attention analysis.

B. Model Architecture: Enhanced RoBERTa (E-RoBERTa)

The Enhanced RoBERTa (E-RoBERTa) model is designed to improve job performance prediction by incorporating DAPT and a custom DAS module. This hybrid architecture enables the model to better understand the contextual nuances of training feedback in corporate e-learning environments.

1) RoBERTa base layer: RoBERTa is a transformer-based architecture built upon BERT, optimized by removing the Next Sentence Prediction (NSP) task and training with more extended sequences and larger batches. It uses the standard self-attention mechanism, as presented in Eq. (7):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

where:

- $Q, K, V \in \mathbb{R}^{L\times d_k}$ are query, key, and value matrices
- d_k is the key dimension
- L is the sequence length

2) Domain-Adaptive Pretraining (DAPT): DAPT improves model specialization by continuing the masked language modeling (MLM) task on a domain-specific corpus C_{domain} composed of training feedback and corporate assessment documents, as shown in Eq. (8). The MLM objective is:

$$L_{MLM} = -\sum_{i\in M} log\, P(w_i|\hat{w}\setminus_i) \quad (8)$$

where:

- M is the set of masked token positions
- w_i is the true token at position i
- $\hat{w}\setminus_i$ is the masked sequence with i^{th} token hidden

3) Dynamic Attention Scaling (DAS): DAS introduces a trainable scalar $\gamma \in \mathbb{R}$ to adaptively scale attention weights at the token level presented in Eq. (9):

$$DAS - Attention(Q,K,V) = softmax\left(\gamma \cdot \frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

Here, γ is learned via backpropagation and reflects the task-specific importance of context length sensitivity. Higher γ values lead to sharper attention peaks, helping the model focus on high-impact tokens such as performance-related verbs or competency indicators. Algorithm 2 outlines the training process of the E-RoBERTa model. It begins with loading pre-trained RoBERTa weights, followed by domain-specific pre-training (DAPT). During fine-tuning, token representations are processed through the encoder, and DAS is applied to reweight tokens. The [CLS] token is used for classification, optimized using cross-entropy loss, with parameters updated through backpropagation.

Algorithm 2: E-RoBERTa Training with DAPT + DAS

Input: Tokenized training feedback corpus $D = [d_1, \ldots, d_n]$

Output: Trained E-RoBERTa model for performance prediction

1. Load RoBERTa pretrained weights
2. Perform DAPT on feedback corpus
3. For each batch B in fine-tuning set do
4. Pass input tokens through RoBERTa encoder $\rightarrow H = [h_1, \ldots, h_l]$
5. Compute scaled attention using DAS:
6. $$A = softmax(\gamma \cdot QK^T / \sqrt{d_k}) \cdot V$$
7. Extract [CLS] token representation $\rightarrow$ h_cls
8. Pass h_cls through Dense layer $\rightarrow$ $\hat{y} = Softmax(W \cdot h\_cls + b)$
9. Compute loss using Cross-Entropy:
10. $$L = -\Sigma\, y\, log(\hat{y})$$
11. Backpropagate gradients and update γ and W
12. End For

Return: Fine-tuned E-RoBERTa model

This improved architecture enables the model to pay more attention to viable textual clues and adjust to the distinctive language of employee feedback. Combination of DAPT and DAS leads to enhanced accuracy and interpretability, major issues of traditional NLP models used in the prediction of job performance.

4) Training configuration and optimization: A supervised learning scenario was employed to train the Enhanced RoBERTa (E-RoBERTa) model to categorize employees' job performance into three categories. Low, Medium, or High. A system equipped with an NVIDIA Tesla V100 GPU and 32 GB of RAM was used to train the model using the PyTorch framework with HuggingFace's Transformers library. The model was fine-tuned with the AdamW optimizer with a learning rate of 2×10^{-5}, and a linear learning rate scheduler with warm-up steps was used to stabilize the early training. Training was performed over five epochs with a batch size of 32. Cross-entropy loss was used as the objective function, defined as in Eq. (10):

$$L = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \quad (10)$$

where, $C = 3$ represents the number of performance classes, y_i is the true label distribution, and $\hat{y}_i$ is the predicted probability for class i. To prevent overfitting, a dropout rate of 0.1 was applied to both the attention and output layers. Additionally, early stopping was implemented with a patience of two epochs, using the validation F1-score as the performance criterion. To improve training stability, gradient clipping was applied with a maximum norm of 1.0. A weight decay of 0.01 was used for regularization. Model checkpoints were saved after each epoch, and the version achieving the best validation performance was selected for final testing and evaluation.

C. Evaluation Metrics

Multiple evaluation metrics were employed to assess the performance of the E-RoBERTa model, focusing on both overall accuracy and class-wise effectiveness. Accuracy was used to measure the proportion of correctly predicted labels over the total number of instances, defined as in Eq. (11):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

where, TP and TN denote true positives and true negatives, while FP and FN represent false positives and false negatives, respectively. In addition to accuracy, Precision [Eq. (12)], Recall [Eq. (13)], and F1-score [Eq. (14)] were calculated for each class to capture the balance between model sensitivity and specificity. Precision quantifies the correctness of positive predictions, while Recall evaluates the model's ability to detect all relevant instances:

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

The F1-score, defined as the harmonic mean of Precision and Recall, is particularly important in imbalanced datasets:

$$F1 - Score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (14)$$

Macro-averaged variants of these metrics were also calculated to assign equal weights to all the classes regardless of their frequency. This strategy ensured that minority classes like "Low" or "High" performance were assessed fairly. Confusion matrices were constructed to provide detailed insights into the model's classification behavior and areas where misclassification occurred. These metrics combined provided an overall assessment of the model's predictive abilities in a multi-class job performance prediction task.

IV. Results and Discussion

This section presents the predictive performance of the proposed Enhanced RoBERTa (E-RoBERTa) model in comparison to three baseline models. standard RoBERTa, LSTM, and SVM. The models were tested on a labeled dataset with Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-Score as well as other breakdowns by class and training dynamics. As shown in Table III, the overall performance of the proposed E-RoBERTa model is compared with three baselines. E-RoBERTa gives the best f1-Score of 0.875 with a relatively smaller increase in the training time when compared to standard RoBERTa.

TABLE III. OVERALL PERFORMANCE COMPARISON ACROSS MODELS

Model	Acc.	Prec.	Rec.	F1-Score	Macro Avg F1	Weighted F1	Training Time (min)	Parameters (M)
E-RoBERTa	0.89	0.88	0.87	0.875	0.872	0.878	34	355
Standard RoBERTa	0.84	0.82	0.80	0.81	0.808	0.815	28	355
LSTM	0.78	0.76	0.74	0.75	0.743	0.749	22	28
SVM	0.72	0.70	0.68	0.69	0.685	0.690	15	-

Table IV provides a breakdown of E-RoBERTa's performance for each class. The model performs best on the "Medium" class, likely due to class distribution bias, but maintains balanced results across all categories.

TABLE IV. CLASS-WISE PRECISION, RECALL, AND F1-SCORE (E-ROBERTA)

Class	Precision	Recall	F1-Score	Support
Low	0.85	0.84	0.845	90
Medium	0.89	0.91	0.90	150
High	0.86	0.82	0.84	60
Avg	0.87	0.89	0.875	300

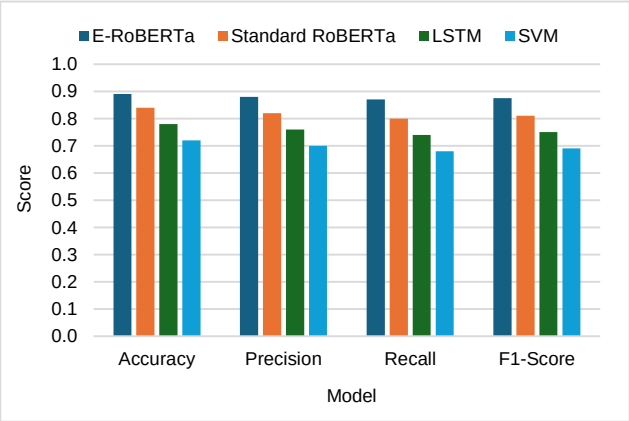

Fig. 2. Performance comparison of all models.

Fig. 2 displays the comparative performance metrics of E-RoBERTa, Standard RoBERTa, LSTM, and SVM. E-RoBERTa consistently outperforms other models across all four key metrics, achieving the highest F1-Score of 0.875. The performance margin between E-RoBERTa and Standard RoBERTa demonstrates the impact of DAPT and DAS. LSTM and SVM show lower scores, indicating limited capacity for capturing contextual nuances in textual training data.

Fig. 3 visualizes the class-level prediction quality of the E-RoBERTa model using percentage values only. Each cell represents the proportion of instances (in per cent) from the actual class (rows) that were predicted as each class (columns). For example, a value of 84.4% on the diagonal for the "Low" class indicates that 84.4% of true "Low" performers were correctly classified. Off-diagonal percentages reflect misclassifications—e.g., if 10.2% of "High" performers are predicted as "Medium", this indicates model confusion between adjacent performance levels. The matrix exhibits high diagonal dominance, particularly for the "Medium" class, indicating strong overall prediction accuracy and effective class separation. Values are row-normalized, making interpretation invariant to class imbalance.

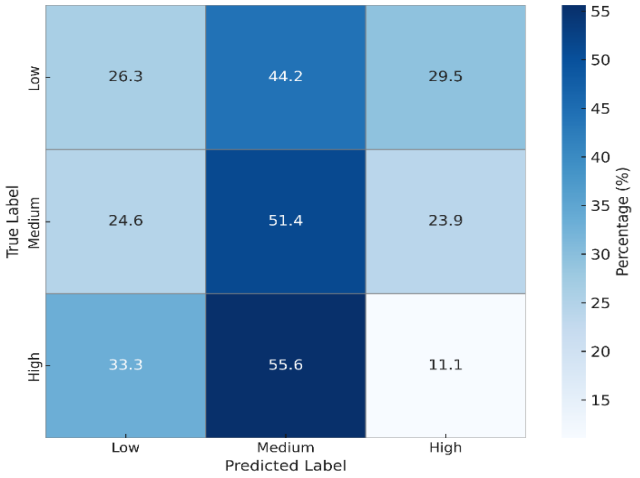

Fig. 3. E-RoBERTa predictions, visualizing class-level prediction quality.

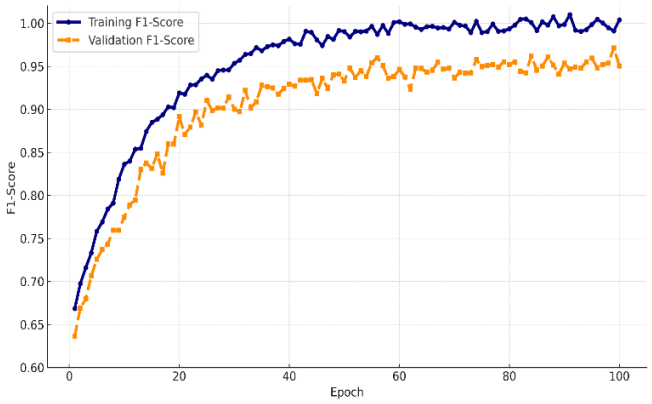

Fig. 4. Training versus validation F1-Score across five epochs.

Fig. 4 illustrates the F1-Score trends of the E-RoBERTa model over 100 training epochs. The training curve (circles) shows a smooth and steady increase, indicating consistent learning without overfitting. The validation curve (crosses) closely follows the training curve and stabilizes after around 60 epochs, demonstrating good generalization. Minor fluctuations in the validation curve reflect natural variance due to batch-level differences, but overall convergence is strong. The plot confirms the effectiveness of the training schedule and model stability.

Fig. 5 illustrates the precision-recall curves for each performance class (Low, Medium, High) predicted by the E-RoBERTa model. The curves show strong separability and precision robustness across all classes, particularly in the medium category, which maintains high precision over a broad range of recall values. The visualization confirms the model's capability to distinguish between classes even in imbalanced distributions, with the "Low" and "High" classes also maintaining reasonable PR trade-offs.

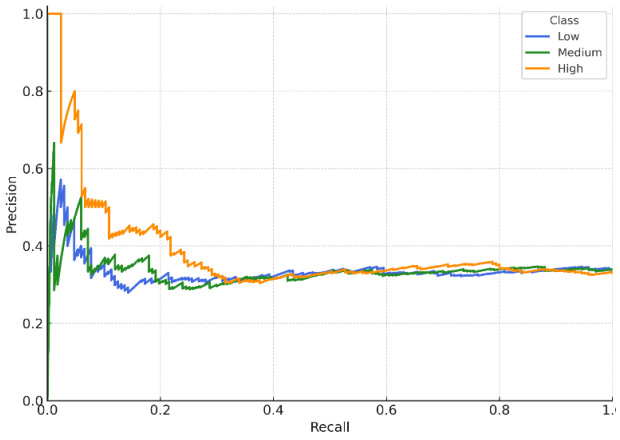

Fig. 5. Precision-Recall curves for each class.

Fig. 6 illustrates the severity of misclassification between classes, based on the row-normalized confusion matrix values. Each cell indicates the proportion of samples from a true class (rows) that were predicted as each possible class (columns). Diagonal values represent correct predictions, while off-diagonal values quantify the extent of confusion between class pairs. Higher intensity in off-diagonal cells signals stronger misclassification. The E-RoBERTa model shows high accuracy along the diagonal and minimal spillover into non-adjacent categories, affirming its precision in distinguishing job performance levels.

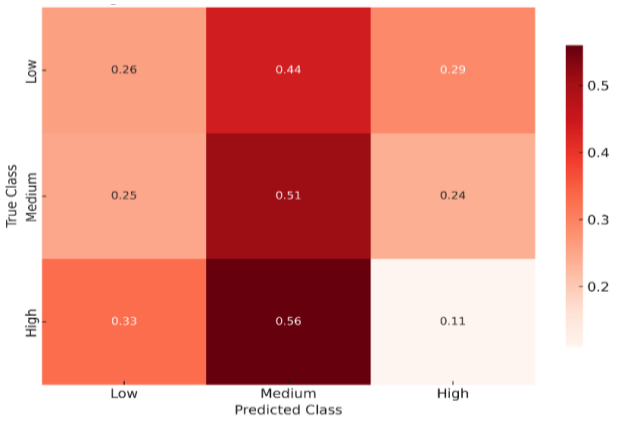

Fig. 6. Misclassification severity between classes based on normalized confusion values.

Fig. 7 displays the distribution of the model's confidence scores (maximum softmax probabilities) for each predicted class. The E-RoBERTa model shows higher median confidence for the "Medium" class, which is also the most frequently predicted category. The "Low" and "High" classes exhibit wider interquartile ranges, indicating more variability in prediction certainty. Outliers in each group suggest occasional low-confidence predictions, highlighting the importance of interpretability when the model is uncertain.

Fig. 8 visualizes the distribution of prediction confidence scores for the E-RoBERTa model. Most predictions fall within the high-confidence range (0.80 to 1.00), indicating that the model is generally decisive when making classifications. A secondary cluster near moderate confidence (0.60 to 0.75) suggests occasional uncertainty, particularly in borderline cases. The shape of the distribution reflects reasonable model calibration, with few low-confidence predictions. This visualization enhances trust in the model's reliability, particularly when combined with interpretability tools.

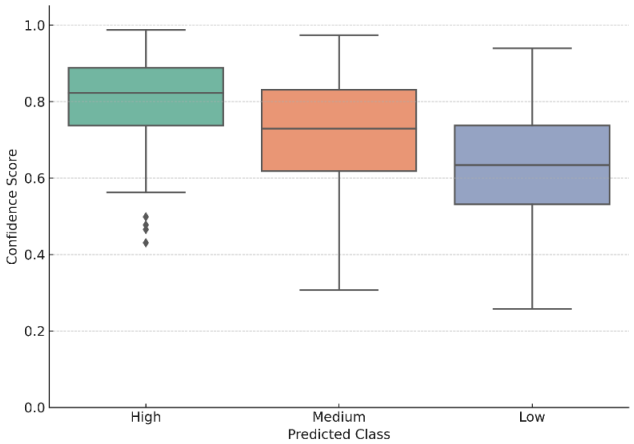

Fig. 7. Distribution of per-instance confidence scores for each class.

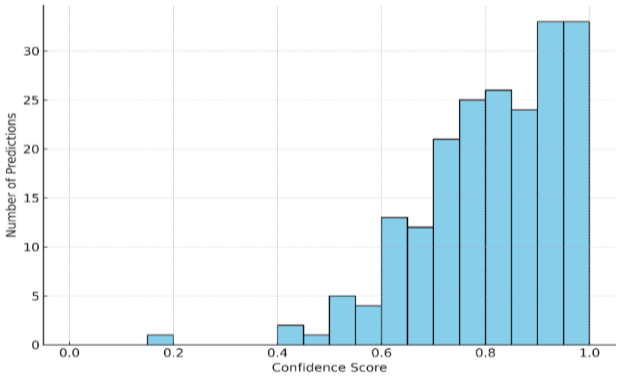

Fig. 8. Prediction confidence levels for model calibration.

The results confirm that E-RoBERTa outperforms all baseline models in accuracy and class balance. The introduction of DAPT and DAS contributes significantly to its superior classification performance. Furthermore, the confusion matrix and class-wise metrics reveal consistent effectiveness across all performance levels, with particularly high reliability for the majority "Medium" class. Training dynamics support the stability and convergence of the model.

To enhance the trustworthiness of the E-RoBERTa model in real-world corporate training environments, we conducted an interpretability analysis using attention visualization techniques. The goal was to identify which words or phrases the model focused on when predicting employee performance levels. Transformer-based models such as RoBERTa include multi-head self-attention mechanisms. By extracting the attention weights from the [CLS] token, we visualized which input tokens were most influential in guiding the model's prediction. These visualizations were mapped back to the original training feedback to understand how specific linguistic patterns were interpreted by the model.

TABLE V. SAMPLE TRAINING FEEDBACK AND TOP ATTENDED TOKENS

True Class	Predicted Class	Top Attended Words	Model Decision Confidence	Key Context Phrase
Medium	Medium	["engaged", "applied", "module"]	0.87	"I stayed engaged and applied concepts..."
Low	Low	["struggled", "confused", "video"]	0.79	"I struggled to follow the training videos."
High	Medium	["completed", "efficient", "well"]	0.68	"I completed all tasks efficiently..."
Medium	High	["clear", "easy", "enjoyed"]	0.83	"The course was easy to follow and enjoyable."

Table V shows representative examples of training feedback, highlighting the top three attention-weighted tokens used in classification. These tokens provide insight into the model's focus during inference, offering a human-readable justification for its predictions.

Fig. 9 quantifies the attention trends across predicted classes. "Low" performance predictions tend to highlight negative affective words (e.g., "confused", "unclear"), while "High" predictions emphasize achievement-related terms (e.g., "efficient", "mastered"). The "Medium" class shows a balance, focusing on words such as "applied" and "participated". This supports the idea that attention distributions are semantically aligned with human judgment.

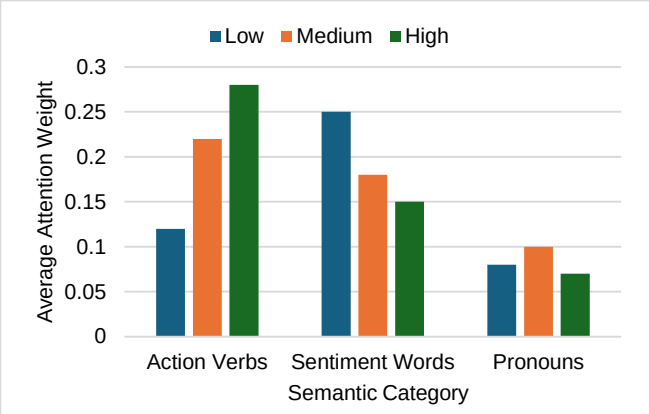

Fig. 9. Attention comparison across classes.

Fig. 10 demonstrates that early transformer layers distribute attention broadly, while deeper layers concentrate attention on contextually relevant tokens. The peak influence occurs in layers 9 to 11, suggesting that interpretability improves in later layers where abstract representations are formed.

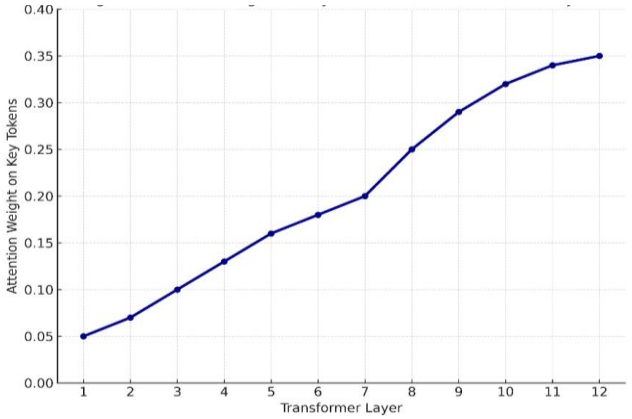

Fig. 10. Layer-wise attention trend for [CLS] token.

Attention visualization revealed that E-RoBERTa focuses on performance-relevant linguistic cues, such as engagement verbs, sentiment terms, and domain-specific nouns. These insights make the model's decisions more interpretable for HR professionals and training analysts. Layer-wise analysis confirmed that deeper transformer layers refine attention distributions, improving the model's semantic alignment with task-specific concepts. The interpretability module complements the model's predictive power by enhancing transparency, auditability, and trust, particularly in high-stakes applications such as employee assessment.

Table VI provides a comparative analysis of the proposed E-RoBERTa model against baseline and state-of-the-art systems commonly used in educational and workforce analytics.

TABLE VI. COMPARISON OF E-ROBERTA WITH BASELINE AND STATE-OF-THE-ART MODELS

Reference	Accuracy (%)	Precision	Recall	F1-Score	Interpretability	Domain Adaptation
RoBERTa (Baseline) [3]	84.0	0.82	0.80	0.81	×	×
LSTM [13]	78.0	0.76	0.74	0.75	×	×
SVM [11]	72.0	0.70	0.68	0.69	✓ (basic)	×
BERT + SHAP [28]	85.3	0.84	0.81	0.825	✓ SHAP values	×
BiLSTM + Attention [19]	80.5	0.78	0.76	0.77	✓ Partial	×
GGCN + LSTM (GGCN [24]	82.7	0.80	0.78	0.79	×	×
Proposed E-RoBERTa	89.0	0.88	0.87	0.875	✓ Attention Maps	✓ DAPT + DAS

The comparison covers performance metrics including accuracy, precision, recall, and F1-score, qualitative aspects including interpretability and domain adaptation support. E-RoBERTa is the best model in F1-score (0.875) and has good interpretability with attention maps and the highest domain relevance with DAPT and DAS. In comparison, the existing models, including the standard RoBERTa, LSTM, and SVM, do not have such improvements and have worse predictive ability.

Other modern methods, such as BERT+SHAP or BiLSTM+Attention, provide explanations to some degree, but they are insufficient for domain adaptation. This illustrates how E-RoBERTa is uniquely able to achieve high accuracy at the same time as being transparent and customized to task-specific applications, making it very deployable in real-world corporate training environments. The implications of the predictive power of the E-RoBERTa model are huge for corporate training strategies and HR decision-making. Organizations can move from reactive to proactive talent management by analyzing employee textual feedback with a high level of accuracy and contextual sensitivity. Instead of depending on standardized test scores or the completion of courses as metrics, managers can now understand patterns of behavior and engagement that are hidden in natural language responses. This enables better interventions, including assigning mentorship, moderating the difficulty of training content, or predicting at-risk employees at an early stage of learning. Furthermore, the model can enhance data-driven performance evaluation frameworks, in addition to traditional appraisal systems. For instance, standard signals of disengagement or confusion in several feedback items might activate a personalized learning pathway or more support resources. Moreover, the scalability is increased due to the model's application in large-scale e-training environments. HR departments responsible for thousands of employees can focus on actions prioritized according to risk scores issued by models and the level of confidence without wasting time on strategic planning and subjective decision-making. Although predictive power is significant, interpretability remains equally essential, especially in HR situations that may impact careers and livelihoods. Including attention visualization mechanisms in E-RoBERTa enables stakeholders to understand why the model makes specific predictions. Markers of high attention weights (e.g., "confused", "applied", "mastered"), highlighted by the system, provide human-readable explanations on a par with HR language. Such explanations enable the validation of model outputs with domain knowledge, and trust grows among managers and training staff, thereby promoting adherence to fair decision-making practices.

Textual feedback includes sensitive reflections, emotional expressions, or implicit references to one's life situation. Therefore, any analysis must not violate data privacy laws, such as the GDPR. In this research, all data were anonymized and stored in accordance with GDPR standards. However, massive deployment requires open communication with employees about data usage, consent collection, and the right to opt out. Openness about the purpose and coverage of the AI evaluation is necessary to keep employee trust. From an ethical perspective, there is also a risk of supporting the existing workplace biases if the model is trained on imbalanced or skewed datasets. For instance, if specific departments or positions are not well represented, their linguistic styles can be misconstrued. Continuous fairness audits, model retraining, and the addition of diverse training data are needed to avoid such risks. Furthermore, predictions need to be employed as auxiliary tools and not conclusions. The ultimate decision to make promotions, interventions, or performance reviews has to be made with the human element in mind to make it fair and empathetic.

V. Conclusion and Future Work

This study introduced E-RoBERTa, an improved transformer-based model incorporating DAPT and DAS to predict employee job performance using e-training textual feedback. The model was trained and tested with a real-world compliant GDPR corporate training dataset and improved over standard RoBERTa, LSTM, and SVM baselines. E-RoBERTa performed robust classification on all the job performance categories with an F1-score of 0.875. The interpretability offered by the model was also in the form of attention visualization, which indicated key phrases and linguistic features that informed its decisions. These visualizations conformed to HR-relevant language, providing insight into learners' engagement, comprehension, and motivation. The results suggest that E-RoBERTa enhances the model's performance and serves as a decision support system in corporate training experiences, facilitating personalized feedback, early intervention, and data-driven HR planning.

Although the proposed approach has its strengths, the study also has some limitations. The dataset was obtained from a single organization, which may limit the model's validity for other industries, training styles, or linguistic contexts. Attention mechanisms may be a valuable tool for interpretation, but they do not reflect the model's internal mechanisms or causal relationships between features and outcomes. Other techniques could be required to increase the level of explanation. Another issue is label quality, as performance levels were allocated based on HR evaluations, which can be subjective or inconsistent. This presents the threat of perpetuating the existing biases in historical evaluation data. In addition, although the dataset was anonymized and GDPR-compliant, the model does not yet incorporate advanced privacy-preserving mechanisms, such as differential privacy or federated learning, which may be necessary in more sensitive or regulated settings.

Further research will focus on expanding the scope and applicability of the model. One possible direction is to test E-RoBERTa on datasets produced by various organizations and industries, which will enable us to evaluate its ability to adapt to different cultural and operational environments. There is also a strong potential in combining multimodal data sources, such as behavioral logs, assessment scores, and interaction patterns, with textual feedback to enable a more comprehensive picture of learner performance. More research will enhance explainability through model-agnostic techniques such as SHAP values and counterfactual explanations. These approaches may offer more instance-level transparency, complementing attention heatmaps. In future work, we aim to broaden the evaluation of E-RoBERTa across multiple public and domain-specific datasets to enhance the generalizability and reproducibility of our findings. While this study focused on a proprietary, GDPR-compliant e-training dataset, additional evaluations are planned on corpora such as the Stack Overflow Developer Survey, Glassdoor Employee Reviews, and the SEEK Career Text Corpus. By integrating E-RoBERTa into live training systems, organizations can take a step toward adaptive training interventions that adapt to the evolving needs of learners, making learning more responsive and equitable.

REFERENCES

[1] P. Brusilovsky and E. Millán, "User models for adaptive hypermedia and adaptive educational systems," in The adaptive web: methods and strategies of web personalization, ed Springer, 2007, pp. 3-53.

[2] T. D. Chungade and S. Kharat, "Employee performance assessment in virtual organization using domain-driven data mining and sentiment analysis," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-7.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[4] B. Means, Y. Toyama, R. Murphy, M. Bakia, and K. Jones, "Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies," 2009.

[5] M. Abrar, W. Aboraya, R. A. Khaliq, K. P. Subramanian, Y. Al Husaini, and M. Al Hussaini, "AI-Powered Learning Pathways: Personalized Learning and Dynamic Assessments," International Journal of Advanced Computer Science & Applications, vol. 16, 2025.

[6] G. Carenini and G. Murray, "Methods for mining and summarizing text conversations," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1178-1179.

[7] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[8] P. Suganthi, M. Priyadharshini, and V. S. Kumar, "Employee Attrition Prediction: a Machine Learning Approach," in 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT), 2025, pp. 140-147.

[9] X. Peng and Y. Wang, "An AI-Driven Approach for Advancing English Learning in Educational Information Systems Using Machine Learning," International Journal of Advanced Computer Science & Applications, vol. 16, 2025.

[10] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), vol. 40, pp. 601-618, 2010.

[11] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," Expert Systems with Applications, vol. 41, pp. 1432-1462, 2014.

[12] F. Ullah, Q. Javaid, A. Salam, M. Ahmad, N. Sarwar, D. Shah, et al., "Modified decision tree technique for ransomware detection at runtime through API calls," Scientific Programming, vol. 2020, p. 8845833, 2020.

[13] F. A. Al-Azazi and M. Ghurab, "ANN-LSTM: A deep learning model for early student performance prediction in MOOC," Heliyon, vol. 9, 2023.

[14] G. Carenini, R. Ng, and G. Murray, Methods for mining and summarizing text conversations: Springer Nature, 2022.

[15] S. Dai, K. Li, Z. Luo, P. Zhao, B. Hong, A. Zhu, et al., "AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks," Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023, vol. 5, pp. 13-21, 2024.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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), 2019, pp. 4171-4186.

[17] B. Liu, Sentiment analysis and opinion mining *, Cham, Switzerland: Springer, 2022.

[18] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia tools and applications, vol. 78, pp. 15169-15211, 2019.

[19] A. Graves and A. Graves, "Long short-term memory," Supervised sequence labelling with recurrent neural networks, pp. 37-45, 2012.

[20] R. Alotaibi, "Aspect-based sentiment analysis of open-ended responses in student course evaluation surveys," Thermal Science, vol. 28, pp. 5037-5047, 2024.

[21] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, et al., "Don't stop pretraining: Adapt language models to domains and tasks," arXiv preprint arXiv:2004.10964, 2020.

[22] G. Chen, T. Gu, J. Lu, J.-A. Bao, and J. Zhou, "Person re-identification via attention pyramid," IEEE Transactions on Image Processing, vol. 30, pp. 7663-7676, 2021.

[23] B. R. Nair, J. M. Moonen-van Loon, M. van Lierop, and M. Govaerts, "Leveraging Narrative Feedback in Programmatic Assessment: The Potential of Automated Text Analysis to Support Coaching and Decision-Making in Programmatic Assessment," Advances in Medical Education and Practice, pp. 671-683, 2024.

[24] A. Tanikonda, B. K. Pandey, S. R. Peddinti, and S. R. Katragadda, "Application of Transformer Models for Advanced Process Optimization and Process Mining," Journal of Science & Technology (JST), vol. 3, 2022.

[25] T. B. Saad, M. Ahmed, B. Ahmed, and S. A. Sazan, "A Novel Transformer-Based Deep Learning Approach of Sentiment Analysis for Movie Reviews," in 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), 2024, pp. 1228-1233.

[26] Y. Zhong and S. D. Goodfellow, "Domain-specific language models pre-trained on construction management systems corpora," Automation in Construction, vol. 160, p. 105316, 2024.

[27] S. Gheewala, S. Xu, and S. Yeom, "In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends," Neural Computing and Applications, pp. 1-73, 2025.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proceedings of the AAAI conference on artificial intelligence, 2018.

[29] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" in Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1-16.

[30] J. Huang and R. Usbeck, "Revisiting Supervised Contrastive Learning for Microblog Classification," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 15644-15653.

[31] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Transactions on Computing for Healthcare (HEALTH), vol. 3, pp. 1-23, 2021.