



Leakage-Aware Arabic Emotion Classification for E-Learning Using Domain-Adaptive Transformer Models and Prototype Feedback Support

Samar A. Jabbar^{1*}, Suhad A. Yousif²

¹ University of Information Technology and Communications (UOITC), Baghdad 10001, Iraq

² Department of Computer Science, College of Science, Al-Nahrain University, Baghdad 64074, Iraq

Corresponding Author Email: MS202430019@uoitc.edu.iq

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310911>

ABSTRACT

Received: 14 March 2026

Revised: 22 July 2026

Accepted: 1 September 2026

Available online: 30 September 2026

Keywords:

Arabic emotion classification, domain-adaptive pretraining, e-learning, Iraqi dialect, label leakage, MARBERTv2

Emotion-aware learning systems require reliable text-based affect detection, particularly for low-resource dialects in which linguistic variation and dataset bias remain major challenges. This study presents a leakage-aware Arabic emotion classification framework for Iraqi e-learning environments that combines domain-adaptive transformer modeling with a prototype feedback support mechanism. A Student-Emotion dataset comprising 1,176 valid Iraqi Arabic responses from university learners was constructed and annotated into six emotion categories. To investigate the effect of shortcut learning, two preprocessing settings, RAW and CLEAN, were compared. The CLEAN setting removed explicit emotion-related lexical cues and duplicate responses before model evaluation. Classical term frequency-inverse document frequency (TF-IDF)-based classifiers and four transformer architectures were evaluated under a unified experimental protocol. Domain-adaptive pretraining (DAPT) was subsequently applied using an unlabeled Iraqi Arabic corpus to reduce dialect mismatch before supervised fine-tuning. The experimental results showed that MARBERTv2 with DAPT achieved the highest performance, with an accuracy of 0.9153 and a Macro-F1 score of 0.9167 under the leakage-controlled test setting. The predicted emotional states were then integrated into a prototype rule-and-retrieval feedback mechanism to demonstrate the potential use of emotion-aware learner support. Overall, the findings highlight the importance of leakage-controlled evaluation and dialect-adaptive language modeling for developing reliable Arabic affective computing applications in educational environments.

1. INTRODUCTION

Emotional states influence student engagement, motivation, satisfaction, performance, and learning outcomes. Affect is therefore an important component of the educational experience [1]. Self-regulated learning also contributes substantially to academic performance in online and blended learning environments [2]. In these settings, emotional states may reveal difficulties that are not immediately apparent from academic performance alone. For example, anxiety may reflect assessment pressure or technical uncertainty, whereas frustration may result from usability problems or unclear instructional requirements. Identifying such states can support the design of more appropriate educational interventions.

Emotion classification is particularly relevant to e-learning because many behavioral and facial cues available in face-to-face classrooms are limited or absent in online environments. Previous studies have examined emotion recognition as a basis for adaptive educational interventions [3, 4]. Student-written responses, reflections, course reviews, and feedback therefore provide accessible and relatively non-intrusive sources of information about learners' experiences.

Arabic affective natural language processing (NLP) remains challenging because of morphological complexity,

orthographic variation, and dialectal diversity [5]. These challenges are particularly pronounced in informal Iraqi Arabic, which includes colloquial vocabulary, non-standard spelling, and expressions that differ from Modern Standard Arabic. Although several Arabic sentiment and emotion resources are available, resources specifically targeting Iraqi Arabic remain limited. Existing resources are also derived mainly from social-media discourse rather than learner-generated educational text [6, 7].

Pretrained transformer models provide contextual representations that can address some forms of linguistic variability in Arabic text. Bidirectional Encoder Representations from Transformers (BERT) and Arabic-specific models, including ARBERT and MARBERT, have demonstrated effective transfer learning across Arabic varieties [8, 9]. However, transferring models from general-domain or social-media text to educational text may introduce domain mismatch. Domain-adaptive pretraining (DAPT) can reduce this mismatch by further adapting a pretrained model to relevant unlabeled text before supervised fine-tuning [10, 11].

Reliable evaluation is also important because self-reported emotion datasets may contain explicit emotion names or lexical cues that directly indicate the target label. Classifiers

may exploit these patterns and produce overly optimistic performance estimates without learning the intended contextual relationships. This behavior represents a form of shortcut learning [12]. To examine this risk, the present study compares RAW and CLEAN preprocessing settings. The CLEAN setting removes predefined label-indicative lexical cues and exact duplicate responses to provide a more conservative evaluation.

Existing Iraqi Arabic resources illustrate both the linguistic and domain gaps addressed in this study. Askar and Sjarif [6] developed a Mesopotamian-Iraqi social-media corpus for sentiment and emotion classification. Almahdawi and Teahan [7] introduced an Iraqi Facebook dataset annotated according to Ekman's six basic emotions. In the present study, the text from the latter dataset was used without its original labels for Iraqi-dialect DAPT. Although both resources are valuable for Iraqi Arabic affective NLP, they consist of general social-media content rather than educational responses produced by students.

Arabic educational NLP has also received increasing attention. Recent studies have analyzed student feedback and course evaluations using sentiment-analysis and transformer-based methods [13]. However, this work has focused primarily on sentiment polarity or general course evaluation. Fine-grained emotion classification of Iraqi Arabic educational text under leakage-aware evaluation therefore remains insufficiently investigated.

The present study addresses these limitations through a six-class emotion-classification framework based on responses written by Iraqi university students about their e-learning experiences. The study compares RAW and CLEAN preprocessing settings and evaluates classical term frequency-inverse document frequency (TF-IDF)-based classifiers together with four transformer models. The transformer models are assessed before and after Iraqi-dialect DAPT. Their predicted emotions are then integrated into a prototype rule-and-retrieval feedback module. The novelty of the study therefore does not lie in introducing transformer models or DAPT as new techniques. Rather, it lies in evaluating and integrating these methods within a leakage-aware framework designed specifically for Iraqi Arabic educational text.

The main contributions of this study are as follows:

- (a) A domain-specific Iraqi Arabic Student-Emotion dataset was constructed from university students' descriptions of their e-learning experiences.
- (b) A leakage-aware evaluation strategy was developed to examine the effects of predefined lexical-cue removal and exact duplicate removal on emotion-classification performance.
- (c) Classical TF-IDF-based classifiers and four transformer models were systematically evaluated, with the transformer models assessed before and after DAPT.
- (d) The predicted emotions were integrated into a prototype hybrid rule-and-retrieval feedback module for emotion-aware learner support.
- (e) The model confidence score was incorporated as an auxiliary signal for downstream feedback selection.

2. RELATED WORK

Arabic educational NLP has increasingly examined learner-generated text to identify students' experiences and attitudes.

Ali [14] analyzed Arabic tweets about online learning using sentiment- and emotion-mining techniques. Aljabri et al. [15] investigated Arabic tweets related to distance learning in Saudi Arabia during the COVID-19 pandemic. Both studies demonstrated the applicability of Arabic NLP to online-learning contexts. However, their datasets were collected from public social media rather than directly from students in structured educational settings.

More recent studies have focused on educational feedback. Alzaid and Fkih [16] analyzed Arabic e-learning feedback from social-media posts using a hybrid fuzzy and bidirectional long short-term memory (BiLSTM) model. Louati et al. [17] applied a Support Vector Machine (SVM) to Arabic course reviews. Alamoudi et al. [13] evaluated classical machine-learning models and AraBERT using Arabic student-evaluation text. Collectively, these studies demonstrate the growing use of Arabic NLP for educational feedback analysis. However, they primarily address sentiment polarity or general course evaluation rather than fine-grained emotion classification and emotion-aware learner support.

Arabic emotion recognition has also advanced in recent years. Khalil et al. [18] applied deep-learning methods to emotion analysis of Arabic tweets. Alqahtani and Alothaim [19] reviewed Arabic emotion-analysis models, datasets, and pretrained language models. Althobaiti [20] introduced a fine-grained Arabic emotion dataset, while subsequent research proposed a stacking ensemble with self-training for low-resource Arabic emotion recognition [21]. Together, these studies demonstrate the increasing use of contextual and transfer-learning approaches for fine-grained Arabic emotion classification.

Research on Iraqi Arabic has focused primarily on sentiment and emotion analysis in social-media text. Almahdawi and Teahan [7] introduced an Iraqi Facebook corpus annotated with six basic emotions. Alfarhany and Abdullah [22] evaluated deep-learning methods for Iraqi sentiment and emotion analysis using social-media datasets. Hussein and Lakizadeh [23] compared machine-learning, deep-learning, and AraBERT-based models for Iraqi-dialect sentiment analysis. In their evaluation, AraBERT achieved the strongest overall performance. Although these studies provide evidence for Iraqi-dialect modeling, their datasets are based predominantly on social-media content rather than educational text.

Emotion recognition in learning environments has also received increasing attention. Yu et al. [3] reviewed 117 studies on automated emotion recognition in online learning. They reported that previous research focused mainly on technical system development, with comparatively less attention to educational integration and deployment. Research has also expanded from emotion detection to automated educational feedback. Parker et al. [24] used large language models (LLMs) to analyze 2,500 end-of-course survey comments through classification, information extraction, thematic analysis, and sentiment analysis. Morris et al. [25] examined transformer-based methods for providing automated formative feedback on student-authored summaries. However, these studies did not combine Iraqi-dialect emotion classification, leakage-aware evaluation, and a rule-and-retrieval feedback prototype within a single framework.

Three research gaps emerge from the reviewed literature. First, Iraqi Arabic affective resources are based predominantly on social-media discourse rather than education-specific student responses. Second, Arabic educational NLP studies

focus mainly on sentiment polarity or general feedback analysis rather than fine-grained emotion classification. Third, previous studies rarely integrate Iraqi-dialect emotion modeling, leakage-aware evaluation, DAPT, and downstream emotion-aware feedback within a single framework. The present study addresses these gaps through a purpose-built Iraqi Arabic Student-Emotion dataset, RAW and CLEAN preprocessing settings, transformer evaluation before and after DAPT, and a prototype rule-and-retrieval feedback module.

3. METHODOLOGY

Figure 1 illustrates the architecture of the proposed emotion-aware Arabic e-learning framework. The framework was designed to provide reproducible and interpretable emotion classification together with a prototype feedback module. In this study, emotion classification refers to the supervised prediction of a single dominant emotion from student-written text. Emotion recognition refers to the broader research field, whereas emotion-aware support refers to the use of predicted emotions to guide learner feedback.

Student-written Iraqi Arabic text was processed using two configurations: RAW and CLEAN. As described in Section 3.2, RAW retained validity-filtered text after whitespace removal, whereas CLEAN additionally removed exact duplicate responses and predefined label-indicative lexical cues to reduce shortcut learning and label leakage.

The processed text was evaluated using classical and transformer-based classifiers. The transformer models were also evaluated after DAPT on a separate Iraqi Arabic corpus. The predicted emotion and its confidence score were then passed to the feedback module. The primary pathway used a hybrid rule-and-retrieval strategy, whereas an optional Gemini-based pathway generated short emotion-conditioned feedback. Emotion classification was evaluated using a fixed development/test split and stratified cross-validation (CV). The feedback module was demonstrated qualitatively as a proof-of-concept prototype.

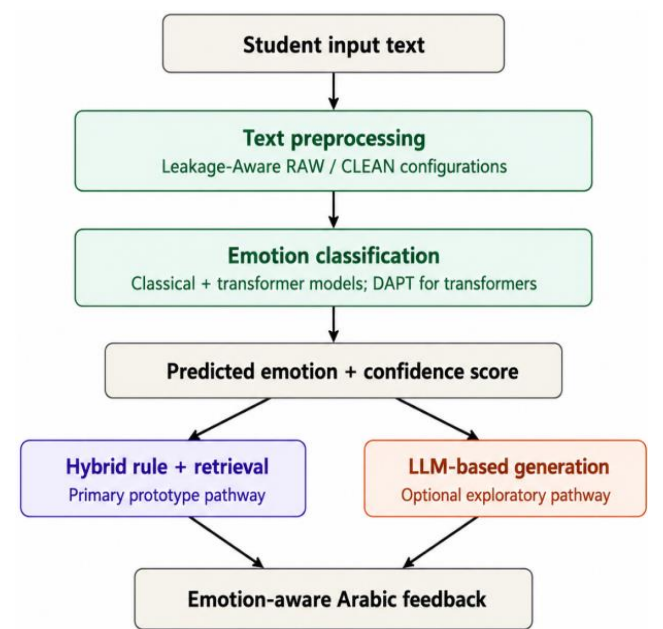


Figure 1. Overall architecture of the proposed emotion-aware Arabic e-learning framework

3.1 Description and collection of the dataset

The Student-Emotion dataset was collected through a structured Google Forms questionnaire distributed across public and private Iraqi universities. Questionnaire distribution was supported by an official facilitation letter issued by the University of Information Technology and Communications (UOITC). Data collection lasted approximately five months, from September 2025 to February 2026. Participation was voluntary and based on informed consent, and no personally identifying information was retained.

The questionnaire contained eight items covering academic status, gender, age group, e-learning experience, platform-use frequency, and the main difficulty encountered during e-learning. Participants also selected one dominant emotion from six predefined categories: Joy, Anger, Anticipation, Trust, Neutral, and Anxiety. An open-ended item asked participants to describe their emotional experience in Iraqi Arabic. Each labeled instance consisted of the selected emotion and its corresponding free-text response.

The six-category scheme was adopted as a compact, task-specific taxonomy rather than as a comprehensive representation of academic emotions. Joy and Trust represented broadly positive responses, whereas Anger and Anxiety represented broadly negative responses. Neutral represented a non-polar state, while Anticipation represented a future-oriented state. Trust reflected confidence in technology-mediated learning, whereas Anticipation reflected expectations regarding future learning experiences or improvements. The categories were selected to capture affective states that could be directly interpreted from students’ self-reported e-learning experiences and subsequently mapped to different forms of learner support.

The selected taxonomy does not imply that Trust and Anticipation are more important than other learning-related emotions. Frustration, confusion, and boredom are also discussed in educational and affective-learning research [1, 4]. However, these states were not included in the predefined questionnaire. The six-class scheme should therefore be interpreted as a task-specific representation of students’ e-learning emotions rather than as a comprehensive taxonomy of academic affect. Its limited coverage is acknowledged as a dataset limitation.

After initial validation, empty and invalid responses were removed, resulting in 1,176 valid instances. Because no persistent personal identifier was retained, the number of unique participants could not be verified. The records are therefore reported as valid responses rather than as unique participants. Further leakage-control procedures, including duplicate removal, are described in Section 3.2. The six classes were relatively balanced, with proportions ranging from 15.5% to 18.7%.

Although background information was collected, aggregate demographic statistics were not retained in the anonymized modeling corpus. Institution-specific distributions were also unavailable because the questionnaire did not require an institution identifier. These limitations restrict the assessment of dataset representativeness and generalizability. Table 1 presents representative responses from the six emotion categories.

The final modeling corpus contained three fields: a numerical identifier, an Iraqi Arabic free-text response (*Post*), and its corresponding emotion label (*Label*).

Table 1. Representative Iraqi Arabic examples from the six emotion categories

Emotion	Example Iraqi Arabic Response	Literal English Translation
Joy	مناسب لتجنب الازدحامات خصوصا في بغداد بسبب الازدحامات	Suitable for avoiding traffic congestion, especially in Baghdad
Trust	احسن شي صار وكلش عندي ثقة بي	It is one of the best things that happened, and I have great confidence in it
Neutral	ما زال التعليم بشكل حضوري افضل واكثر تأثيرا من التعليم الالكتروني	Face-to-face education is still better and more effective than e-learning
Anger	اشعر بالغضب لأن هذا النظام ممل ويشعروني بالنعاس عند استخدامه	I feel angry because this system is boring and makes me sleepy when I use it
Anticipation	بالاستفادة ونتأمل المزيد من التطور	We benefit from it and hope for further development
Anxiety	التعليم الالكتروني حلو لكن المشكلة هي الالتزام وضعف الانترنت	E-learning is good, but the problem is commitment and weak internet

3.2 Preprocessing and leakage control (RAW vs. CLEAN)

All student responses underwent validity filtering before the two preprocessing configurations were generated. Records with missing or empty *Post* or *Label* values were excluded. Leading and trailing whitespace was also removed from all textual fields.

Two configurations were evaluated: RAW and CLEAN. RAW retained the validity-filtered text without additional leakage-control procedures. CLEAN removed predefined lexical cues that explicitly indicated the six target emotions.

The cue list included مسعود، سعيد، فرح for Joy; غاضب، غضب for Anger; واثق، ثقة for Trust; توقع، انتظار for Anticipation; قلق، خوف، خائف، متوتر for Anxiety; and محايد، حياد for Neutral. The corresponding English class names were also included.

After cue removal, empty responses were discarded. Exact duplicate records were then removed using the processed *Post* and *Label* fields. These operations directly defined the CLEAN configuration used in the experiments. Table 2 summarizes the differences between the RAW and CLEAN configurations.

Table 2. Comparison of RAW and CLEAN preprocessing settings

Aspect	RAW	CLEAN
Basic validity filtering and whitespace stripping	✓	✓
Missing/empty Post–Label filtering	✓	✓
Predefined label-cue removal	✗	✓
Empty-text filtering after cue removal	✗	✓
Exact duplicate removal based on Post + Label	✗	✓
Fixed random seed and evaluation protocol	✓	✓
Explicit lexical shortcut risk	Higher	Reduced

The cue list was manually defined rather than derived from the evaluation data. The first author selected explicit emotion names and common Arabic and Iraqi Arabic variants associated with the six categories. The second author subsequently reviewed the list. This procedure was intended to reduce direct lexical shortcuts and encourage the classifiers

to rely more strongly on contextual information.

Because the procedure was lexicon-based, expressions not included in the predefined list could remain in the processed text. These expressions included unseen Iraqi Arabic terms, morphological variants, synonyms, and indirect paraphrases. Additional terms could be incorporated into future versions of the lexicon. Therefore, CLEAN reduced explicit lexical label leakage and shortcut learning but did not eliminate all possible semantic cues.

3.3 Modeling setup: Classical baselines and transformer fine-tuning

Classical machine-learning baselines and transformer models were evaluated using the same CLEAN preprocessing configuration, fixed development/test split, random seed, and evaluation protocol. This setup enabled a controlled comparison between non-contextual lexical representations and contextual pretrained representations.

For the classical baselines, Iraqi Arabic text was represented using TF-IDF word unigrams and bigrams. Classification was performed using a Calibrated Linear Support Vector Classifier (SVC) and Complement Naive Bayes. These models served as lightweight, non-contextual reference baselines. However, their reliance on lexical-frequency features limited their ability to capture contextual information, negation, and subtle distinctions among related emotions.

Four pretrained transformer models were evaluated. MARBERTv2 was selected because it was pretrained on Arabic social-media and dialectal text, which made it suitable for informal Iraqi Arabic. AraBERTv2 was included as an Arabic-specific baseline. Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) were included as multilingual baselines with different pretraining regimes. Each transformer model was fine-tuned for six-class, single-label emotion classification by adding a classification head and updating the model parameters using the labeled Student-Emotion dataset.

The selected models represented complementary modeling approaches, including conventional lexical, Arabic-specific, dialect-oriented, and multilingual representations. This model set was considered appropriate for the comparative objectives of the study because all approaches could be evaluated under the same preprocessing and evaluation protocol. The comparison was designed to address two experimental questions: whether contextual pretrained representations outperformed TF-IDF-based baselines and whether Iraqi-dialect DAPT improved transformer performance. The objective was not to provide an exhaustive benchmark of recent Arabic emotion-classification architectures. Instead, the study aimed to conduct a controlled comparison aligned with its specific experimental objectives.

3.4 Domain-adaptive pretraining

A two-stage training strategy was adopted to address Iraqi-dialect variation and the limited amount of labeled educational data. In the first stage, DAPT was performed using the Iraqi Arabic Emotion Corpus introduced by Almahdawi and Teahan [7]. The original corpus was manually collected from Facebook between December 2016 and August 2018 using emotion-related seed words and Facebook search filters, including a Baghdad-Iraq location constraint. Manual collection was used to reduce noise from non-textual content, such as links and images.

The original study reported 1,365 posts and evaluated the

corpus using four external annotators, with an average inter-annotator agreement of 0.751. The version used in the present experiments contained 1,361 Iraqi Arabic Facebook posts covering the same six Ekman emotion categories: Anger, Disgust, Fear, Happiness, Sadness, and Surprise.

Only the text field was used during DAPT. The original emotion labels were discarded because adaptation was performed through masked language modeling (MLM) rather than supervised emotion classification. Missing and empty entries were removed, and the remaining posts were treated as unlabeled Iraqi Arabic text. This stage adapted the transformer representations to Iraqi-dialect vocabulary, colloquial expressions, informal writing styles, and non-standard linguistic patterns.

The Facebook corpus was selected because it shared important linguistic characteristics with the Student-Emotion dataset. However, the two corpora represented different domains. The Facebook posts covered broad social and personal topics, whereas the Student-Emotion responses focused on academic, technical, and organizational aspects of e-learning. Therefore, the Facebook corpus was used only for dialect-oriented adaptation and not as a substitute for educational training data.

In the second stage, the adapted models were fine-tuned on the labeled Student-Emotion dataset for six-class emotion classification. Task-specific learning was based exclusively on the educational dataset, while the Facebook corpus provided only additional exposure to Iraqi Arabic. The held-out Student-Emotion test data were excluded from both DAPT and fine-tuning and were used only for final evaluation. Figure 2 illustrates the two-stage training protocol and the separation of the adaptation, fine-tuning, and test data.

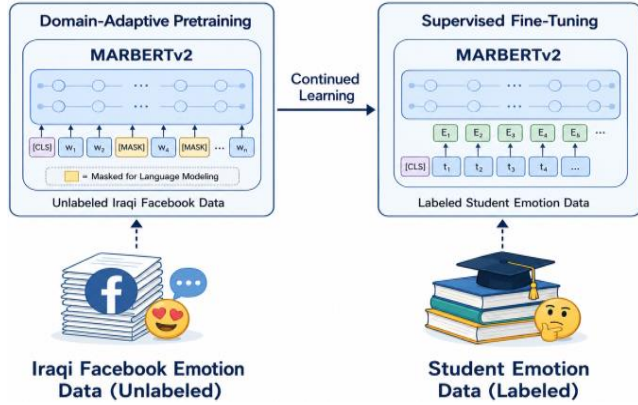


Figure 2. Leakage-aware data separation protocol

3.5 Labeling strategy, annotation reliability, and confidence-based feedback control

Each Student-Emotion instance consisted of a short text paired with one self-selected emotion category from the predefined questionnaire. To assess annotation reliability, 200 responses were randomly selected and independently labeled by a second annotator who was blinded to the original labels. Inter-rater agreement was measured using Cohen's kappa, as shown in Eq. (1):

$$\kappa = (P_o - P_e) / (1 - P_e) \quad (1)$$

where, P_o denotes the observed agreement between the two raters, and P_e denotes the agreement expected by chance.

Annotation reliability and model confidence served different purposes within the proposed framework. Cohen's kappa was used during dataset validation to assess agreement between human label assignments. At inference time, the classifier produced a predicted emotion label $\hat{e}$ and an uncalibrated model-confidence score c . The score was defined as the maximum Softmax probability across the emotion classes, as shown in Eq. (2):

$$c = \max(\text{Softmax}(z)) \quad (2)$$

where, z denotes the model logits.

The confidence score was used as an auxiliary control signal rather than as a calibrated measure of predictive reliability. When $c < 0.40$, the rule-and-retrieval pathway produced a general clarification-oriented response instead of making a strong emotion-specific assumption. Predefined greeting, gratitude, and congratulatory intents were excluded from this rule. When $c < 0.60$ and the probability assigned to a secondary emotion exceeded 0.25, the prototype could also include the secondary predicted emotion in the response.

These thresholds were heuristic prototype settings and were not derived through formal probability calibration. Accordingly, the confidence score should be interpreted as an operational control signal rather than as a calibrated estimate of predictive reliability. Formal probability calibration should be evaluated before confidence scores are used in higher-stakes educational decisions.

3.6 Hybrid emotion-aware feedback delivery via rule and retrieval, with an optional large language model-based alternative

The feedback module operated after emotion classification, as illustrated in Figure 3. Its primary pathway combined rule-based decision logic with the retrieval of previously stored responses. A large language model (LLM) pathway based on Gemini was included only as an optional exploratory alternative.

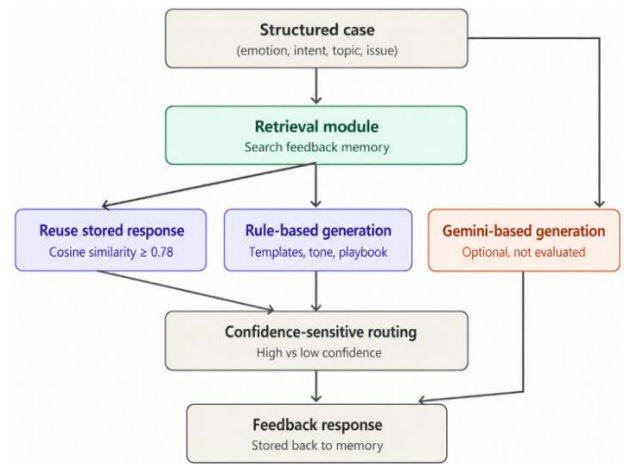


Figure 3. Hybrid emotion-aware feedback architecture with primary rule-and-retrieval and optional Gemini-based pathways

Given a student message x , the best-performing saved classifier predicted the emotion $\hat{e}$ and its associated confidence score c . In parallel, a lightweight context-detection module used keyword-based heuristics to infer the message intent or

topic and to identify the likely issue category, such as internet connectivity, platform access, time constraints, or deadlines. A short key phrase was also extracted using TF-IDF over sentence segments. The system then represented the case as the tuple $(\hat{e}, \text{intent}, \text{topic}, \text{issue})$.

The retrieval module searched the stored feedback memory, *feedback_database.pkl*, using multilingual sentence embeddings. A stored response was reused when its cosine similarity reached the predefined threshold of 0.78. Otherwise, the rule-based pathway constructed a response using curated templates, tone configurations, and issue-specific playbooks. Newly constructed rule-based responses could subsequently be stored for reuse in similar cases.

For example, a message expressing anxiety about weak internet connectivity could be assigned the emotion *Anxiety*, the intent *venting*, and the issue category *internet/platform*. If no stored response reached the similarity threshold, the system applied the corresponding rule-based blueprint, such as *Anxiety + venting*.

Gemini was used only to generate or paraphrase short Arabic responses based on the student message, predicted emotion, and confidence score. This optional pathway was not formally evaluated and was not required for prototype operation. The primary rule-and-retrieval pathway therefore operated independently of Gemini.

3.7 Evaluation metrics

Model performance was evaluated using accuracy, Macro-F1, and confusion matrices. Accuracy represents the proportion of correctly classified instances. Macro-F1 assigns equal weight to all emotion classes by averaging their class-specific F1 scores. Confusion matrices were used to examine class-level errors and recurring misclassification patterns.

For the six-class, single-label classification task, accuracy was calculated as shown in Eq. (3):

$$\text{Accuracy} = \frac{\sum_{i=1}^K C_{ii}}{N} \quad (3)$$

where, C_{ii} denotes the number of correctly classified instances for class i , and N denotes the total number of evaluated instances.

Precision and recall were calculated using Eqs. (4) and (5), respectively:

$$\text{Precision} = TP / (TP + FP) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

where, TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

The F1-score was calculated using Eq. (6):

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Finally, Macro-F1 was calculated using Eq. (7):

$$\text{Macro-F1} = \frac{1}{K} \sum_{i=1}^K F1_i \quad (7)$$

where, K denotes the number of emotion classes and $F1_i$

denotes the F1 score for class i .

4. EXPERIMENTS AND RESULTS

This section presents the experimental setup and evaluates the effects of preprocessing, classical and transformer-based emotion classification, DAPT, and the prototype feedback module.

4.1 Experimental setup

The RAW and CLEAN preprocessing configurations were used to evaluate the combined effects of lexical-cue removal and duplicate removal. Subsequent comparisons between the classical and transformer-based models used the CLEAN configuration. Fixed data splits and a random seed of 42 were applied to support reproducibility and comparability across models.

Table 3. Experimental settings and hyperparameters used for reproducibility

Component	Setting
Random seed	42 (Hugging Face set_seed, Python, NumPy, PyTorch, and CUDA)
Data split	Stratified development/test split = 85/15 (random_state = 42)
Cross-validation	Five-fold StratifiedKFold on the development set (shuffle = True, random_state = 42)
Models	MARBERTv2, AraBERTv2, mBERT, XLM-R, TF-IDF + Calibrated Linear SVC, and TF-IDF + Complement Naive Bayes
Tokenization	max_length = 128, truncation enabled, padding = "max_length"
Fine-tuning	3 epochs, learning rate = 2e-5, batch size = 16/32 (train/eval), weight decay = 0.01
DAPT corpus size	1,361 Iraqi Arabic Facebook posts
DAPT (MLM)	2 epochs, learning rate = 5e-5, batch size = 16, max_length = 128, mlm_probability = 0.15; final adapted checkpoint saved after training
Optimization consistency	The same fine-tuning and DAPT configurations were applied across MARBERTv2, AraBERTv2, mBERT, and XLM-R
Hardware	Google Colab; NVIDIA Tesla T4 GPU; CUDA 12.8
Training time	Final fine-tuning: approximately 1.12–1.20 min/model; five-fold CV: approximately 4.59–4.93 min/model; DAPT: approximately 1.52–3.37 min/model
Evaluation	Accuracy, Macro-F1, and confusion matrices RAW and CLEAN; CLEAN additionally removes predefined label-indicative cues
Preprocessing	removes predefined label-indicative cues, texts rendered empty after cue removal, and exact duplicate (<i>Post</i> , <i>Label</i>) records

Note: DAPT = domain-adaptive pretraining, BERT = Bidirectional Encoder Representations from Transformers, mBERT = multilingual BERT, XLM-R = XLM-RoBERTa, TF-IDF = term frequency-inverse document frequency.

Five-fold stratified cross-validation (CV) was performed on the development set to assess performance variability. The held-out test set was excluded from both training and CV and was reserved for final evaluation. The same fine-tuning configuration was applied to MARBERTv2, AraBERTv2, mBERT, and XLM-R. The controlled settings included the

number of epochs, learning rate, training and evaluation batch sizes, weight decay, maximum sequence length, and random seed.

A common DAPT configuration was also applied to all four transformer models. The DAPT corpus contained 1,361 Iraqi Arabic Facebook posts, as described in Section 3.4. The reproducibility runs were conducted in Google Colab using an NVIDIA Tesla T4 GPU with CUDA 12.8. Wall-clock training times were recorded during these runs. Final fine-tuning required approximately 1.12–1.20 minutes per model. Five-fold CV required approximately 4.59–4.93 minutes per model, whereas DAPT required approximately 1.52–3.37 minutes per model, depending on the transformer architecture.

Unless otherwise stated, all reported results correspond to the held-out test set. Table 3 summarizes the experimental configuration and hyperparameters.

4.2 Effects of label leakage (RAW vs. CLEAN)

MARBERTv2 was evaluated before DAPT under the RAW and CLEAN preprocessing configurations. The same data splits, fine-tuning configuration, and random seed were used in both conditions. Macro-F1 decreased from 0.9390 under RAW to 0.8830 under CLEAN. Accuracy similarly decreased from 0.9380 to 0.8810.

This decline indicated that the combined removal of explicit lexical cues and duplicate records produced a more conservative performance estimate. The higher RAW scores were consistent with the model benefiting from label-indicative words and repeated response patterns. However, because both operations were applied simultaneously, their individual effects could not be isolated. The CLEAN configuration was therefore used as the primary basis for subsequent model comparisons.

4.3 Classical machine-learning baselines

Two classical machine-learning baselines were evaluated under the CLEAN configuration: TF-IDF with a Calibrated Linear SVC and TF-IDF with Complement Naive Bayes. Both models used word unigram and bigram features and served as non-neural reference baselines for comparison with the transformer models.

As shown in Table 4, TF-IDF + Calibrated Linear SVC outperformed Complement Naive Bayes. It achieved a test accuracy of 0.8475 and a Macro-F1 of 0.8528, whereas Complement Naive Bayes achieved 0.8079 and 0.8104, respectively. Figures 4 and 5 present the corresponding confusion matrices.

Table 4. Performance of classical machine-learning baselines under the CLEAN configuration

Model	Feature Representation	Test Accuracy	Test Macro-F1
TF-IDF + Calibrated Linear SVC	TF-IDF word unigrams and bigrams	0.8475	0.8528
TF-IDF + Complement Naive Bayes	TF-IDF word unigrams and bigrams	0.8079	0.8104

Note: SVC = Support Vector Classifier, TF-IDF = term frequency-inverse document frequency.

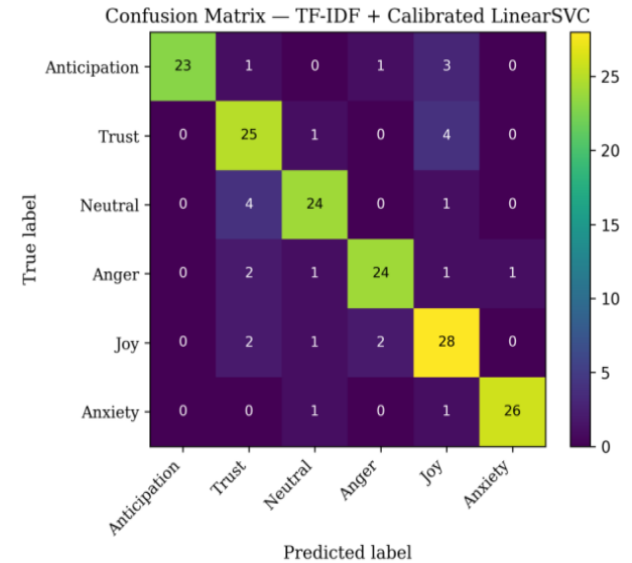


Figure 4. Confusion matrix for term frequency–inverse document frequency (TF-IDF) + Calibrated Linear Support Vector Classifier (SVC)

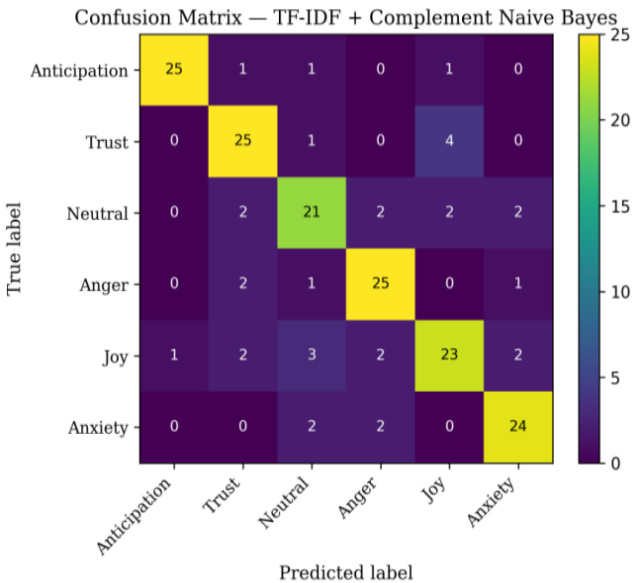


Figure 5. Confusion matrix for term frequency–inverse document frequency (TF-IDF) + complement Naive Bayes

The TF-IDF + Calibrated Linear SVC produced a greater concentration of predictions along the main diagonal, consistent with its higher test performance. Misclassifications occurred mainly among Trust, Joy, Neutral, and Anticipation. These categories can share similar lexical patterns despite representing different emotional states. Sparse TF-IDF features may therefore be insufficient to capture subtle contextual distinctions among related emotions. This limitation motivated the subsequent evaluation of contextual transformer models.

4.4 Fine-tuning transformer models before domain adaptation

MARBERTv2, AraBERTv2, mBERT, and XLM-R were fine-tuned under the CLEAN configuration. The same data splits, random seed, and fine-tuning configuration were used

to ensure a controlled comparison. As shown in Table 5, MARBERTv2 achieved the highest held-out test performance before DAPT, with an accuracy of 0.8814 and a Macro-F1 of 0.8832.

Table 5 also reports the five-fold CV Macro-F1 as the mean

and sample standard deviation (SD). Approximate 95% confidence intervals (CIs) were calculated across the five folds using the t -distribution. These intervals describe performance variability across the development-set partitions and complement the results obtained from the held-out test set.

Table 5. Transformer baselines before domain-adaptive pretraining (DAPT) under the CLEAN configuration

Model	Test Accuracy	Test Macro-F1	5-Fold CV Macro-F1 (Mean $\pm$ Sample SD)	Approx. 95% CI of Mean
MARBERTv2	0.881356	0.883190	0.8408 $\pm$ 0.0250	[0.8097, 0.8718]
AraBERTv2	0.802260	0.807848	0.7131 $\pm$ 0.0988	[0.5903, 0.8359]
mBERT	0.785311	0.790542	0.7355 $\pm$ 0.0451	[0.6796, 0.7915]
XLM-R	0.728814	0.735140	0.6847 $\pm$ 0.0456	[0.6281, 0.7414]

Note: BERT = Bidirectional Encoder Representations from Transformers, mBERT = multilingual BERT, XLM-R = XLM-RoBERTa, TF-IDF = term frequency-inverse document frequency, CV = cross-validation, CI = confidence interval.

4.5 Domain-adaptive pretraining

Table 6 presents the performance of the four transformer models after DAPT under the CLEAN configuration. On the held-out test set, DAPT increased the Macro-F1 scores of MARBERTv2, AraBERTv2, and XLM-R, whereas mBERT showed a slight decrease. XLM-R exhibited the largest observed improvement. However, MARBERTv2 with DAPT

achieved the highest overall test performance, with an accuracy of 0.9153 and a Macro-F1 score of 0.9167.

Table 6 also reports the five-fold CV Macro-F1 as the mean and sample standard deviation (SD). Approximate 95% CIs were calculated across the five folds. The observed effects of DAPT varied across models and should therefore be interpreted cautiously because of the limited dataset size and the variability observed across folds.

Table 6. Transformer performance after domain-adaptive pretraining (DAPT) under the CLEAN configuration

Model	Test Accuracy	Test Macro-F1	5-Fold CV Macro-F1 (Mean $\pm$ Sample SD)	Approx. 95% CI of Mean
MARBERTv2 + DAPT	0.915254	0.916676	0.8457 $\pm$ 0.0287	[0.8100, 0.8813]
AraBERTv2 + DAPT	0.847458	0.851216	0.7806 $\pm$ 0.0184	[0.7578, 0.8034]
mBERT + DAPT	0.779661	0.780826	0.7624 $\pm$ 0.0502	[0.7000, 0.8247]
XLM-R + DAPT	0.836158	0.840398	0.7183 $\pm$ 0.0613	[0.6423, 0.7944]

Note: BERT = Bidirectional Encoder Representations from Transformers, mBERT = multilingual BERT, XLM-R = XLM-RoBERTa, CV = cross-validation, CI = confidence interval.

4.6 Before vs. after domain-adaptive pretraining: Absolute gains

Table 7 presents the changes in held-out test performance after DAPT relative to the corresponding pre-DAPT baselines. XLM-R showed the largest increase in both accuracy and Macro-F1, followed by AraBERTv2 and MARBERTv2. In contrast, mBERT showed a slight decrease in both measures.

Tables 5 and 6 provide complementary estimates of cross-validation variability. The approximate 95% CIs overlap substantially across several pre- and post-DAPT conditions. Therefore, the observed test-set changes were interpreted descriptively rather than as evidence of statistically significant effects.

Table 7. Absolute gains after domain-adaptive pretraining (DAPT)

Model	Accuracy Gain	Macro-F1 Gain
MARBERTv2	+0.033898	+0.033487
AraBERTv2	+0.045198	+0.043367
mBERT	-0.005650	-0.009716
XLM-R	+0.107345	+0.105257

Note: BERT = Bidirectional Encoder Representations from Transformers, mBERT = multilingual BERT, XLM-R = XLM-RoBERTa.

4.7 Prototype deployment and qualitative demonstration

A Gradio-based prototype was used to demonstrate the end-to-end operation of the feedback module. Figure 6 presents the prototype interface, including the student input, preprocessed text, selected response style, predicted emotion, confidence

score, and Arabic feedback. Table 8 provides representative examples of prototype outputs.

Two experts in e-learning and Arabic language informally reviewed the interface and selected feedback examples. However, the review did not involve a structured rubric, independent ratings, inter-rater agreement analysis, student evaluation, or comparison with baseline feedback methods. It should therefore be interpreted as a qualitative walkthrough rather than as objective evidence of educational effectiveness.

The interface is a web-based form with the following elements:

- Student Input:** نص الطالب (بعد المعالجة)
- Preprocessed Text:** نص الطالب
- Selected Response Style:** أسلوب الرد (عراقي لطيف)
- Predicted Emotion:** التصنيف + الثقة (فرح (Joy) | confidence=0.662)
- Arabic Feedback:** الرد (مبارك لك، بداية زينة! نريد نشوفك تخطو الخطوة الجاية سوا)
- Buttons:** إرسال (Send) and مسح (Clear)

Figure 6. Gradio-based prototype interface for qualitative demonstration of the emotion-aware Arabic feedback generation module

Table 8. Qualitative examples of student inputs, detected emotions, and emotion-aware Arabic feedback with English translations

Student Input (Arabic)	Literal English Translation (Student Input)	Emotion	Generated Feedback (Arabic)	Literal English Translation (Generated Feedback)
أتمنى ما ينقطع الإنترنت	I hope the internet does not disconnect.	Anxiety	لا تقلق، هذه مشكلة شائعة وسنحاول دعمك وتحسين تجربتك التعليمية	Do not worry; this is a common issue, and we will try to support you and improve your learning experience.
هذا النظام مزعج	This system is annoying.	Anger	نعتذر عن الإزعاج، رأيك مهم وسنعمل على معالجة هذه المشكلة	We apologize for the inconvenience; your feedback is important, and we will work on addressing this issue.
التعليم الإلكتروني مريح	Online learning is comfortable.	Joy	يسعدنا سماع ذلك، نتمنى لك تجربة تعلم ناجحة ومستمرة	We are glad to hear that, and we wish you a successful and continuous learning experience.

The prototype demonstrated the functional integration of emotion classification and feedback delivery. However, it did not establish the pedagogical effectiveness of the generated responses. Establishing such effectiveness would require structured expert assessment, baseline comparisons, and evaluations involving students and instructors.

4.8 Discussion

The comparison between the RAW and CLEAN configurations demonstrated the importance of leakage-aware evaluation. Performance decreased after predefined lexical cues and duplicate records were removed. This decline was consistent with the higher RAW performance being partly associated with surface-level lexical cues and repeated response patterns. The CLEAN configuration therefore provided a more conservative basis for subsequent model comparisons.

Performance also varied across model families. TF-IDF + Calibrated Linear SVC provided a strong classical baseline and outperformed several transformer configurations. However, MARBERTv2 achieved higher held-out test performance both before and after DAPT. These results are consistent with the suitability of MARBERTv2 for modeling informal Iraqi Arabic in the present task, without implying that transformer models universally outperform classical baselines.

The effects of DAPT were model-specific. MARBERTv2, AraBERTv2, and XLM-R showed improved held-out test performance after DAPT, whereas mBERT showed a slight decline. XLM-R exhibited the largest test-set improvement, while MARBERTv2 with DAPT achieved the highest overall test performance. Nevertheless, the overlapping cross-validation CIs indicate uncertainty regarding the magnitude of these changes. The observed differences should therefore be interpreted descriptively rather than as statistically significant effects.

The predicted emotions were integrated into a prototype based primarily on rule-and-retrieval feedback. Gemini was retained only as an optional exploratory pathway and was not required for prototype operation. The prototype demonstrated the technical feasibility of integrating emotion classification with feedback delivery; however, it did not establish educational effectiveness.

5. CONCLUSIONS

This study presented an emotion-aware Arabic e-learning framework that combined leakage-aware emotion classification with a prototype feedback module. The combined removal of explicit lexical cues and duplicate

records resulted in a more conservative estimate of model performance. MARBERTv2 achieved the strongest overall performance, whereas the effects of DAPT varied across transformer models. These findings indicate that Iraqi-dialect adaptation may improve performance for some model architectures under the present experimental setting, but its effect is model-dependent.

Several limitations should be considered when interpreting these findings. The Student-Emotion dataset was based on self-reported responses from Iraqi university students and used a fixed six-category emotion taxonomy. Its relatively limited size, together with the absence of detailed demographic and institution-level information, restricted the assessment of representativeness and generalizability. The dataset also reflects the cultural and linguistic characteristics of Iraqi students, which may limit the transferability of the findings to other Arabic-speaking contexts and dialects. Possible self-report bias should also be considered.

Additional limitations relate to domain adaptation and prototype evaluation. The DAPT corpus originated from social media rather than an educational domain, and the confidence thresholds were heuristic rather than formally calibrated. Moreover, the feedback module was neither evaluated with real students nor deployed within an operational learning platform. Accordingly, the proposed framework should be regarded as a research prototype rather than as a validated educational tool.

Future research should examine larger and more diverse Arabic educational datasets and evaluate generalization across Arabic dialects and cultural contexts. The emotion taxonomy could also be extended to include learning-related states such as frustration, confusion, and boredom. Further work should include structured expert assessment, probability calibration, comparisons with baseline feedback methods, and evaluations involving students and instructors in real educational settings.

ACKNOWLEDGMENT

The authors thank the participating students and the Iraqi universities that supported questionnaire distribution and data collection. The authors also acknowledge the Google Colab computational resources used in the experiments.

REFERENCES

- [1] Wu, R., Yu, Z. (2022). Exploring the effects of achievement emotions on online learning outcomes: A systematic review. *Frontiers in Psychology*, 13: 977931. <https://doi.org/10.3389/fpsyg.2022.977931>

- [2] Xu, Z., Zhao, Y., Liew, J., Zhou, X., Kogut, A. (2023). Synthesizing research evidence on self-regulated learning and academic achievement in online and blended learning environments: A scoping review. *Educational Research Review*, 39: 100510. <https://doi.org/10.1016/j.edurev.2023.100510>
- [3] Yu, S., Androsov, A., Yan, H., Chen, Y. (2024). Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments. *Computers & Education*, 220: 105111. <https://doi.org/10.1016/j.compedu.2024.105111>
- [4] Anwar, A., Rehman, I.U., Nasralla, M.M., Khattak, S.B.A., Khilji, N. (2023). Emotions matter: A systematic review and meta-analysis of the detection and classification of students' emotions in STEM during online learning. *Education Sciences*, 13(9): 914. <https://doi.org/10.3390/educsci13090914>
- [5] Aladeemy, A.A., Alzahrani, A., Algarni, M.H., et al. (2024). Advancements and challenges in Arabic sentiment analysis: A decade of methodologies, applications, and resource development. *Heliyon*, 10(21): e39786. <https://doi.org/10.1016/j.heliyon.2024.e39786>
- [6] Askar, A.K.A.J., Sjarif, N.N.A. (2021). Annotated corpus of Mesopotamian-Iraqi dialect for sentiment analysis in social media. *International Journal of Advanced Computer Science and Applications*, 12(4): 101-105. <https://doi.org/10.14569/ijacsa.2021.0120413>
- [7] Almahdawi, A.J., Teahan, W.J. (2019). A new Arabic dataset for emotion recognition. *Advances in Intelligent Systems and Computing*, 998: 200-216. https://doi.org/10.1007/978-3-030-22868-2_16
- [8] 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)*, Minneapolis, USA, pp. 4171-4186. <https://doi.org/10.18653/v1/n19-1423>
- [9] Abdul-Mageed, M., Elmadany, A., Nagoudi, E.M.B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7088-7105. <https://doi.org/10.18653/v1/2021.acl-long.551>
- [10] Gururangan, S., Marasović, A., Swayamdipta, S., et al. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342-8360. <https://doi.org/10.18653/v1/2020.acl-main.740>
- [11] Alshehri, W., Al-Twairish, N., Alothaim, A. (2023). Affect analysis in Arabic text: Further pre-training language models for sentiment and emotion. *Applied Sciences*, 13(9): 5609. <https://doi.org/10.3390/app13095609>
- [12] Geirhos, R., Jacobsen, J.H., Michaelis, C., et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665-673. <https://doi.org/10.1038/s42256-020-00257-z>
- [13] Alamoudi, H., Aljojo, N., Munshi, A., et al. (2023). Arabic sentiment analysis for student evaluation using machine learning and the AraBERT transformer. *Engineering, Technology & Applied Science Research*, 13(5): 11945-11952. <https://doi.org/10.48084/etasr.6347>
- [14] Ali, M.M. (2021). Arabic sentiment analysis about online learning to mitigate COVID-19. *Journal of Intelligent Systems*, 30(1): 524-540. <https://doi.org/10.1515/jisys-2020-0115>
- [15] Aljabri, M., Chrouf, S.M.B., Alzahrani, N.A., et al. (2021). Sentiment analysis of Arabic tweets regarding distance learning in Saudi Arabia during the COVID-19 pandemic. *Sensors*, 21(16): 5431. <https://doi.org/10.3390/s21165431>
- [16] Alzaid, M., Fkih, F. (2023). Sentiment analysis of students' feedback on e-learning using a hybrid fuzzy model. *Applied Sciences*, 13(23): 12956. <https://doi.org/10.3390/app132312956>
- [17] Louati, A., Louati, H., Kariri, E., Alaskar, F., Alotaibi, A. (2023). Sentiment analysis of Arabic course reviews of a Saudi university using support vector machine. *Applied Sciences*, 13(23): 12539. <https://doi.org/10.3390/app132312539>
- [18] Khalil, E.A.H., Houbay, E.M.F.E., Mohamed, H.K. (2021). Deep learning for emotion analysis in Arabic tweets. *Journal of Big Data*, 8: 136. <https://doi.org/10.1186/s40537-021-00523-w>
- [19] Alqahtani, G., Alothaim, A. (2022). Emotion analysis of Arabic tweets: Language models and available resources. *Frontiers in Artificial Intelligence*, 5: 843038. <https://doi.org/10.3389/frai.2022.843038>
- [20] Althobaiti, M.J. (2023). An open-source dataset for Arabic fine-grained emotion recognition of online content amid COVID-19 pandemic. *Data in Brief*, 51: 109745. <https://doi.org/10.1016/j.dib.2023.109745>
- [21] Althobaiti, M.J. (2023). Arabic emotion recognition in low-resource settings: A novel diverse model stacking ensemble with self-training. *Applied Sciences*, 13(23): 12772. <https://doi.org/10.3390/app132312772>
- [22] Alfarhany, A.A.R., Abdullah, N.A.Z. (2023). Iraqi sentiment and emotion analysis using deep learning. *Journal of Engineering*, 29(9): 150-165. <https://doi.org/10.31026/j.eng.2023.09.11>
- [23] Hussein, H.H., Lakizadeh, A. (2025). A systematic assessment of sentiment analysis models on Iraqi dialect-based texts. *Systems and Soft Computing*, 7: 200203. <https://doi.org/10.1016/j.sasc.2025.200203>
- [24] Parker, M.J., Anderson, C., Stone, C., Oh, Y. (2025). A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*, 35(2): 444-481. <https://doi.org/10.1007/s40593-024-00414-0>
- [25] Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M., McNamara, D. (2025). Formative feedback on student-authored summaries in intelligent textbooks using large language models. *International Journal of Artificial Intelligence in Education*, 35(3): 1022-1043. <https://doi.org/10.1007/s40593-024-00395-0>