MarsadLab at TAQEEM 2025: Prompt-Aware Lexicon-Enhanced Transformer for Arabic Automated Essay Scoring

Mabrouka Bessghaier[1], Md. Rafiul Biswas[2], Amira Dhouib[3], Wajdi Zaghouani[1]
[1]Northwestern University in Qatar, Qatar
[2]Hamad Bin Khalifa University, Qatar,
[3] LaTICE Lab, University of Kairouan
mbiswas@hbku.edu.qa
{mabrouka.bessghaier,wajdi.zaghouani}@northwestern.edu

Abstract

We present the MarsadLab submission to TAQEEM 2025 Shared Task A on Automated Essay Scoring (AES) in Arabic. Our system extends AraBERT with a prompt-type embedding and lexicon-based features. The lexicon captures statistical associations between word usage and essay quality under each prompt type, providing prompt-aware, interpretable signals that complement semantic embeddings. Our system achieved an average QWK of 0.438, highlighting both the promise and the challenges of incorporating prompt-sensitive lexical knowledge into AES. This work represents a first attempt at leveraging a task-aware lexicon for Arabic AES, showing that lexical features provide educational value through interpretability but also require more sophisticated integration. Future improvements could combine these lexical indicators with discourse-, syntax-, and content-level features, as well as explore richer fusion strategies to better exploit their potential.

1 Introduction

Automated Essay Scoring (AES) aims to predict human-assigned scores for student essays, offering applications in large-scale assessment and educational feedback. While AES has been widely studied for English, progress in Arabic remains limited due to scarce datasets, morphological complexity, and diverse rhetorical styles.

The TAQEEM 2025 Shared Task introduces the first large-scale benchmark for Arabic AES (Bashendy et al., 2025), evaluating systems on holistic score prediction across two writing prompts: explanatory and persuasive. This dual requirement makes the task particularly challenging, as effective systems must capture not only semantic meaning but also prompt-specific discourse and stylistic features. Furthermore, our submission was evaluated under the cross-prompt setting, where systems must generalize across different prompts, further increasing task difficulty.

Our submission explores a hybrid design that integrates AraBERT semantic embeddings with lexicon-based features. The lexicon captures statistical correlations between words and essay scores within each prompt type, offering interpretability and potentially complementing contextual embeddings. Although our results did not surpass the baseline, the analysis provides valuable insights into the difficulties of feature fusion and the role of lexical cues in Arabic AES.

2 Background

TAQEEM 2025 Shared Task A focuses on Arabic AES, where the goal is to predict a continuous holistic score for essays written in response to specific prompts. Each essay is linked to a prompt text, a prompt type (either explanatory or persuasive), and a human-assigned holistic score ranging from 0 to 32.

The dataset is structured around three components: (i) prompts that define the writing task and its type, (ii) student essays written in response to these prompts, and (iii) holistic scores provided by human raters. All essays are written in Modern Standard Arabic (MSA), covering academic writing across the two genres. Each instance thus consists of the essay text, the prompt information, and a holistic score. As shown in table 1, the training dataset of this task includes 425 essays written in response to two different prompts: one explanatory (215 essays) and one persuasive (210 essays). Each essay is annotated with a holistic human score ranging from 2 to 31, indicating a broad spread of writing quality. The distribution of essays across prompts is relatively balanced, ensuring that models are exposed to both explanatory and persuasive writing.

Automated essay scoring has been heavily studied, with early approaches relying mainly on regres-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 998–1002
November 8-9, 2025 ©2025 Association for Computational Linguistics

sion models and hand-crafted linguistic features to capture aspects of writing quality. More recent research has increasingly focused on ensemble methods and deep learning models, which aim to better capture lexical, syntactic, and discourse-level characteristics of student writing (Ramnarain-Seetohul et al., 2025). However, AES has not advanced as rapidly due to linguistic complexity and the scarcity of large-scale annotated resources. One of the few early attempts is the work of Alqahtani (2019), who proposed a rule-based system for evaluating Arabic essays based on surface-level criteria such as spelling, punctuation, essay structure, coherence, and style (Alqahtani et al., 2019).

Several efforts have attempted to lay the groundwork for advancing AES in Arabic by providing resources that target key aspects of writing quality. For example, (Zaghouani et al., 2024) built the Qatari Corpus of Student Argumentative Writing. The proposed corpus presents a bilingual (Arabic/English) resource that captures discourse structure, coherence signals, and learner-writing phenomena. Complementary resources have focused on error annotation for learner Arabic, offering normalization and correction procedures, inter-annotator agreement metrics, and foundations for assessing grammar and fluency (Zaghouani et al., 2014). Similarly, gold-standard corrections for learner errors have been proposed in (Zaghouani et al., 2015), covering orthographic, morphological, syntactic, and punctuation mistakes, thereby enabling benchmarks for automated error correction and linguistic quality assessment. In addition, auxiliary resources such as Arabic diacritization guidelines provide conventions for orthography and phonology consistency, supporting disambiguation tasks relevant for spelling- and diacritic-aware quality assessment (Zaghouani et al., 2016). Research on punctuation and sentence-boundary annotation has also introduced resources for mechanics and readability, contributing cues for punctuation restoration and coherence modeling (Zaghouani and Awad, 2016). Together, these initiatives provide the linguistic and annotation foundations necessary for advancing Arabic AES, complementing scoring models by supplying resources on grammar, fluency, coherence, and overall writing quality.

In order to drive further progress in this domain, the TAQEEM 2025 Shared Task (Bashendy et al., 2025) presents the first extensive dataset for Arabic AES. Unlike previous small-scale or resource-specific efforts, it provides a balanced dataset of persuasive and explanatory essays for comprehensive scoring, allowing for systematic examination under cross-prompt settings.

In fact, the role of lexical features has been emphasized in assessing text quality. Such features describe the surface characteristics of textual responses, including single words, stemmed or lemmatized forms, prefixes, suffixes, or n-grams. Their extraction is relatively simple, and many algorithms have been proposed for Automatic Short Answer Grading (ASAG) tasks based on lexical similarity, overlap measures, or lexical statistics (Haller et al., 2022). These approaches laid an important foundation for later AES systems, especially in contexts where more sophisticated syntactic or semantic models were not available. In this work, we aim to further investigate the contribution of lexical features in the context of Arabic AES.

Prompt ID	Prompt Type	Essays
1	Explanatory	215
2	Persuasive	210
Total	–	425

Table 1: Distribution of essays and score ranges across prompts

3 System Overview

Our system extends a transformer-based regressor with a prompt-aware lexicon that captures lexical signals of essay quality. The overall workflow involves (i) building the lexicon from training data, (ii) extracting aggregated lexical features for each essay, and (iii) integrating these features with AraBERT embeddings in a hybrid architecture.

3.1 Task-Aware Lexicon Construction

We created a custom lexicon (1–3-grams) designed to reflect how word usage relates to essay quality under different prompt types (explanatory vs. persuasive). This process involved three main steps:

Merging resources. Essay texts, human-assigned scores, and prompt metadata were combined using shared identifiers (essay_id, prompt_id).

Preprocessing. Essays were normalized (removing diacritics and unifying variants of alif, ya, and taa marbuta), cleaned of non-Arabic characters, digits, and punctuation, and then tokenized into words.

Stopwords were deliberately retained, as function words such as connectives, discourse markers, and particles can vary systematically across explanatory and persuasive writing and thus provide useful discriminative signals. To avoid lexical leakage, any tokens appearing in the corresponding prompt text were excluded, ensuring the lexicon reflects only the language of student essays rather than the instructions.

Computing lexical statistics. For each unique `(word, prompt_type)` pair, we calculated:

- **Frequency:** how often the word occurs in essays of that prompt type.
- **Mean score:** average holistic score of essays containing the word.
- **Score variability:** the standard deviation of scores associated with the word.
- **Richness:** the number of unique score values linked to the word.
- **Z-score:** For each token, we compared the average score of essays containing that token with the mean score of all essays written under the same prompt type. The difference was normalized by the standard deviation of scores across the entire prompt type, yielding a classic z-score:

  $$z = \frac{\text{mean\_score(token)} - \text{mean\_score(prompt\_type)}}{\sigma_{prompt\_type} + 10^{-5}}$$

  This measures how far above or below the prompt-type average the token's essays tend to score, relative to the overall variability in that prompt type. Tokens occurring mainly in stronger essays have positive z-scores, while those associated with weaker essays receive negative z-scores.
- **Importance:** defined as frequency $\times$ |z-score|, highlighting words that are both frequent and strongly associated with higher or lower quality. defined as the logarithm of the token's document frequency, multiplied by its positive z-score:

  $$\text{importance} = \log(1 + \text{count}) \times \max(0, z)$$

  This formulation ensures that tokens are ranked higher when they are both frequent and associated with above-average essay scores, while logarithmic scaling prevents extremely common tokens from dominating the lexicon. Only tokens with positive importance were retained.

The result is a lexicon table where each row corresponds to a word conditioned on a prompt type, enriched with its statistical profile. This lexicon provides interpretable insight into vocabulary patterns rewarded or penalized by human raters.

3.2 Lexicon Feature Integration

While our lexicon construction relies on associations between words and essay scores, we do not assume that words directly cause higher or lower scores. Instead, certain lexical items tend to co-occur with patterns of stronger writing and can therefore serve as useful signals. In explanatory prompts, higher-scoring essays frequently include causal and elaborative markers such as الأسباب ("the reasons"), بشكل ("in a way"), أهم ("most important"), which help writers clarify causes, emphasize significance, or indicate conditions. In persuasive prompts, stronger essays often use العديد ("many") to generalize claims, إلا ("except/but") to introduce concessions or contrasts, and مما ("which/thereby") to connect evidence with conclusions. These examples illustrate that while no single word determines essay quality, their systematic distribution provides interpretable clues about how students construct explanations or persuasive arguments. The task-aware lexicon is thus employed not as a causal determinant of scores but as a descriptive resource that highlights lexical tendencies associated with stronger or weaker essays under different prompt types. Such words can be markers of reasoning and structure, and their use often reflects the essay's quality. So the created lexicon was used to derive numerical features for each essay: **(i) Total importance:** the cumulative weight of all matched tokens in an essay. This reflects how much the essay overall makes use of words that are associated with higher importance scores. **(ii) Maximum importance** is the highest importance value among the essay's matched tokens, capturing the strongest single lexical signal present. **(iii) Average z-score (weighted)** specifies the central tendency of lexical associations in the essay, computed as the importance-weighted mean of token z-scores. These features were appended as auxiliary variables to each essay instance.

System	Prompt	QWK	MSE	RMSE	Avg. QWK	Avg. RMSE
Baseline	9	0.608	33.148	5.76	0.639	5.37
	10	0.670	24.862	4.99		
MarsadLab	9	0.447	40.431	6.36	0.438	7.07
	10	0.428	60.679	7.79		

Table 2: Comparison of Baseline and MarsadLab submissions

3.3 Model Architecture

We extended AraBERT-v2 with an additional branch for lexicon-based features, building a hybrid architecture that combines deep contextual embeddings with interpretable lexical signals. The system is based on the following steps:

1. **Essay encoding with AraBERT.** The essay text is encoded using AraBERT-v2 (encoder-only).

2. **Prompt-type signal.** A learned embedding representing the prompt type is added element-wise to the pooled essay vector. This provides the model with an explicit indication of whether the essay is explanatory or persuasive, helping it adapt its representations to genre-specific expectations.

3. **Lexicon feature extraction.** In parallel, each essay is mapped to a three-dimensional vector derived from the lexicon: (i) total importance, (ii) maximum importance, and (iii) weighted average z-score.

4. **Feature concatenation.** The pooled AraBERT vector (dimension 768, after prompt-type addition) is concatenated with the lexicon feature vector (dimension 3), yielding a combined representation of size 771. This joint representation ensures that both semantic and lexical signals are captured in a shared feature space.

5. **Regression head.** The combined vector is passed through a projection block consisting of a linear transformation, layer normalization, and dropout. A final linear layer produces a single logit, which is mapped to the valid score range $[0, 32]$ using a sigmoid and affine scaling. Training is optimized with Mean Squared Error (MSE) loss against the human-provided holistic scores.

This design allows the model to capture both deep semantic information (through AraBERT) and prompt-sensitive lexical cues (through the lexicon features). The concatenation step explicitly fuses these two types of signals, ensuring that the model considers not only meaning and discourse but also interpretable markers of explanation or persuasion that human raters often reward.

4 Experimental Setup

We trained models using AraBERTv2 with AdamW optimizer (learning rate 2e-5), batch size 8, max length 512, and early stopping on dev QWK. Evaluation follows official test protocol with QWK as the primary metric and RMSE as a secondary metric.

5 Results

Table 2 compares our submissions with the official baseline. The baseline achieved an average QWK of 0.639 (RMSE = 5.37), with consistent performance across both prompts. In contrast, our system obtained an average QWK of 0.438 (RMSE = 7.07). The drop was observed across both expository (Prompt 9) and persuasive (Prompt 10) essays. A likely reason for underperformance is the simplistic concatenation of features with AraBERT embeddings, which may not allow the model to weigh contextual versus lexical information dynamically. Another factor may be the small size of the dataset, which restricts the coverage of the constructed lexicon.

Compared to the baseline, our system underperformed in both QWK and RMSE. While the baseline achieved higher agreement with human raters, our hybrid AraBERT+lexicon approach demonstrated stable but lower performance. This suggests that our current fusion strategy does not fully exploit the complementary strengths of contextual and lexical features. Future work should explore attention-based fusion or prompt-adaptive weighting.

6 Conclusion

We presented the MarsadLab system for TAQEEM 2025 Task A, extending AraBERT with a prompt-type embedding and a task-aware lexicon for Arabic AES. The lexicon offered interpretable features—total importance, maximum importance, and weighted average z-score—that capture prompt-sensitive lexical tendencies. Our system achieved an average QWK of 0.438, showing that lexical features can be successfully integrated into AES, but also highlighting the need for more advanced methods to fully exploit their potential.

While the lexicon provides transparency and insight into genre-sensitive vocabulary, it remains correlational and incomplete. Future work should expand the lexicon across more prompts, combine it with discourse- and syntax-level features, and explore richer integration strategies such as attention-based fusion or prompt-adaptive regression.

7 Limitations

The task-aware lexicon we created gives useful and interpretable signals for essay scoring, but it is not enough on its own to capture the full complexity of writing quality. It reflects correlations between words and scores, yet essay quality also depends on broader aspects such as coherence, organization, and depth of reasoning, which cannot be reduced to lexical patterns. Another limitation is that the lexicon was built from only two prompts, one explanatory and one persuasive. This means some of the word associations may be domain-specific and tied to the topics of these prompts rather than general markers of writing quality. Finally, the way we integrated lexical features with AraBERT relied on simple concatenation, which likely limited the model's ability to make effective use of both contextual and lexical information. These points show that while the lexicon is a helpful resource, it should be seen as a first step. Future work should expand it to more prompts, add discourse- and syntax-level features, and test more advanced fusion methods to improve both generality and performance.

Acknowledgements

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Hmoud Alqahtani, Sabri Mahmoud, and Shadi Al-Saqqa. 2019. Automated essay scoring for arabic essays using content and text features. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–6.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor, and Tamer Elsayed. 2025. TAQEEM 2025: Overview of the First Shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.

Vidasha Ramnarain-Seetohul, Yasmine Rosunally, and Vandana Bassoo. 2025. Ensemble and hybrid models in automated essay scoring: A literature review. *SN Computer Science*, 6(6):729.

Wajdi Zaghouani, Abdelhamid Ahmed, Xiao Zhang, and Lameya Rezk. 2024. QCAW 1.0: Building a qatari corpus of student argumentative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13382–13394, Torino, Italia. ELRA and ICCL.

Wajdi Zaghouani and Dana Awad. 2016. Building an arabic punctuated corpus. 2016(1).

Wajdi Zaghouani, Houda Bouamor, Abdelati Hawwari, Mona Diab, Ossama Obeid, Mahmoud Ghoneim, Sawsan Alqahtani, and Kemal Oflazer. 2016. Guidelines and framework for a large scale Arabic diacritized corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3637–3643, Portorož, Slovenia. European Language Resources Association (ELRA).

Wajdi Zaghouani, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider, and Kemal Oflazer. 2015. Correction annotation for non-native Arabic texts: Guidelines and corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 129–139, Denver, Colorado, USA. Association for Computational Linguistics.

Wajdi Zaghouani, Nizar Habash, Behrang Mohit, Abeer Heider, Alla Rozovskaya, and Kemal Oflazer. 2014. Annotation guidelines for non-native arabic text in the qatar arabic language bank. 2014(1).