MorphoArabia at BAREC Shared Task 2025: A Hybrid Architecture with Morphological Analysis for Arabic Readability Assessment

Fatimah Emad Eldin
Department of Computer and Information Sciences
Faculty of Graduate Studies for Statistical Research, Cairo University
12422024441586@pg.cu.edu.eg

Abstract

This paper presents MorphoArabia, a system developed for the BAREC Shared Task 2025 on Arabic Readability Assessment. The approach is centered on the hypothesis that deep morphological analysis is fundamental for modeling the complexity of the Arabic language. A regression model was fine-tuned on AraBERTv2 with morphologically-aware tokenization via CAMeL Tools. Various configurations were explored for the strict, constrained, and open tracks, including a hybrid model with seven engineered lexical features. The system demonstrated highly competitive performance, securing top-10 rankings in all six subtasks and achieving a peak Quadratic Weighted Kappa (QWK) of 84.2% on the strict sentence-level task. All code and models are publicly available to facilitate future research.

1 Introduction

Automatic Readability Assessment for Arabic is a challenging task, primarily due to the language's rich and complex morphology (Liberato et al., 2024). Consequently, traditional readability formulas that rely on surface-level features are often insufficient for capturing the nuanced difficulty of Arabic text (Al-Tamimi et al., 2014). The BAREC Shared Task 2025 (Elmadani et al., 2025a) addresses this by providing a large-scale, fine-grained dataset annotated on a 19-level readability scale. This paper introduces MorphoArabia, a system designed to address this challenge by explicitly modeling Arabic morphology. The core hypothesis is that a model's performance can be significantly improved by providing it with text analyzed at the morpheme level. This hypothesis is tested across the three competition tracks:

- **Strict Track:** A fine-tuned AraBERTv2 (Antoun et al., 2020) regression model using only the official BAREC corpus.
- **Constrained Track:** A hybrid architecture augmenting the base model with seven engineered lexical features derived from the SAMER corpus (Alhafni et al., 2024).
- **Open Track:** The base regression model trained on a combination of the BAREC and DARES corpora (El-Haj et al., 2024).

The system achieved competitive results across all tracks, notably securing 2nd place in both the strict and open document-level tasks, validating the effectiveness of the morphologically-aware approach. Key findings include the superior performance of regression over classification for this task, along with the challenges of harmonizing datasets with disparate annotation scales. To ensure reproducibility, all code and models are available on GitHub[1] and Hugging Face[2].

2 Background and Related Work

2.1 Task Description

The BAREC Shared Task 2025 utilizes the Balanced Arabic Readability Evaluation Corpus (BAREC), a dataset exceeding 1 million words annotated for readability assessment (Elmadani et al., 2025b). The task's 19-level annotation scheme is detailed in the official guidelines (Habash et al., 2025). The task comprises two primary goals:

- **Task 1: Sentence-level Readability Assessment:** Predict a readability score (1-19) for a given Arabic sentence.
- **Task 2: Document-level Readability Assessment:** Predict an overall document readability score, defined by the highest score of any sentence within it.

[1] https://github.com/astral-fate/barec-Arabic-Readability-Assessment
[2] https://huggingface.co/collections/FatimahEmadEldin/barec-shared-task-2025-689195853f581b9a60f9bd6c

Proceedings of The Third Arabic Natural Language Processing Conference, pages 286–296

Participation was offered across three tracks: Strict, Constrained, and Open, each with distinct data constraints.

2.2 Related Work and Available Datasets

The landscape of Arabic NLP resources is extensively documented by initiatives like **Masader** (Alyafeai et al., 2021; Altaher et al., 2022). For readability assessment, several key resources include:

- **DARES** (El-Haj et al., 2024): A corpus of school textbooks with fine-grained (G1-G12) and coarse-grained labels.
- **OSMAN** (El-Haj and Rayson, 2016): A readability metric providing a continuous 0-100 score.
- **ARC-WMI** (AL-Dayel et al., 2018): A medical corpus with three difficulty levels.
- **SAMER Project**: This project introduced a lexicon with a 5-level scale (L1-L5) (Elmadani et al., 2025b). A related Google Docs add-on was also developed for word-level readability visualization (Hazim et al., 2022).
- **SAMER Corpus** (Alhafni et al., 2024): A text simplification corpus with parallel texts across multiple readability levels, used for comprehensive modeling approaches ranging from rule-based methods to pretrained language models (Liberato et al., 2024).

3 System Overview

The system employs two main architectures: a base regression model for the Strict and Open tracks, and a hybrid model for the Constrained track, which incorporates engineered features.

3.1 Morphological Analysis

The preprocessing pipeline utilized the CAMeL Tools `d3tok` analyzer (Obeid et al., 2020) for external datasets such as SAMER and DARES. This tool performs deep morphological analysis by disambiguating words in context and segmenting them into constituent morphemes, capturing complexities often missed by standard tokenization.

3.2 Feature Engineering

For the Constrained track, the system was enhanced with a hybrid architecture integrating engineered lexical features with the Transformer model's contextual understanding. Seven numerical features were engineered for each sentence using the SAMER lexicon to provide explicit signals about text complexity. A detailed description of these features is provided in Table 3 in Appendix A. The final sentence representation is created by concatenating the Transformer's '[CLS]' token embedding with this 7-dimensional feature vector, which is then passed to a regression head for prediction.

3.3 Level Mapping for External Datasets

To augment training data for the Constrained and Open tracks, external corpora were incorporated, necessitating mapping their distinct annotation scales to the 19-level BAREC scale.

- **DARES Corpus**: For the Open track, "G1-G12" labels were directly mapped to BAREC levels 1-12.
- **SAMER Corpus**: For the Constrained track, SAMER's 5-level scale was harmonized with BAREC's 19-level scale. A heuristic mapped SAMER levels L3, L4, and L5 to BAREC values 4, 10, and 16, respectively.

This heuristic mapping process was identified as a potential source of noise and variance, potentially impacting model performance by introducing inconsistencies.

4 Experimental Setup

4.1 Datasets

Datasets and distributions were defined by each competition track. All data was preprocessed into the `d3tok` format before training.

- **Strict Track**: Limited to the official **BAREC** corpus.
  - **Sentence-level**: BAREC training (54,845 sentences) and development (7,310 sentences) and (7,286) test records, for the development phase, and (3,417) for the blind testing phase.
  - **Document-level**: Official document splits for the development phase is: 1,518 training, 194 development, 210 testing, and (100) for the tblind esting phase.
- **Constrained Track**: **BAREC** training data augmented with the **SAMER Corpus**.
  - Combined **sentence-level training set**: 97,874 sentences.

Track	Task	Dev (QWK)	Public Test (QWK)	Blind Test (QWK)	Hugging Face Model
Strict	Sentence	82.64	83.61	**84.2**	[Link]
	Document	71.07	65.91	79.90	[Link]
Constrained	Sentence	80.07	80.71	82.9	[Link]
	Document	75.60	62.70	75.5	[Link]
Open	Sentence	83.10	82.06	83.9	[Link]
	Document	72.85	57.11	**79.2**	[Link]

Table 1: Final QWK scores and Hugging Face models for each task. For document-level tasks, scores were derived from the sentence-level model by assigning each document the highest readability score found among its sentences.

  - Original BAREC development set (7,310 sentences) used for validation.

- **Open Track**: This track permitted the use of external data, with experiments primarily focusing on combining the **BAREC** and **DARES** datasets. Different data configurations were explored to optimize performance for both sentence-level and document-level tasks. More details on the data distributions for the Open track can be found in Appendix B.

4.2 Training and Hyperparameters

Models were fine-tuned with varied hyperparameters, primarily adjusting learning rate (2e-5-5e-5) and epochs (6-20). All models used the AdamW optimizer and an early stopping callback monitoring validation QWK score. A detailed summary of the hyperparameter values for the best performing models can be found in Appendix E.

4.3 Evaluation Metrics

The primary metric for evaluation is the Quadratic Weighted Kappa (QWK) (Cohen, 1968), as defined in Equation 1.

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \quad (1)$$

In this formula, O is the matrix of observed agreement, E is the matrix of expected agreement, and $w_{ij} = (i-j)^2$ is the quadratic weight matrix that penalizes larger disagreements more severely.

5 Results

The system demonstrated strong and consistent performance across all competition tracks. As summarized in Table 1, the top performance achieved was a sentence-level Quadratic Weighted Kappa (QWK) of 84.2% in the Strict track and a document-level QWK of 79.2% in the Open track. Full configurations for the best-performing models are detailed in Appendix E (Table 7).

5.1 Sentence-Level Analysis

The system achieved a highly competitive QWK score of 84.2% on the Strict track, earning a 7th-place rank on the official leaderboard. This performance is nearly identical to the official BAREC benchmark score of 84.4% (Elmadani et al., 2025b). The minor difference is attributed to variations from the custom morphological analysis used in this work, as opposed to the official pre-processed dataset provided by the organizers.

In the Constrained track, the hybrid model yielded a QWK of 82.9%, which earned a 3rd-place ranking. For the Open track, a QWK of 83.9% was attained by augmenting the training data with the DARES corpus, resulting in a 2nd-place ranking. It was noted, however, that neither of these results exceeded the performance observed in the Strict track. This observation reinforces the notion that the difficulties inherent in mapping different annotation scales can introduce label and domain variance, which may temper the performance improvements expected from additional data.

5.2 Document-Level Analysis

For the document-level task, no models were directly fine-tuned on full documents. Instead, the assessment was derived from the corresponding sentence-level models by assigning each document the maximum readability score predicted among all its sentences. A substantial increase in the QWK was observed between the development phase and the final blind test evaluation. The document-level QWK for the Strict track increased from a development score of 62.37% to a final blind test score of 79.9%, achieving a 2nd-place

rank in the official evaluation (Tables 1, 7). A 3rd-place ranking was secured in the Constrained track with a score of 75.5%. In the Open track, the score increased from a development QWK of 60.48% to 79.2%, which also ranked 2nd (Tables 1, 7). This considerable improvement suggests the blind test set featured a different distribution of document complexity, one where difficulty was determined by a few outlier sentences.

5.3 Comparison with Official Baseline

A direct comparison with the final baseline results released by the shared task organizers reveals that the MorphoArabia system demonstrated a significant performance improvement across all six subtasks. The official baseline scores are sourced from the final leaderboards published on the BAREC Shared Task website.

‘‘‘

As shown in Table 2, MorphoArabia outperformed the baseline by a notable margin in every category. The most substantial gains were observed in the document-level tasks, where the system's max-score aggregation strategy proved highly effective, leading to improvements of +17.9, +13.5, and +17.2 QWK points for the Strict, Constrained, and Open tracks, respectively. The sentence-level tasks also showed consistent improvements, confirming the robustness of the morphologically-aware approach.

5.4 Hyperparameter and Data Ablation Analysis

The optimal configuration for the sentence-level task did not yield the best performance for the document-level task. The model from Experiment 2 achieved the highest sentence-level QWK on the blind test set (83.9%), whereas the model from Experiment 5 yielded the top document-level score (79.2%). Notably, the best sentence-level performance on the validation set (83.6%) was achieved in Experiments 3 and 4, not Experiment 2 (Table 6). This suggests that the document-level task, being highly sensitive to single-sentence errors, benefits from a validation set that better mirrors the complexity distribution of the augmented training data. Furthermore, experiments combining all three datasets (BAREC, SAMER, and DARES) did not lead to superior results, highlighting that more data is not always beneficial when significant label and domain variance is introduced, as detailed in Appendix C.2 (Table 6).

5.5 Ablation on Task Formulation

To validate the problem formulation, an ablation study was conducted comparing the primary regression approach (predicting a continuous score) against a multi-class classification alternative (predicting one of 19 discrete levels). For this comparison, the classification models were tested using a custom, non-morphological data normalization pipeline in place of the d3tok Morphological Analyzer (see Appendix C.1). The classification approach consistently yielded inferior performance compared to the morphologically-aware regression model, as detailed in Appendix C.2 (Table 5). This result confirmed that the regression framework was the more effective formulation for this task.

5.6 Morphological Error Analysis

A key source of error was identified as preprocessing artifacts. Appendix F (Table 8) provides examples where the d3tok analyzer failed to produce a morphological analysis, instead inserting a NOAN (No Analysis) token. This occurs for words not in its vocabulary or for words with valid but less frequent morphological forms. This noise, introduced during data augmentation, can degrade model reliability, especially for the document-level task.

6 Discussion

The performance of the MorphoArabia system, summarized in Table 1, validates the core hypothesis that a morphologically-aware model is highly effective for Arabic readability assessment. The top score achieved in the Strict sentence-level track (84.2% QWK) was highly competitive, nearly matching the official BAREC benchmark of 84.4% and underscoring the success of this foundational approach.

A key observation from the results is that models augmented with external data (Constrained and Open tracks) did not surpass the baseline model trained exclusively on the BAREC corpus. This suggests that the benefits of additional data were negated by noise introduced when harmonizing disparate datasets. Heuristically mapping different annotation scales (e.g., SAMER's 5-level and DARES's 12-level) to BAREC's 19-level schema likely introduced significant label and domain variance, highlighting that annotation quality and consistency are paramount.

Track	Task	MorphoArabia (QWK)	Official Baseline (QWK)
Strict	Sentence	**84.2%**	81.5%
	Document	**79.9%**	62.0%
Constrained	Sentence	**82.9%**	81.5%
	Document	**75.5%**	62.0%
Open	Sentence	**83.9%**	81.5%
	Document	**79.2%**	62.0%

Table 2: Comparison of final blind test QWK scores between MorphoArabia and the official shared task baseline.

The document-level assessment strategy, which assigned the maximum sentence score to the document, proved effective, securing second-place rankings in two tracks. The significant QWK score increase in the blind test suggests its distribution contained documents whose difficulty was driven by a few outlier sentences, a characteristic well-suited to the chosen max-score approach. Additionally, ablation studies confirmed that formulating the task as regression consistently outperformed a multi-class classification approach (Appendix C.2, Table 5).

Despite the system's success, two primary limitations were identified. First, reliance on the d3tok analyzer made the system susceptible to preprocessing artifacts. It failed to parse out-of-vocabulary or infrequent words, inserting a NOAN (No Analysis) token (see Appendix F). This introduced noise by depriving the model of crucial morphological information, a particularly detrimental issue for the document-level task where a single unanalyzed word can determine the overall score. Second, the simplistic, direct mapping used for data augmentation presents a significant challenge, as it fails to account for subtle differences in annotation criteria between corpora, leading to label noise.

7 Conclusion

This paper presented MorphoArabia, a system developed for the BAREC Shared Task 2025 on Arabic readability assessment. The system was centered on the hypothesis that a deep morphological approach is fundamental to modeling the nuances of Arabic text complexity. It employed two main architectures: a fine-tuned AraBERTV2 regression model and a hybrid model enhanced with seven engineered lexical features for the Constrained track. The results demonstrated highly competitive performance, securing 2nd place in the strict and open document-level tasks and achieving a peak QWK of 84.2% on the strict sentence-level task, thereby validating the effectiveness of the core approach. Despite its success, this work identified two primary limitations.

First, the process of harmonizing external datasets (SAMER and DARES) with different annotation scales introduced label noise, which may have tempered performance gains on the augmented tracks. Second, the system's reliance on the d3tok analyzer made it susceptible to preprocessing artifacts, where out-of-vocabulary or morphologically infrequent words were not analyzed, potentially degrading model reliability.

These limitations inform several directions for future work. Research could focus on more sophisticated domain adaptation techniques to better integrate external corpora and mitigate the effects of label variance. Another avenue is to improve the robustness of the morphological preprocessing pipeline, either by fine-tuning the analyzer on a broader vocabulary or by developing strategies to handle analysis failures.

Finally, exploring architectures that are directly fine-tuned on the document-level task, rather than relying on sentence-level aggregation, presents a promising path toward further performance improvements. In the interest of open science and to facilitate future research, all code for preprocessing, training, and evaluation, alongside the final fine-tuned models for each track, have been made publicly available. The experimental code is accessible on GitHub, and the models are hosted on the Hugging Face Hub, providing a strong and reproducible baseline for future work in Arabic readability assessment.

Acknowledgments

Acknowledgment is made to the organizers of the BAREC Shared Task 2025, the developers of CAMeL Tools, and the creators of the DARES and SAMER corpora for their invaluable resources.

References

Abeer AL-Dayel, Hend Al-Khalifa, Sinaa Alaqeel, Norah Abanmy, Maha Al-Yahya, and Mona Diab. 2018. ARC-WMI: Towards building Arabic readability corpus for written medicine information. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Abdel-Karim Al-Tamimi, Manar Jaradat, Nuha Aljarrah, and Sahar Ghanim. 2014. AARI: Automatic Arabic Readability Index. *The International Arab Journal of Information Technology*, 11(4):384–391.

Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. The SAMER Arabic text simplification corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16079–16093, Torino, Italia. ELRA and ICCL.

Yousef Altaher, Ali Fadel, Mazen Alotaibi, Mazen Alyazidi, Mishari Al-Mutairi, Mutlaq Aldhbuiub, Abdulrahman Mosaibah, Abdelrahman Rezk, Abdulrazzaq Alhendi, Mazen Abo Shal, Emad A. Alghamdi, Maged S. Alshaibani, Jezia Zakraoui, Wafaa Mohammed, Kamel Gaanoun, Khalid N. Elmadani, Mustafa Ghaleb, Nouamane Tazi, Raed Alharbi, and 2 others. 2022. Masader plus: A new interface for exploring +500 arabic nlp datasets. *Preprint*, arXiv:2208.00932.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2021. Masader: Metadata sourcing for arabic text and speech data resources. *Preprint*, arXiv:2110.06744.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Mahmoud El-Haj and Paul Rayson. 2016. OSMAN — a novel Arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255, Portorož, Slovenia. European Language Resources Association (ELRA).

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. DARES: Dataset for Arabic readability estimation of school materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 103–113, Torino, Italia. ELRA and ICCL.

Khalid N. Elmadani, Bashar Alhafni, Hanada Taha, and Nizar Habash. 2025a. BAREC shared task 2025 on Arabic readability assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*, Suzhou, China. Association for Computational Linguistics.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025b. A large and balanced corpus for fine-grained Arabic readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16376–16400, Vienna, Austria. Association for Computational Linguistics.

Nizar Habash, Hanada Taha-Thomure, Khalid N. Elmadani, Zeina Zeino, and Abdallah Abushmaes. 2025. Guidelines for fine-grained sentence-level Arabic readability annotation. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 359–376, Vienna, Austria. Association for Computational Linguistics.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Juan Liberato, Bashar Alhafni, Muhamed Khalil, and Nizar Habash. 2024. Strategies for Arabic readability modeling. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 55–66, Bangkok, Thailand. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

A Engineered Lexical Features

This appendix provides a detailed description of the seven engineered lexical features that were integrated into the hybrid model for the Constrained track. These features were designed to provide the model with explicit, interpretable signals about text complexity, complementing the deep contextual understanding from the Transformer architecture. Table 3 lists these features.

B Data Distribution for the Open Track

This appendix outlines the specific data configurations used for the Open track experiments. Since this track allowed for external data, different strategies were employed to combine the BAREC and

No.	Description	No.	Description
1	Sentence length (total characters)	5	Maximum word difficulty in the sentence
2	Number of words in the sentence	6	Count of 'hard' words (difficulty score > 4)
3	Average word length in characters	7	Fraction of Out-of-Vocabulary (OOV) words
4	Mean word difficulty (from lexicon)		

Table 3: Description of the seven engineered lexical features used in the hybrid model.

DARES datasets to optimize performance for both the sentence-level and document-level tasks.

- **Sentence-Level Task Data Configuration**:
  - **Training Set**: A combination of the BAREC dataset and the official train/development splits of the DARES dataset, totaling 64,548 records. This consisted of:
    * 9,703 records from the DARES training set.
    * 1,380 records from the DARES development set.
    * Remaining records from the BAREC dataset.
  - **Validation Set**: 8,690 records.
- **Document-Level Task Data Configuration**: A more specific data splitting strategy was employed.
  - **Training Set**: Consisted of the entire BAREC dataset combined with 85% of the complete DARES dataset (merging its train, development, and test splits). This resulted in a total of 66,634 records for training.
  - **Validation Set**: The remaining 15% of the combined DARES data, totaling 9,391 records.

C Ablation Studies and Alternative Approaches

This appendix details the ablation studies that were conducted to validate the final system design. These experiments explored alternative approaches to key aspects of the pipeline, including preprocessing methods, Open Track data configurations, and the fundamental task formulation (classification vs. regression).

C.1 Ablation on Preprocessing: Morphological vs. Custom Pipeline

The core hypothesis of this work posits that deep morphological analysis surpasses simple surface-level normalization for Arabic readability. To test this, an alternative, custom preprocessing pipeline was implemented and evaluated. This custom method simplifies input rather than providing the linguistic enrichment of the d3tok analyzer. Its key steps include:

- **Aggressive Normalization:** Standardizes different forms of characters (e.g., Alef (أ, إ, آ) to ا, Taa Marbuta (ة) to Haa (ه)).
- **Diacritic Removal:** Strips all short vowel markings (تَشْكِيل).

This custom approach consistently yielded inferior results compared to the morphologically analyzed text. This suggests that linguistic information, such as morpheme boundaries and diacritics, preserved and added by the d3tok method, is vital for accurate text complexity assessment. Table 4 provides a direct comparison.

C.2 Ablation on Task Formulation: Classification vs. Regression

The fundamental framing of the readability assessment task was also explored. The problem can be approached as either a multi-class classification problem (predicting one of 19 discrete levels) or a regression problem (predicting a continuous score). An ablation study was conducted to evaluate the efficacy of a classification approach. As shown in Table 5, several pre-trained models were fine-tuned for sequence classification on the sentence-level task, but the regression approach ultimately yielded superior performance for this shared task.

Original Sentence	Custom Preprocessing (Tested)	d3tok Analysis (Used)
أَلَيْسَتْ هَذِهِ الْعَاطِفَةُ؟	اليست هذه العاطفه؟	أَ+ لَيْسَتْ هَذِهِ ال+ عَاطِفَةُ ؟
حَوَّلَ السَّائِقُ وَجْهَةَ فَرَسَيْهِ.	حول السائق وجهه فرسيه.	حَوَّلَ ال + سَائِقُ + وَجْهَةَ NOAN

Table 4: Comparison of the d3tok analysis (used in the final system) and the custom normalization pipeline (tested in an ablation study).

Track	Model Used	Dev (QWK)	Test (QWK)
Strict Sentence	CAMeL-Lab/readability-arabertv02-word-CE	73.31	78.20
	aubmindlab/bert-base-arabertv02	81.0	82.60
	CAMeL-Lab/readability-arabertv2-d3tok-reg	74.95	69.7
	CAMeL-Lab/bert-base-arabic-camelbert-mix-sentiment	81.60	82.70
Constrained Sentence	aubmindlab/bert-base-arabertv02	78.50	79.60
	CAMeL-Lab/readability-arabertv02-word-CE	69.0	72.20
Open Sentence	CAMeL-Lab/readability-arabertv02-word-CE	78.0	79.60

Table 5: Ablation study results for the classification approach on the sentence-level task, using the custom, non-morphological preprocessing pipeline.

D Ablation on Open Track Data Configuration

For the Open track, multiple experiments were conducted to determine the optimal mix of BAREC and DARES data, alongside ideal hyperparameters. The results, summarized in Table 6, indicate that the best configuration for the sentence-level task differed from that for the document-level task.

- **Best Sentence Performance (Exp 2):** The highest sentence-level QWK (83.9) was achieved with a lower learning rate (2e-5) over 18 epochs, using a simple concatenation of BAREC and DARES datasets for training and validation.
- **Best Document Performance (Exp 5):** The highest document-level QWK (79.2) was achieved with a higher learning rate (5e-5) and a more careful data splitting strategy. The validation set was explicitly augmented with a stratified 15% sample of the DARES data to better reflect the training distribution. This highlights the document task's sensitivity to validation set composition for robust model selection.

E Detailed Best Performing Models

This appendix presents a comprehensive breakdown of the final configurations used to achieve the best reported results. Table 7 details the specific models, training hyperparameters, data distributions, and corresponding QWK scores on both development and test sets for each track and task.

F Morphological Analysis Errors

This appendix illustrates a key challenge encountered during data augmentation: morphological analysis errors. Table 8 provides concrete examples where the CAMeL Tools d3tok analyzer failed to parse a word, inserting a NOAN (No Analysis) token. This issue arises with words not in the analyzer's vocabulary, such as uncommon proper nouns, or with words having valid but infrequent morphological forms. This introduced noise that could degrade model reliability, especially for the document-level task where a single mis-analyzed word can impact the score of an entire sentence.

ID	Training Parameters	Data Distribution	Sent. QWK	Doc. QWK
Exp 1	LR=5e-5, Epochs=6	**Train:** BAREC (54.8k) + DARES train (9.7k). **Total: 64.5k**. **Dev:** BAREC (7.3k) + DARES dev (1.4k). **Total: 8.7k**.	83.5	73.8
Exp 2	LR=2e-5, Epochs=18	Same as Exp 1.	**83.9**	76.1
Exp 3	LR=3e-5, Epochs=18	**Train:** All combined: BAREC + SAMER + DARES. **Total: 107.6k**. **Dev:** BAREC (7.3k).	83.6	74.6
Exp 4	LR=3e-5, Epochs=10	Same as Exp 1.	83.6	78.6
Exp 5	LR=5e-5, Epochs=20	**Train:** BAREC (54.8k) + 85% of merged DARES (11.8k). **Total: 66.6k**. **Dev:** BAREC (7.3k) + 15% of merged DARES (2.1k). **Total: 9.4k**.	83.0	**79.2**

Table 6: Summary of ablation experiments for the Open Track, detailing hyperparameters, data configurations, and final test set performance for both sentence and document tasks. LR refers to Learning Rate.

Track	Task	Model Used	Training Parameters	Data Distribution	Dev QWK	Test QWK
Strict	Sentences	aubmindlab/bert-base-arabertv2	Epochs=20 Learning Rate=3e-5 Warmup Ratio=0.1 Weight Decay=0.01	54845 training 7310 validation	82.30	84.20
	Document	aubmindlab/bert-base-arabertv2	Epochs=20 Learning Rate=3e-5 Warmup Ratio=0.1 Weight Decay=0.01	54845 training 7310 validation	62.37	79.90
Constrained	Sentences	CAMEL-Lab/readability-arabertv2-d3tok-reg	Epochs=8 Learning Rate=3e-5 Warmup Ratio=0.1 Weight Decay=0.01 Logging Steps=100	97874 training 7310 validation	81.0	82.9
	Document	CAMEL-Lab/readability-arabertv2-d3tok-reg	Epochs=15 Learning Rate=3e-5 Warmup Ratio=0.1 Weight Decay=0.01	97874 training 7310 validation	64.30	75.5
Open	Sentences	CAMEL-Lab/readability-arabertv2-d3tok-reg	Epochs=18 Learning Rate=2e-5 Warmup Ratio=0.1 Weight Decay=0.01	64548 training 8690 validation	82.70	83.90
	Document	CAMEL-Lab/readability-arabertv2-d3tok-reg	Epochs=20 Learning Rate=5e-5 Warmup Ratio=0.1 Weight Decay=0.01	66634 training 9391 validation	60.48	79.20

Table 7: Full details of the best performing models across all tracks and tasks.

ID	Original Sentence	D3tok Analysis
SAMER_13172	«كيف تنصرفين يا شوكار؟!»	« كيف تنصرفين يا NOAN
SAMER_15232	حول السائق وجهة فرسيه	حول ال + سائق + وجهة NOAN

Table 8: Examples of preprocessing artifacts from the SAMER corpus, where the CAMeL Tools analyzer failed to produce a morphological analysis, inserting a NOAN token instead. Failures occurred on an uncommon proper noun (شوكار), and a morphologically complex noun (فرسيه).