STBW at BAREC Shared Task 2025: AraBERT-v2 with MSE-SoftQWK Loss for Sentence-Level Arabic Readability

Saoussan Trigui
Independent Researcher
triguisaoussan51@gmail.com

Abstract

Automatic Readability Assessment estimates how hard a text is for its target readers, using features such as vocabulary, spelling, morphology, etc. Based on this premise, we evaluate our experiments on Arabic language under the BAREC 2025 shared task protocol. This paper addresses the sentence-level readability assessment task with strict track, that allows only the use of BAREC train set to predict Arabic readability on a fine-grained 19-level scale. Our solution is based on a two-phase fine-tuning of AraBERT-v2 on a custom feature set of the BAREC corpus. In the blind test set, the system achieves a QWK of 85.6%.

1 Introduction

Automatic Readability Assessment (ARA) is the task of computationally modeling the reading and comprehension difficulty of a text for a specific target audience. Its applications are diverse and impactful, spanning human-facing scenarios such as selecting appropriate educational materials for language learners, supporting readers with learning disabilities, and facilitating self-directed learning. In machine-facing contexts, ARA is instrumental in ranking search results by complexity, controlling the reading level of machine-translated output, and evaluating the efficacy of automatic text simplification systems (Vajjala, 2021). For Arabic, ARA is particularly challenging due to rich morphology, orthographic variation (e.g., diacritics and normalization) and dialectal/code-switching phenomena.

In this work, we aim to build a strong yet simple Arabic ARA system for the BAREC Shared Task (Elmadani et al., 2025a). We participate in the strict track of the sentence-level readability assessment where models must be trained exclusively on the training set of the Balanced Arabic Readability Evaluation Corpus (BAREC)[1] (Habash et al., 2025).

We cast sentence readability as scalar regression and then explicitly align training with the evaluation metric Quadratic Weighted Kappa (QWK). Concretely, we fine-tune AraBERT in two phases: an MSE warm-up followed by a differentiable QWK objective (SoftQWKLoss) that converts the scalar prediction into soft, distance-aware probabilities over the 19 levels and optimizes $(1 - \mathrm{QWK})$ on a soft confusion matrix. On the input side, we inject text metadata and statistics as well as linguistic cues (D3Tok) derived from CAMeL tools (Obeid et al., 2020). Empirically, this combination reaches QWK test results of 84.88, improving slightly over the 83.9 QWK score yielded by MSE-only phase. On the blind test leaderboard, we ranked in the 4th position out of 16 participations, with a 85.6 QWK for the sentence readability level subtask. The code of this solution is publicly available[2].

In summary, the proposed solution is composed of: (1) a compact AraBERT pipeline for Arabic readability that requires minimal feature engineering yet remains competitive, and (2) a metric-aligned training recipe (MSE $\rightarrow$ SoftQWK) that is architecture-agnostic and easy to reproduce.

Next, we present some background and we formalize the task and its input/output setup; then we present our method and training objectives, describe the experimental setup, and report results. We follow with error analyses, discuss limitations, and conclude.

2 Background

2.1 History of Automatic Readability Assessment (ARA)

The origins of ARA date back nearly a century to the development of manually computed readability formulas. These formulas are characteristically simple, often expressed as weighted linear func-

[1] https://barec.camel-lab.com/sharedtask2025

[2] https://github.com/Saoussan/BAREC_Arabic_Readability_Assessment

Proceedings of The Third Arabic Natural Language Processing Conference, pages 362–366
November 8-9, 2025

tions of easily quantifiable, surface-level textual features (Vajjala, 2021). Among the most influential and enduring of these is the Flesch Reading Ease formula (Flesch, 1948), which calculates a score on a 0-100 scale, based on average sentence length and average word length in syllables, and the Dale-Chall formula which uses a predefined list of common words to identify "difficult" vocabulary (Dale and Chall, 1948).The shift towards supervised machine learning, which reframes readability assessment as a classification or regression problem, allowed for the integration of richer sets of linguistic features. Algorithms like Support Vector Machines (SVM) or Random Forests, demonstrated superior performance compared to traditional formulas (Imperial and Kochmar, 2023). While researchers have progressed to complex neural network architectures, the traditional, simpler formulas continue to exist, especially in fields like education or healthcare, that value prediction interpretability (Vajjala, 2021).

2.2 Challenges of Arabic Language

Applying ARA methodologies to the Arabic language presents a set of challenges that are not adequately addressed by models developed primarily for English. These challenges stem from the inherent linguistic characteristics of Arabic, which impact text complexity (Cavalli-Sforza et al., 2018). First, Arabic is a morphologically rich language, characterized by a highly inflectional system. This means that surface-level metrics like average word length in characters, may be poor indicators of difficulty for Arabic. Second, Arabic orthography is marked by an ambiguity due to the optionality of diacritics (short vowel markings) in most written texts. A single undiacritized word form can correspond to multiple distinct words with different meanings and pronunciations, which can only be resolved through context. Third, no one speaks the Modern Standard Arabic (MSA) as a native mother tongue, a language that can differ substantially in lexicon, phonology, and grammar from the daily dialect. This complicates the very definition of a "target reader" and may make the task of assessing readability for L1 speakers challenging (Cavalli-Sforza et al., 2018).

2.3 ARA for Arabic

The research community has recently focused on Arabic text readability providing scientific resources (Al Khalil et al., 2020; Alhafni et al., 2024; Elmadani et al., 2025b; Habash et al., 2025; Hazim et al., 2022). The trajectory of Arabic ARA has largely mirrored that of English, beginning with attempts to adapt or create formulas tailored for Arabic (El-Haj and Rayson, 2016; Cavalli-Sforza et al., 2018; Liberato et al., 2024) and machine learning techniques (Cavalli-Sforza et al., 2018; Bessou and Chenni, 2021). Recently, we witnessed the development of pre-trained language models (PLMs), pre-trained specifically for the Arabic language (Inoue et al., 2021; Liberato et al., 2024; Antoun et al., 2020). Upon its release, AraBERT established new state-of-the-art results across various Arabic NLP benchmarks. In this work, we propose a two-phase fine-tuning of AraBERT PLM to predict Arabic text readability levels.

3 System Overview

For our experiments, we build upon the AraBERT-v2 baseline (Elmadani et al., 2025b), but extend it with additional features and a two-phase optimization strategy. Specifically, we fine tune AraBERT-v2 (Antoun et al., 2020) as the backbone encoder, while enriching its input with surface-level statistical indicators (e.g., word count and word-length statistics) and a morphologically segmented representation generated using the D3tok segmenter (Obeid et al., 2020), alongside the raw text. On top of the encoder, we employ a single-neuron regression head to predict continuous readability scores. In phase one, we fine-tune the model using mean squared error (MSE) loss to capture the ordinal nature of readability classes. Since the main evaluation metric in the challenge is QWK, we continue the fine-tuning of the model in phase two, with a differentiable Soft Quadratic Weighted Kappa (SoftQWK) loss (de la Torre et al., 2018), directly aligning optimization with the official evaluation metric (Cohen, 1968).

For this purpose, we take inspiration from (Diaz and Marathe, 2019) and turn the scalar prediction into a soft class distribution to be compatible with the SoftQWK loss. We clamp the real prediction value to a $[1 \dots 19]$ vector and spread its mass over the 19 readability levels with a Gaussian window centered at the predicted class. That yields a soft label probability vector $P \in \mathbb{R}^K$. With one-hot gold labels T, we construct a soft confusion matrix

$$O = T^t P \tag{1}$$

and the chance agreement E. Using the standard

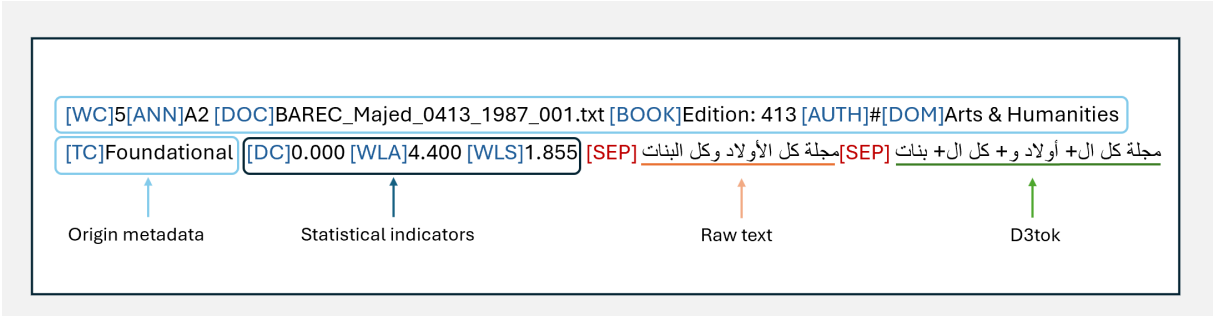

Figure 1: Example of an enriched data sample

quadratic weight matrix W, the QWK is defined as

$$\kappa = 1 - \frac{\langle W, O \rangle}{\langle W, E \rangle}, \tag{2}$$

During the training, we optimize the loss L

$$\mathcal{L} = 1 - \kappa. \tag{3}$$

4 Experimental Setup

4.1 BAREC Corpus

All our experiments are running on the Balanced Arabic Readability Evaluation Corpus (BAREC) dataset (Elmadani et al., 2025b; Habash et al., 2025). BAREC comprises over 69K sentences (around 1M words) covering three domains: Humanities, Social Sciences, and STEM and aimed at three readership groups (Foundational, Advanced, Specialized). Each sentence is annotated for readability on a fine-grained 19-level scale using guidelines developed by the authors. BAREC is considered as the largest Arabic corpus for readability assessment.

The authors establish baseline readability models (Elmadani et al., 2025b), at multiple granularities (19, 7, 5, 3 levels). We use the existing split of the BAREC Corpus which is ≈80% for training, ≈10% for validation, and ≈10% for testing (see Table 2).

Regarding feature preparation, we extend the input representation with additional components beyond the BAREC baseline setup. Each training instance is formatted as a single sequence that concatenates (i) origin metadata provided by the corpus, (ii) surface-level statistical indicators such as word count and word-length statistics, (iii) the raw text, and (iv) its D3tok-based morphological segmentation. The adopted features are listed below:

- **Word count**: the number of words in the raw sentence
- **Document**: the name of the source document
- **Book**: the name of the document's book
- **Author**: the name of the document's author
- **Domain**: the document's domain (one of *Arts & Humanities*, *STEM* or *Social Sciences*)
- **Text class**: the document's readership group (one of *Foundational*, *Advanced*, or *Specialized*)
- **Diacritics coverage**: frequency of diacritics in the raw sentence
- **Average word length**: the mean number of characters per word in the raw sentence
- **Word length standard deviation**: the standard deviation of the number of characters per word in the raw sentence
- **Sentence**: the raw text
- **D3tok**: morphologically segmented representation of the sentence

To ensure the model can differentiate between these heterogeneous sources of information, the components were separated by the special delimiter token `[SEP]`. We add a list of field separators as special tokens to the tokenizer (`[WC]`, `[ANN]`, `[DOC]`, `[BOOK]`, `[AUTH]`, `[DOM]`, `[TC]`, `[DC]`, `[WLA]`, `[WLS]`) in order to prevent them from being broken into subwords. This enriched representation provides the encoder with both shallow statistical cues and deeper morphological structure, while maintaining a structured and learnable input format. The model's input will look like the example in Figure 1.

4.2 Our experiments

We treat readability level prediction as a regression problem. We use a two-phase training schedule with distinct losses. In **Phase 1**, We fine-tune the AraBERT-v2 pretrained model in mixed-precision mode for 6 epochs with a batch size of 64. An

Loss	Acc^{19}	±1 Acc^{19}	Acc^{7}	Acc^{5}	Acc^{3}	QWK	Dist
SoftQWK	34.8%	74.0%	64.8%	72.0%	86.6%	**84.8%**	1.19
Baseline	43.1%	73.1%	61.1%	67.8%	75.9%	84.0%	1.13

Table 1: Results of our system compared to the baseline on the shared BAREC Test set

	#Documents	#Sentences	#Words
Train	1,518 (79%)	54,845 (79%)	832,743 (80%)
Dev	194 (10%)	7,310 (11%)	101,364 (10%)
Test	210 (11%)	7,286 (10%)	105,264 (10%)
All	**1,922 (100%)**	**69,441 (100%)**	**1,039,371 (100%)**

Table 2: BAREC splits

AdamW optimizer minimizes the Mean Squared Error (MSE) between the scalar prediction $\hat{y}$ and the gold label y using a learning rate of 2×10^{-5} with linear warm-up over 10% of the total updates. We consider the best checkpint on validation set, which is the third epoch, then, in **Phase 2**, we switch to the differentiable QWK objective (SoftQWKLoss): each $\hat{y}$ is converted to a Gaussian-smoothed distribution over the 19 levels, a soft confusion matrix is accumulated, and we minimize $1 - \kappa$ so that optimization is aligned with the leaderboard metric.

For evaluation, from the raw scalar $\hat{y}$ we report *MAE*. For ordinal metrics, we round and clip $\hat{y}$ to the range $[1, 19]$ and compute QWK, tolerance-1 accuracy (AdjAcc19), exact 19-way accuracy (Acc19), and coarse-bin accuracies (Acc7/Acc5/Acc3) obtained by collapsing the 19 levels into $7/5/3$ groups (Elmadani et al., 2025b).

5 Results

In **Phase 1:**. We train an AraBERTv2-based system on inputs combining origin metadata, statistical indicators, raw text, and D3Tok features, using an MSE loss. This yields an evaluation QWK of 83.9. In **Phase 2:** We then fine-tune the Phase-1 model with the SoftQWK loss, reaching almost a QWK of 84.9 which is above the shared task baseline (Table 1).

Error analysis: The per-level MAE plot (Figure 2) shows the largest errors at the highest readability levels. To understand the origins of these errors, we analyse the confusion matrix (Figure 3) which indicates that many true level-18/19 items are predicted as level 16. This is clearly due to the class imbalance that affects the boundary at the top of the scale. In the future, we will investigate employing data and loss weighting techniques to tackle this problem.

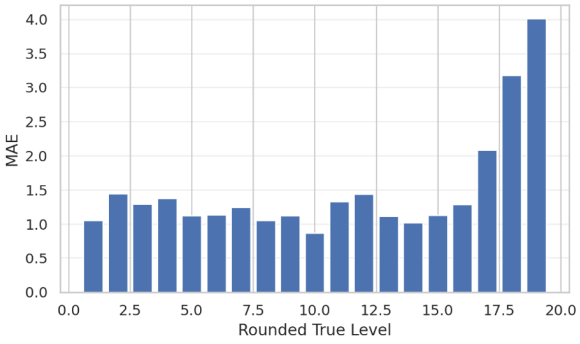

Figure 2: Per-level evaluation MAE vs. true level

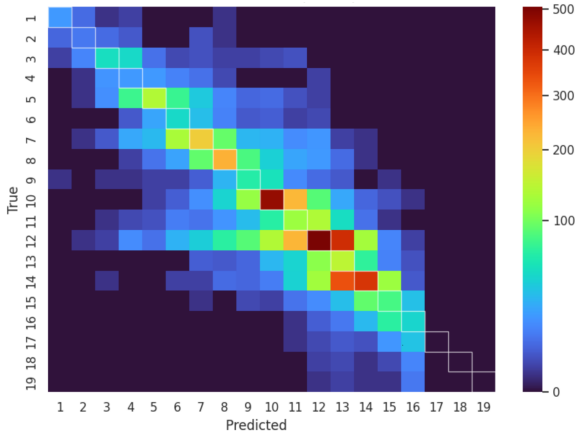

Figure 3: Confusion matrix (evaluation) of predictions cluster along the diagonal, with underestimation at high true levels (17–19).

6 Conclusion

In this work, we presented a competitive system for Arabic readability in the BAREC shared task. Using AraBERTv2 with lightweight metadata/statistics and CAMeL-derived D3Tok features from the BAREC dataset, we trained in two phases: an MSE warm-start followed by a metric-aligned SoftQWK loss. This increased QWK from 84.0% to 84.88% on the test set. Error analysis shows that most remaining mistakes occur at the highest levels (17–19), likely due to class imbalance. Going forward, we plan to mitigate this by augmenting training with the SAMER dataset (Alhafni et al., 2024) and other related resources.

References

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for Standard Arabic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France. European Language Resources Association.

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.

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.

Sadik Bessou and Ghozlane Chenni. 2021. Efficient measuring of readability to improve documents accessibility for arabic language learners. *arXiv preprint arXiv:2109.08648*.

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: current state and future directions. *Procedia computer science*, 142:38–49.

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

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154. Machine Learning and Applications in Artificial Intelligence.

Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

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.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

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.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. *arXiv preprint arXiv:2305.13478*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). 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.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.