Qais at BAREC Shared Task 2025: A Fine-Grained Approach for Arabic Readability Classification using a Pre-trained Model.

Samar Ahmed[1]
samar.sass6@gmail.com

[1]NAMAA, Riyadh, Saudi Arabia

Abstract

In this paper, the results are presented within the context of the BAREC 2025 Shared Task (Elmadani et al., 2025a; Habash et al., 2025; Elmadani et al., 2025b) for Arabic text readability prediction. Participation in both the strict and open tracks achieved QWK scores of 82.5% and 83%, respectively. The proposed approach employs a 19-level fine-grained classification framework at the sentence level, leveraging the BAREC dataset (Elmadani et al., 2025a; Habash et al., 2025; Elmadani et al., 2025b) and transformer based $AraBERT$ models. To address class imbalance, underrepresented levels were augmented with additional samples. By incorporating rich linguistic and structural features, including morphology, syntax, and vocabulary, the system surpasses less fine-grained methods in precision and reliability.

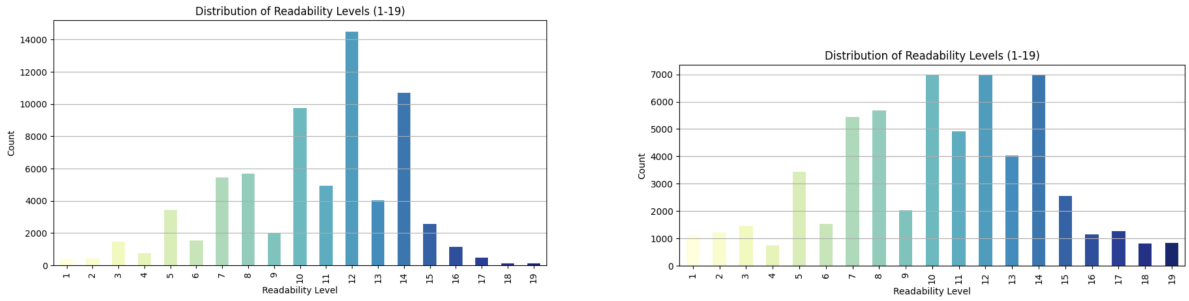
1 Introduction

Readability indicates to the ease with which someone can understand a particular text or sentence (Nassiri et al., 2022). Although the majority of early research in this area focused on English due to the abundance of rich and extensive datasets, readability evaluation for Arabic and other languages has gained attention in recent years. However, the Arabic language presents unique challenges, such as the lack of annotated datasets and the complexities of its syntactic and morphological structure. Readability is therefore a critical aspect of NLP, with practical implications across domains like education, public communication, and digital platforms, where improving text clarity enhances understanding for both native speakers and language learners. A number of studies have attempted to assess Arabic readability using various metrics and linguistic levels. For example, (El-Haj and Rayson, 2016) counts stressed, long, and short syllables to measure readability. This technique is a good starting point, but it falls short of effectively expressing the complex details of Arabic syntax and morphology. The BAREC dataset (Elmadani et al., 2025a; Habash et al., 2025) addresses some of these limitations by incorporating a wider range of linguistic and structural features such as spelling, word count, morphology, syntax, vocabulary, and content to improve classification accuracy. Previous studies have examined readability at various linguistic levels, including the sentence, document, word, and token levels, and have employed different scales, ranging from 3 to 7 levels, such as (Al Khalil et al., 2020) and (Hazim et al., 2022). However, no prior work has systematically investigated the potential of fine-grained readability levels for Arabic, particularly when combined with advanced transformer based language models at the sentence level. In this study, the limitation of broad-scale readability measures is addressed by employing a fine-grained 19-level classification system derived from the BAREC dataset. This framework is applied at the sentence level using large-scale Arabic language models based on $AraBERT$. By combining a fine-grained readability scale with advanced transformer based language models, the proposed approach aims to produce more accurate and reliable readability estimates for Arabic texts. This contribution not only expands the methodological landscape of Arabic readability assessment but also provides a scalable foundation for educational, institutional, and technological applications requiring precise control over text complexity. The rest of this paper is organized as follows: Section 2 reviews related work on Arabic readability, Section 3 describes the methodology, Section 4 presents the model results, Section 5 offers a discussion, and Section 6 provides an error analysis.

2 Related work

Focusing on Arabic readability research, researchers have made significant efforts to address

Proceedings of The Third Arabic Natural Language Processing Conference, pages 306–311



(a) Distribution before balancing

(b) Distribution after balancing

Figure 1: Distribution of dataset before and after balancing

the scarcity of data by building datasets to measure text difficulty, supporting the Arabic NLP community. One notable contribution is Al Khalil et al. (2020) and Alhafni et al. (2025), the Simplification of Arabic Masterpieces for Extensive Reading (SAMER) project, which presents a five-level readability lexicon for Modern Standard Arabic, manually annotated by language professionals from three Arab regions. The lexicon was built from news articles and literary texts. Following this, Al-Twairesh et al. (2016) introduced MADAD, a tool based on collecting readability annotations on Arabic texts at the sentence and paragraph levels using pairwise and direct rating methods, helping to fill the gap in Arabic readability data. Subsequent research, such as Elmadani et al. (2025a); Habash et al. (2025), developed a large and reliable dataset for assessing Arabic text readability at multiple granularities, fine-tuning $AraBERT$ to establish a baseline for sentence-level classification. Studies leveraging the SAMER dataset have used varied approaches:Liberato et al. (2024) assessed readability with methods ranging from rule-based to pre-trained language models, and Hazim et al. (2022) presented a Google Docs add-on for automatic Arabic word-level readability visualization, providing difficulty assessment, substitution suggestions, and foundational resources such as a graded readability lexicon and a parallel corpus.

3 Methodology

3.1 BAREC Dataset

The BAREC dataset (Elmadani et al., 2025a; Habash et al., 2025; Elmadani et al., 2025b) contains 69,441 Arabic sentences (more than 1 million words) from various genres and audiences, annotated across 19 readability levels from kindergarten to postgraduate, following the Inspired by the Taha/Arabi21 (Taha-Thomure, 2017). Annotations are performed manually, with high agreement between annotators (Quadratic Weighted Kappa = 81.8%), ensuring data quality. It is openly available and benchmarked using multiple readability assessment methods, supporting research and educational applications in Arabic readability.

3.1.1 Dataset for Strict track

For this track, the original BAREC dataset was used without any modifications, and no data augmentation was applied. The dataset consisted of 69,441 rows, with 80% allocated for training, 10% for development, and 10% for testing. Furthermore, BAREC provided a blind test set of 3,420 cases.

3.1.2 Dataset for Open track

In the Open track, we extended the training data by generating synthetic sentence-level examples using ChatGPT-based augmentation to improve model generalization across underrepresented readability levels. The original dataset was highly imbalanced, with some classes significantly overrepresented while others had very few samples. To address this, both up-sampling and down-sampling techniques were applied. Specifically, levels 1, 2, 17, 18, and 19 were up-sampled using GPT-generated data, whereas levels 10, 12, and 14 were down-sampled to 7,000 instances. This number was selected because it closely matches the size of the dataset's largest class after removing the three over-represented categories, thereby helping to balance the data distribution, as illustrated in Figure1a and Figure 1b. The final dataset consisted of 59,236 rows.

3.2 Model

$AraBERT$, a BERT-based pre-trained language model developed specifically for Arabic (Antoun et al., 2020), was introduced to address

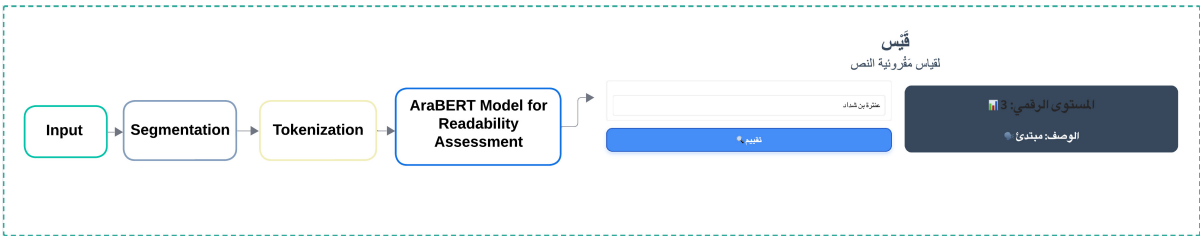


Figure 2: Flow of Qais model

the limitations of multilingual models by providing an architecture optimized for Arabic NLP tasks. It has achieved notable improvements in tasks such as Sentiment Analysis, Named Entity Recognition, and Question Answering. In the present work, two variants were utilized: aubmindlab/bert-arabertv2 and aubmindlab/bert-large-arabertv2, with aubmindlab/bert-arabertv2 selected as the primary model due to its stronger performance. Figure 2 illustrates the flow diagram for classifying Arabic texts according to readability levels. The process starts with entering the sentence, followed by segmentation. The segmented data is then tokenized and fed into the model, which has been trained on segmented data to enhance accuracy. Finally, the model produces a readability classification for the input sentence.

3.3 Hyperparameters

As part of hyperparameter optimization, the models were trained using NVIDIA A100 and T4 GPUs in Google Colab. The learning rates were set to either 2e-5 or 5e-5, with a weight decay of 0.01 to mitigate overfitting. Batch sizes were configured to 4 or 8, depending on the model's complexity and resource requirements. Maximum number of epochs was set to 20, and the AdamW optimizer, which is used by default in $AraBERT$, was employed during training.

4 Results

This task include a readability assessment, which evaluates both tracks using multiple metrics (Elmadani et al., 2025a; Habash et al., 2025; Elmadani et al., 2025b). Quadratic weighted Kappa (QWK) measures the agreement between predictions and accurate labels, with higher penalties for larger errors. It is the primary evaluation metric. Accuracy (Acc) is the percentage of exact matches between predictions and accurate labels using a 19-level scale (Acc19). Simplified versions include Acc7, Acc5, and Acc3, where the 19 levels are grouped into 7, 5, or 3 categories. Adjacent Accuracy (±1 Acc19) counts predictions as correct if they are exactly right or within ±1 level of the actual label. The average distance (dist) or mean absolute error (MAE) measures the average absolute difference between the predicted and actual labels. In the Readability Assessment task, results are reported for two evaluation tracks: Sentence-level Strict and Sentence-level Open. In the first track, the original BAREC dataset was used without any modifications, a QWK score of 82.5%. In the second track, data augmentation techniques, including up-sampling and down-sampling, were applied to the BAREC dataset, resulting in a slightly improved QWK score of 83.0%, as shown in the Table 1. These findings underscore the significant impact of data balancing on model performance. Two variants from the $AraBERT$ series were experimented with: aubmindlab/bert-base-arabertv2 and aubmindlab/bert-large-arabertv2. Initial experiments with arabertv2-base consistently yielded high performance, with QWK scores ranging between 80 and 83 across both tracks. In contrast, improving to the larger $AraBERTv2$ model resulted in reduced accuracy, with QWK scores ranging from 70% to 78%. Different learning rates (2e-5 and 5e-5) were also investigated, with 2e-5 consistently yielding better outcomes. Coral Loss, which preserves the ordinal character of labels by penalising predictions based on their distance from the actual label, was also investigated. However, when Coral Loss was applied, accuracy decreased.

5 Discussion

In this study, it was observed that the base-arabertv2 model, when trained on the BAREC dataset in both tracks, outperformed the large-arabertv2 model. This is likely because the large-arabertv2 model requires a larger dataset and greater computational resources. In the open track, a slight improvement over the strict track was recorded, which can be attributed to the signifi-

Track	QWK	±1 Acc	Acc19	Dis	Acc 7	Acc 5	Acc 3
Sentence-level Strict	82.5	54.8	71.8	1.1	65.1	69.5	75.3
Sentence-level Open	83.0	54.2	71.8	1.1	66.0	70.0	75.8

Table 1: Performance results Readability Assessment for both tracks

cant class imbalance in the dataset. Furthermore, there was substantial variance across the 19 readability levels: the initial and final levels contained far fewer samples, While the middle levels had relatively more data. This uneven distribution hurt the model's overall performance. To address this issue, class weights were applied in the strict track to reduce the impact of the severe imbalance. In the open track, the data was manually balanced and reduced, with adjustments made to extreme classes. However, the improvement achieved was not substantial, likely because class weights had already been applied to mitigate the imbalance. Upon examining the dataset, it was also found that some rows were duplicated and contained unfamiliar words, such as كونغ فو *kwn fw* which appeared frequently. Although written in Arabic script, كونغ فو *kwn fw* is a foreign term, and its repetition could potentially hinder the model's ability to interpret and classify inputs accurately. For example, the term might be classified as a high difficulty word, while in reality, it is simply a proper noun commonly used in western contexts. It was also noted that some sentences contained non-Arabic words written in Arabic characters. Such issues may reduce the clarity of the dataset's texts and hinder the overall performance of the model. Although the last few levels (17, 18, and 19) are highly similar, this did not cause significant confusion for the model, as their difficulty is very close. Merging these levels into a single unified level might have yielded slightly better results than keeping them separate. In contrast, differences between other levels appeared more distinct and beneficial, and it is likely that levels containing more data were classified with greater accuracy.

6 Error analysis

To better understand the model's performance in the strict and open tracks, a manual analysis was conducted on more than ten randomly selected sentences with divergent readability labels. The analysis revealed that both tracks produced competitive results, with minimal differences in overall performance. However, specific error patterns were observed. When an Arabic word contained or was attached to numbers, the model occasionally generated inconsistent readability predictions. For example, in the sentence 208 ماجد *mAjd*, which represents only a name, the expected classification was level 1; the trained model, however, assigned it level 3 in both tracks. Although such cases may not significantly affect performance, numbers can sometimes alter the contextual interpretation. In the sentence 10 جمادى الأولى *jmAdý AlÂwlý* 1440 هـ *h*, the model assigned a score of 13 in both tracks, likely due to numerical elements introducing classification confusion. This misclassification is particularly problematic as the actual difficulty level of the sentence is beginner-level. Another difficulty that has been noticed is that words written in Arabic script are derived from English. These phrases frequently obtained excessively high readability scores, despite their difficulty correlating more closely with the first or second levels of difficulty. Such misclassification can result in content incompatibilities with the intended audience. At higher levels, notably above level 10, the model demonstrated improved classification accuracy, with errors becoming less common and severe. This improvement can be attributed to the use of better syntactic and lexical patterns in larger phrases, which are less likely to contain numbers or symbols that could interfere with the model's classification process.

7 Conclusion and Future Work

In this work, a fine-tuned $AraBERT$ model was presented for the BAREC shared task in the strict and open tracks, targeting Arabic sentence readability assessment. The results, while satisfactory, indicate potential for further improvement. Future work will begin with traditional machine learning approaches and progress towards deep learning methods, ultimately leveraging pre-trained models, alongside enhanced data cleaning, class balancing, and class merging. The system is envisioned to be deployed as a web-based tool for the Arabic.

Limitations

Resource constraints on Google Colab Pro limited experimentation with larger datasets and models, with restricted RAM causing occasional training crashes. To mitigate this issue, batch sizes were reduced; however, future experiments will require access to larger computing resources to fully realize the model's potential.

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.

Nora Al-Twairesh, Abeer Al-Dayel, Hend Al-Khalifa, Maha Al-Yahya, Sinaa Alageel, Nora Abanmy, and Nouf Al-Shenaifi. 2016. Madad: a readability annotation tool for arabic text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4093–4097.

Bashar Alhafni, Reem Hazim, Juan David Pineros Liberato, Muhamed Al Khalil, and Nizar Habash. 2025. 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.

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).

Khalid Elmadani, Nizar Habash, and Hanada Taha. 2025a. 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.

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

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2022. Arabic l2 readability assessment: Dimensionality reduction study. *Journal of King Saud University-Computer and Information Sciences*, 34(6):3789–3799.

Hanada Taha-Thomure. 2017. *Arabic Language Text Leveling*. Educational Book House.

Appendix A: Readability Dimensions Used for Sentence Generation

In my experiments, I provided GPT with the six dimensions from the BAREC readability framework (Elmadani et al., 2025a; Habash et al., 2025) and asked it to follow them when generating sentences at different readability levels. These dimensions are briefly described below:

1. **Word Count:** Measured by counting unique printed words (punctuation and diacritics ignored). This feature is constrained to a maximum of 20 words up to level 11 (Kaf).

2. **Orthography and Phonology:** Focused on word length (syllable count) and special letters such as hamzas. Final diacritics are ignored (words are read in pause form).

3. **Morphology:** Included derivation and inflection (e.g., tense, aspect, number). Simpler forms (e.g., present tense before past, singular before plural) appear at lower levels. This feature is used up to level 13 (Mim).

4. **Syntactic Structures:** Tracked sentence complexity, ranging from single words (level 1 –

Alif) to more complex structures. Applied up to level 15 (Seen).

5. **Vocabulary:** Central across all levels. Shared words across dialects and Modern Standard Arabic appear in easier levels, while technical terms are introduced in higher levels.

6. **Ideas and Content:** Evaluated required prior knowledge, symbolic decoding, and conceptual connections. Progression moves from familiar ideas to specialized knowledge, and from literal meanings to abstract concepts.

These dimensions guided the construction of sentence examples used in our readability experiments.