ANLPers at BAREC Shared Task 2025: Readability of Embeddings: Training Neural Readability Classifiers on the BAREC Corpus

Serry Sibaee[1*] **Yasser Alhabashi**[1] **Omer Nacar**[2] **Adel Ammar**[1] **Wadii Boulila**[1]
[1]Prince Sultan University, Riyadh, Saudi Arabia
[2]Tuwaiq Academy – Tuwaiq Research and Development Center
{yalhabashi, ssibaee, aammar , wboulila}@psu.edu.sa
{o.najar}@tuwaiq.edu.sa
[*]Corresponding author: ssibaee@psu.edu.sa

Abstract

This paper presents a neural approach to Arabic readability assessment using the BAREC corpus for fine-grained classification across 19 readability levels. Our two-stage system combines embeddings from multiple pre-trained Arabic transformer models (ARBERTv2, MARBERTv2, AraBERT) with a Multi-Layer Perceptron classifier. We achieve competitive performance with Quadratic Weighted Kappa scores of 73.00-76.35, accuracy of 44.73%, and adjacent accuracy of 61.40%, within 8% of baseline models. The system offers significant practical advantages including rapid training time (10 minutes per experiment), compact architecture (12-15 million parameters), and efficient inference, making it suitable for resource-constrained deployment. Our analysis identifies dataset quality challenges including inconsistent diacritization and annotation issues that impact performance. This work provides a foundation for practical Arabic readability assessment tools in educational applications.

1 Introduction

Automatic readability assessment has become increasingly important in educational technology, content adaptation, and accessibility applications in many languages including Arabic (Liberato et al., 2024). Traditional readability metrics rely heavily on surface-level features such as sentence length and syllable counts (Uçar et al., 2024), which often fail to capture the nuanced linguistic complexity that affects human comprehension. Recent advances in neural language models and contextual embeddings offer new opportunities to develop more sophisticated readability classifiers that can better model the relationship between text characteristics and reading difficulty (Hazim et al., 2022).

This work investigates the application of modern neural architectures and embedding techniques to readability classification using the BAREC corpus. We address key challenges in current modeling approaches including the need for better representation of semantic complexity, syntactic structures, and discourse coherence. Our novel approach combines multiple embedding strategies with attention mechanisms to create interpretable readability predictions. The contributions of this work include empirical analysis of embedding effectiveness for readability tasks and a comprehensive evaluation framework for neural readability classifiers.

2 Background

Text readability plays a vital role in ensuring comprehension, retention, and engagement, especially in educational and medical (Venturi et al., 2015) contexts where aligning reading material with student proficiency is critical. Fine-grained readability frameworks, such as Fountas and Pinnell (Ransford-Kaldon et al., 2010) for English and the 19-level system for Arabic (Elmadani et al., 2025b), and some researchers used RL to develop readability assessment systems (Mohammadi et al., 2023) are widely used to support literacy development.

In this work, we participate in the **BAREC Shared Task 2025 on Arabic Readability Assessment:Sentence-level-Open** (Elmadani et al., 2025a), which focuses on sentence-level classification into one of 19 Taha-Thomure levels.(Taha-Thomure, 2017), from kindergarten to postgraduate proficiency. We use the newly released **BAREC corpus** (Elmadani et al., 2025b), a large, balanced dataset (splitted as: 54845 training sample, 7310 validation, 7286 test and 3420 blind-test) annotated according to the fine-grained guidelines outlined by (Habash et al., 2025). The task is a challenging multi-class classification problem requiring precise sentence-level prediction.

The corpus is derived in part from the **SAMER Arabic Text Simplification Corpus** (Alhafni et al., 2025) and (Al Khalil et al., 2020), and Figure 1

Proceedings of The Third Arabic Natural Language Processing Conference, pages 269–273

illustrates the distribution of sentences across the 19 levels. Our system aims to automatically predict the correct level for each input sentence from raw Arabic text, enabling more effective support for educational applications and adaptive reading technologies.

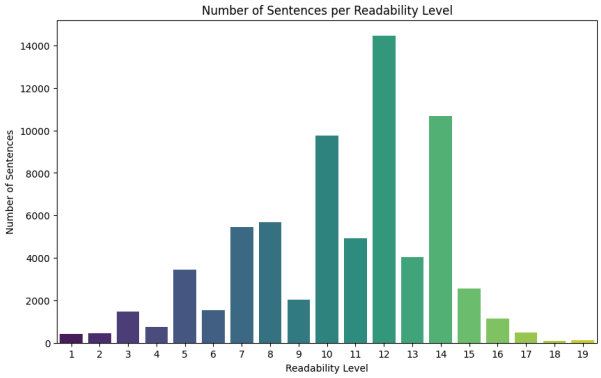

Figure 1: Distribution of sentences across the 19 readability levels in the BAREC corpus

Below are example sentences (randomly selected from the dataset from each level) along with their assigned readability levels:

- Level 1: ضعيفٌ - "Weak"
- Level 2: مرحباً. - "Hello."
- Level 3: الْقُرْآنُ الْكَرِيمُ - "The Noble Quran"
- Level 4: الْحَمْدُ لله - "Praise be to Allah"
- Level 5: أُصدرُ حُكْمًا: - "I issue a judgment:"
- Level 6: قلت الأمل - "I said hope"
- Level 7: أرجوك تخلص فورًا من ... الخ - "Please get rid of immediately... etc."
- Level 8: أَتَأَمَّل المَشْهَدَ الآتِي، ثُمَّ ... الخ - "I contemplate the following scene, then... etc."
- Level 9: كيف أهدأ وأنا بالأمس ... الخ - "(How can I calm down when yesterday... etc."
- Level 10: ثالثًا: سريعة جدًا، تتحرك ... الخ - "Third: very fast, it moves... etc."
- Level 11: لست متأكدًا أي أفق ... الخ - "I'm not sure which horizon... etc."
- Level 12: وأشهر مدنه: أرجيش، بَدْلِيس أو ... الخ - "And its most famous cities: Erciş, Bitlis or... etc."
- Level 13: فإذا جاءتْ بشدّتها وغمراتِها، عندَ ... الخ - "When it comes with its intensity and overwhelming force, when... etc."
- Level 14: وَقَالَ مَعْهَدُ بُحُوثِ السَرطَانِ ... الخ - "And the Cancer Research Institute said... etc."
- Level 15: ويَنطوي هذا التَوْزيع في الوظائف ... الخ - "And this distribution in functions involves... etc."
- Level 16: ولكن قيمتَه لا تُقدَرُ في ... الخ - "But its value cannot be estimated in... etc."
- Level 17: تَسمَعُ للحَلِي وَسْوَاساً إذا ... الخ - "You hear a whisper of jewelry when... etc."
- Level 18: تَرَى اللَحزَ الشَحيحِ إِذَا ... الخ - "You see the meager flesh when... etc."
- Level 19: يَتْبَعْنَ قُلَّةَ رَأْسِهِ وَكَأَنَّهُ ... الخ - "They follow the crown of his head as if he... etc."

The complexity progression across readability levels is also reflected in the sentence length characteristics. Figure 2 demonstrates the distribution of word counts per sentence across different readability levels, showing how sentence complexity generally increases with higher readability levels.

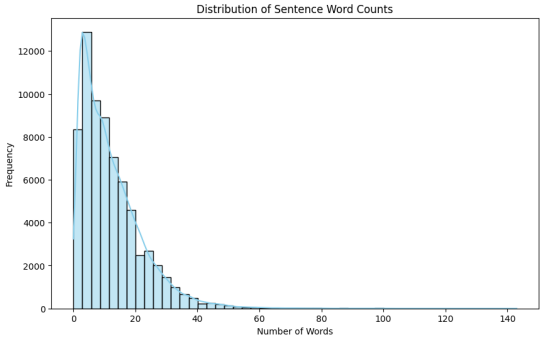

Figure 2: Distribution of word counts per sentence

Developing accurate automatic readability assessment models for Arabic is essential for advancing literacy education, supporting language learning applications, and improving academic performance evaluation. This task plays a vital role in standardizing Arabic text complexity assessment and contributes to the broader goal of enhancing Arabic language education through technology-driven tools.

3 System Overview

Our system for automatic readability assessment is a two-stage pipeline designed to first extract deep

linguistic features from Arabic text and then predict a readability score (Sibaee et al., 2024). This architecture addresses the core challenge of capturing the complex interplay of semantic and syntactic features that determine text difficulty[1].

The entire process can be conceptualized as a composition of two functions. First, an embedding function, E, maps the input text T to a fixed-size vector representation $\mathbf{e}$. This vector is then processed by a prediction model, M, parameterized by trainable weights W, to produce the final readability score, $\hat{y}$.

1. **Text Embedding:** $\mathbf{e} = E(T)$, where $\mathbf{e} \in \mathbb{R}^d$
2. **Readability Prediction:** $\hat{y} = M_W(\mathbf{e})$

Stage 1: Multi-Model Text Embedding (E). To create a robust feature vector, we generate embeddings from an ensemble of pre-trained Arabic transformer models: ARBERT, AraBERT, and MARBERT. Our design decision to use multiple models is to ensure the final representation is rich and generalized. For each model, the input text is tokenized, and the model outputs contextualized embeddings for every token. We compute a single sentence-level vector for each model by taking the mean of its token output embeddings. The final embedding, $\mathbf{e}$, is the element-wise average of the vectors from all three models. This averaging technique smooths the representation space and captures a broader range of linguistic nuances critical for readability assessment.

Stage 2: Readability Prediction Model (M). The resulting embedding vector $\mathbf{e}$ serves as the input to our prediction model, M, which is a Multi-Layer Perceptron (MLP). This feed-forward neural network is configured with several hidden layers and is trained to learn the complex, non-linear mapping from the dense text features to a continuous readability score. The model's parameters, W, are optimized using a regression loss function to minimize the error between its predicted scores and the ground-truth labels.

4 Experimental Setup

4.1 Dataset and Preprocessing

We evaluate our approach on the CAMeL-Lab/BAREC-Shared-Task-2025-sent dataset (Elmadani et al., 2025a) from Hugging Face (we did not evaluate on the validation split so we added them to the training to expand the samples)[2], a benchmark for Arabic readability assessment. The preprocessing pipeline consists of two steps: (1) text normalization by removing non-Arabic letters and numbers, and (2) lemmatization using Sina Tools (Hammouda et al., 2024) to reduce text nosiness and data sparsity. The methodology consist of trying multiple combination of the pre-processing techniques in the expirements which showed a very closed results either with them or direct training without pre-processing.

4.2 Model Architecture

Our system combines pre-trained embedding models with a Multi-Layer Perceptron (MLP) classifier, implemented in PyTorch using Hugging Face libraries. We evaluate two embedding categories: general multilingual models (LaBSE (Reimers and Gurevych, 2020), all-MiniLM-L6-v2, Matryoshka-based (Nacar et al., 2025)) and Arabic-specific BERT models (ARBERTv2, MARBERTv2 (Abdul-Mageed et al., 2021), AraBERTv2 (Antoun et al.)).

4.3 Training Configuration

The MLP architecture uses 3-4 hidden layers in descending configuration (e.g., [4096, 2048, 1024, 512]). Training employs AdamW optimizer with learning rates of 10^{-4} or 10^{-5}, batch sizes up to 65,536 (using A100-80GB), and 800-2000 epochs with early stopping. Regularization includes dropout (0.3-0.5) and weight decay (10^{-5}). All experiments use random seed 42 for reproducibility.

5 Results

We conducted extensive experiments across multiple configurations, achieving consistent performance on key metrics (QWK, Accuracy, Adjacent Accuracy) with QWK scores ranging from 65 to 76. This section presents our most promising results on both test and blind-test datasets provided by the shared task.

5.1 Experimental Configurations

After conducting numerous experiments, we observed that the results were highly similar; therefore, we selected the two best configurations, tak-

[1]The system is open-sources on github https://github.com/riotu-lab/readability_library_training

[2]note:The system is not directly comparable to other participants' systems because it uses the development set for training.

ing into account their differences in specific aspects, as shown in Table 1.

Parameter	Exp-1	Exp-2
Emb. Model	ARBERTv2	MARBERTv2
Input Size	768	768
Hidden Layers	SY	[SY, 512]
Dropout Rate	0.2	0.4
Learning Rate	10^{-5}	10^{-2}
Epochs	800	1200
Weight Decay	10^{-5}	$3 * 10^{-5}$
Early Stop	25	100
Scheduler	5	25

Table 1: Training configurations for best-performing experiments. Note: the default hidden layer is [4096, 2048, 1024] symbolized as 'SY'

5.2 Main Results

The primary findings of our experiments are presented in Table 2, which provides a comparative overview of model performance across different evaluation settings. The results indicate that both experiments achieved nearly identical accuracy and adjusted accuracy, with only slight variations in QWK. This consistency demonstrates the robustness of the approach across test and blind test datasets.

Exp.	Accuracy (%)	Adj Accuracy (%)	QWK
Exp-1	44.73	61.35	76.35
Exp-2	44.70	61.40	73.00

Table 2: Performance results on test dataset (exp-1) and blind test (exp-2)

5.3 Analysis and Discussion

Through extensive experimentation and dataset analysis (Sibaee et al., 2025), we identify two key observations:

5.3.1 Dataset Characteristics

Our analysis reveals several data quality issues that impact model performance: (1) inconsistent word diacritization across texts, (2) irregular punctuation usage patterns, (3) incomplete or fragmented sentences containing irrelevant symbols and noise, and (4) incorrect readability classifications for certain sentence types, particularly poetry verses and literary excerpts. These inconsistencies introduce noise that affects the reliability of readability predictions[3].

5.3.2 Model Architecture Performance

While our approach did not achieve state-of-the-art results, it demonstrates competitive performance compared to the baseline model (Elmadani et al., 2025b), achieving QWK scores within 8% of the baseline. However, our pipeline offers significant practical advantages: (1) substantially faster training time (approximately 10 minutes per experiment), (2) compact model size (12-15 million parameters). These characteristics make our approach particularly suitable for fast training in resource-constrained.

6 Conclusion

This research demonstrates that efficient neural architectures can achieve competitive performance for Arabic readability assessment while offering substantial practical advantages. Our two-stage system achieved QWK scores of 73.00-76.35 on the BAREC corpus, performing within 8% of baseline models with significantly faster training time and compact model size. The approach successfully addresses deployment considerations critical for educational technology applications in resource-constrained environments. Our analysis identified important dataset quality issues including inconsistent diacritization and annotation challenges that affect model performance. While not achieving state-of-the-art results, this work establishes a practical foundation for Arabic readability classification and highlights key areas for future corpus development and model improvement.

Acknowledgments

We would like to thank Prince Sultan University for their generous support in enabling this research.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

[3]Also as shown in figure 1, there is small amout of high level sentences so we expanded it using more Arabic poems and some teaching manzomat (more than 13K sample) on the link https://huggingface.co/datasets/JadwalAlmaa/Expand_BAREC. Note: we did not used this new dataset in our training.

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. 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. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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 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.

Tymaa Hammouda, Mustafa Jarrar, and Mohammed Khalilia. 2024. Sinatools: Open source toolkit for arabic natural language processing. *Preprint*, arXiv:2411.01523.

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.

Hamid Mohammadi, Seyed Hossein Khasteh, Tahereh Firoozi, and Taha Samavati. 2023. Text as environment: A deep reinforcement learning text readability assessment model. *Preprint*, arXiv:1912.05957.

Omer Nacar, Anis Koubaa, Serry Sibaee, Yasser Al-Habashi, Adel Ammar, and Wadii Boulila. 2025. Gate: General arabic text embedding for enhanced semantic textual similarity with matryoshka representation learning and hybrid loss training. *arXiv preprint arXiv:2505.24581*.

Carolyn R Ransford-Kaldon, E Sutton Flynt, Cristin L Ross, Louis Franceschini, Todd Zoblotsky, Ying Huang, and Brenda Gallagher. 2010. Implementation of effective intervention: An empirical study to evaluate the efficacy of fountas & pinnell's leveled literacy intervention system (lli). 2009-2010. *Center for Research in Educational Policy (CREP)*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Serry Sibaee, Abdullah Alharbi, Samar Ahmad, Omer Nacar, Anis Koubaa, and Lahouari Ghouti. 2024. ASOS at KSAA-CAD 2024: One embedding is all you need for your dictionary. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 697–703, Bangkok, Thailand. Association for Computational Linguistics.

Serry Sibaee, Omer Nacar, Adel Ammar, Yasser Al-Habashi, Abdulrahman Al-Batati, and Wadii Boulila. 2025. From guidelines to practice: A new paradigm for arabic language model evaluation. *Preprint*, arXiv:2506.01920.

Hanada Taha-Thomure. 2017. *Arabic Language Text Leveling* (معايير هنادا طه لتصنيف مستويات النصوص العربية). Educational Book House (دار الكتاب التربوي للنشر والتوزيع).

Suna-Şeyma Uçar, Itziar Aldabe, Nora Aranberri, and Ana Arruarte. 2024. Exploring automatic readability assessment for science documents within a multilingual educational context. *International Journal of Artificial Intelligence in Education*, 34(4):1417–1459.

Giulia Venturi, Tommaso Bellandi, Felice Dell'Orletta, and Simonetta Montemagni. 2015. NLP–based readability assessment of health–related texts: a case study on Italian informed consent forms. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 131–141, Lisbon, Portugal. Association for Computational Linguistics.