GNNinjas at BAREC Shared Task 2025: Lexicon-Enriched Graph Modeling for Arabic Document Readability Prediction

Passant Elchafei[*]
Ulm University, Germany
passant.elchafei@uni-ulm.de

Mayar Osama[*]
German University in Cairo, Egypt
mayar.osama@guc.edu.eg

Mohamed Rageh
German University in Cairo, Egypt
mohamad.rageh@student.guc.edu.eg

Mervat Abuelkheir
German University in Cairo, Egypt
mervat.abuelkheir@guc.edu.eg

Abstract

We present a graph-based approach enriched with lexicons to predict document-level readability in Arabic, developed as part of the Constrained Track of the BAREC Shared Task 2025. Our system models each document as a sentence-level graph, where nodes represent sentences and lemmas, and edges capture linguistic relationships such as lexical co-occurrence and class membership. Sentence nodes are enriched with features from the SAMER lexicon as well as contextual embeddings from the Arabic transformer model. The graph neural network (GNN) and transformer sentence encoder are trained as two independent branches, and their predictions are combined via late fusion at inference. For document-level prediction, sentence-level outputs are aggregated using max pooling to reflect the most difficult sentence. Experimental results show that this hybrid method outperforms standalone GNN or transformer branches across multiple readability metrics. Overall, the findings highlight that fusion offers advantages at the document level, but the GNN-only approach remains stronger for precise prediction of sentence-level readability.

1 Introduction

Accurately assessing the readability of Arabic documents is essential for educational technologies, language learning platforms, and adaptive content delivery systems. The task poses significant linguistic challenges due to the diglossic nature of Arabic, rich morphology, and the scarcity of large-scale annotated corpora (Imperial and Kochmar, 2023). The BAREC Shared Task 2025 (Elmadani et al., 2025a) addresses this by providing a fine-grained classification benchmark: assigning one of 19 readability levels to Arabic texts at both the sentence and document level.

Previous work on Arabic NLP has applied deep contextual models such as BERT variants for various classification tasks, including readability prediction (Al-Tamimi et al., 2014; Antoun et al., 2020). Although effective, these approaches typically operate only on text sequences and often overlook explicit structural and lexical relationships that can influence readability. In contrast, graph-based methods make it possible to encode document-level structure and linguistic relationships directly (Sun et al., 2023). In this work, we explicitly incorporate such relationships by leveraging the SAMER lexicon for lexical difficulty features and constructing a heterogeneous sentence-lemma graph with multiple edge types (e.g., HAS_LEMMA, OCCUR_WITH, IN_CLASS, IN_DOMAIN). This allows our model to combine the strengths of contextual embeddings with explicit lexical and structural graph modeling, which we show experimentally to improve both sentence-level and document-level readability prediction.

We propose a hybrid approach that represents each document as a graph, where nodes correspond to sentences and lemmas, and edges represent linguistic relationships such as `HAS_LEMMA`, `OCCUR_WITH`, and `IN_CLASS`. Each sentence node is enriched with difficulty signals from the SAMER lexicon (Al Khalil et al., 2020) and contextual sentence embeddings from the `readability-arabertv2-d3tok-CE` model, a fine-tuned variant of AraBERTv2 optimized for Arabic readability classification (Antoun et al., 2020).

To integrate both modalities, we train the GNN (graph modality) and the transformer (text modality) independently and use late fusion to merge their readability predictions at the end of inference. This approach combines the strengths of structured lexical-graph features and contextual

[*] Equal contribution.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 261–265
November 8-9, 2025 ©2025 Association for Computational Linguistics

text embeddings, without mixing intermediate features. Document-level labels are then obtained by pooling the sentence-level predictions, using max-pooling to reflect the most difficult sentence.

Our experiments demonstrate that this lexicon-enriched, confidence-aware, graph-based approach significantly improves prediction performance over individual branches. The results emphasize the importance of combining structured lexical knowledge with neural contextualization and fusion to better capture Arabic document readability.

2 Related Work

Automatic readability assessment has become a key area in NLP due to its applications in education, text simplification, and adaptive content delivery. In English, early studies relied on surface-level features such as sentence length and word frequency, followed by statistical models and, more recently, neural methods that capture semantic and discourse-level information (Imperial and Kochmar, 2023).

For Arabic, early research was constrained by resource scarcity and linguistic complexity. One of the first efforts was the AARI index (Al-Tamimi et al., 2014), which used handcrafted lexical and syntactic features derived from academic curricula. Later, the SAMER Lexicon (Al Khalil et al., 2020) introduced a large-scale graded vocabulary resource. Subsequently, it was showcased in a word-level readability visualization system designed for assisted text simplification (Hazim et al., 2022). More recently, the SAMER Corpus (Alhafni et al., 2024) provided the first manually annotated Arabic parallel dataset for text simplification targeting school-aged learners. These resources provided the foundation for subsequent work.

In recent years, several datasets have advanced Arabic readability modeling. The BAREC corpus (Elmadani et al., 2025b) provides a large-scale benchmark with 19 readability levels at both the sentence and document level, while the DARES dataset (El-Haj et al., 2024) focuses on Saudi school textbooks. Complementary approaches, such as AraEyebility (Baazeem et al., 2025), integrate eye-tracking signals to connect human cognitive processing with readability prediction. In addition, (Liberato et al., 2024) explored strategies for Arabic readability modeling, highlighting the need to combine lexical resources with modern learning-based approaches. A survey by (Cavalli-Sforza et al., 2018) provides an overview of the challenges and future directions for Arabic readability assessment.

Overall, most Arabic readability models have focused on surface features or contextual embeddings in isolation, with limited integration of structured lexical knowledge. To our knowledge, no prior work has combined lexicon-enrichment with graph-based modeling for Arabic document readability. Our work addresses this gap by integrating the SAMER Lexicon into a heterogeneous sentence-lemma graph, capturing both vocabulary difficulty and structural relations to improve fine-grained readability prediction.

3 System Overview

The purpose of our approach is to capture the linguistic characteristics and the relationships between the features of two datasets: BAREC (Elmadani et al., 2025b) and SAMER (Al Khalil et al., 2020). The BAREC dataset consists of **sentences** annotated with their corresponding **readability levels**. The SAMER dataset consists of **lemmas**, each associated with an average **readability level** across different dialects, along with additional features such as frequency of occurrence and part-of-speech (POS) tags for each *(lemma, readability level)* pair.

We integrate the two datasets by extracting lemmas from the sentences while preserving their POS tags and recording the count of diacritics. The extraction of lemmas was performed using the CAMeL Tools Morphology Analyzer (Obeid et al., 2020). Each extracted lemma is then matched against the SAMER lexicon to enrich it with statistical attributes such as average readability, frequency, and POS. This alignment ensures that the SAMER lexicon contributes directly to the graph as node features rather than as isolated entries.

The combined data is reformulated into a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of multiple node and edge types. The node set $\mathcal{V}$ includes:

- **Sentences**: Represented by 768-dimensional embeddings obtained from the CAMeL-Lab Arabic readability model (readability-arabertv2-d3tok-CE), augmented with linguistic features.
- **Lemmas**: Characterized by statistical attributes such as average readability and frequency.

- **Classes**: Educational difficulty levels, one-hot encoded as Foundational, Advanced, or Specialized.
- **Domains**: Subject domains encoded as Arts & Humanities, STEM, or Social Sciences.

The main objective of our approach is to leverage both classical linguistic features from the two datasets and deep Graph Neural Networks (GNNs) to capture hidden patterns within the data. The edge set $\mathcal{E}$ contains the following directed relations:

- `sentence` $\rightarrow$ `lemma` (*HAS_LEMMA*): Indicates lexical composition.
- `lemma` $\leftrightarrow$ `lemma` (*OCCUR_WITH*): Represents lemma co-occurrence in context.
- `sentence` $\rightarrow$ `class` (*IN_CLASS*): Connects each sentence to its labeled difficulty class.
- `sentence` $\rightarrow$ `domain` (*IN_DOMAIN*): Links sentences to their broader academic domain.

Once the data is structured into the graph format, the first step is to apply input feature transformation, where we transform the node features into the model's hidden dimensions, for which we used a linear layer between the original dimensions to the target dimension to find optimal projection.

$$h_v^{(0)} = W_{\text{in}}^{(\tau)} x_v, \quad \text{for } v \in \mathcal{V}_\tau$$

where $h_v^{(0)}$ is the initial hidden representation of node v after projection, $W_{\text{in}}^{(\tau)}$ is the trainable weight matrix for input transformation for node type, $\mathcal{V}_\tau$ denotes nodes of type τ, and x_v is the raw feature vector.

The core of the model consists of a stack of SAGEConv (Hamilton et al., 2017) hidden layers. Each is used to learn the graph embeddings over the heterogeneous graph. It uses neighbor sampling and aggregation. Each layer applies a learnable linear transformation to the combined features; this transformation allows the model to learn complex feature interaction while maintaining consistent dimensions across the layers.

The model consistes of 4 GNN layers, for which we use ReLU activation function σ and layer normalization to avoid linearity and improve the gradient flow.

$$h_v^{(k)} = \sigma\left(\text{AGGREGATE}_{\text{type}}\left(\left\{h_u^{(k-1)} : u \in \mathcal{N}_{\text{type}}(v)\right\}\right)\right)$$

where $h_v^{(k)}$is the hidden representation of node v at layer k, $h_u^{(k-1)}$ is the hidden representation of neighbor node u from the previous layer, and $\mathcal{N}_{\text{type}}(v)$ denotes the set of neighboring nodes of v connected via a specific edge type. Additionally, we use a residual connection per layer. This preserves the features and provides more stable training, especially for the sentence nodes.

$$h_v^{(k)} \leftarrow \text{LayerNorm}\left(h_v^{(k)} + h_v^{(k-1)}\right)$$

Finally, an MLP layer used for the classification.

$$y_v = \text{MLP}(h_v^{(L)})$$

4 Experimental Results

We conduct experiments on both sentence-level and document-level readability prediction tasks, as defined in the BAREC Shared Task 2025. For sentence-level classification, each sentence is represented as a node in the graph and labeled with one of 19 readability levels. For document-level prediction, we reuse the same model architecture and apply aggregation over sentence-level predictions. Specifically, we take the most difficult predicted sentence level (i.e., max pooling) as the document's predicted readability level based on the intuition that the most complex sentence may determine the document's comprehensibility floor.

We evaluate two configurations:

- **Late Fusion:** Combining weighted outputs from the GNN and transformer-based sentence encoder.
- **GNN Only:** Using the graph-based model without fusion.

The results in Table 1 show distinct trends between sentence-level and document-level tasks. For document-level prediction, Late Fusion outperforms the GNN-only baseline in both Quadratic Weighted Kappa (QWK; 76.9% vs. 75.6%) and exact accuracy (42.0% vs. 40.0%), while maintaining similar scores in the other metrics. QWK is a standard evaluation metric for ordinal classification that accounts for the degree of disagreement between predicted and true labels, making it particularly relevant for readability level prediction.

In contrast, for sentence-level prediction, the GNN-only model achieves substantially higher accuracy (50.0% vs. 41.4%) and better results in most metrics, despite both models having the same

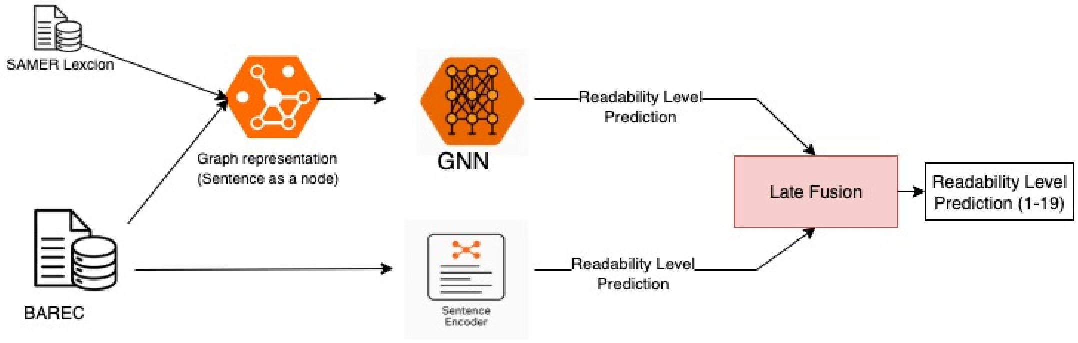

Figure 1: Overview of our proposed hybrid architecture for Arabic readability prediction. Sentence-level graphs are constructed using lexical relations from the SAMER lexicon and structural information from BAREC data. The GNN branch processes the graph to produce a softmax probability distribution over 19 readability levels, while the sentence encoder branch generates parallel probabilities from contextual embeddings. Inference uses late fusion, where both probability vectors are combined at the prediction stage using a tunable weight, yielding the final readability level for each sentence or aggregated document.

Task Level	Model Variant	QWK	Acc	Acc +/-1	Dist	Acc 7	Acc 5	Acc 3
Document-Level	GNN Only	75.6	40.0	83.0	0.8	60.0	60.0	90.0
Document-Level	Late Fusion	**76.9**	**42.0**	82.0	0.8	60.0	61.0	90.0
Sentence-Level	GNN Only	**78.5**	**50.0**	67.2	1.3	61.2	66.1	74.9
Sentence-Level	Late Fusion	**78.5**	41.4	65.9	1.4	55.4	62.6	72.7

Table 1: Performance of the GNN-based model and Late Fusion on sentence-level and document-level readability prediction, evaluated with Quadratic Weighted Kappa (QWK), accuracy, accuracy within ±1, distribution score, and accuracy at multiple granularity levels (7, 5, and 3).

QWK (78.5%). This indicates that, at the finer sentence granularity, the graph-based model alone is more effective, while the fusion approach may dilute some of the GNN's discriminative power for exact classification.

Overall, the findings highlight that fusion offers advantages at the document level, but the GNN-only approach remains stronger for precise sentence-level readability prediction.

5 Conclusion

In this paper, we proposed a hybrid approach for Arabic document readability prediction by combining graph-based reasoning with contextual transformer-based modeling. Our architecture integrates lexical difficulty knowledge from the SAMER lexicon, sentence embeddings from a fine-tuned AraBERTv2 variant, and a structured graph representation of each document.

For sentence-level prediction, we demonstrated the benefits of lexicon-enriched heterogeneous graph modeling using a weighted GNN. For document-level prediction, we reuse the same graph setup and infer the document's label by selecting the maximum difficulty among sentence-level predictions. This design aligns with the task's objective of identifying the highest comprehension barrier within a document.

By applying late fusion between the GNN and transformer predictions, we achieved stronger performance across both levels. Our results highlight the complementary nature of structural and contextual signals and the promise of fusion-based systems for fine-grained Arabic readability tasks.

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.

Abdel-Karim Al-Tamimi, Manar Jaradat, Nuha Aljarrah, and Sahar Ghanim. 2014. Aari: Automatic arabic readability index. *International Arab Journal of Information Technology*, 11:370–378.

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.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2025. Araeyebility: Eye-tracking data for arabic text readability. *Computation*, 13(5).

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: Current state and future directions. *Procedia Computer Science*, 142:38–49. Arabic Computational Linguistics.

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.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pages 1024–1034. Curran Associates, Inc.

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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. 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 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Qi Sun, Kun Zhang, Kun Huang, Tiancheng Xu, Xun Li, and Yaodi Liu. 2023. Document-level relation extraction with two-stage dynamic graph attention networks. *Knowledge-Based Systems*, 267:110428.