# PalNLP at BAREC Shared Task 2025: Predicting Arabic Readability Using Ordinal Regression and K-Fold Ensemble Learning

**Mutaz Ayesh**
Cardiff University / Cardiff, Wales, UK
AyeshMA@cardiff.ac.uk

## Abstract

PalNLP addressed Arabic readability level prediction as a fine-grained ordinal classification problem by strictly using the Balanced Arabic Readability Evaluation Corpus (BAREC). The approach treats the 19-class ordinal classification problem as a regression task with post-hoc threshold optimization, leveraging a BERT-based model and an ensemble strategy. The system achieved a Quadratic Weighted Kappa (QWK) score of 81.1 in the blind test dataset, indicating an almost perfect agreement between the system's classifications and the true labels, and placing 18th out of 24 teams. The findings show that the model effectively learned broad readability patterns, with a competitive ±1 accuracy, but faced challenges in accurately predicting readability levels of most sentences.

## 1 Introduction

The overlap between automatic readability assessment (ARA) and other NLP tasks highlights its importance. In summarization, for example, readability frameworks and ARA may complement classic summarization metrics by evaluating the output of audience-aware or level-controlled summarization models by predicting the level of the generated summary against the original text input. Controlling summaries for readability levels can help these models generate summaries that are more suitable for their targets, as was done by Luo et al. (2022) for biomedical texts.

Similarly, Plain Language (PL) and Easy-to-Read[1] (E2R) initiatives have been gaining traction in Europe (Espinosa-Zaragoza et al., 2023; Martínez et al., 2024; Madina et al., 2024). They aim to make governmental texts more accessible for non-native speakers, people with reading limitations, and people with cognitive, intellectual, or learning disabilities. As part of the CLEARS Shared Task in IberLEF-2025 (Botella-Gil et al., 2025), Ayesh et al. (2025) attempted to transform Spanish texts in accordance with PL and E2R guidelines and used the Fernández Huerta Readability Index as one of the main metrics of evaluating the results. This index shows the importance of readability levels as an evaluation metric for a successful summary. Such alignment to reader proficiency supports better comprehension and learning outcomes (Elmadani et al., 2025b).

This task is particularly challenging for Arabic due to its morphological richness and orthographic ambiguity, and the *diglossia* that exists between Modern Standard Arabic and spoken dialects (Suwaiyan, 2018; Liberato et al., 2024; Elmadani et al., 2025b). The scarcity of large, fine-grained, and publicly available Arabic readability resources has further limited the development of robust modeling approaches. Existing resources like the word-level SAMER Lexicon (Al Khalil et al., 2020) and word- and document-level SAMER Corpus (Alhafni et al., 2024) are valuable but often domain-specific or coarse in granularity.

The new Balanced Arabic Readability Evaluation Corpus (BAREC) (Elmadani et al., 2025b) offers an opportunity to explore readability prediction with high granularity by providing over 69 thousand sentences[2] labeled across 19 readability levels, enabling modeling that captures lexical, morphological, and syntactic variation.

This paper presents the system that was submitted to the BAREC 2025 Shared Task (Elmadani et al., 2025a) for predicting BAREC readability levels, which can be summarized as a regression-then-discretization approach that is optimized for Quadratic Weighted Kappa (QWK). This formulation directly accounts for the ordinal nature of the labels and prioritizes proximity to the real level over exact level matches. The contributions of

[1] Easy-to-read is also referred to as “easy reading".

[2] *Sentence* here is used broadly to mean a standalone text.

*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 343–349
November 8-9, 2025 ©2025 Association for Computational Linguistics

this system can be summarized as follows: (1) a regression-based approach with coordinate descent threshold optimization for ordinal classification, (2) the integration of a BERT-based model with class imbalance handling and ensemble aggregation, and (3) the analysis of performance across granularities, showing strong ordinal capture but challenges in fine-grained separation.

## 2 Background

Readability assessment in Arabic has benefited from recent advances in corpus creation and lexical resource development. The Taha/Arabi21 framework (Taha-Thomure, 2017) provides a 19-level scale for educational text leveling, which BAREC adapts to the sentence level through refined annotation guidelines encompassing lexical, morphological, syntactic, and semantic features (Habash et al., 2025). Complementary resources include the SAMER readability lexicon (Al Khalil et al., 2020), which contains over 26,000 lemmas annotated with five difficulty levels by language experts from multiple Arab regions, and the SAMER reading corpus (Alhafni et al., 2024), which spans 1.4 million tokens from UAE curriculum materials and 5.6 million tokens from literary works. These resources support both lexical- and document-level readability modeling. Tools such as the word-level readability visualization add-on (Hazim et al., 2022) demonstrate practical applications in assisted text simplification and highlight the potential of integrating lexical difficulty features into automatic assessment systems.

In this shared task, participants predicted readability levels of texts from the BAREC dataset, with evaluation based on QWK. PalNLP participated in the *strict*, *sentence-level* track, meaning no additional external data was used in the development of the system alongside the sentence-level version of the BAREC dataset.

## 3 System Overview

The system addresses Arabic readability prediction as a continuous regression problem with post-hoc threshold optimization, treating the 19-class ordinal classification task through a regression-then-discretization approach optimized for Quadratic Weighted Kappa (QWK). This is due to the ordinal nature of the readability levels. The system used CAMeL-Lab's readability-arabertv2-d3tok-CE, which was used in the dataset's paper (El-madani et al., 2025b), as the foundation model. Although the model was originally fine-tuned as a classification model with cross-entropy loss, this system adapted its architecture for regression to leverage the strong readability-sensitive features learned in the CE setup while optimizing for continuous predictions. It was then combined with a threshold optimization algorithm, and later, an ensemble methodology.

| Hyperparameter | Value |
|---|---|
| Input processing | Padding to 512 tokens |
| Batch size | 16 |
| Epochs | 6 with early stopping |
| Learning rate | 2e-5 with adaptive scheduling |

Table 1: Hyperparameters used in the system. Early stopping also includes patience of 3 epochs.

The system used the sentence-level BAREC dataset, loaded from HuggingFace, without any additional data. Instead of using the default training and validation splits, these two sets were combined, and 5-fold stratified cross-validation was applied to the merged dataset. This was due to a sustained plateau in validation loss throughout the initial experiments. As a result, the system is not directly comparable to other participants' systems. The test split remained unchanged. To address class imbalance, PyTorch's WeightedRandomSampler was used with inverse class frequency weighting during training to ensure that rare readability levels were adequately represented.

## 4 Experimental Setup

The core architecture consists of a BERT-based regressor with a single continuous output head where ordinal class labels are treated as continuous values for training. MSE loss was employed with AdamW optimization, and a combination of linear warm-up and ReduceLROnPlateau scheduling based on validation QWK performance. Table 1 shows the specific hyperparameter values in the system.

Throughout the tens of experiments that were run before this final one was adopted, the systems under-performance on the validation dataset was observed despite achieving good scores in the training. A key innovation in this approach is the coordinate descent algorithm for threshold optimization; rather than using simple rounding to discretize continuous predictions, the model iteratively optimizes the thresholds associated with each class to maximize QWK on validation data through grid-based coordinate descent with multiple passes. This strategy consistently provided 1-2% improvements over naive rounding during the training.

Final results are derived by thresholding the continuous predictions into class labels. The best model of each fold gets saved and the different folds are used to predict the readability level by aggregating the predictions using each fold's threshold weights.

## 5 Results and Error Analysis

### 5.1 On the provided datasets

The results of cross-validation, found in Table 4 in Appendix A, showed consistency, with a QWK range of 79.85-80.21, indicating robust generalization.

After the training was done, and the system concluded with a QWK score of 79.66 with global thresholds, the predictions on the provided test dataset were obtained by ensembling all different folds, where predictions from the best model of each fold were combined using a weighted ensemble approach. This means that fold-specific threshold weights were applied before aggregating to final discrete readability predictions. The final QWK score on the test set was 77.7.

Table 2 summarizes the system's performance on the test dataset after ensembling. The results show that the model certainly learned the ordinal structure of BAREC well and that its misclassified labels were close to the correct level, as evidenced by the ±1 level accuracy. The model, however, struggled with exact classification. An illustration of this can be found in Figure 1.

**Impact of domain and word count.** After a curious look into the top 100 sentences with the predicted levels furthest from the true levels[3], it was apparent that those that were underestimated (i.e., the true readability levels were higher than the predicted ones) were short, with 94% of those being fewer than 5 words long. 76% of those short sentences are specialized or advanced texts; 32% are specialized and advanced texts from the Emirati curriculum, while 22% come from the Quran. Detecting the true readability level of these specific sentences might have required a model that also considers qualitative features, such as the source of the text and its class. Examples of such texts can be found in Appendix D.

A similar pattern can be seen among sentences whose readability levels were overestimated (i.e., their true readability levels were lower than the predicted ones) where 46% were 5 words long or fewer, and 68% were 7 words long or fewer. The length of these sentences might have had an impact, but the impact of the type of the text (foundational, specialized, or advanced) was not as significant, as there was somewhat an equal distribution between specialized and advanced (52%) and foundational (48%) texts. A deeper look into why the model overestimated their levels is required.

**Impact of diacritics.** Despite using an Arabic-specific BERT model, it seems that the system continuously misclassified texts with diacritics as ones with high readability levels. While the reasons behind why that happened make sense, it was not an outcome that was expected at all. The sentence with one of the greatest differences from the true readability level was a diacritized proper name[4] with no inherent difficulty. It had a readability level of 3 but was misclassified as having a readability level of 15. Another example[5] had a readability level of 8 but was classified as 15 due to the diacritics. These stark differences reflect the importance of pre-processing Arabic texts to allow the trained models to capture real features that reflect the readability levels of texts, rather than superficial ones such as diacritics that do not necessarily entail a difficult or advanced level.

After this error was detected, the test set was passed through the system to generate predictions, however, this time the diacritics were stripped using PyArabic's[6] `strip_diacritics` method beforehand. The performance on the de-diacritized test set can be found in Table 2, alongside the original scores before stripping diacritics. The new results better resemble those of PalNLP's on the blind test set, and an improvement can be seen in all metrics, especially a +3.5 improvement in the QWK score and both the exact and ±1 level accuracy scores.

Additionally, the ranges of difference between the true readability and predicted levels dropped from (-15, 12) to (-11, 8)[7]: the drop in each com-

[3] 50 sentences in each direction (positive and negative differences) were considered in this analysis.

[4] شِهابُ الدّينِ أَحْمَدُ بْنُ ماجِدٍ
Sentence ID: 10400320088

[5] عَجَبًا مِن النظَّارةِ السوداءِ لم تَحجُبِ المعنى عن الرُقباءِ
Sentence ID: 30100250057

[6] https://pypi.org/project/PyArabic/

[7] The highest negative difference is on the left, and the highest positive difference is on the right.

| Metric | Before SD | After SD |
|---|---|---|
| QWK | 77.7 | 81.25 |
| Exact Accuracy | 29.99% | 34.37% |
| ±1 Level Accuracy | 65.88% | 69.48% |
| 7-Class Accuracy | 50.96% | 54.78% |
| 5-Class Accuracy | 51.85% | 53.17% |
| 3-Class Accuracy | 66.58% | 67.94% |

Table 2: The system's performance on the test set, before and after stripping diacritics (SD).

ponent indicates reduced error bounds, reflecting fewer extreme under- and over-estimations and more tightly aligned predictions with the true readability levels. Table 5 in Appendix B further solidifies the improvement in performance; it shows a great improvement in exact predictions (+319) coupled with consistently less differences after stripping diacritics.

The heat maps in Appendix D further illustrate the improvement: after stripping diacritics, the confusion matrix becomes more diagonal, with noticeably fewer misclassifications concentrated in the upper readability levels.

### 5.2 On the blind test dataset

The system achieved 18th place out of 24 teams with an official QWK score of 81.1. The score is close to the organizers' baseline of 81.5. Table 3 contains a summary of the performance of PalNLP's system on the blind test set. Overall, the consistency between cross-validation (79.96-80.21), test set (77.7), and competition results (81.1) demonstrates the effectiveness of the validation strategy, and the system performing better in the blind test set shows that the model did not overfit on the training dataset.

It can be safely said that the ordinal structure of the BAREC dataset was effectively captured by the system, as evidenced by the much smaller gap in ±1 accuracy between the system (69.8%) and the organizers' (72.0%). This indicates that the model learned the ordinal structure well and that its misclassified labels are mostly close to the correct level. Additionally, performance gaps between PalNLP's system and the baseline decreased dramatically as classification granularity was reduced, from 25% difference in 19-class accuracy to only 4% in 3-class accuracy. This shows that the model successfully learned broad readability patterns.

The system, however, struggled with fine-grained distinctions between adjacent levels. The significant gap in exact accuracy between this system (33.1%) and the organizers' (58.1%) contrasted with the minimal QWK difference is expected as the regression framework was optimized for rank correlation rather than precise classification.

| Metric | PalNLP | Baseline |
|---|---|---|
| Avg. Absolute Distance | 1.3 | 1.0 |
| QWK | 81.1 | 81.5 |
| Exact Accuracy | 33.1% | 58.1% |
| ±1 Level Accuracy | 69.8% | 72.0% |
| 7-Class Accuracy | 57.2% | 67.7% |
| 5-Class Accuracy | 63.6% | 71.4% |
| 3-Class Accuracy | 72.5% | 76.5% |

Table 3: The system's performance on the blind test set, provided by the prediction log on CodaBench. The baseline scores were taken from the competition's leaderboard on CodaBench.

## 6 Conclusion

This paper presented a regression-then-discretization system for Arabic readability prediction on the BAREC dataset, with a focus on maximizing QWK. By modeling the task as a continuous regression problem with post-hoc threshold optimization, the results showed that the system captured the ordinal nature of readability levels in BAREC, favoring proximity to the true label over exact agreement. The BERT-based model with stratified cross-validation, class imbalance handling, and ensemble aggregation produced results that consistently generalized across validation, test, and competition evaluations.

Several key observations emerged. (1) The system broadly understood readability patterns but found fine-grained separation between adjacent levels challenging. (2) The regression formulation, combined with threshold optimization, consistently outperformed naive rounding strategies and improved alignment with the dataset's ordinal structure. (3) The error analysis highlighted systematic weaknesses, such as underestimation of short, specialized sentences, and misclassification of diacritized text as advanced-level material. (4) The consistent alignment between cross-validation, test, and blind test results prove that PalNLP's strategy was robust with minimal overfitting.

The role of pre-processing and text-specific features point toward future refinements, such as training the model *after* handling diacritics and possibly other pre-processing techniques. Further work may explore the performance of this system after integrating additional resources such SAMER that can alleviate the effect of class imbalances in BAREC.

## 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.

Mutaz Ayesh, Nicolás Gutiérrez-Rolón, and Fernando Alva-Manchego. 2025. CardiffNLP at CLEARS-2025: Prompting large language models for plain language and easy-to-read text rewriting.

Beatriz Botella-Gil, Isabel Espinosa-Zaragoza, Alba Bonet-Jover, Margot Madina, Lucas Molino Piñar, Paloma Moreda, Itziar Gonzalez-Dios, María Teresa Martín Valdivia, and Ureña. 2025. Overview of CLEARS at IberLEF 2025: Challenge for Plain Language and Easy-to-Read Adaptation for Spanish Texts. *Procesamiento del Lenguaje Natural*, 75.

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.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Paloma Moreda, and Manuel Palomar. 2023. Automatic text simplification for people with cognitive disabilities: Resource creation within the ClearText project. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2024. Towards reliable e2r texts: A proposal for standardized evaluation practices. In *Computers Helping People with Special Needs*, pages 224–231, Cham. Springer Nature Switzerland.

Paloma Martínez, Lourdes Moreno, and Alberto Ramos. 2024. Exploring large language models to generate easy to read content. *Preprint*, arXiv:2407.20046.

Laila Suwaiyan. 2018. Diglossia in the arabic language. *International Journal of Language  Linguistics*, 5.

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

## A Cross-validation scores

Table 4 presents the five-fold cross-validation results. Across folds, the optimized thresholding strategy (QWK$_{opt}$) consistently outperformed fixed rounding (QWK$round$) by about 1–2 points, confirming the benefit of post-hoc threshold optimization. Training generally converged within 3–6 epochs, with early stopping triggered in three out of five folds. These results indicate stable model performance and reduced overfitting across folds.

| Fold | QWK$_{opt}$ | QWK$_{round}$ | Epochs | ES |
|---|---|---|---|---|
| 1 | 79.85 | 78.05 | 3 | At epoch 1 |
| 2 | 80.21 | 78.82 | 5 | At epoch 3 |
| 3 | 79.96 | 78.56 | 6 | No |
| 4 | 79.96 | 78.85 | 6 | No |
| 5 | 79.90 | 78.77 | 6 | At epoch 5 |

Table 4: Cross-validation results, with 5 folds. QWK$_{opt}$ refers to the QWK score using the threshold optimization strategy detailed earlier, as opposed to the score using fixed rounding shown in QWK$_{round}$. Early stopping (ES) was included here to show when the QWK results on the (custom) validation dataset plateaued.

## B Differences between predicted and true levels in the test set, before and after SD

Table 5 shows the distribution of differences between predicted and true levels before and after stripping diacritics. The results show a reduction in large deviations (e.g., no cases at ±15 or ±12 after SD, and consistent decreases from ±11 to ±5), alongside an increase in exact matches (0 difference rose from 2185 to 2504). This indicates that SD reduces the frequency of extreme cases while improving overall alignment with the gold labels.

| Difference | F$_{beforeSD}$ | F$_{afterSD}$ |
|---|---|---|
| ±15 | 1 | 0 |
| ±12 | 1 | 0 |
| ±11 | 5 | 2 |
| ±10 | 6 | 3 |
| ±9 | 11 | 6 |
| ±8 | 27 | 18 |
| ±7 | 84 | 62 |
| ±6 | 94 | 59 |
| ±5 | 185 | 163 |
| ±4 | 344 | 306 |
| ±3 | 600 | 519 |
| ±2 | 1128 | 1086 |
| ±1 | 2615 | 2558 |
| 0 | 2185 | 2504 |

Table 5: The frequencies of differences between the predicted and true levels in the test set, before and after stripping diacritics (SD). The 0 difference in the last row is synonymous with the frequency of exact predictions made by the system.

## C Heat maps

Figures 1 and 2 show the normalized confusion matrices before and after stripping diacritics. The post-SD heat map exhibits a clearer diagonal pattern, reflecting reduced over-prediction of lower readability levels and stronger agreement between true and predicted labels.

## D Examples of extreme differences

- نُمُوُ السُكَّانِ “Population growth"
  predicted RL: 6, true RL: 14
  (ID: 20400120059)

- إِضاءَةٌ “Lighting"
  predicted RL: 3, true RL: 11
  (ID: 20400200031)

- .القانونُ “The law."
  predicted RL: 4, true RL: 12
  (ID: 20400360004),

- :أُناقشُ “I discuss"
  predicted RL: 4, true RL: 12
  (ID: 20400550017)

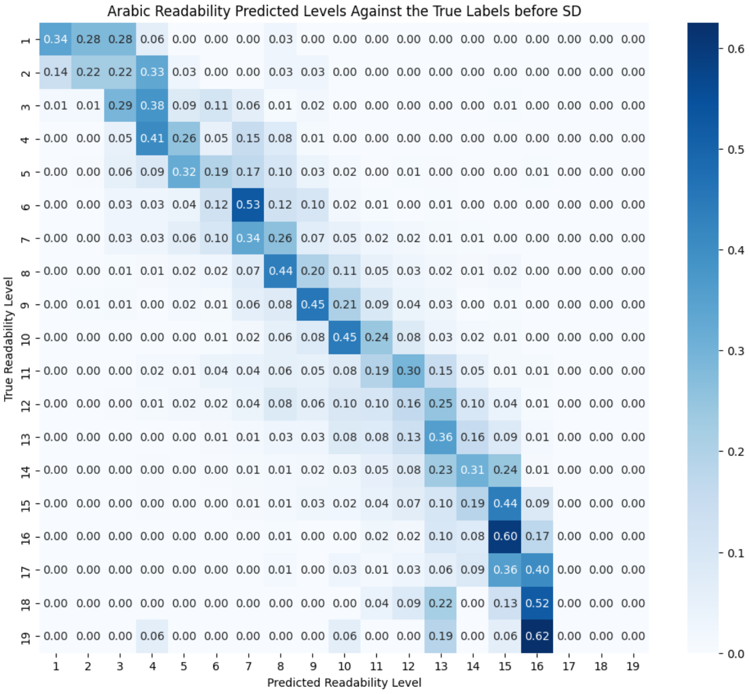

Figure 1: A normalized confusion matrix (heat map) of predicted levels against the true levels of texts in the test dataset **before** stripping diacritics.

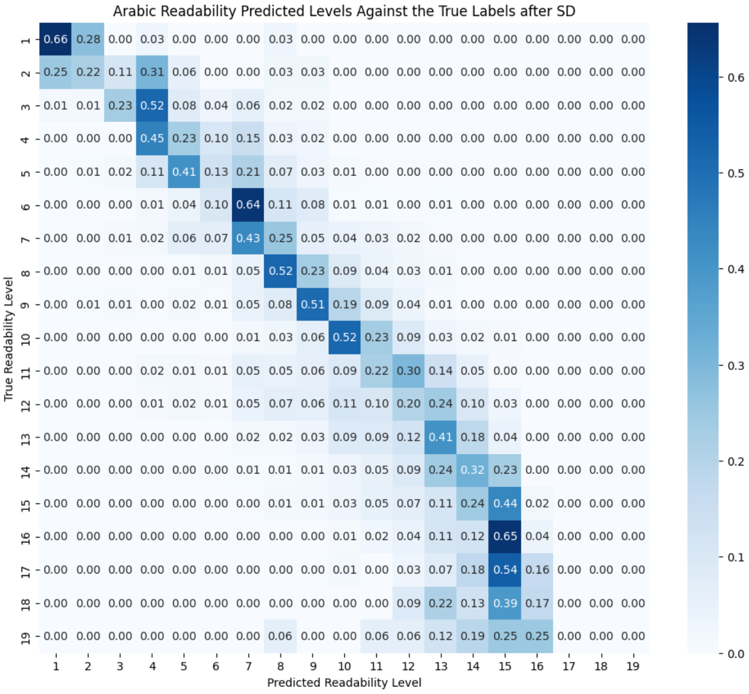

Figure 2: A normalized confusion matrix (heat map) of predicted levels against the true levels of texts in the test dataset **after** stripping diacritics.