LIS at BAREC Shared Task 2025: Multi-Scale Curriculum Learning for Arabic Sentence-Level Readability Assessment Using Pre-trained Language Models

Anya Amel Nait djoudi, Patrice Bellot, Adrian-Gabriel Chifu
Aix-Marseille Université, CNRS, LIS
{anya-amel.NAIT-DJOUDI, patrice.bellot, adrian.chifu}@univ-amu.fr

Abstract

Sentence-level readability assessment, which measures how easily individual sentences can be understood, has seen significant advances in English. However, Arabic readability assessment remains underexplored, primarily due to the language's morphological complexity and the scarcity of fine-grained annotated datasets. To address this gap, we leveraged the BAREC corpus, which provides 69K sentences annotated across 19 readability levels, enabling us to develop and compare five different modeling strategies ranging from lightweight classifiers to fine-tuned Arabic language models. Our experiments revealed that task-specific pretraining with CamelBERT yielded substantial performance gains, while curriculum learning offered benefits in specific scenarios. Ultimately, direct fine-tuning achieved state-of-the-art performance (QWK = 82.4). Through detailed error analysis, we identified that models struggled most with distinguishing between the lower readability level 2 and higher readability levels (15-19), highlighting the inherent challenges in fine-grained Arabic readability modeling across the full spectrum of proficiency levels.

1 Introduction

Readability assessment measures how easily a text can be read and understood by its target audience. It is crucial in various contexts including pedagogical settings, foreign language learning, health literacy (Djoudi et al., 2025), and content accessibility (Xia et al., 2016; Vajjala and Meurers, 2012; Collins-Thompson and Callan, 2004; Fox and Duggan, 2013). To enable automatic assessment, many resources have been made available ranging from datasets with annotated readability levels to readability assessment models. While progress in English readability assessment has been extensive (Azpiazu and Pera, 2019; Deutsch et al., 2020; Qiu et al., 2021; Devlin et al., 2019), Arabic readability assessment remains less studied. Arabic presents unique challenges due to its morphological richness, complex derivational patterns, and limited availability of fine-grained annotated resources. The available datasets predominantly employ binary (Soliman and Familiar, 2024) or ternary classification schemes (Al-Khalifa and Al-Ajlan, 2010), which train models with an oversimplified view of reading proficiency. Fine-grained readability levels offer the potential to better capture the continuous spectrum of literacy levels across diverse readers and provide more nuanced assessments that align with real-world reading abilities.

To address these limitations, we utilize the newly introduced BAREC corpus (Balanced Arabic Readability Evaluation Corpus) [1] (Elmadani et al., 2025b), a large-scale, fine-grained corpus containing over 69,000 sentences from 1,922 documents. The corpus spans 19 readability levels, from kindergarten (1) to postgraduate (19), covering diverse genres and domains (Arts & Humanities, Social Sciences, STEM) across three readership groups (Foundational, Advanced, Specialized).

Our contributions include systematic evaluation of five distinct model architectures, spanning lightweight MLP classifiers over pre-trained embeddings to full progressive and direct fine-tuning of Arabic language models. We demonstrate that:

1. task-specific pretraining is essential, with readability-focused CamelBERT substantially outperforming general-purpose models;
2. curriculum learning provides situational benefits in fine-tuning settings;
3. direct fine-tuning achieves state-of-the-art performance (QWK = 82.4);
4. comprehensive error analysis reveals difficulty in distinguishing lower readability level 2 and higher readability levels (15 to 19).

[1] BAREC corpus: https://huggingface.co/datasets/CAMeL-Lab/BAREC-Shared-Task-2025-sent

Proceedings of The Third Arabic Natural Language Processing Conference, pages 367–375

2 Background

Text difficulty evaluation traditionally used surface-level formulas like DCRS (Dale and Chall, 1948), FKGL (Kincaid et al., 1975), Dawood and El-Heeti (Al-Dawsari, 2004), AARI (Al Tamimi et al., 2014), and OSMAN (El-Haj and Rayson, 2016). The development of readability corpora enabled richer statistical and neural modeling approaches.

These corpus-driven advances have been most prominent in English, with resources including WeeBit (Vajjala and Meurers, 2012), Newsela (Xu et al., 2015), Cambridge (Xia et al., 2016), OneStopEnglish (Vajjala and Lučić, 2018) (document-level), S1131 (Štajner et al., 2017), CEFR-SP (Arase et al., 2022) (sentence-level), and cross-lingual corpora MDTE (De Clercq and Hoste, 2016), CompDS (Brunato et al., 2018). For Arabic, efforts such as (Hazim et al., 2022) have facilitated Arabic readability annotations, leading to the development of resources spanning multiple granularities: documents (Arability (Al-Khalifa and Al-Ajlan, 2010), DLI (Forsyth, 2014), Taha/Arabi21 (Taha-Thomure, 2017), ZAEBUC (Habash and Palfreyman, 2022), QAES (Bashendy et al., 2024)), sentences (README++ (Naous et al., 2024), DARES (El-Haj et al., 2024), BAREC (Elmadani et al., 2025b)), and words (KELLY (Kilgarriff et al., 2014), SAMER (Al-Khalil et al., 2020), Arabic Vocab Profile (Soliman and Familiar, 2024), extended SAMER (Alhafni et al., 2024)).

Building on these corpus development efforts, broader research has focused on Arabic readability modeling using diverse strategies (Liberato et al., 2024). To advance this field further and provide a standardized evaluation framework, the BAREC Shared Task 2025 (Elmadani et al., 2025a) introduces 19-level fine-grained readability prediction with three tracks: Strict (BAREC corpus only), Constrained (BAREC & SAMER corpora), and Open (any public data). We participate in the Strict Track for sentence-level assessment, predicting Arabic sentence difficulty on a 19-point scale (1 = easiest, 19 = hardest) using models trained exclusively on BAREC training data.

3 System Overview

We experiment with `CAMeLBERTMix_MLP` and `CAMeLBERTWCE_MLP`, which combine contextual embeddings from Arabic BERT models with a lightweight multilayer perceptron (MLP) classifier. `CAMeLBERTMix_MLP` employs bert-base-arabic-camelbert-mix [2] (Inoue et al., 2021), a general-purpose encoder trained on a mix of modern standard, dialectal, and classical Arabic, while `CAMeLBERTWCE_MLP` uses readability-camelbert-word-CE [3] (Elmadani et al., 2025b), fine-tuned on the same dataset as this shared task. Sentences are tokenized with a maximum length of 256 tokens, encoded, and aggregated via mean pooling over attention-masked hidden states to yield 768 dimensional embeddings. These embeddings are normalized using RobustScaler and fed into the MLP, which was selected via 5-fold stratified cross-validation. The best-performing configuration is a two-layer architecture (256-128 neurons), ReLU activation, L2 regularization ($\alpha = 0.01$), trained with Adam optimization (initial learning rate 10^{-4}, adaptive scheduling), batch size 64, and a maximum of 300 epochs.

`P_CAMeLBERTWCE_MLP` architecture implements a progressive multilayer perceptron (MLP) trained with a curriculum learning strategy. The model is trained sequentially through multiple stages of increasing granularity (3-5-7-19 levels), utilizing the same 768 dimensional embeddings as `CAMeLBERTWCE_MLP`. The 19-level ground truth labels are collapsed into intermediate targets using custom binning strategies to create more balanced distributions: 3-level bins `[0, 7, 13, 19]`, 5-level bins `[0, 4, 8, 12, 16, 19]`, and 7-level bins `[0, 3, 6, 9, 12, 15, 17, 19]`. Training begins with a simple 3-level classifier (2-layer architecture with 256 and 128 neurons), then progresses to 5-level and 7-level classifiers, and finishes with a 19-level classifier (3-layer architecture with 512, 256, and 128 neurons). The stage-specific architecture scales with task complexity. A weight transfer mechanism preserves learned hidden layer representations between stages. The model is optimized using Adam with an adaptive learning rate (initial $\eta = 10^{-3}$, reduced to 5×10^{-4} during transfer phases), a batch size of 64, and L2 regularization ($\alpha \in [0.008, 0.01]$) and a maximum of 100-300 epochs per stage depending on complexity.

`PFT_CAMeLBERTWCE` implements a Progressive CamelBERT (Inoue et al., 2021) Fine-tuning approach that fine-tunes the CAMeL-Lab/readability-camelbert-word-CE transformer model through curriculum learning stages `[3, 5, 7, 19]`. The ap-

[2] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix

[3] https://huggingface.co/CAMeL-Lab/readability-camelbert-word-CE

proach learns dynamic label mappings from dataset annotations rather than using fixed binning strategies like we did in P_CAMeLBERTWCE_MLP. Training begins with 3-level classification using a dropout-regularized classification head (dropout=0.3), progressively transferring the fine-tuned BERT encoder weights to subsequent stages while initializing fresh classification heads for each target granularity. Training uses AdamW optimization with linear warmup scheduling, adaptive learning rates (2e-5 initial, 1e-5 for transfer stages). The intuition is that learning coarse readability distinctions (3-5-7 levels) first provides foundational representations that will improve fine-grained (19 levels) classification performance.

FT_CAMeLBERTWCE serves as an ablation study to test the core hypothesis behind PFT_CAMeLBERTWCE. It eliminates the progressive curriculum learning stages to perform direct, end-to-end fine-tuning of the CAMeL-Lab/readability-camelbert-word-CE model on the full 19-level classification task. The motivation for this simpler approach is twofold. First, it questions whether a powerful pre-trained transformer inherently possesses the latent linguistic understanding to discern fine-grained readability distinctions without being guided through coarser labels. Second, it tests if the considerable computational and architectural overhead of multi-stage progressive training is justified, or if a single-stage model can achieve comparable performance more efficiently. To ensure a fair comparison, FT_CAMeLBERTWCE retains the same optimization strategy (AdamW, linear warmup) and regularization (dropout=0.3) as the final stage of PFT_CAMeLBERTWCE. Thus allowing us to directly attribute any performance differences to the presence or absence of the curriculum learning framework, rather than other hyperparameters.

4 Experimental Setup:

4.1 Dataset

We used the Balanced Arabic Readability Evaluation Corpus (BAREC)[4], a large-scale, fine-grained corpus containing over 69,000 sentences from 1,922 documents. As illustrated in Table 1, the corpus spans 19 readability levels, from kindergarten (1) to postgraduate (19), which can also be collapsed into coarser 7, 5, or 3 readability levels. For more details on sentence readability annotation, refer to the BAREC readability annotation guidelines by (Habash et al., 2025).

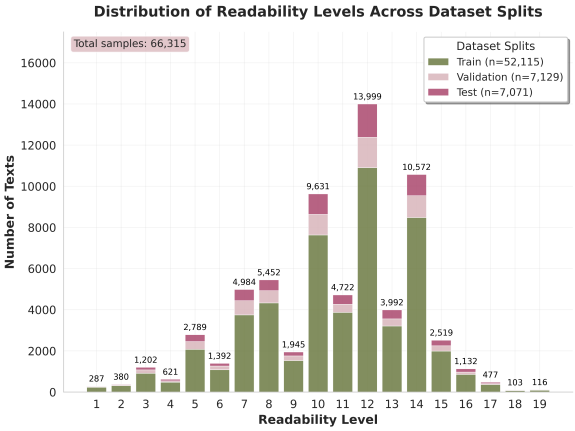

Figure 1: Distribution of the train/validation/test split in the BAREC corpus by level of readability

Level	Arabic	Translation	Reasoning
1	'نعم'	"Yes"	Simple single-word
19	'أو صورة مثلت في النفس من أملي'	"Or an image represented in the soul from my hope."	Complex poetic expression, rich in imagery

Table 1: Comparison of linguistic complexity between beginner and advanced Arabic expressions.

4.2 Preprocessing

Before training, we applied a multi-stage cleaning pipeline to ensure text consistency and quality. This involved removing sentences with missing values or placeholders (e.g., #NAME?), collapsing excessive whitespace, and dropping exact duplicates. For Arabic processing, we utilized CAMeL Tools (Obeid et al., 2020), a comprehensive suite of NLP resources for morphological analysis, disambiguation, dialect identification, normalization, and tokenization [5]. A key step was dediacritization to remove short vowels and phonetic marks witch are infrequent in modern Arabic and can introduce noise in NLP tasks. Dediacritization allowed us to focus on underlying lexical forms (e.g., كِتَابٌ جَمِيلٌ was converted to كتاب جميل, or "A beautiful book"). The pipeline removed 2.5-5% of rows, retaining 95-97% of the dataset across splits (52k train, 7.1k validation, 7k test). The final distribution of these readability levels is shown in Figure 1.

[4]BAREC corpus: https://huggingface.co/datasets/CAMeL-Lab/BAREC-Shared-Task-2025-sent

[5]CAMeL Tools:https://github.com/CAMeL-Lab/camel_tools

4.3 Metrics

The BAREC shared task [6] organizers define the readability prediction problem as an ordinal classification task and adopt the following evaluation [7] measures: **Quadratic Weighted Kappa (QWK) (Cohen, 1968):** The primary evaluation metric for the shared task. It's an extension of Cohen's Kappa that measures agreement between predicted and reference labels, applying a quadratic penalty to larger misclassifications. **Accuracy (Acc):** The percentage of exact matches between predicted and gold labels on the 19-level scale scale Acc^{19}. Variants Acc^{7}, Acc^{5}, and Acc^{3} are computed on collapsed 7-, 5-, and 3-level versions of the scale, respectively. **Adjacent Accuracy ($\pm$1 Acc^{19}):** counts predictions as correct if they are either exact matches or differ by at most one level from the true label. **Average Distance (Dist):** Also referred to as Mean Absolute Error (MAE), it captures the mean absolute difference between predicted and reference labels.

5 Results

Our team's runs (LIS in (Elmadani et al., 2025a)), were evaluated on Codabench [8]. Table 2 shows results on the blind test set, using the QWK and Acc^{19} metrics (Section 4.3). Initial experiments with MLP classifiers on embeddings demonstrated the importance of domain-specific pretraining. The general-purpose `CAMeLBERTMix_MLP` performed poorly, while `CAMeLBERTWCE_MLP` utilizing embeddings from a readability focused model demonstrated marked improvement, confirming the efficacy of task-specific pretraining. Introducing curriculum learning (`P_CAMeLBERTWCE_MLP`) provided only a marginal gain, indicating that while progressive binning can stabilize training, its impact is limited with a frozen encoder. In contrast, curriculum learning proved more beneficial in the fine-tuning setting. `PFT_CAMeLBERTWCE`, which progressively fine-tunes the encoder through increasingly granular label spaces, outperformed both MLP-based models. Finally, direct fine-tuning without curriculum (`FT_CAMeLBERTWCE`) achieved the best overall QWK (82.4) and second-best Acc^{19} (57.5), slightly surpassing the progressive strategy.

To better understand model behavior, we conducted an error analysis of the two fine-tuning approaches (`PFT_CAMeLBERTWCE` and `FT_CAMeLBERTWCE`) on the preliminary test set. We reported conditional error rates by readability level, calculated as the percentage of incorrect predictions within each readability level. Additionally, we provided a confusion matrix in Figures 3 and 4 (Appendix A) for both the best-performing model (`FT_CAMeLBERTWCE` and second-best model (`PFT_CAMeLBERTWCE`) to illustrate prediction tendencies and error patterns in greater detail. While `FT_CAMeLBERTWCE` achieved superior overall performance, the error rate analysis (Figure 2, Appendix A) reveals that `PFT_CAMeLBERTWCE` demonstrates lower error rates for specific readability levels (3-6, 9, 11, 12, 16, 17), suggesting that curriculum learning provides targeted improvements for certain readability levels despite lower aggregate performance. This level-specific analysis complemented the aggregate metrics by revealing where each model struggled most in distinguishing between readability levels, providing insights into the model's systematic biases and failure modes.

6 Conclusion

We evaluated neural approaches for Arabic sentence readability assessment, comparing MLP classifiers using CamelBERT embeddings with transformer fine-tuning methods. Task-specific pretraining proved to be essential, general embeddings failed while readability-focused ones improved performance substantially. Curriculum learning provided marginal gains with frozen encoders but helped stabilize fine-tuning. Direct CamelBERT fine-tuning (`FT_CAMeLBERTWCE`) achieved best results (QWK = 82.4, Acc = 57.5), surpassing baselines and slightly outperforming progressive fine-tuning. Our experiments highlight three key insights. First, task-specific pretraining is crucial for Arabic readability assessment, with domain-aligned representations significantly outperforming general-purpose embeddings. Second, curriculum learning offers modest but situational benefits. Third, direct fine-tuning remains both efficient and effective, achieving state-of-the-art performance without complex training strategies. Error analysis revealed systematic biases across difficulty levels, as illustrated in Figure 2 (Appendix A). Future work will explore hybrid architectures combining transform-

[6] BAREC shared task: https://barec.camel-lab.com/sharedtask2025

[7] Evaluation metrics: https://github.com/CAMeL-Lab/barec_analyzer/tree/main

[8] Codabench: https://www.codabench.org/

Model	Description	QWK	Acc^{19}
Baseline	Competition baseline	81.5	**58.1**
CAMeLBERTMix_MLP	Word Embedding + MLP	41.2	21.2
CAMeLBERTWCE_MLP	Word Embedding + MLP	80.5	55.7
P_CAMeLBERTWCE_MLP	Word Embedding + Progressive training of MLP	80.7	55.9
PFT_CAMeLBERTWCE	Progressive Fine-tuning	<u>82.0</u>	56.7
FT_CAMeLBERTWCE	Standard Fine-tuning	**82.4**	<u>57.5</u>

Table 2: Model performance on the blind test measured using Quadratic Weighted Kappa (QWK) and Accuracy (Acc^{19}). P = progressive training strategy; PFT = progressive fine-tuning; FT = standard fine-tuning; MLP = multi-layer perceptron. CAMeLBERTWEC = CAMeLBERT-Word-CE; CAMeLBERTMix = CAMeLBERT-Mix. **Bold** indicates the best result; <u>underlined</u> indicates the second-best.

ers with linguistic features and multi-agent frameworks.

Limitations

This study was limited to transformer-based approaches and word embeddings. Incorporating explicit linguistic features (such as syntactic complexity, lexical diversity, and discourse markers) could complement these neural representations and potentially improve both readability prediction accuracy and model explainability.

References

M Al-Dawsari. 2004. The assessment of readability books content (boys-girls) of the first grade of intermediate school according to readability standards. *Sultan Qaboos University, Muscat*.

Hend S Al-Khalifa and Amani A Al-Ajlan. 2010. Automatic readability measurements of the arabic text: An exploratory study. *Arabian Journal for Science and Engineering*, 35(2 C):103–124.

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.

Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. 2014. Aari: automatic arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):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.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. Qaes: First publicly-available trait-specific annotations for automated scoring of arabic essays. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 337–351.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, Giulia Venturi, and 1 others. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2690–2699. Association for Computational Linguistics.

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

Kevyn Collins-Thompson and Jamie Callan. 2004. Information retrieval for language tutoring: An overview of the reap project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–545.

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

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*, 42(3):457–490.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the*

North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186.

Anya Amel Nait Djoudi, Patrice Bellot, and Adrian-Gabriel Chifu. 2025. Bioreadnet: A transformer-driven hybrid model for target audience-aware biomedical text readability assessment. In *Proceedings of the 2025 ACM Symposium on Document Engineering*, DocEng '25, page 1–10. ACM.

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.

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.

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.

Jonathan Neil Forsyth. 2014. *Automatic readability prediction for modern standard Arabic*. Ph.D. thesis, Brigham Young University. Department of Linguistics and English Language.

Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Health*, 2013:1–55.

Nizar Habash and David Palfreyman. 2022. Zaebuc: An annotated arabic-english bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88.

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. *arXiv preprint arXiv:2210.10672*.

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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Juan Piñeros Liberato, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2024. Strategies for arabic readability modeling. *arXiv preprint arXiv:2407.03032*.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 12230.

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.

Xinying Qiu, Yuan Chen, Hanwu Chen, Jian-Yun Nie, Yuming Shen, and Dawei Lu. 2021. Learning syntactic dense embedding with correlation graph for automatic readability assessment. 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 3013–3025, Online. Association for Computational Linguistics.

Rasha Soliman and Laila Familiar. 2024. Creating a cefr arabic vocabulary profile: A frequency-based multi-dialectal approach. *Critical Multilingualism Studies*, 11(1):266–286.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th international joint conference on artificial intelligence, ijcai*, volume 17, pages 4096–4102.

Hanada Taha-Thomure. 2017. Arabic language text leveling(). *Educational Book House ()*.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

A Model performance analysis

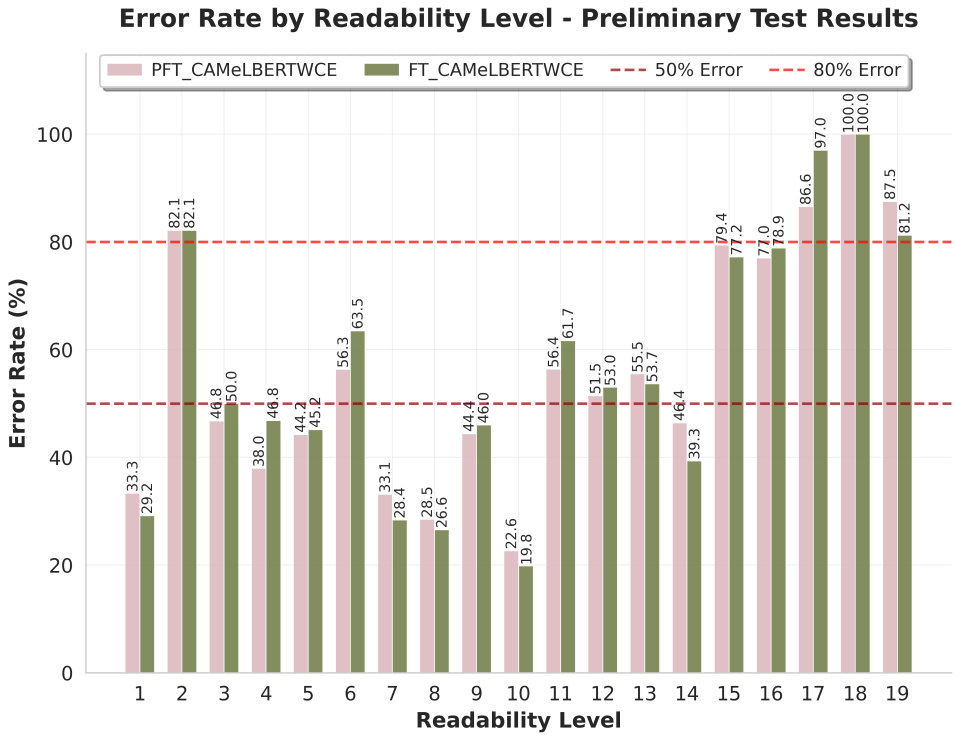

Figure 2: Error Rate by readability level on the preliminary test set

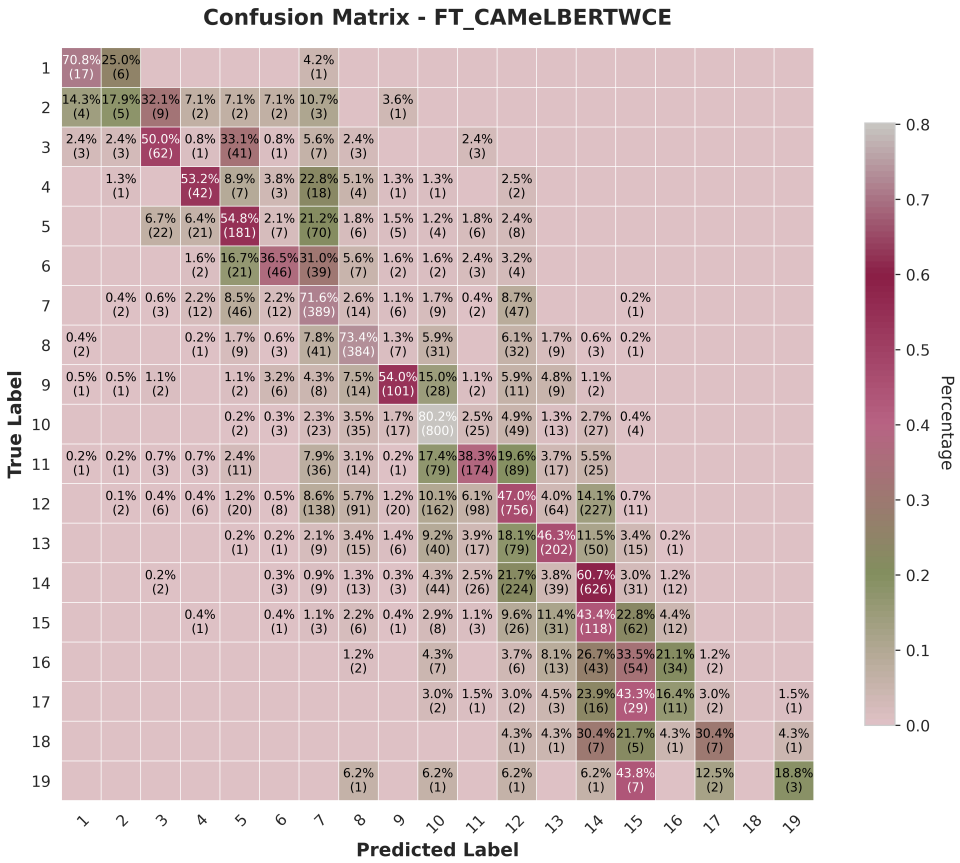

Figure 3: Prediction of FT_CAMeLBERTWCE on the preliminary test set

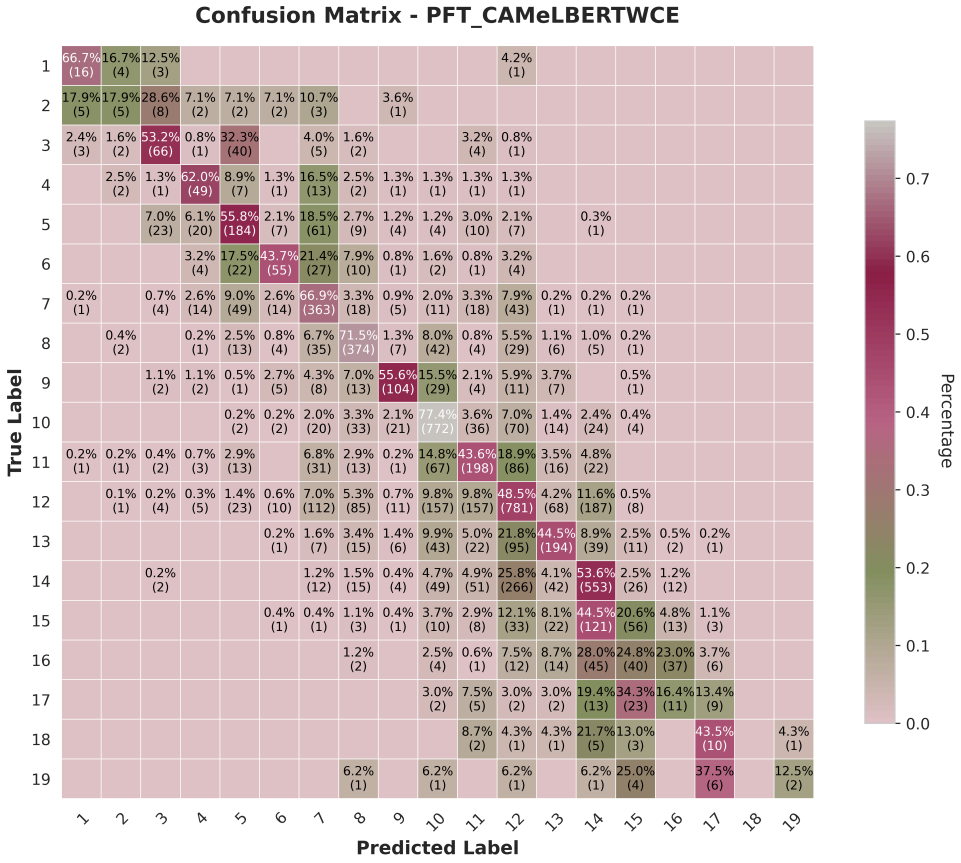

Figure 4: Prediction of PFT_CAMeLBERTWCE on the preliminary test set