DARES: Dataset for Arabic Readability Estimation of School Materials

Mo El-Haj[1], Sultan Almujaiwel[2], Damith Premasiri[1], Tharindu Ranasinghe[3], Ruslan Mitkov[1]
[1]School of Computing and Communications, Lancaster University, UK
[2]College of Humanities and Social Sciences, King Saud University, KSA
[3]School of Computer Science and Digital Technologies, Aston University, UK
{m.el-haj, d.dolamullage, r.mitkov}@lancaster.ac.uk, salmujaiwel@ksu.edu.sa, t.ranasinghe@aston.ac.uk

Abstract

This research introduces DARES, a dataset for assessing the readability of Arabic text in Saudi school materials. DARES compromise of 13,335 instances from textbooks used in 2021 and contains two subtasks; **(a)** Coarse-grained readability assessment where the text is classified into different educational levels such as primary and secondary. **(b)** Fine-grained readability assessment where the text is classified into individual grades. We fine-tuned five transformer models that support Arabic and found that `CAMeLBERTmix` performed better in all input settings. Evaluation results showed high performance for the coarse-grained readability assessment task, achieving a weighted F1 score of 0.91 and a macro F1 score of 0.79. The fine-grained task achieved a weighted F1 score of 0.68 and a macro F1 score of 0.55. These findings demonstrate the potential of our approach for advancing Arabic text readability assessment in education, with implications for future innovations in the field.

Keywords: Arabic, text readability, LLMs, NLU, Saudi school textbooks.

1. Introduction

Text readability refers to the measure of how easily a piece of text can be understood by its readers (Dale and Chall, 1949). Assessing text readability is important for both educators and learners, as it helps improve the readability levels of educational materials (Zamanian and Heydari, 2012). As a result, automatic readability assessment tools have been developed in recent years to automate the process of selecting reading materials and assessing reading ability. Furthermore, automatic readability assessment tools have proven useful in other natural language processing (NLP) applications such as machine translation (Alva-Manchego and Shardlow, 2022) and text simplification (Aluisio et al., 2010; North et al., 2022, 2023, 2024).

Earlier automatic readability assessment tools depended on classical formulas incorporating values such as average word length and average sentence length (Flesch, 1948). However, supervised machine learning (ML) methods have recently proved successful in assessing readability (Imperial, 2021; Qiu et al., 2021). ML-based methods can consider a broader range of text features than classical formulas, such as sentence complexity, vocabulary difficulty, and the cohesion and consistency of the texts. Very recently, deep learning-based ML models have helped automate feature extraction and loosen the dependence on language specificities of automatic readability assessment (Martinc et al., 2021; Imperial, 2021).

Supervised ML models that we described before typically require a training dataset to train the models. Particularly, deep learning models would require a more extensive training set as these models fine-tune thousands of parameters in the training process (Devlin et al., 2019). To address this need, the NLP community has shown significant interest in constructing readability datasets that can be used to train the ML models (Imperial, 2021). Several datasets have been developed for high-resource languages such as English (Xia et al., 2016), Spanish (Morato et al., 2021), German (Naderi et al., 2019) and Portuguese (Leal et al., 2018). For Arabic also, there exist several datasets and methods which aim to develop readability estimation applications (Baazeem et al., 2021; Berrichi et al., 2022).

In this research, we revisit the task of Arabic readability assessment in school textbooks. While there exist several datasets for readability assessment in Arabic, we argue that these datasets have limited practical relevance in real-world scenarios. For example, Al Khalil et al. (2018) introduced a large corpus consisting of texts randomly selected from the school textbooks of the United Arab Emirates and trained different ML models to predict the grade given a text. While this approach can produce a large number of training instances, the texts do not contain information about which concepts a particular text is trying to describe in the textbook. Therefore, with such a corpus, it is challenging to discern whether a certain description of a concept is readable and consequently understandable for a given grade level, limiting its practical relevance. In this research, we address this limitation by introducing DARES, which diverges from the practice of

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 103–113

randomly collecting text from school textbooks. Instead, DARES only consists of texts that describe certain concepts. As far as we know, this is the first readability dataset that contains information about concepts. We also introduce novel neural network architectures that incorporate concepts in the readability measure.

The main contributions of this research are;

1. We introduce DARES: A dataset for Arabic readability estimation based on Saudi school material. DARES has two subtasks; **(a)** Coarse-grained readability assessment where the text is classified into different educational levels such as primary and secondary. **(b)** Fine-grained readability assessment where the text is classified into individual grades.

2. We trained multiple transformer models on both subtasks of DARES that support Arabic with different input settings and evaluated the results. We also conducted a detailed error analysis.

3. We released DARES[1], as an open-access dataset alongside the trained machine-learning models.

2. Related Work

Text readability assessment has been an active area of research across various languages for the past decade, with initial methods proposing metric formulas based on factors like sentence length and word syllable count (Crossley et al., 2011; Pitler and Nenkova, 2008). Subsequently, machine learning approaches emerged, leveraging features extracted from the text at different levels, such as words, phrases, and sentences (François and Miltsakaki, 2012). The advent of Transformer models, particularly those stemming from the BERT architecture, in the last five years revolutionised the field by employing self-attention mechanisms to grasp word context, thereby advancing the state-of-the-art in various NLU tasks (Devlin et al., 2019). Despite advancements, the development of more sophisticated techniques and language models tailored for Arabic NLU is ongoing, necessitating greater attention to custom data to accommodate the diversity of Arabic text-level readability (El-Haj and Rayson, 2016).

However, it is still not as efficient as the state-of-the-art models built for English (El-Haj et al., 2018). The work of (Tanaka-Ishii et al., 2010) sorted the readability using SVM with insufficient training data. François (2015) conducted a study on the intersection of readability and computational linguistics, applying NLP-based historical readability research. That same year, (Saddiki et al., 2015) researched Arabic as a Foreign Language using a public corpus and NLP techniques. The focus on Arabic continued with (Alotaibi et al., 2016) work on the readability of medicine leaflets and (Malik et al., 2019; El-Haj et al., 2018) introduction of an Arabic-specific readability assessment. The experiments on readability assessment in Arabic have been growing, with a number of studies published in recent years and reviewed by some studies (Cavalli-Sforza et al., 2018; Nassiri et al., 2023; El-Haj and Rayson, 2016; Bessou and Chenni, 2021; Khallaf and Sharoff, 2021). Al Khalil et al. (2018) describe a reading corpus in Modern Standard Arabic where the authors select random texts for each grade to compile a corpus.

Previously, readability assessments have been conducted using various approaches. (Bessou and Chenni, 2021; Saddiki et al., 2015; Khallaf and Sharoff, 2021) categorised documents into different readability levels, ranging from 'easy' to 'very difficult'. The study by (Vajjala, 2022) addressed the scarcity of resources for readability assessment across languages, including Arabic (Vajjala, 2022; Vajjala and Lučić, 2018). (Cavalli-Sforza et al., 2018) emphasised the need for more tools and resources in Arabic readability research. Additionally, (Dalvean and Enkhbayar, 2018) proposed a new readability measure for fiction texts, while (Al Khalil et al., 2018) introduced a levelled reading corpus for Arabic text readability estimation based on the UAE curriculum and fiction. (Malik et al., 2019) highlighted the necessity for improved Arabic readability tools in patient educational materials, and (Benzahra and Yvon, 2019) examined readability and comprehension in journalistic texts.

Machine learning techniques have also been applied in Arabic text classification. (Bessou and Chenni, 2021) explored this area, while (Khallaf and Sharoff, 2021) utilised Arabic-BERT and XLM-R for Arabic sentence difficulty classification. Furthermore, (Vajjala, 2022) provided a comprehensive review of readability assessment trends, focusing on traditional readability formulas.

In 2023, significant advancements were made. (Nassiri et al., 2023) delved into Arabic readability approaches, while (Crossley et al., 2023; Vajjala, 2022) investigated the use of transformers for readability assessment and highlighted open challenges in the field, respectively. Finally, (Hazim et al., 2023) introduced a practical application: a Google Docs add-on for Arabic readability, featuring lemmatisation and a readability lexicon.

Our approach diverges from prior research. We emphasise the extraction of texts based on con-

[1] https://github.com/DamithDR/arabic-readability-assessment

cepts (a specific word accompanied by descriptive text that explicates its meaning), a departure from traditional methods as it enables us to gauge readability in relation to specific concepts and assess comprehension levels across different grade levels, a capability lacking in previous studies, e.g. (El-Haj and Rayson, 2016).

3. DARES Dataset

The DARES dataset is sourced from the books from the Saudi Education school system. The dataset includes schoolbooks from grades 1 to 12, aligning with the educational framework set by the Ministry of Education in Saudi Arabia[2]. This dataset is derived from the new literacy plan introduced in 2021 by the Saudi Ministry of Education, incorporating the latest educational content updates for students across these grades. The curriculum covers a wide range of subjects, including religious and social studies, languages, sciences, technology, physical education, life skills, activity classes, and artistic pursuits.

3.1. Dataset Preparation

We first selected 307 books authored in Arabic for the 1-12 grades in Saudi schools for 116 subjects. Out of them, 48 were from the early elementary level (Grades 1-3), 62 were taken from the upper elementary level (Grades 4-6), 86 were from the intermediate level (Grades 7-9), and 111 were from the high school level (Grades 10-12). Some schoolbooks are published in English, and we did not include them in this research. The statistics about subjects and number of books are shown in Table 1.

Grade	Books	Words	Subjects
1	18	64,590	7
2	15	71,594	5
3	15	104,357	5
4	21	294,704	7
5	23	387,750	7
6	18	337,551	7
7	28	619,777	8
8	24	488,841	8
9	34	885,880	11
10	65	2,106,350	26
11	33	1,237,985	16
12	13	572,478	10
Total	307	7,171,857	116

Table 1: Dataset Statistics for each tier and grade with respect to number of books, words and subjects.

[2] https://moe.gov.sa/

Subject	Books	Words
AI	1	53,314
Arabic Language	67	645,855
Artistic Education	21	516,448
Arts	1	19,208
Athletics	1	44,164
Biology	7	519,878
Business	2	101,787
Chemistry	9	407,466
Computer Science	4	105,997
Critical Thinking	7	110,260
Data Science	1	32,843
Decision Making	1	113,728
Digital Skills	15	570,145
Ecology	3	95,797
Economics	1	27,013
Finance	4	93,315
Geography	3	64,671
Geology	1	60,799
Hadith	1	14,909
Health	3	112,699
History	3	68,976
IoT	2	44,966
Islamic Studies	37	574,684
Law	1	23837
Life and Family Skills	23	290,802
Life Skills	4	49,303
Management	1	101,516
Math	5	137,955
Physics	7	551,156
Professional Skills	2	25,711
Psychology	1	46,683
Quran Sciences	1	23,648
Research Skills	5	139,526
Science	28	637,783
Sociology	18	538,794
Software Eng	1	27,870
Tech	6	178,341
Total	307	7,171,857

Table 2: Dataset statistics for each subject with respect to number of books and words.

3.2. PDF to Text Conversion, OCR Processing, and Post-Editing

As the first step, we converted the original educational materials, provided in PDF format, into plain text files. We utilised tools specifically designed for PDF-to-Text conversion. In order to handle instances where the text was embedded within images, we used the open-source Arabic-trained OCR from Tesseract OCR[3]. Table 2 lists the names of the subjects, the number of textbooks, and the count of running tokens in each.

The process of extracting accurate texts proved

[3] https://github.com/tesseract-ocr/tessdata

to be less efficient than anticipated due to the variety of Arabic fonts used in the PDF files, such as AXtManal, GESSTwoLight, Helvetica, and Lotus. These fonts introduced an added complexity for the OCR. Therefore, the text obtained through OCR and subsequent conversion underwent a post-editing phase. This step was conducted by an Arabic language linguist (also a co-author of this paper) who meticulously reviewed and refined the dataset, ensuring that the 13,335 extracted key words, along with their corresponding texts, were accurately represented. This process guaranteed both syntactic accuracy and semantic coherence within the dataset, which was derived from the 307 textbooks.

3.3. Text Pre-processing

As the first pre-processing step, we used sentence segmentation to divide the text into discrete sentences. We also used the Arabic tokenisation framework[4] to perform text tokenisation and Part-of-Speech (POS) tagging.

As we mentioned before, the DARES dataset focused only on the sentences that describe concepts. Therefore, we selected sentences beginning with a 'DET NOUN' POS tag and grouped them by grade level, focusing specifically on sentences that start with the Arabic definite article ' ' at the beginning of texts in the post-processed dataset. This technique was employed because words starting with ' ' are often keywords that are defined and explicated in the curriculum. Subsequently, we carefully reviewed the extracted words, along with their corresponding texts and subjects, and removed instances where the context did not serve to define the concepts of the words. This refinement process ensured that our dataset was not only accurately tagged but also contextually coherent and relevant to the concepts and subjects under consideration. The final dataset had 13,335 instances describing concepts. Several samples of the dataset is available on Table 3.

3.4. Tasks

In the DARES dataset, we used a hierarchical labelling schema that contains two tasks, which we describe below.

(I) Coarse-grained readability assessment In this task, we grouped the grades into four levels: early elementary level (Grades 1-3), upper elementary level (Grades 4-6), intermediate level (Grades 7-9), and high school level (Grades 10-12) aligning with the Saudi school's system and used them as the labels. Figure 1 shows the number of concepts and the token distribution of each level.

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

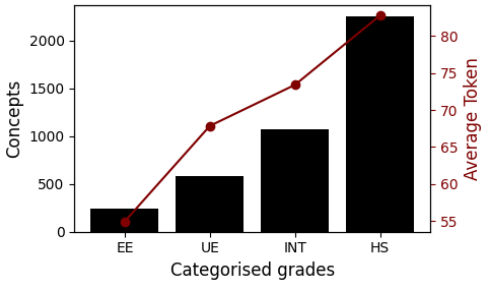


Figure 1: Concept and token distribution for the Coarse-grained level in DARES dataset. The labels are early elementary (EE), upper elementary (UE), intermediate (INT), and high school (HS).

(II) Fine-grained readability assessment For this task, we employed the original grades as the labels, resulting in a total of 12 distinct labels. Figure 2 shows the number of concepts and the token distribution of each level.

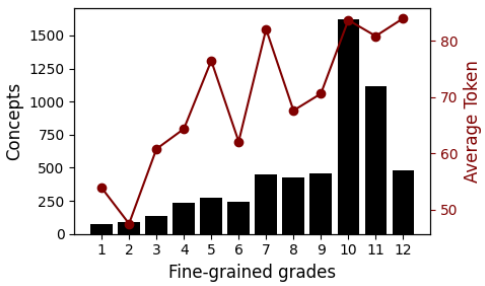


Figure 2: Concept and token distribution for the Coarse-grained level in DARES dataset.

Our methodology is based on neural transformers, which have provided state-of-the-art results in many NLP tasks, including readability assessment. We experimented with several transformer models that support Arabic; `XLM-R Large` (Conneau et al., 2020), `mBERT` (Devlin et al., 2019), `AraELECTRA` (Antoun et al., 2021), `AraBERTv2` (Antoun et al., 2020) and `CAMeLBERTmix` (Inoue et al., 2021). These models have performed well in different Arabic NLP tasks (Premasiri et al., 2022).

With each transformer model, we experimented with three input settings.

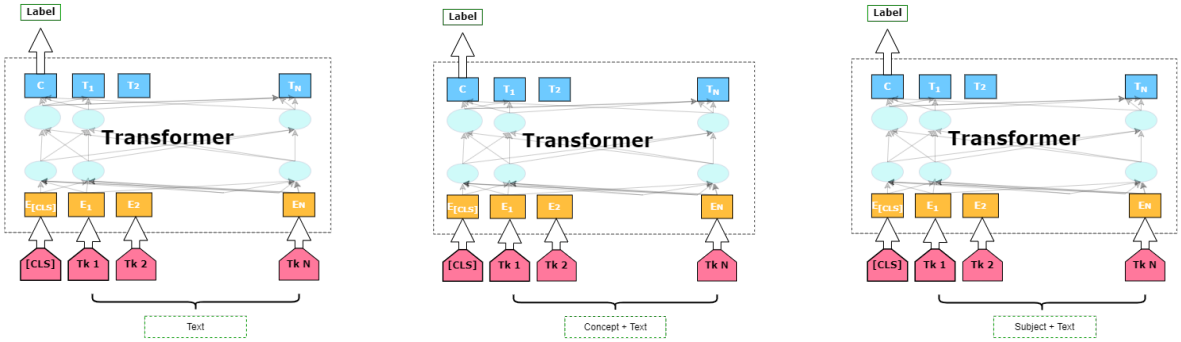
1. **text**; where we only feed the text as the input to the transformer model.

2. **concept + text**; where we concatenate the concept to the text and provide as the input to the transformer model.

3. **subject + text**; where we concatenate the subject to the text and provide as the input to the transformer model.

Subject	Concept	Arabic Text	Label(s) CG	FG
الأحياء (Biology)	الغذاء (Food)	الغذاء من الطاقة ، وهو كمية الحرارة اللازمة لرفع درجة العضلات القلبية عضلات لا إرادية حرارة الماء درجة سيليزية واحدة (Food is a form of energy, which is the amount of heat needed to raise the temperature of the involuntary cardiac muscles by one Celsius degree of water heat.)	HS	G10
العلوم (Science)	الخلية (Cell)	الخلية المجهرية تتكون جميع المخلوقات الحية من خلايا ، انظر الشكل وتعد البكتيريا أصغر المخلوقات الحية . ويتكون جسمها من خلية واحدة فقط (All living creatures are composed of microscopic cells, see the figure. Bacteria are the smallest of living organisms and consist of only one cell.)	INT	G7
العلوم (Science)	البذرة (Seed)	البذرة جزء النبات الذي ينمو ليعطي نباتا جديدا . البذرة داخل ثمرة الخوخ يمكن أن تنمو فتصير شجرة خوخ (A seed is a part of the plant that grows to produce a new plant. The seed inside a peach fruit can grow into a peach tree.)	EE	G1

Table 3: Example data instances. The column Subject represents the relevant subject the text was extracted from, and the column Concept indicates the sub-area in the subject which the text was extracted from while Text shows the extracted text. CG shows the course-grained label. The labels are early elementary (EE), upper elementary (UE), intermediate (INT), and high school (HS). FG shows the fine-grained label to the text. English translations are in green.



(a) Text (b) Concept + Text (c) Subject + Text

Figure 3: The input setting used for experiments

From an input sentence, transformers compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we build a classifier for the task. For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \text{softmax}(W\boldsymbol{h} + b)$, where $W \in \mathbb{R}^{k \times d}$ is the softmax weight matrix, and k is the number of labels.

For all the experiments, we used a batch size of eight, Adam optimiser with learning rate $2\text{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the softmax layer, were updated. All the models were trained for five epochs.

4. Results and Evaluation

We evaluated all of our models and their variations in both tasks in DARES separately. We first divided the dataset into training sets (70%), testing sets (20%) and validation sets (10%). We trained the model on the training set and fine-tuned it on the validation set. Finally, we evaluated the performance on the testing set. For both subtasks, we used Macro F1 and Weighted F1 as the evaluation metrics to compare different models. We ran each experiment five times with five different random seeds and reported the mean. We also report the standard deviation.

4.1. Coarse-grained Readability Assessment

Table 4 shows the results for coarse-grained readability assessment. The CAMeLBERTmix model with the 'Subject+Text' input setting provided the best result, achieving a Weighted F1 score of 0.91 and a Macro F1 score of 0.79. AraELECTRA and AraBERTv2 with the same input setting followed closely to the best result, providing 0.89 Weighted F1 scores.

Input Setting	Model Name	Weighted F1	Macro F1
Text	XLM-R Large	0.53±0.13	0.32±0.15
	mBERT	0.66±0.17	0.47±0.21
	AraELECTRA	0.82±0.01	0.69±0.01
	AraBERTv2	0.81±0.00	0.70±0.01
	CAMeLBERTmix	0.84±0.00*	0.74±0.01*
Concept + Text	XLM-R Large	0.56±0.15	0.36±0.18
	mBERT	0.70±0.14	0.52±0.17
	AraELECTRA	0.82±0.00	0.70±0.01
	AraBERTv2	0.74±0.16	0.59±0.21
	CAMeLBERTmix	0.84±0.00*	0.75±0.01*
Subject + Text	XLM-R Large	0.80±0.02	0.59±0.04
	mBERT	0.85±0.03	0.65±0.06
	AraELECTRA	0.89±0.01	0.72±0.05
	AraBERTv2	0.89±0.00	0.75±0.01
	CAMeLBERTmix	**0.91±0.00***	**0.79±0.01***

Table 4: Test set results for coarse-grained readability assessment. We report **Weighted F1** and **Macro F1** for all the models and input settings. The best result from all the experiments are highlighted in **bold**.

It is also noticeable that the multilingual models such as XLM-R Large and mBERT are outperformed by Arabic specific transformer models such as AraBERTv2, AraELECTRA and CAMeLBERTmix in all the input settings. This highlights the effectiveness of language-specific transformer models in readability assessment tasks.

Overall, the 'Subject+Text' setting improved the results of all the transformer results. However, it should be noted that the 'Text' setting also provides close results, especially for Arabic-specific transformer models.

4.2. Fine-grained Readability Assessment

Table 5 presents the results for coarse-grained readability assessment. As shown in the results, the 'Subject+Text' settings with the CAMeLBERTmix model also provided the best results for the fine-grained readability assessment task, achieving a Weighted F1 score of 0.68 and a Macro F1 score of 0.55. Similar to the previous task, all the models demonstrated high performance in the 'Subject+Text' setting. Furthermore, Arabic-specific transformer models produced superior results than the multilingual transformer models.

Input Setting	Model Name	Weighted F1	Macro F1
Text	XLM-R Large	0.29 ±0.12	0.15 ±0.10
	mBERT	0.51 ±0.06	0.37 ±0.06
	AraELECTRA	0.56 ±0.01	0.42 ±0.01
	AraBERTv2	0.40 ±0.20	0.28 ±0.20
	CAMeLBERTmix	0.59 ±0.01	0.49 ±0.01
Concept + Text	XLM-R Large	0.25 ±0.13	0.12 ±0.11
	mBERT	0.53 ±0.02	0.39 ±0.03
	AraELECTRA	0.56 ±0.01	0.41 ±0.02
	AraBERTv2	0.56 ±0.01	0.44 ±0.01
	CAMeLBERTmix	0.60 ±0.01	0.51 ±0.01
Subject + Text	XLM-R Large	0.51 ±0.02	0.30 ±0.03
	mBERT	0.59 ±0.02	0.41 ±0.04
	AraELECTRA	0.63 ±0.00	0.44 ±0.01
	AraBERTv2	0.61 ±0.02	0.44 ±0.02
	CAMeLBERTmix	**0.68 ±0.00**	**0.55 ±0.01**

Table 5: Test set results for fine-grained readability assessment. We report **Weighted F1** and **Macro F1** for all the models and input settings. The best result for each input setting is marked as *, and the best result from all the experiments are highlighted in **bold**.

It should also be noted that the F1 scores for the fine-grained task are lower than the coarse-grained task. However, this is expected since the fine-grained task has more classes compared to the coarse-grained task.

5. Error Analysis

In this section, we provide a detailed error analysis of the two tasks. For the error analysis, we only use the best model and the input setting from the previous section, CAMeLBERTmix, with the 'Subject+Text' setting. The error analysis is conducted with the confusion matrix and the misclassified instances in the test set.

5.1. Coarse-grained Readability Assessment

Figure 4 illustrates the confusion matrix for coarse-grained readability assessment. Overall, the testing dataset comprises 2681 instances, among which only 252 were misclassified, indicating a relatively low error rate.

As shown in Figure 4, notable misclassifications happen between close levels such as UE and EE, where 36 UE texts were occasionally mistaken as EE. However, misclassification between distant levels such as EE and HS, are very rare.

In the following list, we show some misclassified instances with their translations in the coarse-grained task.

1. True label: EE, Predicted label: UE

Sample texts:

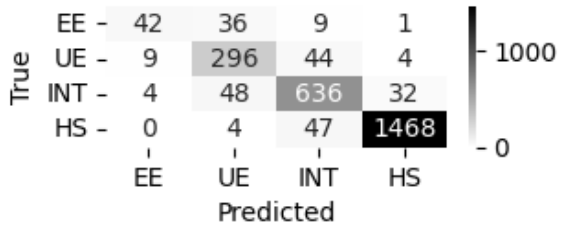


Figure 4: Confusion matrix for coarse-grained text readability estimation. The labels are early elementary (EE), upper elementary (UE), intermediate (INT), and high school (HS).

المهارات الرقمية : القواعد التي عليك اتباعها أثناء استخدام وسائل التواصل الاجتماعي يجب ألا تشارك المعلومات الشخصية مطلقا مع الأشخاص الذين تتعرف عليهم عبر الإنترنت ، ويشمل ذلك اسمك وعنوانك ورقم هاتفك ، وكذلك بريدك الإلكتروني وكلمات المرور .

Translation: Digital skills: The rules you must follow while using social media include never sharing your personal information with people you meet online. This includes your name, address, phone number, as well as your email and passwords.

2. True label: HS, Predicted label: INT

Sample texts:

الرياضيات : الاهتمام بالمهارات الرياضية ، والتي تعمل على ترابط المحتوى الرياضي وتجعل منه كلا متكاملا ومن بينها مهارات التواصل الرياضي ، ومهارات الحس الرياضي ، ومهارات جمع البيانات وتنظيمها وتفسيرها ، ومهارات التفكير العليا .

Translation: Mathematics: It is important to pay attention to mathematical skills, which interconnect mathematical content, making it an integrated whole. These skills include mathematical communication, sense of maths, data collection, organisation and interpretation skills, and higher-order thinking skills.

3. True label: UE, Predicted label: HS

Sample texts:

اللغة العربية : الترادف هو ما اختلف لفظه واتفق معناه ، أو هو إطلاق عدة كلمات على مدلول واحد ، كالأسد والليث وأسامة التي تعني مسمى واحدا ، والحسام والسيف والمهند معنى .

Translation: Arabic Language: Synonymy is when different words have the same meaning, or when several words refer to the same signified thing, such as "أسد", "ليث", and "أسامة" which all mean 'lion, and "حسام", "سيف", "مهند" which carry the same meaning for 'sword'.

4. True label: UE, Predicted label: INT

Sample texts:

العلوم : البلاستيدات الخضراء ، وهي مملوءة بمادة خضراء تسمى الميتوكندريا يحرق الغذاء في هذا الجزء . أما الخلية الحيوانية فلا تحتوي على البلاستيدات أو الكلوروفيل . الخلايا النباتية لها جدار خلوي هناك جدار صلب يحيط بالخلية النباتية يسمى الجدار الخلوي ، يعطيها شكلا يشبه الصندوق . أما الخلايا البلاستيدات الخضراء تعد مصانع الغذاء في الخلية ، وتحتوي على مادة الكلوروفيل .

Translation: Science: Green plastids are filled with a green substance called mitochondria that burns food in this part. As for animal cells, they do not contain plastids or chlorophyll. Plant cells have a cell wall, there is a hard wall surrounding the plant cell called the cell wall, which gives it a box-like shape. As for the green plastids, they are the food factories in the cell and contain chlorophyll.

5. True label: HS, Predicted label: UE

Sample texts:

المهارات الرقمية : الإنترنت شبكة عالمية تتيح لأي حاسب متصل بها الاتصال بالحاسبات الأخرى . تقدم خدمات منها الشبكة العنكبوتية العالمية تعد أحد خدمات الإنترنت وهي نظام من المستندات المترابطة تسمى صفحات الويب ويمكن لكل صفحة ويب الارتباط بواحدة أو أكثر من الصفحات الأخرى . للوصول إلى صفحات الويب نستخدم برامج تسمى متصفحات الويب تتيح لنا تصفح هذه الصفحات والضغط على الروابط للانتقال إلى صفحات أخرى . تسمى هذه الروابط ارتباطات تشعبية . تعد كل صفحة ويب فريدة ويمكن التعرف عليها من خلال عنوان يسمى محدد مواقع الويب . لاحظ أن العنوان هنا يحتوي على اسم المضيف . بالإضافة إلى معلومات أخرى تستخدم للوصول إلى مستند معين لدى مضيف محدد .

Translation: Digital Skills: The internet is a global network that allows any computer connected to it to communicate with other computers. It offers services, one of which is the World Wide Web, a system of interlinked documents called web pages. Each web page can link to one or more other pages. To access web pages, we use programs called web browsers that allow us to browse these pages and click on links to go to other pages. These links are called hyperlinks. Each web page is unique and can be identified by an address called a URL. Note that the address here contains the host name, as well as other information used to access a specific document on a specific host.

6. True label: INT, Predicted label: UE

Sample texts:

المهارات الرقمية : البحث عن مجلد أو ملف عندما يكون لديك الكثير من الملفات على جهاز الحاسب الخاص بك ، فمن الطبيعي أن تنسى المكان الذي حفظتها فيه ، لذلك إذا كنت بحاجة إلى ملف ، فيمكنك البحث عنه .

Translation: Digital Skills: When you have many files on your computer, it is normal that you might forget where you saved them. Therefore, if you need a file, you can search for it.

7. True label: INT, Predicted label: HS

Sample texts:

علوم الحاسب : أشكال اللبنات : القبعات بدء المقاطع البرمجية واقتناص الأحداث . اللبنات القابلة للتكديس تكوين الخطوات البرمجية عبر صفها (تكديسها) مع بعضها . الكل حاوية للبنات الأخرى لتطبيق التأثير (تكرار ، تحقق) على محتوياتها من اللبنات . الشروط : تعيد قيم

منطقية (صواب / خطأ) يمكن استخدامها في كل الاختيار والتكرار . القيم: الحصول على البيانات بعد إجراء العمليات عليها . مثلاً : ضم سلسلتين من النصوص . توليد رقم عشوائي ، مدخلات المستخدم بعد إجابته على سؤال ما ، إلخ .

Translation: Computer Science: Types of building blocks: Start blocks for software pieces and capturing events. Stackable blocks compile the programming steps by lining them up (stacking) together. Container blocks apply effects (repeat, check) on their contained blocks. Conditions: Return logical values (true / false) that can be used in choice and repetition blocks. Values: Obtain data after performing operations on it. For example: concatenating two strings, generating a random number, user inputs after answering a question, etc.

Misclassifications naturally occur for texts that lie on the boundary between the later stages of EE and the early stages of UE within individual subjects. This is evident in cases 1, 2, 4, and 6 for 'Digital Skills', and in case 7 for 'Computer Science'.

5.2. Fine-grained Readability Assessment

Figure 4 illustrates the confusion matrix for coarse-grained readability assessment. Among the 2681 test instances, 871 instances were misclassified, which is higher than the coarse-grained. According to the confusion matrix, majority of the misclassification occur between the close grades which also belonged to the same label in the coarse-grained level. For example, 146 Grade 10 instances were misclassified as Grade 11 and 133 Grade 10 instances were misclassified as Grade 11. This illustrates that the model may struggle with distinguishing these grades. Furthermore, misclassifications are higher among the Grade 10,11 and 12, suggesting that better models should be deployed when assessing readability in these grades.

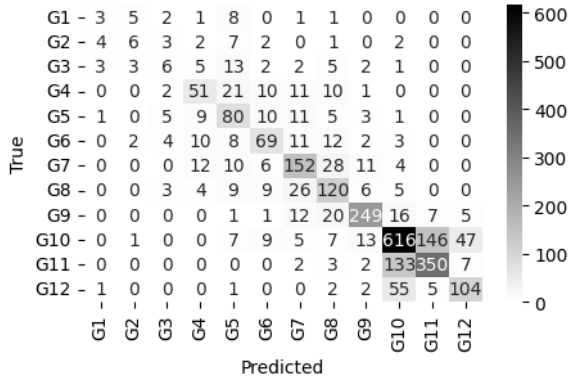


Figure 5: Confusion matrix for fine-grained text readability estimation. The labels are the different grades

6. Conclusions and Future Work

In this paper, we introduced DARES, a dataset for Arabic readability estimation based on Saudi school material. DARES has two subtasks; 1. Coarse-grained readability assessment where the text is classified into different educational levels such as primary and secondary. 2. Fine-grained readability assessment where the text is classified into individual grades.. To the best of our knowledge, DARES is the first readability assessment dataset based on concepts. We trained several transformer models that support Arabic under different input settings. The results showed that `CAMeLBERTmix` model provided the best results in both subtasks under the 'Subject + Text' setting. Furthermore, the results showed that multilingual models do not show competitive results compared to the Arabic specific models. In terms of error analysis, the majority of errors in the coarse-grained set were found in the 'Science' subject, followed by 'Arabic Language', 'Artistic Education', and 'Islamic Studies'. The fine-grained set also showed the highest number of errors in the same subjects, except for 'Artistic Education', with 'Chemistry' and 'Physics' adding to the error count as well.

The outcomes of this research hold significant implications for Arabic language education. DARES dataset can be used to The proposed readability assessment models offer educators a reliable means to prepare appropriate reading materials, enhancing the learning experience. Our research addresses the challenge of making complex concepts accessible to a wider range of students.

In future work, we hope to extend the dataset into more concepts and involve more school material. We would also like to incorporate large language models particularly trained in Arabic, such as Jais (Sengupta et al., 2023) in our methods as they have shown state-of-the-art results in many NLP tasks. Finally, we would like to develop a text summarisation pipeline for Arabic, which will have the capability to summarise the text, which has a high readability for a particular grade.

Limitations

While this study aims to advance Arabic text readability understanding, we have identified the following limitations.

1. Limited dataset size - We accept that DARES only has 13335 instances and is limited in size compared to other readability datasets. However, as we explained before, this is due to the unique nature of the way we collected DARES focusing on concepts.

2. Involvement of other readability datasets - As a

language resources paper, we did not focus on techniques such as transfer learning from other readability datasets that could have improved the results. In this paper, we focus more on the dataset collection.

3. Involvement of large language models - As we mentioned before, we did not experiment with any large language model. The models we experimented will serve as a baseline for the dataset.

Ethical Considerations

This research adheres to strict ethical standards throughout the data collection, analysis, and interpretation processes. We have taken careful measures to ensure compliance with ethical guidelines regarding educational materials, including copyright and intellectual property rights. It is important to note that the curriculum used in this research is not distributed or reused; rather, it is processed and produced solely as a training dataset for research purposes. This approach aligns with the policies outlined by the Saudi Authority for Intellectual Property and ensures the responsible use of educational materials. Additionally, our data preparation procedures prioritise transparency, integrity, spell-checking, and expert review to maintain accuracy and fidelity in our research outcomes.

References

Muhamed Al Khalil, Hind Saddiki, Nizar Habash, and Latifa Alfalasi. 2018. A leveled reading corpus of Modern Standard Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sihaam Alotaibi, Maha Alyahya, Hend Al-Khalifa, Sinaa Alageel, and Nora Abanmy. 2016. Readability of arabic medicine information leaflets: a machine learning approach. *Procedia Computer Science*, 82:122–126.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego and Matthew Shardlow. 2022. Towards readability-controlled machine translation of COVID-19 texts. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 287–288, Ghent, Belgium. European Association for Machine Translation.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ibtehal Baazeem, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2021. Cognitively driven arabic text readability assessment using eye-tracking. *Applied Sciences*, 11(18):8607.

Marc Benzahra and François Yvon. 2019. Measuring text readability with machine comprehension: a pilot study. In *Workshop on Building Educational Applications Using NLP*, pages 412–422, Florence, Italy.

Safae Berrichi, Naoual Nassiri, Azzeddine Mazroui, and Abdelhak Lakhouaja. 2022. Impact of feature vectorization methods on arabic text readability assessment. In *The International Conference on Artificial Intelligence and Smart Environment*, pages 504–510. Springer.

Sadik Bessou and Ghozlane Chenni. 2021. Efficient measuring of readability to improve documents accessibility for arabic language learners.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, David B Allen, and Danielle S McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, 23(1):84–101.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Michael Dalvean and Galbadrakh Enkhbayar. 2018. Assessing the readability of fiction: A corpus analysis and readability ranking of 200 english fiction texts. *Linguistic Research*, 35:137–170.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj, Abdulaziz Malik, and Michael K Paasche-Orlow. 2018. Readability of arabic vs english patient educational materials. In *2018 SGIM Annual Meeting*.

Mahmoud El-Haj and Paul Edward Rayson. 2016. Osman: A novel arabic readability metric. In *Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 2:79–97.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2023. Arabic word-level readability visualization for assisted text simplification. *Computational Approaches to Modeling Language Lab*.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of Arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413.

Abdulaziz Malik, Mahmoud El-Haj, and Michael K Paasche-Orlow. 2019. Readability of patient educational materials in english versus arabic. *HLRP: Health Literacy Research and Practice*, 3(3):e170–e173.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Jorge Morato, Ana Iglesias, Adrián Campillo, and Sonia Sanchez-Cuadrado. 2021. Automated readability assessment for spanish e-government information. *Journal of Information Systems Engineering and Management*, 6(2):em0137.

Babak Naderi, Salar Mohtaj, Karan Karan, and Sebastian Möller. 2019. Automated text readability assessment for german language: a quality of experience approach. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Naoual Nassiri, Violetta Cavalli-Sforza, and Abdelhak Lakhouaja. 2023. Approaches, methods, and resources for assessing the readability of arabic texts. *ACM Transactions on Asian Low-Resource Language Information Processing (TALLIP)*, 22(4):95.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Multils: A

multi-task lexical simplification framework. *arXiv preprint arXiv:2402.14972*.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022. ALEXSIS-PT: A new resource for Portuguese lexical simplification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6057–6062, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, and Ruslan Mitkov. 2022. DTW at qur'an QA 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 88–95, Marseille, France. European Language Resources Association.

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.

Hind Saddiki, Karim Bouzoubaa, and Violetta Cavalli-Sforza. 2015. Text readability for arabic as a foreign language. In *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2).

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377", Marseille, France. European Language Resources Association.

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, New Orleans, Louisiana, USA. Association for Computational Linguistics.

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.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1).