while excluding one of the four main group of features blocks POS, Syntactic-features, CEFR-level lexical features, sentence embedding, along with using only the sentence embeddings (XLM-R since this was shown in comparison of embeddings below). This shows that the sentence embeddings significantly contribute to the classification results Table 8), in spite of the efforts to create hand-crafted features. Nevertheless, the linguistic features are useful in interpreting the results of purely neural classification. This results prove that the transformer models provide a rich representation for the sentences covering linguistic features.

According to the primary results from the feature selection and the ablation experiments, which proved the use of sentence embedding alone could fulfill the task without extensive use of linguistics features. These results encouraged us to continue experimentation with applying only the sentence embedding feature to reduce the number of features which consequently decreases the data analysis and training time.

Feature set	**P**	**R**	**F-1**
Exclude XLM-R	0.49	0.63	0.55
Exclude POS	0.55	0.71	0.62
Exclude Syntactic	0.57	0.69	0.59
Exclude CEFR	0.55	0.71	0.62
Only XLM-R	0.75	0.77	**0.75**

Table 8: SVM Classification ablation experiment on 3-way classification

4.5 Testing on Dataset Two

For the binary classification, the classifier reached F-1 of 0.94 and 0.98 for Arabic-BERT and SVM XML-R respectively. However, when testing the binary classifiers trained from DataSet One on Dataset Two the accuracy drops considerably, see Table 9.

As the confusion matrix in Table 10 shows, both classifiers performed better in identifying the complex instances rather than simple ones, so the F1 measure drops. However, the initial results on dataset two shows that XLM-R classifier performed better than Arabic-BERT, we still consider Arabic-BERT classifiers [both 3-way and binary] as best classifier so far. Our interpretation for these confusions is because of the fictional nature of Dataset Two. First, the fiction is well represented in the training data for the A+B levels in Dataset One, while the C level (Snapshot corpus) contains texts of many different types from the internet, so that the classifiers could not handle the mismatch in genres. The other possible reason is that what is considered as complex sentences which are worth simplification according to developers of Dataset Two does not really seem to be complex as to be only suitable for the C-level students. More research is needed to identify the difference between the two datasets.

	Arabic-BERT	**XLM-R**
P	0.60	0.56
R	0.50	0.53
F-1	0.53	0.54

Table 9: Fine-tuned Arabic-BERT versus SVM XLM-R Classifier's performance on Dataset two

	ArabicBert		**XLM-R**	
Predicted	**A**	**C**	**A**	**C**
A	19	2961	138	2842
C	46	2934	223	2757

Table 10: Confusion Matrix with binary classifier Arabic-BERT versus XLM-R on Dataset Two.

4.6 Dataset Two Sentence Similarity

For the purpose of this experiment we needed to include non-simplified sentence to the Dataset two. So that, we duplicated the 2980 complex sentences without simplification aligned with the exact sentence without modification and labeled them with 0 indicating not paraphrased/simplified. Resulting in a Dataset consist of 2980 sentences with right simplification labeled as 1 and other 2980 non-simplified with label 0, in a total of 5960 sentence. The two models trained on this similarity task (AraBert and Arabic-Bert) achieve the F-1 measure of 0.98, leading to the ability to detect sentences which need simplification according to the Dataset Two standard.

5 Related Work on Arabic

The last two decades have seen enormous efforts (especially for the English language) to develop readability measurement ranging from the traditional readability formulae to ML algorithms. English language researchers have introduced more than 200 readability formulae (DuBay, 2004) as well as hundreds of models (Schwarm and Ostendorf, 2005). In contrast, less research has addressed Arabic language issues and their challenges for robust readability formulae.

Some attempts to formulate statistical formulae for the Arabic language reflected traditional English formulae such as the Flesch–Kincaid Grade. The simplest formulae included the average word length, the average sentence and other surface features. According to Cavalli-Sforza et al. (2018) these simple formulae are Dawood formula (1977), Al-Heeti formula (1984), and the formula presented by Daud et al. (2013) based on a corpus. The more sophisticated formulae represent the syllables and more insights the Arabic sentences, such as AARI Base by Al Tamimi et al. (2014) and OSMAN by El-Haj and Rayson (2016).

Other studies were conducted to measure text readability by targeting either first or second language learners for Arabic language modelled using different ML algorithms. Most of these studies used the previously traditional features along with varying lists of part of speech features (POS) representing the words in each document as in studies by (Al-Khalifa and Al-Ajlan, 2010; Forsyth, 2014; Saddiki et al., 2015; Nassiri et al., 2018a). Forsyth (2014) used the word frequency dictionary by Buckwalter and Parkinson (2014) to classify the words' level against this dictionary frequencies. The dictionary was used later by Nassiri et al. (2018b) along with 133 POS features to achieve an accuracy of 100% with 3-classes. The 'Al-Kitaab' textbook has a word list introduced at the beginning of each chapter in the book. These lists were used by Cavalli-Sforza et al. (2014) for comparing the words appeared in a text against this list and labelling them by (target, known, unknown). Saddiki et al. (2018), highlight adding new syntactic features to their features targeting more indepth analysis. They used two different datasets for both first and second Arabic language learning. This yielded an accuracy of 94.8%, 72.4% for first language learners and second language learners respectively.

6 Conclusions

We present the first attempt to build a methodology for Arabic difficulty classification on the **sentence** level. We have found that while linguistic features, such as POS tags, syntax or frequency lists are useful for prediction, Deep Learning is the most important contribution to performance, but the traditional features can help in interpreting the black box of Deep Learning alone. For this specific task and for the Arabic language, fine-tuned Arabic-BERT offers better performance than other sentence embedding methods. Also, application of the classifiers trained on one dataset to a very different evaluation corpus shows that the classifiers learn some important properties of what is difficult in Arabic, but the transfer is more successful for the feature-based models than for the BERT-based ones.

In the end, our best classifier is reasonably reliable in detecting complex sentences; however, it is less successful in separating between the lower learner levels. Still the binary classifier provides the functionality for filtering out really difficult sentences, not suitable for the learners. If we are thinking of Arabic learners especially in higher education, we are expecting learners to graduate with a BA degree in the case of Arabic as a complex language with confidence in reading B2 texts, which implies that the tool for separating A+B vs C level texts is really useful for undergraduate teaching.

Through our tool providing computational assessment of difficulty, we will be able: i) to select the appropriate texts for students; ii) to access ever-larger volumes of information to find educational material of the right difficulty online; iii) to explore curriculum-based assessment to find what is most effective in finding gaps in a curriculum that can be filled according to students' needs.

Our future work involves building a parallel simple/complex Arabic corpus for sentence simplification. The corpus will be classified on the basis of how difficult the sentences are in a Common Crawl snapshot of Arabic web pages. Using the text difficulty classifier, we can split the corpus into two groups for complex and simple sentences. We also consider the semantic similarity detection on **"Saaq al-Bambuu"** as a benchmark, which could be used in the corpus compilation. In this study we only performed some ablation analysis, but because BERT-like models are more useful as the classifiers, we want to investigate their performance via probing for linguistic features following the BERTology framework (Rogers et al., 2020; Sharoff, 2021). We also want to explore the link between the difficulty assessment on the document vs sentences levels (Dell'Orletta et al., 2014).

7 Acknowledgments

This research is a part of PhD project funded by Newton-Mosharafa Fund. All experiments presented in this paper were performed using Advanced Research Computing (ARC) facilities provided by Leeds University.

References

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.

Saud Al-Sanousi. 2013. *Saaq al-Bambuu*. Arab Scientific Publishers Inc., Lebanon.

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.

A. Alfaifi and Eric S Atwell. 2013. Arabic learner corpus v1: A new resource for arabic language research. In *In proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2)*. Leeds.

Amal Alsaif. 2012. *Human and automatic annotation of discourse relations for Arabic*. University of Leeds.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

K Brustad, M Al-Batal, and A Al-Tonsi. 2015. *Al-Kitaab fii Tacallum al-cArabiyya. A Textbook for Beginning Arabic: Part One Third Edition*. Georgetown University Press.

Tim Buckwalter and Dilworth Parkinson. 2014. *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Violetta Cavalli-Sforza, Mariam El Mezouar, and Hind Saddiki. 2014. Matching an arabic text to a learners' curriculum. In *Proc. 5th Int. Conf. on Arabic Language Processing (CITALA), Oujda, Morocco*, pages 79–88.

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. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Nuraihan Mat Daud, Haslina Hassan, and Normaziah Abdul Aziz. 2013. A corpus-based readability formula for estimate of arabic texts reading difficulty. *World Applied Sciences Journal*, 21:168–173.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics*, 165(2):163–193.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Marilena Di Bari, Serge Sharoff, and Martin Thomas. 2014. Multiple views as aid to linguistic annotation error analysis. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 82–86.

William H DuBay. 2004. The principles of readability. *Online Submission*.

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.

Laila Familiar. 2016. *Saud al-Sanousi's Saaq al-Bambuu: The Authorized Abridged Edition for Students of Arabic*. Georgetown University Press.

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

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) [Online]*.

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.

Gerry Knowles and Zuraidah Mohd Don. 2004. The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9(1):69–81.

Kendall Maurice and Gibbons Jean Dickinson. 1990. Rank correlation methods. *London: Edward Arnold*.

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2018a. Arabic readability assessment for foreign language learners. In *International Conference on Applications of Natural Language to Information Systems*, pages 480–488. Springer.

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2018b. Modern standard arabic readability prediction. In *International Conference on Arabic Language Processing*, pages 120–133. Springer.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for