SATLab at BAREC Shared Task 2025: Optimizing a Language-Independent System for Fine-Grained Readability Assessment

Yves Bestgen
Statistical Analysis of Text Laboratory (SATLab)
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

This paper presents SATLab's participation in the BAREC shared task on estimating the readability of sentences and documents. The proposed system is based on character n-grams fed into a support vector regression. A procedure is then applied to try to optimize the Quadratic Weighted Kappa, the main challenge measure, by tuning the decision thresholds used to transform continuous values into ordered categories. Performance is significantly lower than that of the best systems, but nevertheless superior to that of several deep learning approaches.

1 Introduction

Being able to estimate the level of difficulty of a sentence, paragraph, or text has long been an important goal in education (Dale and Chall, 1948). It has been repeatedly demonstrated that students learn better when the texts they are asked to understand are neither too simple nor too complex for them (Vajjala, 2022). It is also important in our society that documents produced by administrations, journalists, and even generative AI can be understood by their recipients while remaining sufficiently interesting to read. Striking the right balance between uninteresting simplicity and discouraging complexity requires the ability to accurately assess readability[1].

Conducting research in this field requires corpora annotated by experts according to readability level. As highlighted in Vajjala and Lučić (2018), the vast majority of available datasets are composed of texts. This is certainly a relevant level of granularity, but a text that is generally simple may contain very complex sentences, which are well beyond the comprehension of the average reader. Corpora in which sentences have been annotated according to their readability level are very rare, and even more so in languages other than English (Hazim et al., 2022; Liberato et al., 2024; Vajjala and Lučić, 2018). Very recently, Elmadani et al. (2025b) developed a corpus for Arabic: the Balanced Arabic Readability Evaluation Corpus (BAREC), which contains 69,441 sentences classified into 19 readability levels. This corpus is at the heart of the BAREC Shared Task 2025 (Elmadani et al., 2025a), which invites participants to develop an automatic approach to estimating the readability of this material. This paper presents SATLab's participation in this shared task.

As in many areas of NLP, deep learning approaches and the use of pre-computed embeddings have proven to be the most effective for estimating the readability of documents or sentences (Lee et al., 2021; Naous et al., 2024; Martinc et al., 2021). However, Vajjala and Lučić (2018) achieved excellent results with a much simpler system, using character n-grams, a well-established approach in computational linguistics (Damashek, 1995), which are fed into some classical supervised approaches such as logistic regression. The advantage of such an approach is that it is completely language-independent, but also that it is does not requires additional resources. For a number of years, SATLab has specialized in using this type of approach to solve complex tasks such as predicting eye saccades during reading (Bestgen, 2021a) or identifying offensive content and hate speech in languages with few linguistic resources (Bestgen, 2021b). Using such a language-independent system in the BAREC task will allow for at least a partial evaluation of the benefits provided by complementary knowledge such as pre-computed embeddings and by the use of far more complex architectures. However, it should be noted that the experiments in Vajjala and Lučić (2018) were conducted on less than 200 texts obtained by asking teachers to rewrite English newspaper articles at

[1] It also requires the ability to assess the reader's language proficiency, but this is an issue that will not be addressed here (Bestgen, 2017)

Proceedings of The Third Arabic Natural Language Processing Conference, pages 280–285

three levels of ESL learners (elementary, intermediate, and advanced). The criterion was therefore to distinguish between these three levels of complexity. It is far from obvious that character n-grams will be equally effective in accurately assessing the fine-grained readability level of Arabic sentences, as is the case in the BAREC 2025 task described below.

2 The BAREC Shared Task 2025

The BAREC Shared Task 2025 (Elmadani et al., 2025a) is based on the Balanced Arabic Readability Evaluation Corpus (BAREC), which contains 1,922 Arabic documents whose sentences (N = 69,441) have been evaluated by annotators in terms of readability on a 19-point scale, 19 indicating the most difficult sentences to understand (Elmadani et al., 2025b; Habash et al., 2025). The corpus covers many genres and topics intended for different target audiences. Annotated examples from the corpus are presented in the two papers mentioned above.

The goal is to develop an automatic model capable of estimating readability levels. These estimates can be made at the sentence or document level. Since the readability of the documents was not directly annotated, the organizers decided that it was equal to the readability level of its most difficult sentence. The material provided by the organizers for the development of the system consists of the entire BAREC corpus. It is divided into three subcorpora: the Learning subcorpus (L) consisting of 1,518 documents and 54,845 sentences, the Development subcorpus (D) consisting of 194 documents and 7,310 sentences, and the Public Test subcorpus (PT) consisting of 210 documents and 7,286 sentences.

Three tracks are available to participants. For the first track, known as "strict," the only readability annotated data that can be used are those from BAREC corpus. For the second track, participants can also use the training set of SAMER Corpus and the SAMER Lexicon (Alhafni et al., 2024; Al Khalil et al., 2020), while for the third track, any publicly available resource can be used. SATLab participated in both tasks of the "strict" track.

The main metric for the challenge is the Quadratic Weighted Kappa (QWK). Several other metrics were also proposed by the organizers, such as accuracy, the percentage of cases where reference and prediction classes match in the 19-level scheme. These will not be discussed here because, as pointed out by Elmadani et al. (2025b), different approaches are needed to optimize a system for these different metrics. The SATLab system will therefore be optimized for the main metric, QWK.

The baseline proposed by the organizers is described in Elmadani et al. (2025b). It is a highly effective baseline which uses, among other things, fine-tuning the very effective Arabic BERT-based models.

3 System Overview

The system proposed by SATLab for the Sentence-level task is mainly based on the character n-grams of the sentences to be analyzed. Some statistics about the sentences, such as their length in characters, and some variables provided in the corpus, such as the annotator, are also taken into account. All these indices are fed into a very classic supervised learning procedure, support vector regression (SVR). A regression-type approach was chosen because Elmadani et al. (2025b) showed this kind of approaches were particularly effective when the metric was QWK.

SVR produces a continuous value that must be converted to integers in the 19-ordinal category system used for readability annotation. This can be done in a very simple way, by rounding these continuous values to the nearest integer and ensuring that none of the values obtained are less than 1 or greater than 19. However, Beckham and Pal (2017) showed that it was possible to improve the QWK of a predictor by modifying the loss function. Their approach is relatively complex, at least for me. For this reason, a simple procedure was developed to try to optimize the QWK by tuning the decision thresholds used to transform continuous values into ordered categories.

The system developed for the document-level task is also based on a SVR mainly fed with the continuous readability estimates from the sentence-level system described above. The features used include the lowest readability value (highest score on Readability Level 19) returned by the Sentence-level system for the document, a series of features encoding the proportion of sentences in the document that have a predicted value equal to a given value (after rounding), and a few global statistics and variables provided in the corpus. The SVR continuous readability estimates were converted to the 19-ordinal category system by the procedure used for the Sentence-level track.

4 Implementation

Almost all of the analyses were performed using a series of custom SAS programs. The QWK optimization was programmed in C. The supervised learning procedure used is the LibLinear L2-regularized L2-loss SVR dual (Fan et al., 2008).

Both systems were optimized on the combination of L and D sets using a 9-fold cross-validation procedure (CV9). These folds were stratified by document, with all sentences from a given document placed in the same fold. In order to ensure that folds contained similar numbers of sentences, the random distribution also took into account the length of the documents in terms of sentences. This CV9 step led to the following parameters being set for the Sentence-level track.

- Character n-grams with n in [1, 6], which occurs at least 10 times in the dataset. These features were weighted by the Sublinear Tf-Idf and then L2-normalised.
- Two global statistics: the log-transformed number of characters and number of different characters.
- Four one-hot encoded variables provided in the corpus: Annotator, Source, Domain and Text Class.
- The SVR regularization parameter and bias were set to 3 and 0.1, respectively.

As explained above, the SVR continuous readability estimates were converted to integers in the 19-ordinal category system by a handcrafted function that attempts to optimize the QWK. The OptiK function takes the SVR continuous readability estimates as input. The thresholds (T) for rounding are initially set to the usual values for rounding to an integer and the QWK is calculated. This value is provisionally considered to be the maximum QWK. Next, the procedure randomly chooses a threshold (Ti) and searches between Ti-1 and Ti+1 for the value for that threshold that produces the largest QWK, starting in the middle of the range of values to be tested and advancing in each direction in turn. This procedure may seem insignificant. However, it favors values in the middle of the interval when there are multiple occurrences of the maximum value. If this maximum value is greater than the current maximum QWK, it replaces it, and Ti is set to the new threshold. This procedure is repeated 150 times, an arbitrary number chosen after some trial and error.

	L->D	L->PT	L+D->PT
No OptiK	76.9	78.1	78.7
OptiK	78.2	79.6	80.5

Table 1: QWK for the Sentence-level

It is not advisable to apply the OptiK function on the data that has been used to train the predictive model, due to model overfitting for this data. It is therefore preferable to use predicted data. Two scenarios were used:

- Train the predictive model on the L set and apply it to the D and PT sets. Then, 1) optimize the thresholds on the D set and evaluate them on the PT set, and 2) optimize the thresholds on the PT set and evaluate them on the D set.
- Train the predictive model on the combination of the L and D sets in CV9, combine the predictions for the 9 folds into a single dataset, optimize the thresholds on it, and evaluate them on the PT set.

For the Document-level task, the predictive model was built based on the following features and parameters:

- The lowest readability value (highest score on Readability Level 19) returned by the Sentence-level system for the document.
- Twelve features encoding the proportion of sentences in the document that have a predicted integer round score equal to a given value from 8 to 19.
- Two global statistics: the log-transformed number of sentences and number of words in the document.
- Three one-hot encoded variables provided in the corpus: Source, Domain and Text Class.
- The SVR regularization parameter and bias were set to 6 and 0.5, respectively.

The SVR continuous readability estimates were converted to the 19-ordinal category system using the OptiK procedure described above.

	L+D (CV9)			L+D->PT
	Mean	Min	Max	
No OptiK	72.2	67.1	77.5	64.3
OptiK				67.1

Table 2: QWK for the Document-level

	Sentence		Document	
	Final	No OptiK	Final	No OptiK
Best	87.5		87.4	
SATLab	82.3	80.2	77.6	73.3
Baseline	81.5		62.0	

Table 3: QWK for the BT set

5 Results

This section presents the performance of the proposed system, first on the D and PT sets, and then on the real challenge, i.e., the Blind Test set (BT). The latter consists of 100 documents and 3,420 sentences.

5.1 Public evaluation sets

Table 1 presents the QWK for the different public evaluation sets for the Sentence-level task. We can see that the PT set is a little simpler than the D Set and that adding the D to the L set for learning improves performance, which is obviously to be expected. Above all, we observe that optimizing the QWK brings a benefit of 1.3% and 1.8% in QWK, which does not seem negligible. The best performance obtained on the PT set is slightly higher than that obtained by the Baseline system (QWK = 80.2). Exceeding this value was one of SATLab's objectives, since the Baseline system uses, among other things, fine-tuning the very effective Arabic BERT-based models (Elmadani et al., 2025b).

The material for the document-level task is relatively small for supervised learning procedures. For this reason, the conditions evaluated are different from those used for the sentence-level task. Learning was performed on the combination of the L and D sets in CV9, QWK optimization on the 9 predicted folds, and final evaluation on the PT set.

The QWKs are significantly weaker for this task (Table 2). This is likely due to the inaccuracy of the Sentence-level model, which produces an overly imperfect estimate of the readability level of the most difficult sentence in a document. It is particularly noteworthy that CV9 performance varies greatly depending on the fold. There is therefore a high degree of instability in the results, probably due to the relatively small number of documents in each fold (N = 190) and in the PT set (N = 210).

5.2 Challenge results: BT set

The main question that this study attempts to answer is that of the performance level of a system based on indices as simple and as unspecific to the task as character n-grams compared to much more complex systems, such as those using precomputed embeddings. As reference points, I chose the Baseline system, described in Elmadani et al. (2025b), and the top-ranked system, !MSA, assuming that it also uses sophisticated techniques. To analyze this BT set, the complete public material (L+D+PT) was used for learning.

Table 3 shows that the SATLab system is capable of outperforming systems that use fine-tuning of BERT-based models, but that QWK optimization is essential to achieve this result. The difference with the best system is clearly significant (5.2%) and justifies the use of more complex models than an SVR on character n-grams, as proposed by SATLab.

It should be noted that the QWK of the Baseline system for documents is significantly lower than the QWK of all other systems that participated in this task. It seems likely that this system's prediction for a document is simply equal to the highest predicted score for the sentences in that document, without taking into account any other features or new learning. There is no doubt that a higher performance could have been achieved.

5.3 Impact of OptiK on thresholds

The results presented above indicate that QWK optimization is essential for the system to achieve a competitive score. This trick, if not used by other systems, somewhat distorts the comparison with them. Indeed, it is reasonable to assume that they could have improved their QWK in this way.

In order to gain a clearer understanding of the effects of OptiK on the thresholds used to transform SVR scores into categories, Figure 1 shows the thresholds obtained by this procedure for submission to the Sentence-level task. The bottom line shows the range of predicted values from the SVR. The middle line simply indicates the thresholds usually used when rounding a real number to a whole number. The top line shows the thresholds obtained using the OptiK procedure. As can be seen, some thresholds are significantly modified. For example, the range of values corresponding to category 1

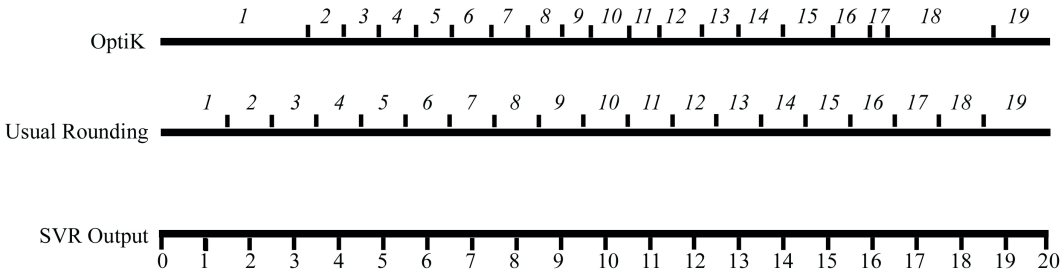

Figure 1: Effects of OptiK on the thresholds

is greatly expanded, while that corresponding to category 17 is smaller than what is obtained using a standard rounding procedure. We can even see that the continuous values corresponding to several categories do not cover the usual range of values (see categories 2 to 6, 9, 10, 17).

6 Conclusion

This paper presents SATLab's participation in the BAREC shared task. The proposed system relies almost exclusively on character n-grams, which are used by an SVR to estimate the readability of Arabic sentences. A post-processing procedure is then applied to the predicted values to optimize the main measure of the challenge: QWK. This system ranks 16th out of 24 in the Sentence-level task when all participating teams are taken into account, and 13th out of 16 in the official ranking composed of participating teams that have published a report about their system. It is 4th out of 8 in the official Document-level ranking, each time for the Strict track. These performances make it more effective than systems using precomputed embeddings, but it is important to remember that a significant part of its effectiveness comes from the QWK optimization procedure and that it is likely that several other systems did not use such a trick.

As for the shared task itself, I think it could be interesting to reevaluate the document-level task. In particular, the analyses conducted in CV9 showed significant variability in performance depending on the fold. The comparison of QWKs on the PT set (SATLab = 67.1) and BT sets (SATLab = 73.3) confirms this significant variability. It could be related to the small size of these samples, which means that changing a few predictions can significantly affect the QWK. It might also be interesting to replace the current procedure for determining the readability level of a document (that of the most difficult sentence) with an annotation made by experts.

Acknowledgments

The author wishes to thank the organizers of this shared task for putting together this valuable event and the reviewers for their very constructive comments. He is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique).

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.

Christopher Beckham and Christopher Pal. 2017. A simple squared-error reformulation for ordinal classification. *Preprint*, arXiv:1612.00775.

Yves Bestgen. 2017. Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 69:65–78.

Yves Bestgen. 2021a. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online. Association for Computational Linguistics.

Yves Bestgen. 2021b. A simple language-agnostic yet strong baseline system for hate speech and offensive content identification. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings. CEUR-WS.org.

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

Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–848.

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*, Vienna, Austria. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

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.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. 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.

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

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 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

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. Association for Computational Linguistics.