Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 09, No. 01, March 2023.

INTERPRETING THE RELEVANCE OF READABILITY PREDICTION FEATURES

Safae Berrichi[1], Naoual Nassiri[2], Azzeddine Mazroui[1] and Abdelhak Lakhouaja[1]

(Received: 10-Nov.-2022, Revised: 3-Jan.-2023 and 31-Jan.-2023, Accepted: 13-Feb.-2023)

ABSTRACT

Text readability is one of the main research areas widely developed in several languages, but it is highly limited when dealing with the Arabic language. The main challenge in this area is to identify an optimal set of features that represent texts and allow us to evaluate their readability level. To address this challenge, we propose in this study various feature selection methods that can significantly retrieve the set of discriminating features representing Arabic texts. The second aim of this paper is to evaluate different sentence-embedding approaches (ArabicBert, AraBert and XLM-R) and compare their performances to those obtained using the selected linguistic features. We performed experiments with both SVM and Random Forest classifiers on two different corpora dedicated to learning Arabic as a foreign language (L2). The obtained results show that reducing the number of features improves the performance of the readability-prediction models by more than 25% and 16% for the two adopted corpora, respectively. In addition, the fine-tuned Arabic-BERT model performs better than the other sentence-embedding methods, but it provided less improvement than the feature-based models. Combining these methods with the most discriminating features produced the best performance.

KEYWORDS

Readability, Feature selection, Sentence embedding, Arabic language, Education.

1. INTRODUCTION

Arabic is one of the most used languages in the world. It is used by more than 400 million people[1] and is the official language of more than 20 countries. Arabic-language processing has attracted much more interest during this century. Many studies have focused on different aspects of Arabic-language processing, such as morphological analysis, resource construction, machine translation, sentiment analysis and readability assessment [1][2][3]. This last field of research (readability assessment), which is widely investigated in several other languages, represents an emerging field of research for Arabic.

Readability refers to the ease with which a reader can understand a written text. It can be assessed using a supervised learning approach, which is one of the most successful branches of Machine Learning (ML). This consists of training a predictive model using a set of data samples that are represented by a number of linguistic features [4]. These vectors are labeled with their classes, which correspond in the case of readability to the difficulty level of the text.

More recently, deep neural networks have been proposed to predict readability of texts in languages with large linguistic resources, such as English [5]. Although the model performance improves when the training corpora are enriched with additional data, this approach is not always adopted given that the readability annotation is time-consuming and has a high cost. An alternative approach to improve the prediction-model performance is to represent the text as embedded vectors using neural models. The latter learn semantics at the sequence level by considering all words in the document.

On the other hand, despite the advantages of ML-classification models that use a variety of linguistic features motivated by the language, they still suffer from several challenges [6]. Indeed, the use of relevant features to represent a text represents a great challenge [7]. Researchers are currently interested in developing solutions that extract the most discriminating features from the vectors that

[1] https://www.internetworldstats.com/stats7.htm (last visited on 09/10/2022)

1. S. Berrichi, A. Mazroui and A. Lakhouaja are with Department of Computer Science, Faculty of Sciences, Mohammed First Uni., Oujda, Morocco. Emails: berrichi.safae@gmail.com, azze.mazroui@gmail.com and abdel.lakh@gmail.com
2. N. Nassiri is with Engineering and Sustainable Development Team, Ibn Zohr Uni., Higher School of Technology, Dakhla. Morocco. Email: naoual.nassiri@gmail.com

represent the original texts. Feature selection is one of the methods used to overcome such a problem. This method consists in selecting the relevant features and eliminating the irrelevant or redundant ones [8]. Representing a text with the most discriminating features can improve the performance by increasing the generalization ability and the classification accuracy.

The contributions of this research can be summarized as follows. First, we identify subsets of the most relevant linguistic features used in Arabic text-readability measurement. Thus, we examine different feature-selection methods, individually and in combination, to determine their impacts on readability prediction models and to identify the optimal feature vector that achieves good performance. A second contribution aims to evaluate various sentence-embedding approaches, such as AabicBert, AraBERT and XLM-R. So, we first develop a readability-measurement model based entirely on one of these embedding approaches. Then, we develop a model that combines these embedded vector representations with different linguistic feature sets selected and judged relevant in this study.

To validate whether the obtained conclusions are independent of the used corpus, we evaluated our first experiments on two corpora. The first one, composed of 321 texts, was collected from the Aljazeera- Learning website[2] for learning Arabic and the second, consisting of 278 texts, was collected from the GLOSS platform[3] whose texts were developed by the Defense Language Institute Foreign Language Center[4] considered to be one of the top foreign language schools.

The remainder of this paper is structured as follows. Section 2 provides the general background related to readability-measurement techniques and feature-selection methods. Section 3 describes the basic concepts of feature selection and the most popular associated techniques adopted in this work. In the same section, the dataset used in this study, as well as the initial feature vector, are described. Section 4 presents the feature-selection results based on an ML approach, while deep-learning-based results are represented in Section 5. The last section is devoted to the conclusion and some thoughts on future work.

2. RELATED WORK

In the field of text-readability classification, many features have been used. However, few works have focused on identifying relevant features representing a document. In this section, we review recent work on measuring the readability of Arabic foreign language (L2) texts that focuses on linguistic features, as well as those that instead of linguistic features use the raw embedding of the input text. We also review studies related to techniques for selecting the most discriminating features.

2.1 Readability Assessment Early Approaches

In 2014, Forsyth described in his thesis [9] a system that consists of automatically predicting the readability of Modern Standard Arabic (MSA). The study is based on a corpus retrieved from the GLOSS platform and consisting of 179 texts. He incorporated lexical and morphological features in the model-generation process. In total, he generated a high-dimensional vector containing 162 features. Based on a cross-validation, he reported a maximum F-score value of 0.78 for three classes (easy, medium and difficult). In the same year, the authors of [10] described a study to evaluate whether a given text is suitable for an MSA learner as L2 using their own corpus. They focused on the vocabulary content of learners' programs and texts, as well as other word-related features. The model achieved an accuracy of about 60%. This study adopted a vector that was significantly less voluminous than Forsyth's (10 features).

Saddiki et al. conducted a study in 2015 [11] in which they evaluated the usefulness of lexical and morphological features for predicting readability. They gathered a corpus from the GLOSS platform consisting of 251 texts and compiled 35 low-complexity features in order to establish a baseline for future research on readability assessment. Their findings suggest that a small set of easily calculable features might be indicative of the reading level of a text. They reported a maximum accuracy of 73.31% and a maximum F-score of 0.73. Their F-score value is close to the value obtained by Forsyth (0.78) using only 35 features instead of the 165 features used by Forsyth.

[2] https://learning.aljazeera.net/en (last visited 09/10/2022)

[3] https://gloss.dliflc.edu/ (last visited 09/10/2022)

[4] https://www.dliflc.edu/ (last visited 09/10/2022)

The same authors conducted a study in 2018 [12] in which they used 146 features based on a GLOSS corpus composed of 576 MSA texts. Their best results reached 72.4% and 0.61 in terms of accuracy and F-score, respectively. These performances are lower than those obtained by the two previous works [9], [11]. In the same period, Nassiri et al. presented a study [13] in which they gathered a GLOSS corpus comprising 230 texts. They introduced a set of 170 features to represent a text. They reported an F-score of 0.9 when testing on the training data. This study suffers from a lack of generality, since the authors report results obtained by testing on the training data. Thus, these results are not comparable to the performances of the previously mentioned studies that reported evaluation results based on the use of conventional training and testing corpora practices. Finally, in 2021, Nassiri et al. presented a study [3] in order to identify a smaller set of features that could provide good readability-prediction accuracy. They eliminated features having low predictive weights, to end up with 76 relevant features and obtain an accuracy score of 86.15% on the test data.

Concerning the evaluation of readability in other fields besides education, we have some works on health texts' readability. In 2018, Al-Aqeel et al. [31] presented a study to assess the readability of written medicine information in terms of both vocabulary use and sentence structure. They assessed readability according to three difficulty levels (easy, intermediate and difficult) for 4,476 sentences. Looking to assess the quality and readability of the Arabic health information about COVID-19, in 2021, Halboub et al. [32] evaluated a set of websites. They concluded that practically all of the Arabic health information available on COVID-19 is not easily readable and understood by the general Arabic-speaking population. In 2022, Jasem et al. [33] conducted a study with the goal of evaluating Arab websites dedicated to breast cancer and recommended ways of improving engagement and access to health information. The results led them to conclude that, in general, the readability scores indicate that the websites are above the recommended reading level.

Most of these health-data dedicated studies are using traditional readability formulae (calculating reading scores) to evaluate the difficulty levels instead of machine-learning approaches and this is due to the limitation of unavailability of annotated data in this field to train supervised models.

2.2 Deep-learning-based Readability Assessment

In most state-of-the-art studies, researchers continue to opt for statistical ML techniques. These approaches are the most appropriate ones given the lack of large annotated corpora for readability and since large amounts of data are generally required to successfully use deep-learning architectures that use text embedding as input. As a result, studies based on these techniques are scarce for Arabic. In this sub-section, we will review some recently published works based on these new techniques.

In languages for which large amounts of data exist, readability-prediction techniques based on deep-learning models are emerging, unlike other languages such as Arabic, focusing on the English language, Deutsch et al. [5]. More recently, Lee et al. reported a study [15] based on the same concept of augmented deep-learning by combining linguistic features with transformers. They reported results supporting the hypothesis that the use of hand-crafted features improves the performance of deep-learning models on smaller datasets.

Concerning the Arabic language, the works are even scarcer. In 2021, Khallaf et al. [16] presented an approach to predict the difficulty of MSA sentences. They compared the performance of different types of sentence embedding (fastText, mBERT, XLM-R and Arabic-BERT) and compared them to traditional linguistic features, such as PoS tags, dependency trees, readability scores and frequency lists for language learners. Their best results were obtained using Arabic-BERT. They reported results with macro-averaged F-1 of about 0.80 and 0.75 for the Arabic-Bert and XLM-R readability classification, respectively.

2.3 Feature Selection

The complexity of natural languages calls for textual feature vectors with high dimensionality. The latter makes the classification process considerably difficult [17] and is therefore considered one of the main challenges in this field.

Feature selection has been addressed in many studies. The objective of these studies is to improve the performance of the models and to provide faster and less complex models. This is performed by selecting subsets of features from high-dimensionality feature sets. The authors in [18] examined the

effect of some feature-selection methods; namely, Information Gain [19], Correlation [20], SVM selection, Gini index and Chi-Square [21] and their combinations on the performance of the SVM classifier for sentiment analysis in dialectal Arabic. The results of this study show that the SVM classifier achieved the best performance when the SVM selection method was used. On the other hand, the SVM classifier performed better when the two methods "correlation and SVM selection" were combined consecutively.

Similarly, Elhassan et al. [22] studied the impact of using the Information Gain and the Chi-Square selection techniques on the performance of Arabic-text classification models based on the Naïve Bayes and the SVM classifiers. The tests performed showed that both feature-selection techniques improve the performance of the models and that the Information Gain technique produces the best results.

A comparative study between some feature selection methods adopted to categorize Arabic texts was reported by [23]. In addition, an adjustment was carried out to the feature-selection approaches by grouping the selected methods (Term Frequency Inverse Document Frequency "TF-IDF", Chi-Square and Information Gain). Furthermore, a new method was proposed to select the most appropriate features. This method is based on semantic fusion and multiple words (SF-MW) to construct the features. The combination of the adopted feature-selection methods yields better results than the individually selected methods. Thus, the proposed SF-MW feature-selection method is promising, because it reduces features and achieves a better classification accuracy.

Based on the review of many previous studies related to readability prediction of Arabic texts (sub-section 2.2), we notice that most of these studies have focused on the evaluation of different classifiers on MSA texts, using a determined set of features. The relevance of the features that compose the readability-prediction vectors in Arabic has been addressed, to our knowledge, only recently [3], [16].

Table 1 summarizes the set of Arabic readability-assessment studies that we have reviewed in this sub-section. The summary contains the distribution of the features categories (discussed in sub-section 3.3) in each study, the feature-selection method adopted (if it exists) and the used classifier(s). From these sets, we have chosen 70 features that we can implement based on the available tools and resources.

Table 1. Arabic readability-assessment studies.

Study	Features	RTF	MF	PDF	FLF	PBF	Selection method	Classifier
(Forsyth, 2014) [9]	162	✓	✓	✓	✓	✓	–	TiMBL [34]
(Cavalli-Sforza et al., 2014)[10]	10	✓	✓	✓			–	K-means
(Saddiki et al., 2015) [11]	35	✓	✓	✓	✓		–	Random Forest, SVM,...etc.
(Nassiri et al., 2017) [13]	170	✓	✓	✓	✓	✓	–	Random Forest, SVM,...etc.
(Saddiki et al., 2018) [12]	146	✓	✓	✓	✓	✓	–	Random Forest, SVM,...etc.
(Nassiri et al., 2021) [3]	76	✓	✓	✓	✓	✓	Correlation	Random Forest, SVM,...etc.
(Khallaf et al., 2021) [16]	5	✓		✓			Recursive Feature Elimination	Random Forest, SVM,...etc.

3. METHODS AND TOOLS

In this section, we will present the methods that we have adopted to address the problem of the suitability of the linguistic features used to assess the difficulty of a text and the impact of using sentence embedding with/without incorporating these features. We will also present the algorithms used to build the readability-prediction models and the used evaluation metrics. The data we have used in this study is also described in this section.

3.1 Relevant Feature-selection Methods

To overcome the problem of the usefulness of feature vectors used as input in a classification task, it is

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 09, No. 01, March 2023.

usually recommended to use specific methods. This leads to enhanced training performance in some instances. These methods are usually divided into feature-extraction methods and feature-selection methods. The main distinction between these two categories is that feature-extraction methods combine the original features to generate new feature sets, while future-selection methods extract feature subsets from the original one. In this study, we will focus on the second category which is based on feature-selection methods. These methods are generally divided into three sub-categories: filter, embedded and wrapper methods. As illustrated in Figure 1, these methods select a subset of features in different manners.

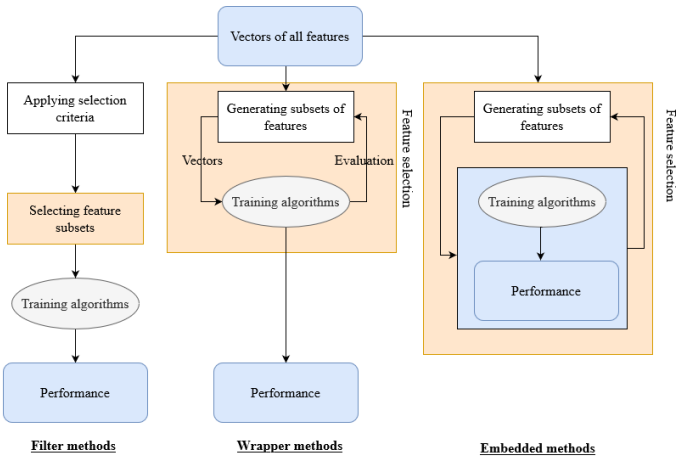

Figure 1. Process of feature-selection methods.

Given the wide range of existing feature-selection methods, choosing the best method for a given problem is a challenge. Furthermore, using these three different variations in feature selection allows us to take advantage of each algorithm that has a particular property and identify the method that captures the most discriminating features. In the following, we describe some reference methods that have shown successful performance and that we adopted in this study:

1) **Filter methods:** these methods are fast, scalable, computationally simple and independent of the classifier. They allow to ignore the dependencies between the feature selection and the classification algorithm. Among these methods, we cite:

   a) Information Gain (IG): a univariate filter that calculates the mutual information for each attribute and each class. Each feature will be ranked according to its information-gain value. Basically, the higher is the value, the more informative the feature is.
   b) Chi-square: a statistical technique used to determine whether there is a significant dependence between two categorical variables. The Chi-square test of independence attempts to determine which of the null hypothesis (independent variables) and the alternative hypothesis (dependent variables) is valid.

2) **Wrapper methods:** these methods require a training step to perform the feature selection. Although these methods are more expensive than filter methods due to their interactions with the classifier, they tend to perform better. The common wrapper method used in this study is based on recursive feature elimination for support vector machines (SVM-RFE) [24]. This method uses the weights of the SVM as a ranking criterion. The concept of SVM-RFE consists of training an SVM classifier in an iterative process and then exploiting the weights in the SVM solution to select some features and eliminate others.

3) **Embedded methods:** the concept of this last category is based on a combination of filter and wrapper methods. They use classification algorithms with an integrated feature-selection capability. They are less expensive in terms of computation than wrapper methods. Among these methods, the inherent features of the SVM (IF-SVM) and Random Forest (IF-RF) algorithms are adopted in this study.

3.2 Classification Algorithms

The Random Forest (RF) algorithm comprises a set of individual decision trees, each one trained on a

random subset of the training data. For a classification problem, the final prediction of RF is determined by a majority vote of the trees. By applying this method, RF reduces the problem of overfitting while maintaining good performance. In addition, it facilitates the interpretation of the results, since the features influencing the predictions can be identified and ranked according to their importance.

The SVM is a classification algorithm based on the statistical-learning theory [25]. It can be used for both linearly separable and non-linearly separable data. It uses a kernel function to project the input data into a high-dimensional space and subsequently determine a weight vector that represents the normal hyperplane for performing binary classification with minimal error rate. The kernels used in SVMs can be linear or non-linear. Non-linear kernels use the polynomial function or the radial basis function (RBF). To the best of our knowledge, the research community generally uses linear kernels to select features by SVMs.

Given the wide use of SVM and RF classifiers in Natural Language Processing (NLP) applications, especially in readability measurement, we adopted them to test the effectiveness of our proposed approaches. Similarly, since SVM feature selection relies on the linear kernel (IF-SVM), we chose to evaluate our models using the same kernel.

3.3 Data and Feature Description

In this study, we used MSA educational corpora intended for learning Arabic as L2. The texts of the latter are labeled with difficulty levels by experts using the ILR scale [26]. The three levels that we used can be interpreted as follows: easy (level 1 and level 1+), medium (level 2) and difficult (level 2+ and level 3). These texts are freely available in two different platforms, from which we have collected two corpora as follows:

1) GLOSS-Reading (GR): a corpus collected from the GLOSS platform which offers thousands of lessons, for independent learners, in dozens of languages, including Arabic. We collected 278 MSA texts annotated according to the three levels of difficulty described above. This corpus comprises a total of 4,666 sentences and 95,469 tokens.
2) Aljazeera-Learning (AL): Aljazeera website for learning the Arabic language also presents educational texts. We have collected from this site 321 texts annotated according to the three levels of difficulty. This last corpus contains 2,442 sentences and 49,345 tokens.

Tables 2 and 3 present in detail the statistics of these corpora according to the three levels of difficulty.

Table 2. Statistics of GLOSS-Reading corpus.

Level	**Texts**	**Sentences**	**Tokens**	**Average Sentence Length**
Easy	66	1,237	11,462	9.26
Medium	95	1,406	33,264	23.65
Difficult	117	2,023	50,770	25.09
Total	**278**	**4,666**	**95,469**	**20.46**

Table 3. Statistics of Aljazeera-Learning corpus.

Level	**Texts**	**Sentences**	**Tokens**	**Average Sentence Length**
Easy	232	1,277	20,840	16.31
Medium	54	378	7,646	20.22
Difficult	35	787	19,859	25.23
Total	**321**	**2,442**	**49,345**	**20.20**

When analyzing these two tables, we can notice that the used corpora suffer from an imbalance of data, especially for Aljazeera-Learning corpus. Regarding the average sentence length, calculated in terms of the number of sentences divided by the number of words, we conclude that it is a very important indicator, since it increases from one level to another for the two corpora.

Textual features associated with the degree of readability can be simple attributes, such as text length

or average word length; or more complex attributes, such as those related to grammatical categories.

The list of features that we used in this study is inspired from [3]. The authors of this work started with 170 features and found that some ratio features that are combinations of other existing features and some features that have the same value for all the texts are not discriminating for the readability prediction task; so they reduced the set from 170 to 76. So, we recompiled many of these features based on the PoS (Part of Speech) tags of the MADAMIRA analyzer [27]. In total, we have used a set of 70 features. This feature set was organized along two dimensions depending on the depth of processing required to extract them:

1) **Raw Text Features (RTF**): many formulae using raw-text features have been adopted and successfully adapted in English and other languages. Their popularity is due to their simplicity to compute and understand. In this category, we have chosen three features; the number of sentences, the number of words and the number of characters in a text.

2) **Features extracted after morphological analysis:** since readability is strongly influenced by vocabulary and word-level information, providing lexical and morpho-syntactic information at the word level can improve predictions. Features related to this different morpho-syntactic information are grouped into a set of sub-categories:

    a) *Morphological Features (MFs)*: these represent a sub-category composed of 5 features based on the distribution of vocabulary in a text. These five features are: the number of lemmas in the text, the number of stems, the number of frequent lemmas (lemmas that appear more than once in a text), the number of ambiguous lemmas (lemmas that have two different PoS tags in the same text) and the number of closed-class tokens[5].

    b) *PoS Dispersion Features (PDFs)*: these from a sub-category composed of 15 features extracted from PoS tags which examines the presence of different grammatical categories in the text (e.g. the number of verbs and the number of nouns).

    c) *Frequent Lemmas' Features (FLFs)*: this subcategory is composed of 10 features computed on the basis of the frequency dictionary (a dictionary containing the list of the 5,000 most frequent lemmas in the Arabic language with their morphological information); for example, the average dispersion of frequent lemmas or the average rank of frequent lemmas. For this sub-category, a lemma is considered frequent if it appears in the frequency dictionary.

    d) *PoS Based Frequency Features (PBFs)*: this is the largest sub-category in this study, comprising a total of 37 features. It represents the ratio of individual frequent grammatical categories (which appear in the frequency dictionary) to words in the text; for example, the average rank of frequent nouns and the maximum rank of frequent verbs.

3.4 Adopted Process

Figure 2 describes the approach implemented in this study. The process is performed in several steps. First, we represent each input text by a vector of 70 features. Given the importance of using the most discriminating features, we first applied the IG, Chi-square, IF-RF, IF-SVM and SVM-RFE feature-selection methods on the original vectors (based on the 70 first features). Then, we analyzed and compared the different feature sets obtained and subsequently we selected the most discriminating ones in the field of Arabic text-readability prediction. The models obtained using the five feature-selection methods and those obtained using the different combinations were evaluated based on the SVM and RF classifiers. Finally, we evaluated the impact of sentence embedding on the performance of readability predictions and we combined them with the best linguistic features obtained in the ML-based experiments.

3.5 Evaluation Measures

The performance of our models was evaluated based on the calculation of accuracy and F-score.

[5] a closed-class token is a token that belong to a grammatical category having a finite list such as prepositions

These measures are calculated based on the following formulae:

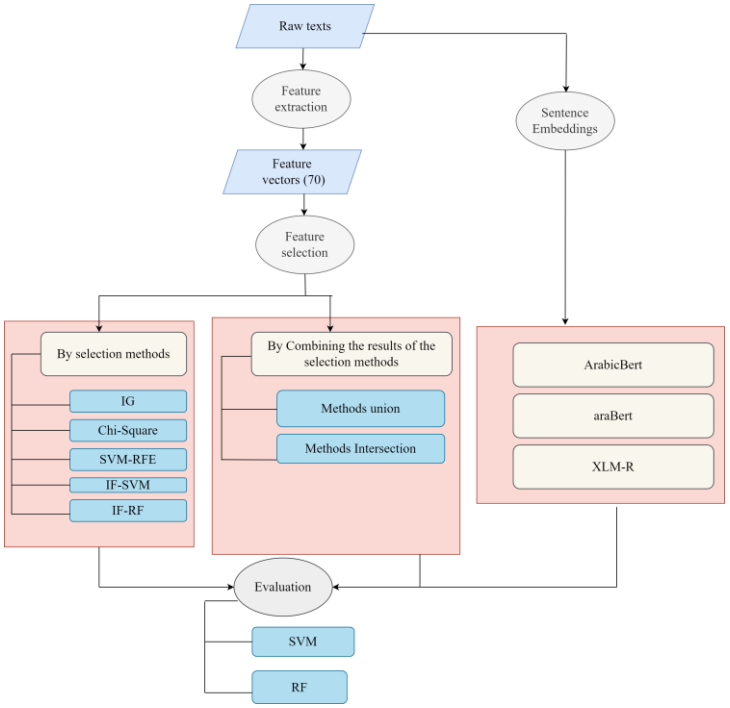

Figure 2. Adopted process for feature extraction.

$$Accuracy = \frac{number\ of\ correctly\ predicated\ documents}{total\ number\ of\ documents\ in\ the\ test\ set} * 100$$

$$Precision_i = \frac{number\ of\ documents\ correctly\ assigned\ to\ the\ class\ C_i}{total\ number\ of\ documents\ assigned\ to\ the\ class\ C_i}$$

$$Precision = \frac{\sum_{i=1}^{n} Precision_i}{n}$$

$$Recall_i = \frac{number\ of\ documents\ correctly\ assigned\ to\ the\ class\ C_i}{number\ of\ documents\ belonging\ to\ the\ class\ C_i}$$

$$Recall = \frac{\sum_{i=1}^{n} Recall_i}{n}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where $Precision_i$ and $Recall_i$ represent, respectively, the precision and the recall of the class C_i and n represents the number of classes.

4. STATISTICAL ML-BASED EXPERIMENTS

4.1 Feature Selection Based on the AL Corpus

The goal of this paper is to select the most relevant features needed to predict the readability of an Arabic text. For this purpose, we conducted a set of experiments to examine the impact of the feature-selection methods discussed in Section 3 on the readability measurement of Arabic texts on the AL corpus. This methods rank the 70 features that we adopted in this study by their relevance degree. We compared the results of the selected features for each selection method with those obtained by the original "BaseLine" classification model (obtained using all the 70 features). Based on several experiments, we found that the best performances of all the selection methods are generally obtained with the 20 most informative features. The subset selected for each selection method is therefore made of the 20 most discriminating features. The parameters use in each selection method are those used by default in the scikit-learn library[6]. For RFE, we used the supervised learning SVM

[6] https://scikit-learn.org/stable/modules/feature_selection.html

estimator with kernel type and a 10-value regularization. To select the discriminative features by the RF machine-learning model that were trained and tested in both RF and SVM (IF-RF and IF-SVM), we kept the same parameters as those used in scikit-learn (the number of trees in the forest is 100 with a gini criterion and max_features to be taken into account when searching for the best split being sqrt). In the IG method, sklearn.feature_selection.mutual_info_classif was used with the default parameters and finally Chi-square sklearn.feature_selection.chi2 was used in the Chi-square method. All of these parameters are also retained in the classifiers and the remainder of the experiments to prevent the impact of the model parameters on the selection.

In the remainder of this experiment, we will evaluate the performance of each selected feature set. Thus, we adopted a random distribution of the AL corpus into 80% for training and the remaining 20% for testing. This distribution is adopted instead of 10-cross validation in order to evaluate the feature performances on the same data sets and this prevents the influence of data distribution on the results. In order to maintain the proportions of the three difficulty levels, we applied stratified sampling. The experiments were performed with both RF and SVM classifiers.

Table 4. Test results of the different feature sets on the AL corpus.

Model	**SVM Classifier**		**RF Classifier**	
	Accuracy	**F-score**	**Accuracy**	**F-score**
BaseLine	63.46%	63.54%	*86.53%*	79.17%
IG	*88.46%*	*83.21%*	*86.53%*	81.49%
Chi-square	73.07%	63.68%	84.61%	70.15%
SVM-RFE	**90.38 %**	**86.77%**	*86.53%*	**83.24%**
IF-RF	76.92%	72.51%	**88.46%**	*83.21%*
IF-SVM	84.61%	79.91%	86.53%	81.49%

Table 4 presents the accuracy and F-score of the "BaseLine" model and those of the models obtained using the most discriminating features selected by each of the five selection methods (IG, Chi-square, SVM-RFE, IF-RF and IF-SVM).

The results clearly show the relevance of using feature-selection methods in the readability prediction of Arabic texts. Indeed, except for the Chi-square method applied with the RF classifier, all the other methods perform better than the "BaseLine" model independently of the classifier used. Moreover, the selection method based on SVM-RFE reported the best results, since it improved the F-score by more than 23% for the SVM classifier and by more than 4% for the RF classifier. The IG and IF-RF methods reported the second best results using the SVM and the RF classifiers, respectively.

Concerning the embedded feature-selection methods IF-SVM and IF-RF, we notice that when these methods are embedded in their classification algorithm (IF-SVM with SVM and IF-RF with RF), the results are better compared to embedding them in another classification algorithm (e.g. IF-RF with SVM). We will exclude, in the rest of this paper, the Chi-square selection method that reported with RF classifier results lower than those reported by the "BaseLine" model.

The four subsets, each one comprising 20 features, generated by the four remaining feature-selection methods (IG, SVM-RFE, IF-RF and IF-SVM) are not identical. We thus analyzed these subsets to identify the most discriminating features. We initially find that only 39 of the 70 features used in the "BaseLine" model appear in these subsets. So, all the selection methods used in this study agreed that the remaining 31 features are irrelevant in the readability-prediction process. Moreover, only 24 features among the 39 are selected by at least two of the four selection methods. The complete list of these 24 features, in addition to three other features that we will discuss later, is presented in the appendix. We also note that all the five feature categories cited in sub-section 3.3 are present in the selected subsets. The dominant category is PBF with nine features, followed by PDF with eight features, as shown in the appendix.

On the other hand, we observe a strong presence of some features, since they appear in at least three of the four subsets. For this reason, we generated combinations of these feature subsets to determine the most informative ones. Table 5 presents the new subsets of features that result from these combinations.

Table 5. New feature subsets obtained on the AL corpus.

Sub-set	Reference	Features
Features that appear in at least one of the subsets	$F_{\cup 4methods}$	39
Features common to all four subsets	$F_{\cap 4methods}$	6
Features that appear in at least three subsets	$F_{\cap 3methods}$	12
Features that appear in at least two subsets	$F_{\cap 2methods}$	24
Union of the top five features from each subset	$F_{\cup 5best}$	12

The six features common to all methods, represented by the $F_{\cap 4methods}$ model, are illustrated in Table 6. Among the six common features, we have the number of lemmas which represents the size of the vocabulary and constitutes a feature strongly related to the difficulty of the text. Indeed, the more important is the vocabulary, the higher the difficulty of the text is. For the other features (numbers of nouns, properNoun and Noun/Open class token), we notice that much of the meaning of a text is in their nominal content.

Table 6. Common features between the four selection methods.

Features	Meaning
Lemma Count	The number of lemmas in a text
Noun Count	The number of nouns in a text
Adverb Count	The number of adverbs in a text
PropNoun Count	The number of proper nouns in a text
ClosedClassTokensCount	The number of closed class words in a text
Noun / Open Class Token	The ratio between nouns and the open class words in a text

After implementing these combinations between subsets of features (Table 5), we measured their impact on the readability prediction using the AL corpus. The performance of the models obtained using these subsets is compared with that of the initial "BaseLine" model, as well as with that of the SVM-RFE model, which is based on 20 features and produced the best results in the previous experiments. The test results are outlined in Table 7.

Table 7. Results of the different classification models on the AL corpus.

Model	SVM Classifier		Features	RF Classifier	
	Accuracy	F-score		Accuracy	F-score
BaseLine	63.46%	63.54%	70	*86.53%*	79.17%
SVM-RFE	90.38%	86.77%	20	86.53%	83.24%
$F_{\cup 4methods}$	71.15%	64.09%	39	**90.38%**	**86.77%**
$F_{\cap 4methods}$	88.46%	83.21%	6	84.61%	79.91%
$F_{\cap 3methods}$	90.38%	86.77%	12	86.53%	84.93%
$F_{\cap 2methods}$	82.69%	78.45%	24	86.53%	81.49%
$F_{\cup 5best}$	**92.30%**	**88.80%**	12	88.46%	84.93%

The results of this second experiment show that some combinations between the subsets obtained by the four feature-selection methods lead to better performance regardless of the classifier used. Indeed, the SVM classifier applied with the 12 most relevant features of the $F_{\cup 5best}$ subset outperforms the SVM- RFE model that uses 20 features by about 2% in terms of F-score. Thus, this leads us to suggest that these features are an indispensable benchmark for predicting readability (for details, see the feature set in the appendix). Similarly, for the RF classifier, the $F_{\cup 5best}$ model performed better than the SVM-RFE model. However, the best performance was recorded by the $F_{\cup 4methods}$ model, which combines all the features of the four selection methods.

4.2 Feature Selection Based on the GR Corpus

In order to determine whether the conclusions of the previous section remain valid regardless of the

corpus used, we examined the performance of the classification models on the GR corpus. The clustering of selected features generated the sets reported in Table 8.

From the selection results on this second corpus, we notice that 47 features from the initial 70 features were selected, unlike AL which only selected 39 features. The selection methods, based on the two corpora, agreed on a subset composed of 29 features. The 24 features that we mentioned in the appendix appear in this last subset obtained (composed of the 29 features).

Table 8. Number of features for the different combinations of the GR corpus.

Sub-set	**Reference**	**Features**
Features that appear in at least one of the subsets	$F_{\cup 4methods}$	47
Features common to all four subsets	$F_{\cap 4methods}$	3
Features that appear in at least three subsets	$F_{\cap 3methods}$	10
Features that appear in at least two subsets	$F_{\cap 2methods}$	20
Union of the top five features from each subset	$F_{\cup 5best}$	12

The performance of the readability-prediction models obtained using the new selected feature sets is reported in Table 9.

Table 9. Results of the different classification models on GR.

Model	**SVM Classifier**		**RF Classifier**	
	Accuracy	**F-score**	**Accuracy**	**F-score**
BaseLine	63.63%	63.27%	65.90%	66.27%
$F_{\cup 4method}$	52.27%	49.91%	79.54%	79.19%
$F_{\cap 4method}$	59.09%	59.49%	65.90%	63.99%
$F_{\cap 3method}$	50.0%	49.38%	72.72%	71.74%
$F_{\cap 2method}$	52.27%	51.49%	77.27%	76.58%
$F_{\cup 5best}$	**63.63%**	**65.91%**	**84.09%**	**83.11%**

Except for the $F_{\cap 4methods}$ model that contains only three features, the RF classifier applied to the other models provided better performance than the "BaseLine" model. Moreover, the $F_{\cup 5best}$ model achieves the best performance outperforming the "BaseLine" model by about 16% and this confirms the results reported on the AL corpus. Regarding the SVM classifier, we notice that the same model ($F_{\cup 5best}$) also improved the performance by about 2%, while the other selection models recorded a degradation compared to the "BaseLine" model.

Table 10 presents a list of the most discriminating features obtained by the $F_{\cup 5best}$ method, which provided the best performance. This list contains a total of 16 features with their meanings, which were identified by this method using the two corpora (AL/GR).

Table 10. List of features identified by $F_{\cup 5best}$ method using GR and AL corpora.

Feature	**Description**
Lemmas Count	Number of lemmas (without redundancy)
Closed Class Tokens Count	Number of tokens having a PoS tag belonging to a finite grammatical category (such as pronouns)
Nouns Count	Number of nouns
Tokens Count	Number of tokens
FreqLemmas Count	Number of frequent lemmas (lemma that appear more than once in the
AdjOpenClassTokens Ratio	The number of adjectives / the number of open class tokens (an open class token is a token that belongs to illimited grammatical category, such us nouns and verbs)
SL1	Number of sentences
Adverb Count	Number of adverbs

Stems Count	Number of stems
Preposition to Token	Number of prepositions / number of tokens
Chars Count	Number of characters
AMb1	Number of ambiguous lemmas (lemmas having two or more different PoS tags in the same text)
Noun Open Class	Number of nouns / number of open class tokens
Median Dispersion of Frequent Types	Frequent types are lemmas that appear in the frequency dictionary, having each one a dispersion value in the Arabic dictionary. This feature consists in calculating the means of all the retrieved values for all the lemmas composing a text.
Range of ranks of frequent adjectives	Frequent adjectives are ranked in the frequency dictionary; we calculate the range of all the ranks of a text adjectives.
The maximum rank of frequent nouns	For this feature, we get the ranks for all the frequent nouns (appearing in the frequency dictionary) and we get the maximum value. This means that we are getting the position of the most frequent noun in a text in the Arabic language represented by the frequency dictionary.

When analyzing the subsets of features obtained from the two corpora, we found differences in the selected features. This leads us to test a new subset C_U obtained by grouping the features identified by the two corpora. The objective of this new combination is to identify the essential features affecting readability regardless the nature of the corpus used. Thus, if we note F_{C1}/M_{C2}, the model using the C1 corpus to select the features and the C2 corpus to build the classification model (training and testing phases), we performed several experiments by choosing C1 and C2 among the AL and the GR corpora. We report in Table 11 the results of the best performing models only, hence the differences observed (last column) between the subsets of features adopted for the different methods.

Table 11. Results of the selected models by AL and GR.

SVM Classifier			
Experiment	**Accuracy**	**F-score**	**Model**
$\boldsymbol{F_{AL}/M_{AL}}$	**92.30%**	**88.80%**	$F_{\cup 5best}$ (12 features)
$\boldsymbol{F_{GR}/M_{AL}}$	88.46%	84.93%	$F_{\cap 4methods}$ (10 features)
$\boldsymbol{C_U/M_{AL}}$	90.38%	88.03%	$F_{\cap 3methods}$ (16 features)
$\boldsymbol{F_{AL}/M_{GR}}$	**68.18 %**	**69.72%**	$F_{\cap 4methods}$ (6 features)
$\boldsymbol{F_{GR}/M_{GR}}$	63.63%	65.91%	$F_{\cup 5best}$ (12 features)
$\boldsymbol{C_U/M_{GR}}$	63.63%	65.17%	$F_{\cap 4methods}$ (7 features)
RF Classifier			
Experiment	**Accuracy**	**F-score**	**Model**
$\boldsymbol{F_{AL}/M_{AL}}$	**90.38%**	**86.77%**	$F_{\cup 4methods}$ (39 features)
$\boldsymbol{F_{GR}/M_{AL}}$	**90.38%**	**86.77%**	$F_{\cap 4methods}$ (3 features)
$\boldsymbol{C_U/M_{AL}}$	88.46%	84.93%	$F_{\cap 3methods}$ (16 features)
$\boldsymbol{F_{AL}/M_{GR}}$	79.54 %	78.99%	$F_{\cup 4methods}$ (39 features)
$\boldsymbol{F_{GR}/M_{GR}}$	**84.09%**	**83.11%**	$F_{\cup 5best}$ (12 features)
$\boldsymbol{C_U/M_{GR}}$	79.54%	79.19%	$F_{\cup 5best}$ (12 features)

We observe that, for the SVM classifier, the best performance was reported by the F_{AL}/M_{AL} model built from the AL corpus and based on the features selected in this same corpus. Indeed, the substitution of these features by those selected for GR does not lead to an improvement of the performance (F_{GR}/M_{AL}), while the implementation of the features selected by AL on the GR corpus (F_{AL}/M_{GR}) outperforms the results provided by the F_{GR}/M_{GR} model by about 4% in terms of F-score. These results confirm the relevance of the features selected in the previous experiment based on the AL corpus. These observations do not hold for the RF classifier. The best model was obtained by applying the three features selected by GR and trained/evaluated on AL

($F_{GR/MAL}$). These features are included in the set of 39 features that gave the same results ($F_{AL/MAL}$). Finally, we notice that the fusion of the features selected by the two corpora did not yield any improvement whatever the corpus used. From these results, we can conclude that the best feature sets are those selected by the AL corpus and reported in the appendix.

5. DEEP-LEARNING-BASED EXPERIMENTS

In parallel with the use of linguistic characteristics to model readability, the representation of a text as an embedding vector using neural models represents another alternative proposed by researchers [16].

We select in this section a number of Arabic transformer models that generate sentence-embedding vectors:

1) AraBERT: is a pre-trained Bert Embeddings model available at "Hugging face transformers"[7]. AraBERT was trained using a combination of 70 million sentences from different resources (Arabic Wikipedia, Arabic corpus [28], …etc.). This model contains both MSA and Dialectal Arabic (DA). We have used the AraBERTv0.1-base model with the same parameters used in [30].
2) ArabicBERT: is another pre-trained BERT model[8] using a concatenation of a filtered subset from "Common Crawl" and a dump of Arabic Wikipedia totaling 8.2 billion words. Different pre-trained BERT models are used. We have selected in this study, the bert-base-Arabic model.
3) XLM-R [29]: is a multi-lingual version of BERT model, trained on "Common Crawl" data instead of Wikipedia with slightly different parameters. The model is trained on 100 different languages and contains approximately 2.5 TB sentences. The main parameters for learning cross-linguistic representations are those mentioned in [29].

These pre-trained models are used in this experiment to represent the text by sentence embedding. Each of them is compared with the models based on the linguistic features studied in the previous subsection. In addition, we investigated whether combining our linguistic feature set with deep-learning models could improve the performance of readability assessment. For this purpose, we have combined the most discriminating feature sets with different sentence-embedding models.

We first measured the performance of prediction models based on information-rich sentence embeddings as a separate feature set (AraBert, ArabicBert and XLM-R) and we compared them with the baseline model based on the 70 linguistic features.

To evaluate these models, we used the GR corpus according to the same distribution adopted in the previous experiments. The classification for the four models was performed with the RF classifier. Table 12 presents the accuracy and F-score results of these models.

Table 12. Results using the initial linguistic features and sentence embedding.

Model	Accuracy	F-score
BaseLine	65.90%	66.27%
AraBert	77.85%	78.34%
ArabicBert	**78.24%**	**79.36%**
XLM-R	71.60%	73.08%

All the three neural models trained using contextual sentence embedding outperformed the baseline model based only on hand-crafted linguistic features. Moreover, the ArabicBert model reported the best performance with an F-score about 13% higher than that of the baseline model. Based on these results, we hypothesize that the semantic and syntactic knowledge implicitly encoded in BERT embeddings can be considerably more informative than traditional linguistic features in predicting reading difficulty. Therefore, this is a likely alternative for low-resource languages that possess limited or no NLP tools, such as a parse tree extraction tool and a semantic parser. On the other hand,

[7] https://huggingface.co/aubmindlab/bert-base-arabert (last visited 09/10/2022)

[8] https://huggingface.co/asafaya/ (last visited 09/10/2022)

the best performance reached an accuracy of 78.24%, given the limitation of the texts used in the training phase, which is also accentuated when using artisanal linguistic features.

We have evaluated the impact of combining linguistic features with the information obtained by the neural models. We thus adopted a simple approach that consists in considering the numerical output of the neural model as a new feature that is incorporated into the linguistic features. We performed tests on the best combinations of the feature subsets identified in the experiments of the previous sections (F∩3methods, F∪4methods and F∪5best). The performances of the readability-prediction models relating to these new combinations are presented in Table 13.

Table 13. Combination of different sets of features and sentence-embedding models.

Model	AraBert		ArabicBert		XLM-R	
	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score
CombinedBaseLine	82.63%	83.39%	81.60%	82.27%	76.00%	77.57%
$F_{\cap 3methods}$	**82.67%**	**83.49%**	**82.32%**	**83.30%**	**77.51%**	**78.96%**
$F_{\cup 5best}$	82.65%	83.28%	81.94%	82.80%	76.75%	78.08%
$F_{\cup 4methods}$	81.17%	82.36%	82.67%	83.18%	75.64%	77.08%

We point out that in general, the combination of linguistic features with the numerical output of the neural model improves the performance of classification based only on linguistic features. Indeed, by comparing the results of Table 9 and Table 13, we find that, in terms of F-score, the performance of the classification based on the combination of linguistic features with the output numerical of one of the two neural models AraBert and ArabicBert has improved for all models.

Similarly, the results of Table 12 and Table 13 suggest that these combined models perform better than those using only the outputs of the neural models. The best performances are obtained by combining the numerical output of the AraBert model with the subset of linguistic features F∩*3methods*.

In order to analyze the results in more detail, we computed the confusion matrix of the best classification model that combines F∩*3methods* with AraBert. It is clear from Table 14 that there is a separation between the prediction levels (easy, intermediate and difficult). Indeed, all the errors are located between the neighboring levels.

Table 14. Confusion matrix of $F_{\cap 3methods}$ model incorporated with AraBert.

	Easy	Intermediate	Difficult
Easy	6	1	0
Intermediate	0	7	2
Difficult	0	1	10

Finally, we conclude that the F∩*3methods* model is composed of the most informative features compared to the other models, which in turn provides an indispensable benchmark for predicting readability.

6. CONCLUSION

In this study, we presented a set of experiments related to the selection of the most informative features for Arabic L2-readability measurement. The obtained test results showed that feature selection is a useful pre-processing tool that not only reduces the number of input features, but also increases the performance of prediction models.

We considered the three main types of selection methods (filter, wrapper and embedded) commonly used to retrieve the most discriminating features. We have thus applied these methods to the two corpora AL and GLOSS. Next, we examined the impact of each selection method on the performance of readability-prediction model and we compared it with the performance of the baseline model using 70 features.

We also demonstrated the relevance of the combination of features selected by the different methods; namely, "F∩*4methods*", "F∩*3methods*", "F∩*2methods*", "F∩*5best*" and "F∪*4methods*". These combinations

were also evaluated on the grouping of the features selected by the two corpora. The performance of all the experiments was evaluated using the two RF and SVM classifiers.

The use of sentence embedding is another approach used to represent a text. The last experiment of this work aimed to investigate the influence of these methods in the field of readability assessment. Thus, we have evaluated the incorporation of the best features that we have selected with these representations. In the case of limited corpus size, we have shown that the best results are obtained by incorporating both types of representation (sentence embedding and hand-crafted linguistic features).

Concerning our future directions of investigations, we are considering the evaluation of readability based on attested/evaluated reading comprehension of human readers. This is due to the fact that the actually available readability corpora are labeled based on experts' opinion, but not on reading comprehension tests. Gathering a large amount of training data, to enhance the performance of sentence embedding-based models and adopt deep-learning classifiers is also one of our future work paths. The readability of Arabic as a native language (L1) is also an emerging field of research; so it will be interesting to examine the impact of feature-selection methods on L1 and compare the results with those on L2, to check whether the L1 and L2 reading learners share common discriminating readability features.

REFERENCES

[1] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh and M. N. Al-Kabi, "A Comprehensive Survey of Arabic Sentiment Analysis," Inf. Processing and Management, vol. 56, no. 2, pp. 320–342, 2019.

[2] S. Berrichi and A. Mazroui, "Addressing Limited Vocabulary and Long Sentences Constraints in English–Arabic Neural Machine Translation," Arabian J. for Science and Engineering, vol. 46, pp. 8245–8259, 2021.

[3] N. Nassiri, A. Lakhouaja and V. Cavalli-Sforza, "Arabic L2 Readability Assessment: Dimensionality Reduction Study," J. of King Saud Uni., Comp. and Inf. Sci., vol. 34, pp. 3789–3799, 2022.

[4] V. Cavalli-Sforza, H. Saddiki and N. Nassiri, "Arabic Readability Research: Current State and Future Directions," Procedia Computer Science, vol. 142, pp. 38– 49, 2018.

[5] T. Deutsch, M. Jasbi and S. Shieber, "Linguistic Features for Readability Assessment," Proc. of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–17, Association for Computational Linguistics, Seattle, USA, 2020.

[6] D. D. Lewis, "Challenges in Machine Learning for Text Classification," Proc. of the 9th Annual Conference on Computational Learning Theory, pp. 1–ff, 1996.

[7] X.-D. Wang, R.-C. Chen, F. Yan, Z.-Q. Zeng and C.-Q. Hong, "Fast Adaptive K-Means Subspace Clustering for High-dimensional Data," IEEE Access, vol. 7, pp. 42639 – 42651, 2019.

[8] I. Guyon, S. Gunn, M. Nikravesh and L. A. Zadeh, Feature Extraction: Foundations and Applications, ISBN: 978-3540354871, Springer, 2008.

[9] J. N. Forsyth, Automatic Readability Prediction for Modern Standard Arabic, Ph.D. Thesis, Department of Linguistics and English Language, Brigham Young University, USA, 2014.

[10] V. Cavalli-Sforza, M. El Mezouar and H. Saddiki, "Matching an Arabic Text to a Learners' Curriculum," Proc. of the 5th Int. Conf. on Arabic Lang. Process. (CITALA), p. 10, Morocco, 2014.

[11] H. Saddiki, K. Bouzoubaa and V. Cavalli-Sforza, "Text Readability for Arabic as a Foreign Language," Proc. of the 2015 IEEE/ACS 12th Int. Conf. of Computer Systems and Applications (AICCSA), pp. 1–8, Marrakech, Morocco, 2015.

[12] H. Saddiki, N. Habash, V. Cavalli-Sforza and M. Al-Khalil, "Feature Optimization for Predicting Readability of Arabic l1 and l2," Proc. of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 20–29, DOI: 10.18653/v1/W18-3703, Melbourne, Australia, 2018.

[13] N. Nassiri, A. Lakhouaja and V. Cavalli-Sforza, "Modern Standard Arabic Readability Prediction," Proc. of the Int. Conf. on Arabic Language Processing (ICALP 2017), Part of the Communications in Computer and Information Science Book Series, vol. 782, pp. 120–133, 2018.

[14] N. Nassiri, A. Lakhouaja and V. Cavalli-Sforza, "Arabic Readability Assessment for Foreign Language Learners," Proc. of the Int. Conf. on Applications of Natural Language to Information Systems (NLDB 2018), vol. 10859, pp. 480–488, 2018.

[15] B. W. Lee, Y. S. Jang and J. H.-J. Lee, "Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features," Proc. of the 2021 Conf. on Empirical Methods in Natural Lang. Process., pp. 10669–10686, DOI: 10.18653/v1/2021.emnlp-main.834, Dominican Rep., 2021.

[16] N. Khallaf and S. Sharoff, "Automatic Difficulty Classification of Arabic Sentences," Proc. of the 6th Arabic Natural Language Processing Workshop, Association for Computational Linguistics (WANLP 2021), pp. 105–114, DOI: 10.48550/arXiv.2103.04386, 2021.

[17] V. N. Vapnik, "Controlling the Generalization Ability of Learning Processes," Chapter 4 in Book: The Nature of Statistical Learning Theory, pp. 89–118, Springer, 2000.
[18] O. Al-Harbi, "A Comparative Study of Feature Selection Methods for Dialectal Arabic Sentiment Classification Using Support Vector Machine," Int. J. of Computer Science and Network Security, vol. 19, no. 1, pp. 167-176, January 2019.
[19] H. U˘guz, "A Two-stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm," Knowledge-based Systems, vol. 24, no. 7, pp. 1024-1032, 2011.
[20] M. A. Hall, Correlation-based Feature Selection for Machine Learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, New Zealand, 1999.
[21] S. Bahassine, A. Madani, M. Al-Sarem and M. Kissi, "Feature Selection Using an Improved Chi-square for Arabic Text Classification," J. of King Saud Uni.-Comp. and Inf. Sci., vol. 32, no. 2, pp. 225–231, 2020.
[22] R. Elhassan and M. Ali, "The Impact of Feature Selection Methods for Classifying Arabic Texts," Proc. of the 2nd Int. Conf. on Comp. App. Inf. Secur. (ICCAIS), pp. 1–6, Riyadh, KSA, 2019.
[23] A. Elnahas, N. Elfishawy, M. Nour and M. Tolba, "Machine Learning and Feature Selection Approaches for Categorizing Arabic Text: Analysis, Comparison and Proposal," The Egyptian Journal of Language Engineering, vol. 7, no. 2, pp. 1–19, 2020.
[24] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, vol. 46, no. 1, pp. 389–422, 2002.
[25] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," Proc. of the 5th Annual Workshop on Computational Learning Theory, pp. 144–152, DOI: 10.1145/130385.130401, 1992.
[26] J. L. D. Clark and R. T. Clifford, "The FSI/ILR/ACTFL Proficiency Scales and Testing Techniques: Development, Current Status and Needed Research," Studies in Second Language Acquisition, vol. 10, no. 2, pp. 129–147, 1988.
[27] A. Pasha, M. Al-Badrashiny, M. T. Diab et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14), vol. 14, pp. 1094–1101, Reykjavik, Iceland, 2014.
[28] I. A. El-Khair, "1.5 Billion Words Arabic Corpus," CoRR abs/1611.04033, arXiv: 1611.04033, DOI: 10.48550/arXiv.1611.04033, 2016.
[29] A. Conneau, K. Khandelwal, N. Goyal et al., "Unsupervised Cross-lingual Representation Learning at Scale," Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451, DOI: 10.18653/v1/2020.acl-main.747, 2020.
[30] A. Safaya, M. Abdullatif and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," Proc. of the 14th Workshop on Semantic Evaluation, pp. 2054–2059, DOI: 10.18653/v1/2020.semeval-1.271, Barcelona, Spain, 2020.
[31] S. Al-Aqeel, N. Abanmy, A. Aldayel, H. Al-Khalifa, M. Al-yahya and M. Diab, "Readability of Written Medicine Information Materials in Arabic Language: Expert and Consumer Evaluation," BMC Health Services Research, vol. 18, DOI: 10.1186/s12913- 018-2944-x, 2018.
[32] E. Halboub, M. S. Al-Ak'hali, H. M. Al-Mekhlafi and M. N. Alhajj, "Quality and Readability of Web-based Arabic Health Information on COVID-19: An Infodemiological Study," BMC Public Health, vol. 21, no. 1, pp. 1–7, 2021.
[33] Z. Jasem, Z. AlMeraj and D. Alhuwail, "Evaluating Breast Cancer Websites Targeting Arabic Speakers: Empirical Investigation of Popularity, Availability, Accessibility, Readability and Quality," BMC Medical Informatics and Decision Making, vol. 22, no. 1, pp. 1–15, 2022.
[34] W. Daelemans, J. Zavrel, K. van der Sloot and A. van den Bosch, "TiMBL: Tilburg Memory Based Learner," Technical Report, Version 6.3, ILK Research Group Technical Report.

APPENDIX: LIST OF FEATURES

Table 15. Common features between IG, IF-RF, IF-SVM, SVM-RFE and $F_{\cup 5best}$.

Category	Features	IF-RF	IF-SVM	IG	SVM-RFE	$F_{\cup 5best}$
RTF	Characters Count	✓		✓		
	Token Count	✓		✓	✓	✓
	Sentences Count		✓	✓		
MF	Stem Count	✓		✓		✓
	Lemma Count	✓	✓	✓	✓	✓
	Ambiguous Lemma Count	✓	✓	✓		
PDF	Noun Count	✓	✓	✓	✓	✓
	Verb Count	✓		✓		
	Adverb Count	✓	✓	✓	✓	✓
	Proper Nouns Count	✓	✓	✓	✓	
	Adjectives Count	✓	✓	✓		

	Closed-Class Tokens Count	✓	✓	✓	✓	✓
	Noun / Open Class Token	✓	✓	✓	✓	
	Adjectives to Open Class Token Ratio		✓	✓	✓	✓
	Preposition to Token Ratio				✓	✓
FLF	Frequent Lemmas Count	✓		✓	✓	✓
PBF	Ratio of frequent subordinating conjunctions to total subordinating conjunctions		✓		✓	
	Ratio of frequent prepositions to total prepositions		✓	✓	✓	
	Ratio of frequent demonstrative pronouns to total demonstrative pronouns			✓	✓	
	Maximum rank of frequent conjunctions		✓		✓	
	Maximum rank of frequent subordinating conjunctions		✓		✓	
	Average rank of frequent subordinating conjunctions		✓	✓		
	Median rank of frequent particles		✓		✓	
	Minimum rank of frequent proper names		✓	✓		
	Range of ranks of frequent adjectives		✓		✓	✓
	The maximum rank of frequent nouns					✓

ملخص البحث:

إنّ مقروئيّة النّصوص تمثّل أحد مجالات البحث التي تطوّرت على نطاقٍ واسع فيما يتعلّق بالعديد من اللُّغات، غير أنها محدودة الى حدّ كبير عندما يتعلّق الأمر باللّغة العربية. ويكمن التّحدي الرئيسي في هذا المجال في تحديد مجموعة مثالية من السِّمات التي تمثّل النّصوص وتسمح لنا بتقييم مُستوى مقروئيّتها. وللتّعامل مع هذا التّحدي، نقترح في هذه الدّراسة طرقاً متنوعة لانتقاء السِّمات تُمكّننا من استرجاع مجموعة السِّمات المميّزة للنّصوص باللّغة العربية. ويتمثّل الهدف الثاني من هذه الدراسة في تقييم طُرق مختلفة لتّضمين الجُمل ومقارنة أداءاتها بتلك التي يتمّ الحصول عليها من السِّمات اللُّغوية المُنتقاة. فقد أجرينا تجارب باستخدام مُصنِّفات (SVM) و (RF) على اثنتين من مجموعات البيانات المخصصة لتعلُّم اللُّغة العربية كلُغة ثانية (L2).

وتكشف النّتائج التي تمّ الحصول عليها أنّ تقليل عدد السِّمات يعمل على تحسين أداء نماذج توقُّع المقروئيّة بنسبةٍ تتجاوز 25% و 16% لمجموعتي البيانات المستخدمتيْن، على الترتيب. بالإضافة الى ذلك، يعمل نموذج (BERT) المزوَّد بإمكانية الضّبط الدّقيق والمخصَّص للُّغة العربية بأداءٍ أفضل مقارنةً بالنّماذج الأخرى المستخدَمة لتضمين الجُمل، لكنّه أدّى الى تحسُّن أقلّ عند مقارنته بالنّماذج القائمة على السِّمات. وقد أدّى الجَمع بين هذه الطّرق والسِّمات الأكثر تمييزاً للنّصوص الى الحصول على أفضل النّتائج.