International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 4, August 2023, pp. 4441~4451
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i4.pp4441-4451

Bayes model for assessing the reading difficulty of English text for English education in Jordan

Yasser Qawasmeh[1], Qasem Al-Radaideh[2], Addy AlQuraan[3], Ahmed Fawzi Otoom[4]
[1]Department of Computer Science and Applications, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, Zarqa, Jordan
[2]Department of Information Systems, Faculty of Information Technology and Computer Science, Yarmouk University, Irbid, Jordan
[3]Department of Basic Sciences, Faculty of Science, The Hashemite University, Zarqa, Jordan
[4]Department of Software Engineering, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, Zarqa, Jordan

Article Info

Article history:

Received Sep 7, 2022
Revised Nov 28, 2022
Accepted Dec 7, 2022

Keywords:

Bayes model
Predicting reading difficulty level
Readability assessment
Reading difficulty
Statistical language model

ABSTRACT

Predicting the reading difficulty level of English texts is a critical process for second language education and assessment. Reading difficulty level is concerned with the problem of matching a reader's proficiency and the appropriate text. The reading difficulty level or readability assessment is the process for predicting the reading grade level required from an input text or document, which corresponds to the reader and to the materials. Students in Jordan at their academic levels find obstacles in finding relevant readable data for any subject at their levels. This paper is intended to introduce a model that foretells the reading difficulty level of a given text in terms of a student's ability to read and understand English as a non-native English speaker in Jordanian schools. In this paper, Jordanian students were classified into four categories according to their knowledge of English. The prediction of the reading difficulty level is achieved by using a modern statistical model that is situated on the Bayes model. The model compares the given text with some standard predefined text that strongly reflects the ability to read and understand English text. The accuracy of the proposed model was tested using the hold-out method. The overall prediction accuracy was 75.9%.

Corresponding Author:

Yasser Qawasmeh
Department of Computer Science and Applications, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology
The Hashemite University, Zarqa, Jordan
Email: yasser@hu.edu.jo

1. INTRODUCTION

A readability or a reading difficulty assessment can be defined as a tool or a model that maps a given text to a statistical value matching a difficulty level. The text difficulty level or readability can be defined as the ability of a reader to understand a written text in natural language [1]–[4]. Klare [5] defines readability as "the ease of understanding or comprehension due to the style of writing". Readability is frequently demented with legibility related to the typeface and layout [6]. Separately, this definition concentrates on the writing style more than the issues and this writing style includes content, coherence, and arrangement. Similarly, Hargis and her colleagues at IBM (1998), as described in [6], it is stated that readability is the "ease of reading words and sentences", as an attribute of clarity.

The main objective of readability is to determine whether the text or reading material matches the reader's reading ability level [7], [8]. The main goal of readability research is to find a method, which

matches the target reader's proficiency with a text or a reading passage in the targeted language. As a result, this will support language learning skills for the readers and the text providers such as web pages and social media applications. In addition, judging the readability of text has many important applications such as when performing text simplification or when sourcing reading material for language learners.

Text difficulty assessment can also provide a basis for assessing whether teaching materials such as language tests are suitable for teaching at the current stage, which is helpful to the evaluation of teaching materials and improves learning efficiency and teaching quality. In text classification, a language model corresponding to a predefined level of difficulty describes each class. The ability to accurately and consistently assess the readability level of texts is crucial to teachers, as it allows them to create and discover content that meets the needs of students with different backgrounds and skill levels, different algorithms are used for text classification, naïve Bayesian algorithm (NB), k-nearest neighbor algorithm (KNN), support vector model (SVM), and artificial neural network (ANN). For the last three decades, natural language processing (NLP) tasks have been intensively studied such as text classification, image caption, and some task-specific translations such as sign language translation, where the need for intelligent systems that precise the visual environment and understand the linguistical information become a demand [9].

Due to the complexity of data on the internet, a high data dimension challenge arises. To simplify the text classification task a data dimension reduction technique is needed with less computing power, shorter processing time, and lower error rate. On the other hand, tasks on smaller datasets are simple and can be processed by linear models such as the naïve bayes classifier [10].

Technically, the reading difficulty level is concerned with the problem of matching between a reader and an appropriate text. As described in [11], [12], the difficulty level of a given text that readers can understand is (in some measure) determined by the reader's vocabulary knowledge. The reading difficulty level or readability assessment is the process for predicting the reading grade level required from an input text or document that corresponds to the reader to understand written materials [13]. Readability evaluation measures only the structural complexity (e.g., new word and sentence length) of the written text. It does not measure other aspects associated with readability, such as the arrangement of material, content, conceptual complexity, or reader characteristics. More than 40 different formulas can be used to carry out a readability assessment. Most of them use counts of language variables, such as word and sentence length, to provide an index of probable reading difficulty.

Access to the internet is growing as a valuable source of information and knowledge gaining. In Jordan, the Ministry of Education (MOE) and universities are motivating students at different academic levels into this new era of knowledge by encouraging students to use the Internet as the main source of information. All schools and universities in Jordan have now become well-equipped with computers and Internet services. Unfortunately, much of the e-text available on the Internet and the web-based resources are in English language. It is well understood that English as a universal language has become the language of choice in terms of knowledge and education regardless of the chosen subjects. Moreover, it is worth mentioning that English is a foreign language in Jordan. However, it has been noticed that students in Jordan, at their various academic levels, find obstacles in finding relevant texts that match their reading skill levels.

Predicting the reading abilities of Jordanian students was really a troublesome issue. Students in Jordan are no longer solely dependent on their books for acquiring the language. Nowadays, almost everyone has access to satellite channels, which broadcast in English as well as access to the Internet. However, those external resources are not reliable in determining their knowledge of English, and they are often a source of hearing rather than reading. This leads to the basis of defining average Jordanian students who tend to rely on their books in the process of learning English.

As in educational environments, one of the top concerns should be the reading difficulty level when instructors have the ability to read the text as a teaching resource. To provide a text that is appropriate to the student's level, instructors are advised to check if a given passage can be easy to read using a specific tool by its intended readers or the targeted students. Several factors may affect the reading difficulty level of a given text. These factors include sentence extent, new vocabulary items, and structural complexity. In addition, text readability can be considered as an important factor that affects students' abilities to read a particular text.

The purpose of this research is to provide students and instructors with a model that could accurately predict and classify the reading difficulty of the retrieved text. This model will allow a search engine to retrieve suitable text that matches the user's reading skills level. Hence, the proper prediction of texts according to a student's level is the major concern of this research. This paper is concerned with providing an easily applicable model that could help students and instructors to find the appropriate text that suits their reading level. The main contribution of this work is the proposal of the dataset of English text from English books in Jordan at different levels. Moreover, we prove the strength of this dataset by experimenting with a Bayesian statistical model and achieving high-performance measures. The rest of this paper is organized as:

In section 2, literature work is provided. Section 3 describes the applied research methodology. Section 4 shows the experimental result. Finally, section 5 presents the project conclusion.

2. RELATED LITERATURE

Several proposed approaches in the literature are used to test and predict the readability difficulty of a given text. The Flesch-Kincaid test is from the traditional readability tests determined on having a sufficiently huge sample size of consistent semantic and syntactic characteristics [14]. Although this measure is viewed to be effective with essays or textbooks, it is not as much of efficient for web pages. Some of the approaches use the average sentence length and average word length to represent the reading difficulty [15]. Other techniques build graphs representation that is used to predict reading difficulty.

Traditional text classification approaches such as NB and SVM are mostly linear or based on bag of words (BoW) representation, which does not consider the position of information of words in sentences. The word position in a sequence can be better handled by the convolutional neural network (CNN) solutions since they process the input data in sequential order such as the char-CNN approach. Recurrent neural network (RNNs) offers another neural network (NN)-based solution for text classification, for a well-designed representational vector, the computing unit (or the memory cell) can exploit the word-level dependency to facilitate the final classification task. An example of NN-based solution called FastText, which offers a performance solution to text classification, while the NN based text classification approaches demand labeled data in training [16]. When the text classification tasks demand low-dimensional and high informative data, a data dimensional reduction technique that should be used must reduce the input data dimension with a small model size at low computational complexity, while maintaining high mutual information between an input and its output, which preserves the positional relationship between input elements and do not demand labeled training data [17].

A data dimension reduction solution proposed by [16] called the tree-structured multi-stage principal component analysis (TMPCA) for text classification task. TMPCA is a multistage principal component analysis (PCA) in a special form with orthonormal rows transform matrix, which maximizes the mutual information between its input and output. A dense network trained on the TMPCA showed a better performance than FastText, char-CNN, and long short-term memory (LSTM) in quite a few text classification datasets. In NLP, dimension reduction is often required to alleviate the so-called "curse of dimensionality" problem. There are many ways to reduce the language data to a compact form, the most popular ones are the NN based techniques, however, they are limited in modeling "sequences of words", which is called the sequence-to-vector (seq2vec) problem. The TMPCA method manages dimension reduction at the sequence level without labeled training data, while it conserves the sequential structure of input sequences, which is beneficial for text classification tasks [17].

The RNN has been evidenced as an efficient solution for NLP, where various models have been proposed such as the bidirectional RNN, the encoder-decoder, the deep RNN, and extended long short-term memory (ELSTM) [18]. Where all models try to solve the sequence-in-sequence-out (SISO) problem. The RNN constructs an interior representation of semantic patterns; the memory of a cell gives it the capacity of mapping an input sequence of a certain length into such a representation, which serves as a function that maps an element in a sequence to the current output. At the cell level of RNN, the LSTM and the gated recurrent unit (GRU) are mostly selected by RNN as their low-level building [19]. Dealing with complex language tasks that require long memory responses such as sentence parsing. LSTM's and GRU's memory decay may have a significant impact on their performance, while ELSTM does not suffer from memory decay and delivers better results [20].

Deep learning multi-modal systems aim to utilize the data in diverse forms on the Internet, which handles data of various modalities, such as image caption generation system where it takes its input in the form of images and generates its output in form of sentences. In image caption processes such as object detection and image segmentation face many challenges, a fast and strong camera's auto exposure (AE) control algorithm was proposed in [21], [22]. In deep learning multi-modal system, a CNN submodule task is to process the input image into a multi-dimensional tensor, which is used as a representative feature of the input image and then fed into an RNN to generate the descriptive sentence that produced a better translation of words in different languages with similar visual appearances, which gives RNNs' the ability to describe the content of images and videos [20].

The simple measure of gobbledygook (SMOG) readability formula is a simple method that can be used to determine the reading level of the written materials. In [23] the SMOG readability formula is used to build a SMOG-based reading level calculator. The SMOG formula starts by counting 10 sentences from the beginning of the given passage, another 10 sentences from the middle of the passage, and then counts 10 sentences from the end. In the second step, the formula counts every word with three or more syllables for each of the 30 sentences and then sums the total number of the counted words. Finally, it uses a conversion

table to estimate the grade or difficulty level of the passage. The Flesch-Kincaid grade level proposed by [14] is similar to what the SMOG readability formula does; however, it gives a number that corresponds to the grade level a person needs to have reached to understand a given text. For example, grade level scores of eight mean that an eighth grader will be able to understand the text.

Fog-Index proposed in [24] is another readability measure, and it is defined as one of the best-known measures that are used to count the struggling level of reading difficulty for documents. Fog Index level indicates the number of education years a reader needs to comprehend a given text material. The perfect score is 7 or 8; anything above 12 is classified to be hard for people to read. The index does not take into consideration if the written texts or paragraphs are too simple or too advanced for a particular reader.

Liu *et al.* [25] proposed and evaluated a SVM technique to automatic recognition of reading levels from user interrogations. As outcome of the proposed technique showed that the SVM performs much fitter than the standard readability index where it has the ability to carry off acknowledgment accuracy close to, or more than, 80% for the 2-category and the 3-category cases. The authors have shown that prospective improvements in retrieval performance, especially when searching over the web, can be accomplished by matching up the interrogations and documents not only by relevant documents but also by the level selected.

Moreover, other researchers used a combination of several features. For instance, in [26] Schwarm and Ostendorf used SVM to combine features from traditional analyzing level measures, statistical language fashions, and different language processing equipment to provide a better approach of assessing studying degrees. Other researchers build prediction models and study how the selection of features affects the version performance, and they used grade degrees, which indicate the wide variety of years of training required to completely recognize a textual content, as a proxy for studying trouble [27].

While Zamanian and Heydari [28] provided the readers in this domain with the most frequently used readability formulas and the background as well as the pros and cons views toward the use of such formulas. An approach was developed by [4] for Arabic language called automatic Arabic readability index (AARI) using factor analysis (average chars per word, average words per sentence, average of difficult words, number of characters, number of words, number of sentences, and number of difficult words), an 1,196 Arabic texts where extracted from the Jordanian curriculum in various subjects for elementary classes (first grade through tenth grade), then compared with automated readability index (ARI) for English language, Lesbarheds index (LIX), and Al-Heeti readability formula for Arabic language.

Nurhamsih [29] conducted a research study of knowing the ability level in reading text in the textbook pronounced whether the reading passages are fitting semantically for the third-year students of senior high school (SMA) or not. The research found that the assigned textbook does not fit linguistically for the SMA students' third year in reason; the students have been learning English for nine years. However, Srisunakrua and Chumworatayee [30] mixed the levels of readability with the linguistic features of reading texts in English textbooks and the Thai National Education English Test, where eight features of linguistic features and three readability methods are used as provided by the computational tool Coh-Metrix. The results revealed a mismatch in the readability levels and linguistic characteristics.

Likewise, Yulianto [7] presented a qualitative study that analyzed the English level readability in reading passages (Pathway to English 2 Textbook) with the Flesch-Kincaid readability formula using the Eighth grade of younger secondary school students. The reported outcomes revealed that only one passage was suitable for seven or eighth grade of younger secondary school students but for elementary school students there were six passages fit for them. New formulas for readability developed by [31] are situated on modern NLP tools for reading comprehension and its rate. The method is rooted on linguistic characteristics that represent the abstract and observable description of the reading process, importantly, classic readability method is outperformed.

Zhang in [32] analyzed the readability of a series of course books volumes 1-4 of new target college English integrated course from the perspective of vocabulary and syntax to verify whether the compilation of the course books obeys the rule of text difficulty development from low to high. The WE Research platform was used to analyze the text readability and calculate the average word length and sentence length of the text using readability indicators namely Flesch reading ease, Flesch-Kincaid grade level, ARI, Coleman-Lian readability score, gunning fog, and the SMOG readability index.

3. METHOD

To achieve the objective of this research, we classified Jordanian students into four categories with respect to their knowledge of English. The prediction of the difficulty is achieved by using a modern statistical model based on bayes model. The model compares the given text with some standard predefined text that strongly reflects the ability to read and understand an English text. The main phases and steps of the proposed approach are depicted in Figure 1.

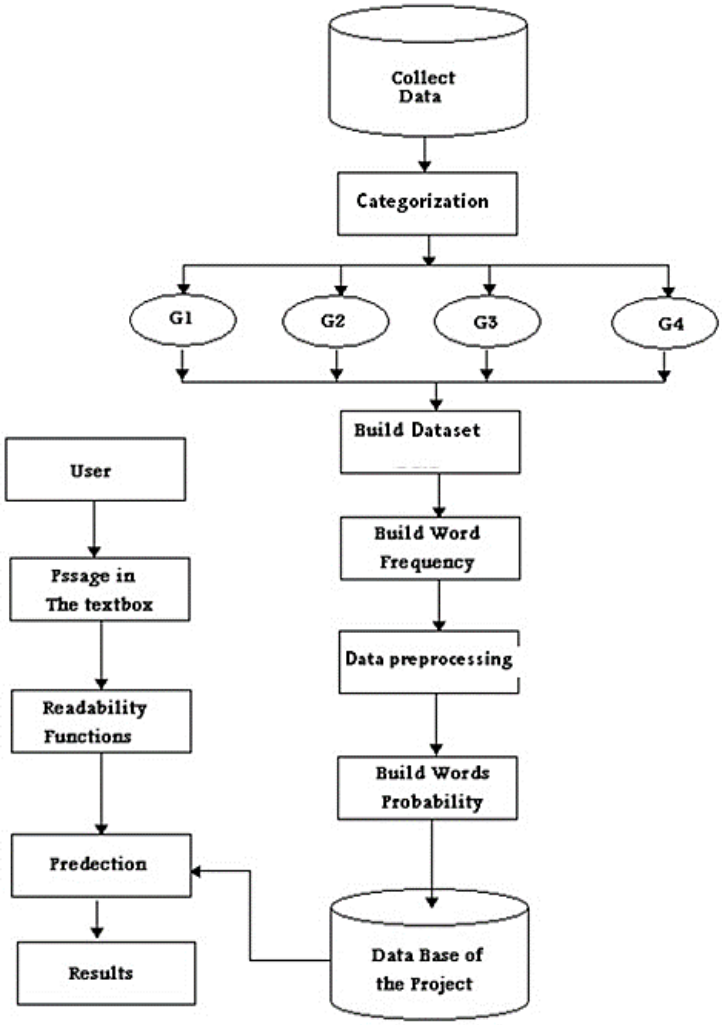

Figure 1. The proposed approach

3.1. Building the dataset phase

The first step of the proposed approach mainly depends on building a relatively acceptable size of standard text datasets to be used in the training phase of the approach. Acquiring standard text datasets was not an easy task since there are no predefined standard text datasets dedicated for assessing reading difficulty level. The problem was in finding some educational resources to be used as a standard dataset that can reflect the capabilities and the performance of most Jordanian students in reading English.

At that time, several questions were raised in mind. For instance, should the research be concerned only with students in public schools? what really reflects the performance of students in English? should the research be concerned with all Jordanian students, or is it enough to consider the average students? in addition, if we are going to do so, how is the proposed method going to define the performance and reading capabilities in English for an average student? should we make two separate models for students of public and private schools? on top of that, are all students in private schools having the same performance in English? the answers to these questions were the key to establish a scheme in building the model.

Mainly, Jordanian schools can be classified into three major categories with respect to the English curricula:

- Public schools: These include schools of the MOE, United Nations Relief and Works Agency for Palestine Refugees (UNRWA) schools, and schools of military forces. All these schools use the curriculum of MOE without using any additional resources in learning English as a second language. Bear in mind that MOE has achieved a considerable leap in teaching English in its schools since it has introduced English curricula into the first grade.
- Private schools use the curriculum of the MOE as the main resource for teaching English, whether alone or with other resources.

- Private schools that consider English as a language of instruction in all subjects like science, math, and history.

Most students in Jordan attend public schools whereas private schools do not contribute a significant percentage of all schools in Jordan. Therefore, the English text of the private schools of the third category is discarded. All the points mentioned above directed us to use the MOE English textbooks as the main source for building the standard text dataset. The documents used in the training phase of this research have been collected from four groups of schools.

- Group 1: This group represents elementary schools, which include the first six classes of the school. In this category, we encountered a few problems; some were related to the books themselves. There were not enough texts or articles to be used in the standard model. Most of the pages are composed of pictures with only a few comments either under the picture or next to it. Short songs and conversations, which were poor in the number of words, have dominated on a significant number of pages. Some of the lessons were also dedicated to letters and simple words and the way of dictation and pronunciation. However, every single word was used even if it was not in the context of a sentence, a conversation, or a song. At the end, 52 documents were collected for this group.
- Group 2: This group represents the mid-schools, which includes the seventh, eighth, and ninth classes of the school. All words that did not fall into a meaningful sentence were eliminated along with headlines and exercises. 40 documents were collected for this group.
- Group 3: This group represents the high schools, which include the tenth, eleventh, and twelfth classes of the school. At the end, 27 documents were collected for this group.
- Group 4: This group represents graduate and undergraduate students at Jordanian universities. The chosen textbooks were randomly selected from the textbooks used by students of different disciplines. At the end, 20 documents were collected for this group.

The above grouping of students is not random; it was found that students at a certain level as in elementary level, or high school perform in a very similar way in reading English texts. For example, one cannot make a clear distinction between the students of the eighth and the ninth class. Therefore, to decrease the bias and increase the objectivity of the research, the above classification was adopted.

Some arguments have been raised regarding classifying university students into the same group, whether they are graduate or undergraduate students. Again, the bulk of students were undergraduates, and there should not be much difference between their English reading capabilities, that is why the decision was to merge them into one category. Jordanian universities may differ in their curricula and teaching methods, this research has focused only on Yarmouk University students as a representative university for all universities of Jordan.

The trend of decreasing the number of documents going from groups 1 to 4 was not random. The number of words in each document was the main determinant factor for such a trend. It was noticed that as the model goes from groups 1 to 4, the number of words in each document increases. Since the proposed approach is concerned with the total number of words and frequency of occurrence for each word, therefore such a trend was followed to decrease the bias of having very large documents containing a large number of words in any group. The total number of words and documents for each group is illustrated in Table 1.

Table 1. The number of words for each group

Group	Number of words	Number of documents
1	8751	52
2	8763	40
3	8460	27
4	12217	20

3.2. The data preparation phase

The data preparation consists of several steps, which can be summarized as:

Step 1: Tables generation. The texts were categorized according to the four different groups and built one table for each group contained the English word extracted from MOE English textbooks. A second table was generated to contain the word and frequency of each word in each grade. A third table was generated to contain every word probability in each grade that will be described in step 4. Then the tables were used as standards to build a dataset that contains the data to be compared with.

Step 2: Counting word frequency: As mentioned in step 1 the project counted the word frequency in each grade and saved it in a new table and counts the total number of words in each group.

Step 3: Data preprocessing. Preprocessing consists of several processes such as tokenization, stop words removal, and stemming. The first and primary data preprocessing step for many text classifiers is to delete the stop words, but in this research in the lower grades (G1, and G2), stop words (like "in", "a", and "an") make up most token occurrences -as you can see in Figure 2 and deleting them may present bias. We, therefore, decided not to delete stop words.

The second step is to remove all words that have a frequency less than 2 because we did a smoothing technique that will be described later. The third step is stemming, but in this project, the root of the words makes the majority of the token's occurrence. For example, when it is said "become" it will refer to be of grade 2 or grade 1 more than grade 4 but when it is said "became" as you can see in Figure 3 it will refer to be of grade 4 more than grade 2 or grade 1, that the simple word would occur in lowest grades models, so stem the word to its root may introduce bias. Table 2 describes the total number of words for each group after the data preprocessing step.

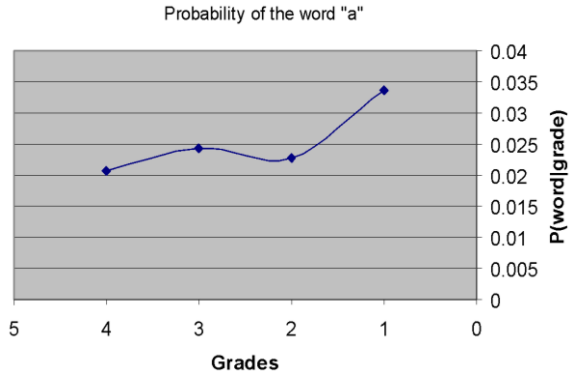

Figure 2. Stop word "a" probability and how it differs over all groups

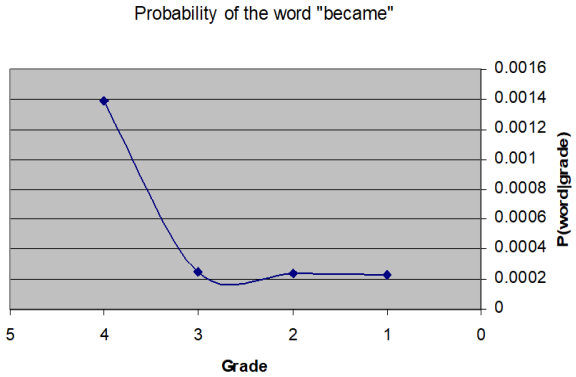

Figure 3. The word "became" probability and how it differs over all groups

Table 2. The number of words for each group after data preprocessing

Group	Number of words
1	8530
2	8428
3	7886
4	11499

Step 4: Computing word probability. A new table was created to save the probability of each word in each group; by dividing the number of occurrences of every word in each group by the total number of all words in each group as in (2).

$$p(w|(G(i)) = \frac{freq}{Total\ number\ of\ all\ words\ in(G(i))} \quad (1)$$

where, $P(w|G\ (i))$ is the possibility of the word for each grade model, $freq$ is the frequency of the word (word counting in each grade), total number of all words in $G(i)$ is the number of tokens (words) in a specific grade, and $G(i)$ is a specific grade model and i varies from 1 to 4.

3.3. Prediction phase

The prediction phase consists of several steps, which can be summarized as:

Step 1: Reading user text: taking the passage (T) from the user.

Step 2: Text tokenization: text tokenization is one of the Text-preprocessing methods that is a sensitive step where text is prepared before the mining process [33], [34]. Text tokenization process eliminating all quotation marks (" '), punctuation marks (?,.,!...) and numbers also converted all capital case letters into small case letters (BOOK to book). The words in the user text are taken separately and the frequency of every word that the user entered is calculated.

Step 3: Computing every word's weight-the approach calculates the weight of every word in every grade according to mathematical equations, which will be discussed in detail later. If the word entered by the user was not found in our training dataset, then we use a smoothing technique. Smoothing is an

unsophisticated method would assign the words that do not exist in the model a possibility of zero that is obviously low if they have even an inaccessible chance of taking place in the underlying language. The project has modified a step that could expand the accuracy of the project estimation, which is called smoothing. In the grades in which the word does not occur, the project will give the word a constant value. The given constant value is calculated by a logical step. The lowest log likelihood of the probability value for words in all grades was found to be -3.75 (Ex. Word "blind" in group 4 has a probability 0.0001739 then the log likelihood is -3.75). The logic step is by multiplying the lowest probability value by two and giving the word not existing in the grade a value near to it, which is -7.

Step 4: Prediction-In this step, we used the Bayesian multinomial NB algorithm. As mentioned in [35], the classifier variant is one of the multinomial NB that is used for multinomial distributed data along the lines of the ones encountered in the text classification or prediction. It is fast, easy to implement with suitable preprocessing and it is competitive with complex models. The model is rooted on an easy classification structure. Assume text passage *T*, the project predicts the semantic complexity of *T* comparative to a specific level of category *Gi* by computing the log-likelihood in which the words of that passage *T* were created from a representative language model of category Gi. The project computes this log-likelihood individually for the four grades, which correspond to the four Jordanian main grades. The reading difficulty of passage *T* is then estimated as the category level of the language model most likely to have generated the passage *T*. As in [15] we define a generative model for a passage text T using the following: i) choose a grade model Gi according to a prior distribution$P(Gi)$; ii) choose a passage length L in tokens according to the distribution $P(L|Gi)$; and iii) suppose a "bag of words" model for the passage, the L tokens from's multinomial distribution based on the "naïve" assumption that each token is independent of all other tokens in the passage, given the language model Gi.

The probability of T given model Gi is therefore:

$$P(T|Gi) = P(L|Gi).L! \prod_{w\epsilon T} \frac{P(w|Gi)^{C(w)}}{C(w)!} \quad (2)$$

where $C(w)$ is the count of type w in T. We want to find the most likely Gi given the passage, or equally, the Gi that maximizes $L(Gi|T) = logP(Gi|T)$. We derived $L(Gi|T)$ from (2) via Bayes' rule, that is:

$$P(Gi|T) = \frac{P(Gi)P(T|Gi)}{P(T)} \quad (3)$$

Then we make two assumptions: i) all grades are equally likely a priori, and therefore $P(Gi) = 1/N$ where N is the number of grades; ii) the passage length probability $P(L|Gi)$ is independent of grade level. By substituting (2) into (3), simplifying, and taking logarithms, we obtain

$$L(G(i)|T) = \sum_{w\epsilon v} C(w)\, logP(w|Gi) + logZ \quad (4)$$

where $logZ$ represents passage length and the prior $P(Gi)$, which, according to our assumptions, do not influence the prediction outcome and may be ignored. The sum of the weight of the words is computed for every grade, and the higher sum value represents the nearest grade to the text.

4. RESULTS AND DISCUSSION

To evaluate the proposed prediction approach, the hold-out method was adopted. It is an approach used to predict the accuracy of the prediction model. In hold-out method, dataset (D) is split into two disjoint datasets. The first dataset (D1) constitutes 70% of the collected text documents and it is used as the training dataset that is used to build the prediction model. The second dataset (D2) constitutes 30% of the documents and it is used as a testing dataset to evaluate the prediction model. The overall accuracy is around 76% and we summarized the accuracy for each group in Table 3. To analyze the result more, we show in Table 4 the confusion matrix of the model that represents counts of words from actual and predicted values.

From this table, we can calculate the precision, recall, and F1 Score for each grade as in (5) to (7):

$$precision(Gi) = \frac{TP}{TP + FP} \quad (5)$$

$$Recall(Gi) = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score(Gi) = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

where, *Gi* is a specific group model and *i* varies from 1 to 4, true positive (TP) is the number of words that were actually in group *i* and predicted to group *i* (to the same group), false negative (FN) is the number of words that were actually in group *i* and predicted as not in group *i* (not the same group), and false positive (FP) is the number of words that were not actually in group *i* and predicted as in group *i* (the same group).

Table 3. The prediction accuracy for each group

The Group	Accuracy
Group 1	65 %
Group 2	76 %
Group 3	75 %
Group 4	81 %

Table 4. The confusion matrix

Actual \ Predicted	Group1	Group2	Group3	Group4	Total number of words
Group1	352	69	55	68	544
Group2	94	726	78	57	955
Group3	86	77	683	62	908
Group4	93	75	72	1041	1281

We show in Table 5 the precision, recall, and F1 score for each group. From this table, we note relatively high measures for different measures. The F1 Score reflects both precision and recall and it reports high values with an average of around 74%. Moreover, there is stability of it across different classes.

Table 5. Precision, recall and F1 score for all groups

Group \ Measure	Precision	Recall	F1 Score
Group1	56 %	65 %	60%
Group2	77%	76%	76%
Group3	77%	75%	76%
Group4	85%	81%	83%
Average	74%	74%	74%

5. CONCLUSION

The main purpose of this research is to build a statistical model that can be used to predict the reading difficulty level for a given text or paragraphs into different grades. The statistical model has been built using the Jordanian English textbooks. The conducted experiments showed that the proposed approach could predict the reading difficulty level of a given text with acceptable accuracy that suits the targeted audience. In addition, in this research, we built and utilized a novel graded corpus of texts collected from the English textbooks used in Jordanian schools. This corpus can be used for further research in reading difficulty level prediction using other methods. From our experiments, we noted high-performance results of the proposed model with an accuracy of 75.9%, precision of 73.7%, recall of 74.2%, and F1 score of 73.7%, and these results prove the strength of our proposed model.

As a future work and to further enrich this area of research, it is recommended that more books and more text types and genres be used in the experiments. Moreover, it is advised to use methods for feature selection. In addition, we plan to experiment with other popular classification models. It would be of much value to embed this work in a web-based search engine so that users can choose both the query and the difficulty level of the results arrived at in their search.

REFERENCES

[1] A.-V. Luong, D. Nguyen, and D. Dinh, "Building a corpus for Vietnamese text readability assessment in the literature domain," *Universal Journal of Educational Research*, vol. 8, no. 10, pp. 4996–5004, Oct. 2020, doi: 10.13189/ujer.2020.081073.

[2] A.-V. Luong and D. Dinh, "Semi-automatic construction of a readability corpus for the Vietnamese language," *Vietnam Journal of Computer Science*, vol. 9, no. 4, pp. 395–415, Nov. 2022, doi: 10.1142/S219688882250021X.

[3] S. Crossley, A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky, "A large-scaled corpus for assessing text readability," *Behavior Research Methods*, Mar. 2022, doi: 10.3758/s13428-022-01802-x.

[4] A. K. Al Tamimi, M. Jaradat, N. Al-Jarrah, and S. Ghanem, "AARI: Automatic Arabic readability index," *The International Arab Journal of Information Technology*, vol. 11, no. 4, pp. 370–378, 2014.

[5] G. R. Klare, *The measurement of readability*. Ames, 1963.

[6] W. H. DuBay, "The principles of readability," *Online Submission*, 2004.

[7] Y. Yulianto, "An analysis on readability of English reading texts with automated computer tool," *J-SHMIC: Journal of English for Academic*, vol. 6, no. 1, pp. 81–91, Mar. 2019, doi: 10.25299/jshmic.2019.vol6(1).2675.

[8] G. Oktavinanda, S. Y. Siska, K. H. Putri, E. A. Rahma, and Alzuhri, "English for mariners coursebook: how does readability formula rate the readability level of texts?," *Esteem Journal of English Education Study Programme*, vol. 5, no. 2, pp. 50–56, 2022.

[9] Y. Su, K. Fan, N. Bach, C.-C. J. Kuo, and F. Huang, "Unsupervised multi-modal neural machine translation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 10474–10483, doi: 10.1109/CVPR.2019.01073.

[10] Y. Su, Y. Huang, and C.-C. J. Kuo, "Efficient text classification using tree-structured multi-linear principal component analysis," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 585–590, doi: 10.1109/ICPR.2018.8545832.

[11] X. Chen and D. Meurers, "Characterizing text difficulty with word frequencies," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 84–94, doi: 10.18653/v1/W16-0509.

[12] A. Widyantoro, J. Jamilah, and A. Purnawan, "Text difficulty vs text readability: Students voices," *EduLite: Journal of English Education, Literature and Culture*, vol. 7, no. 1, Feb. 2022, doi: 10.30659/e.7.1.125-136.

[13] S. Aluisio, L. Specia, C. Gasperin, and C. Scarton, "Readability assessment for text simplification," in *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, 2010, pp. 1–9.

[14] J. P. Kincaid, J. Fishburne, R. P. Rogers, R. L. Chissom, and S. Brad, "Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel," Institute for Simulation and Training, Feb. 1975, doi: 10.21236/ADA006655.

[15] K. Collins-Thompson and J. P. Callan, "A language modeling approach to predicting reading difficulty," in *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, 2004, pp. 193–200.

[16] Y. Su, R. Lin, and C.-C. J. Kuo, "Tree-structured multi-stage principal component analysis (TMPCA): Theory and applications," *Expert Systems with Applications*, vol. 118, pp. 355–364, Mar. 2019, doi: 10.1016/j.eswa.2018.10.020.

[17] Y. Su, R. Lin, and C.-C. J. Kuo, "On tree-structured multi-stage principal component analysis (TMPCA): Theory and applications," *Expert Systems with Applications*, pp. 1–29, Mar. 2018.

[18] Y. Su and C.-C. J. Kuo, "On extended long short-term memory and dependent bidirectional recurrent neural network," *Neurocomputing*, vol. 356, pp. 151–161, Sep. 2019, doi: 10.1016/j.neucom.2019.04.044.

[19] Y. Su, Y. Huang, and C.-C. J. Kuo, "Dependent bidirectional (RNN) with extended-long short-term memory," in *6th International Conference on Learning Representations*, 2018, pp. 1–16.

[20] Y. Su and C.-C. J. Kuo, "Recurrent neural networks and their memory behavior: a survey," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022, doi: 10.1561/116.00000123.

[21] Y. Su, J. Y. Lin, and C.-C. J. Kuo, "A model-based approach to camera's auto exposure control," *Journal of Visual Communication and Image Representation*, vol. 36, pp. 122–129, Apr. 2016, doi: 10.1016/j.jvcir.2016.01.011.

[22] S. Yuanhang and C.-C. J. Kuo, "Fast and robust camera's auto exposure control using convex or concave model," in *2015 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2015, pp. 13–14, doi: 10.1109/ICCE.2015.7066300.

[23] G. H. Mc Laughlin, "SMOG grading: A new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[24] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948, doi: 10.1037/h0057532.

[25] X. Liu, W. B. Croft, P. Oh, and D. Hart, "Automatic recognition of reading levels from user queries," *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, doi: 10.1145/1008992.1009114.

[26] S. E. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics-ACL '05*, 2005, pp. 523–530, doi: 10.3115/1219840.1219905.

[27] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, "A comparison of features for automatic readability assessment," in *Proceedings 23rd international conference on computational linguistics: Posters*, 2010, pp. 276–284.

[28] M. Zamanian and P. Heydari, "Readability of texts: state of the art," *Theory and Practice in Language Studies*, vol. 2, no. 1, Jan. 2012, doi: 10.4304/tpls.2.1.43-53.

[29] Y. Nurhamsih, "The analysis of the readability levels of the reading texts in textbook entitled 'Fast tract to English' for the third year students of SMA based on raygor readability estimate," *International Journal of Language Teaching And Education*, vol. 1, no. 1, pp. 50–57, Feb. 2018, doi: 10.22437/ijolte.v1i1.4598.

[30] T. Srisunakrua and T. Chumworatayee, "Readability of reading passages in English textbooks and the Thai National Education English test: a comparative study," *Arab World English Journal*, vol. 10, no. 2, pp. 257–269, Jun. 2019, doi: 10.24093/awej/vol10no2.20.

[31] S. A. Crossley, S. Skalicky, and M. Dascalu, "Moving beyond classic readability formulas: new methods and new models," *Journal of Research in Reading*, vol. 42, no. 3–4, pp. 541–561, Nov. 2019, doi: 10.1111/1467-9817.12283.

[32] B. Zhang, "Analysis of text readability of college English course books," *OALib*, vol. 8, no. 9, pp. 1–14, 2021, doi: 10.4236/oalib.1107817.

[33] M. Ali Ramdhani, D. S. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 19, no. 2, pp. 1000–1009, Aug. 2020, doi: 10.11591/ijeecs.v19.i2.pp1000-1009.

[34] M. A. Fauzi, A. Z. Arifin, and S. C. Gosaria, "Indonesian news classification using naïve bayes and two-phase feature selection model," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 8, no. 3, pp. 610–615, Dec. 2017, doi: 10.11591/ijeecs.v8.i3.pp610-615.

[35] S. Xu, Y. Li, and Z. Wang, "Bayesian multinomial naïve bayes classifier to text classification," in *Lecture Notes in Electrical Engineering*, Springer Singapore, 2017, pp. 347–352.

BIOGRAPHIES OF AUTHORS

Yasser Qawasmeh received the B.E. degree in computer information systems from the Jordan university of science and technology, Jordan, in 2005, the M.Sc. degree in computer information systems from Yarmouk University, Jordan, in 2007, currently works as an instructor at the Department of Computer Science and Applications, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan. He can be contacted at email: Yasser@hu.edu.jo.

Qasem Al-Radaideh is a Professor of Data Science and Artificial Intelligence at Yarmouk University. He received his Ph.D. in the Data Mining field from the University Putra Malaysia (UPM) in 2005. His research interests include data mining and knowledge discovery in database, rough set-based knowledge reduction and classification, Arabic Language Computation, Natural Language Processing, and Information Retrieval. He has several publications in the areas of data and text mining and Arabic language computation. He can be contacted at email: qasemr@yu.edu.jo.

Addy AlQuraan received the B.E. degree in electrical and computer engineering from The Hashemite University, Jordan, in 2006, the M.Sc. degree in computer science (data communication and computer network) from the University of Malaya, Malaysia, in 2010. Currently works as an instructor in the Department of basic Sciences Faculty of science, The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan. He can be contacted at email: Addy@hu.edu.jo.

Ahmed Fawzi Otoom is currently working as a professor in the Software Engineering department at Hashemite University, Jordan. He has a Ph.D. degree in computer science from the University of Technology, Sydney (UTS), Australia, 2010. He received his BS in Computer Science from Jordan University of Science and Technology, Jordan, and MS in Software Engineering from the University Western Sydney, Australia, in 2002 and 2003, respectively His main research interests include pattern recognition techniques and its application for software engineering and image and video analysis. Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology, The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan. He can be contacted at email: aotoom@hu.edu.jo.