MLAR: Machine Learning based System for Measuring the Readability of Online Arabic News

Mohammed M. Fouad
Faculty of Computing and
Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Marwa A. Atyah
Faculty of Communication and Media
King Abdulaziz University
Jeddah, Saudi Arabia

ABSTRACT

Online news became one of the favorite information sources for most of the people nowadays because of its update rate and availability over the 24 hours rather than the traditional newspapers. Measuring the readability of the news articles gives a clear view for both the readers and the writers about how easily people can read and understand these articles. In this paper, we present MLAR, a new machine learning based system for Arabic text readability, and use it in measuring the readability of the Arabic online news articles from different outlets. The proposed system is able to determine the topic of each article efficiently and calculates its readability score level. The results show that readability of the online Arabic news is affected by the nature of its topic and the source outlet. The writing style of news articles in each topic differs from one outlet to another.

Keywords

MLAR, Machine Learning, Text Mining, Readability, Arabic Online News, Natural Language Processing.

1. INTRODUCTION

With the rapid growth of internet and its services every day, many news outlets, such as newspapers or news agencies, published their updated new online on their official websites. Most of the people use these online services as their source of information and recent news because they are updated frequently and could be reached anytime and anywhere.

The news articles on the website are mainly organized into many sections based on the different news topics as in the printed newspaper. These topics cover most of what the readers like and want to read such as Sports, Economy, Arts, Politics, Religion, etc. For the Arabic region, many news sources outlets, even the non-Arabic ones like BBC and CNN for example, published their Arabic news online because of the variety of the interested topics in this region.

The readability of the text in any news article is an important factor in the relation between the reader and the article contents. It defines the text difficulty level and measures how easily the reader can read and understand the textual information inside this article [1]. To measure the readability of a text, many mathematical formulas were developed, for the English text, based on different text features in the last century. These features may include sentence and word length, percentage of complex words in a sentence, sentence structure and many other factors.

Due to the broad spread of the English language, most of the researchers in this topic use these formulas nowadays as standard tools for measuring the difficulty of the English text. Even, some formulas have been added to the popular word processing programs, such as Microsoft Word®, for automatic measuring of the text readability.

On the other hand, the research in the readability area for the Arabic language is very limited [2]. However, there are some trials to formulate some mathematical formulas that can deal with the Arabic language characteristics. In order to understand and benefit from this readability measure in the online Arabic news, it should be associated with the article topic.

In this paper, a new machine learning based system, called MLAR, is presented to calculate the readability of the Arabic text. In addition, MLAR system is applied on the Arabic online news articles to demonstrate its functionality. The proposed system utilizes an efficient classifier to recognize the topic of the news articles based on their contents. It also uses a recent formula to measure the Arabic text readability for each article.

This paper is organized as follows: Section 2 presents the related work in the readability research area especially in the Arabic language domain. Section 3 illustrates in details the components of the proposed system. The experimental results and discussions are plotted in Section 4. Finally, the conclusions are drawn in Section 5.

2. RELATED WORK

The research of the readability of the English language started in the beginning of the last century. The researchers developed hundreds for mathematical formulas that have been used to measure the difficulty level of certain English text. The main application for these formulas was in the education domain [3]. The readability formulas were used to define the text difficulty with respect to the grade level. In other words, for a certain text, what is the suitable grade level that can easily read and understand it? In addition some researchers were interested in related applications such as book recommendations [4], measuring curriculum difficulty [5] and reading assessment [6].

Table 1 shows some of the most commonly used readability formulas for English text. These formulas incorporate many text factors in order to calculate the equivalent US grade level that matches the readability score of certain text. These factors include average sentence length, average word length, number of syllables per word, and different combinations of these factors which are greatly differ from one formula to another [7].

The readability formulas shown in Table 1 have been used for measuring the difficulty of learning and education materials. For example, Dale-Chall formula in [8] is designed to assess the difficulty of the reading materials for the fourth grade students and above, while Flesh-Kincaid is suitable for upper elementary reading materials through secondary grade [9].

Some formulas were designed very simple, such as SMOG formula [11]. It mainly depends on one factor, which is the number of polysyllable words, i.e. words with three or more syllables, in the text. Also, some formulas depend on specific definitions, such as Spache formula which uses a static list of words that represent the familiar words [12]. The words in the text that are not in this list are considered unfamiliar words.

Table 1. Summary of Well-Known English Readability Formulas

Name	Grade Level Formula
Dale-Chall [8]	Grade = (0.1579 * percent difficult words) + (0.0496 * average sentence length) + 3.6365
Flesch-Kincaid [9]	Grade = (0.39 * #words/#sentences) + (11.8 * average syllables per word) – 15.59
FOG [10]	Grade = 0.4 * (average sentence length + percent difficult words)
SMOG [11]	$\text{Grade} = 3 + \sqrt{\text{\#polysyllable word count}}$
Spache [12]	Grade = (0.121 * average sentence length) + (0.082 * percent unfamiliar words) + 0.659

Other languages do not have the same interest by readability researchers like the English language. Lately, some formulas have been developed to deal with the Arabic text which are the aim of this study.

The researchers started by applying the text mining with natural language processing techniques to build models for Arabic text readability. Table 2 summarizes the presented formulas for Arabic text readability. Al-Heeti in [13] started by a simple formula for Arabic text readability. The Heeti formula only considered the average word length as its main factor. This made the formula too simple to give a good measure for Arabic text readability.

AARI measure is another formula proposed by Al-Tamimi et al. in [14] for Arabic text readability. They tried in their proposed formula to include the main text factors such as number of characters, words, difficult words and sentences in the text. The problem of AARI is that it did not include any features related to the Arabic language which resulted in poor readability measure.

Recently, El-Haj and Rayson in [15] presented OSMAN measure for Arabic text readability. They modified the Flesh-Kinced and FOG formulas for English text to be more suitable for Arabic text. OSMAN measure includes the diacritics in the calculation in order to count the syllables in the words correctly. They proposed a novel approach to count the short, long and stress syllables efficiently. Due to its efficiency, OSMAN measure has been used in this paper as the readability measure for the Arabic online news.

On the other hand, some researchers dealt with the readability issue from text mining perspective. Al-Khalifa and Al-Johani in [2] collected 150 texts from the Saudi Arabia curriculum in the reading books of elementary, intermediate, and secondary schools. These texts were manually labeled into three difficulty levels: easy, medium and difficult. Using natural language processing techniques, they extracted 5 features from this dataset: average word length in letters and syllables, average words per sentence, bigram features and term frequency. These features were used to train a support vector machine (SVM) classifier model. Their model achieved about 73% accuracy rate on the testing dataset. Their model needs to be applied on larger dataset in order to cover most of the keywords of the Arabic text.

Mat Daud et al. in [16] presented a corpus based readability formula. Their proposed formula was simply the average ranks of the words in the sentence. To obtain the word rank, they used King Abdulaziz City for Science and Technology (KACST) Arabic corpus [17]. The formula is very simple but not automated yet which make it not useful for large and online Arabic text.

Table 2. Summary of Available Arabic Readability Formulas

Name	Formula
Al-Heeti [13]	Heeti = (4.414 * average word length) – 13.468
AARI [14]	AARI = (3.28 * #characters) + (1.43 * average word length) + (1.24 * average sentence length)
OSMAN [15]	OSMAN = 200.791 – (1.015 * average sentence length) – 24.181 * (ratio hard words + average syllables per word + average word length + ratio complex words)

3. THE PROPOSED MLAR SYSTEM

In this section, the proposed MLAR system will be described briefly. As shown in Figure 1, the proposed system is composed of two parts. The first one is to build an efficient classifier to detect the topic of the input news articles. The second one is responsible for calculating the readability of these articles using OSMAN formula.

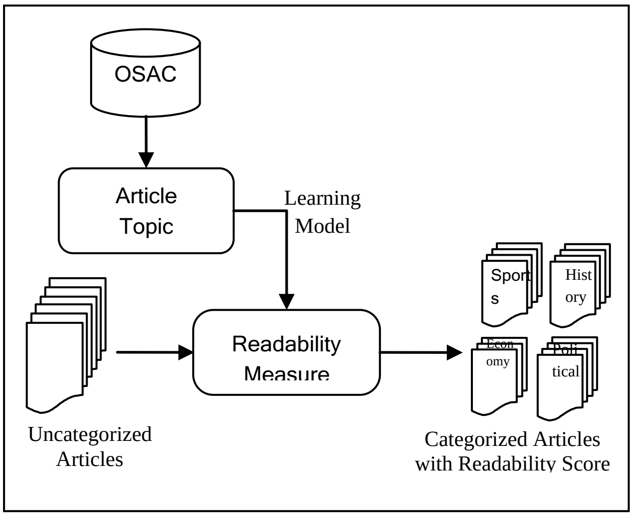

Fig 1: Overview of the Proposed MLAR System

3.1 Article Topic Detection

The topic of the Arabic online news is considered an important factor when analyzing the readability of such news. The readability of each article could be affected by the words and terms used in this topic. For example, the written article describing a political situation may use some words and phrases that are very hard to understand for many people. So, before calculating the readability score for each article, an efficient tool to automatically detect the topic of this article is needed.

In order to build this tool, the Open Source Arabic Corpus (OSAC) provided by Saad and Ashour in [18] had been used. OSAC corpus contains about 22,429 text documents that were automatically collected from the internet and categorized into 10 categories as shown in Table 3. These documents were collected from different sources such as online news, blogs, free electronic books, etc.

The proposed system started by applying the natural language processing techniques in order to process the text documents in OSAC corpus and create the appropriate feature vector for the classification task. The preprocessing step includes some sub-steps such as normalization, tokenization, removing stop-words and light stemming for Arabic language. The output feature vector could be constructed in many ways based on how each word is weighted in the output vector. Different weighting schemas have been studied, as shown in Table 4, to fill the bag-of-words extracted from the OSAC corpus. Consider the following annotations for describing the term weighting schemas. Let (a_{ik}) be the weight of term (i) in document (k), the total number of documents (N), Term frequency (f_{ik}) be the frequency of term (i) in document (k) and Document frequency (df_i) be the number of documents in which term (i) occurs.

Table 3. OSAC Corpus – Topic Distribution

Category/Topic	# Text Documents in OSAC
Economic	3,102
History	3,233
Education & Family	3,608
Religious	3,171
Sport	2,419
Health	2,296
Law	944
Astronomy	557
Stories	726
Cooking Recipes	2,373
Total	**22,439**

Table 4. Term Weighting Schemas

Schema	Description
Binary	$a_{ik} = \begin{cases} 1, & f_{ik} > 0 \\ 0, & \text{otherwise} \end{cases}$
Term Frequency	$a_{ik} = f_{ik}$
Normalized Term Frequency	$a_{ik} = \frac{f_{ik}}{N}$
TF.IDF	$a_{ik} = f_{ik} * \log(\frac{N}{df_i})$

For the classification task, we have applied the Naïve Bayes algorithm because it is known for its good performance. This make it one of the most commonly used algorithms in the similar applications for text classification.

3.2 Online Arabic News Readability

After training the appropriate classifier for topic detection, the next step is to calculate the readability for each article. In our case, OSMAN readability measure presented in [15] have been applied. This metric have many advantages over the other measures for Arabic text readability. First, it can automatically work on Arabic text with or without diacritics which are very important in finding the syllables inside the word. Second, it introduces a new feature, called "Fasseh" that focuses on some aspects related to Arabic script structure.

OSMAN readability measure does not have a specific scale range, but in general, the higher the value the easier the text. The output readability measures are arranged in descending order where easiest text comes first and hardest text comes last.

To achieve the objective of this paper, a dataset with about 10,454 Arabic news articles had been collected from four different online source outlets as shown in Table 5. These articles were collected automatically using open source offline explorer tools with no respect to certain topic, writer nor specific period of time.

Table 5. Dataset Distribution

Source Outlet	# Articles
Al-Masry Al-Youm http://www.almasryalyoum.com/	5,134
Al-Watan www.elwatannews.com/	1,931
BBC Arabic http://www.bbc.com/arabic	1,972

CNN Arabic http://arabic.cnn.com/	1,417
Total	**10,454**

4. RESULTS AND DISCUSSION

The phases of the proposed MLAR system were implemented with different tools. First, the article topic detection classifier was trained and tested using RapidMiner® tool for data mining [19]. Then, we developed a Java application to measure the readability using OSMAN metric. This application used some natural language processing techniques implemented in Stanford CoreNLP Library [20]. All these phases run on a machine with Intel® Core™ i7-4510U 64-bit processer (2.00 GHz), 8.00 GB memory and running Windows 8© operating system.

4.1 Accuracy of Topic Detection Classifier

The proposed model for article topic detection was trained using Naïve Bayes algorithm. In the training phase, different term weighting schemas were applied in the vector generation step. In the testing phase, the accuracy of the model was calculated using 10-fold cross validation. Table 6 shows the accuracy of the topic detection model with different term weighting schemas.

Table 6. Topic Detection Classifier Accuracy

Term Weighting Schema	Accuracy (%)
Binary	93.12 %
Term Frequency	**95.06 %**
Normalized Term Frequency	94.71 %
TF.IDF	94.99 %

As shown in Table 6, the accuracy of the proposed model is very high, which means that the output model can efficiently differentiate between the 10 categories of the OSAC corpus.

4.2 Arabic News Readability Results

After applying OSMAN readability metric on the collected Arabic online news articles, the distribution of the readability of all the collected articles had been tested. As shown in Figure 2, the readability is distributed with nearly Gaussian shape. This means that small number of articles are very hard and also another small portion of these articles are very easy, while the majority of the articles are in the middle range with average difficulty.

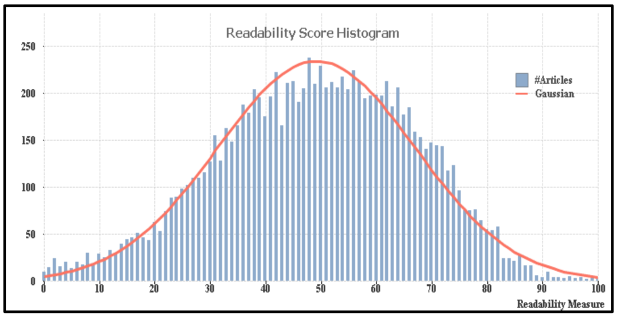

Fig 2: Distribution of News Articles Readability Measure

The objective of this paper is to investigate the readability score of online Arabic news and its relation to the article topic. Figure 3 shows the average readability score for the news articles in each topic ordered in ascending order. As shown in Figure 3, Law and Economy articles have the lowest score which means that they are too difficult to read and have more complex words than other topics. On the other hand, the Stories articles and Cooking Recipes are easy to read because their writers used a lot of simple words. Remaining topics come on the average readability score.

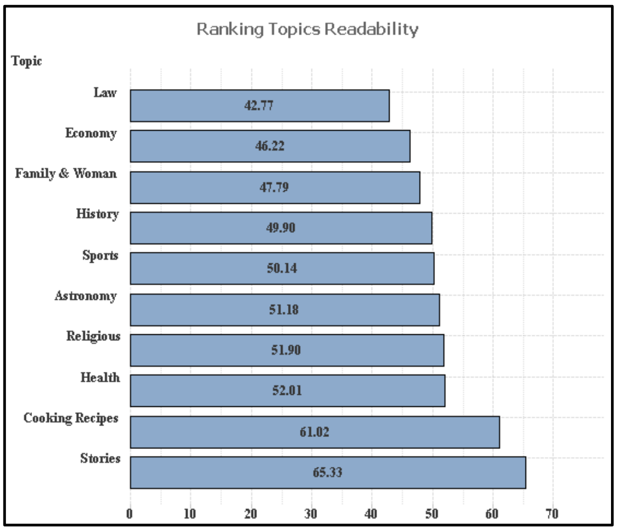

Fig 3: Ranking Topics Readability

From another point of view, the average readability score for the articles belonging to each source outlet is calculated as shown in Figure 4. This gives us a clear view about the writing style of these outlets and the reading difficulty of their published articles.

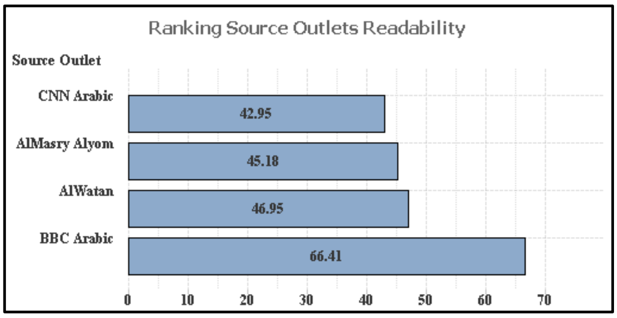

Fig 4: Ranking Source Outlets Readability

As shown in Figure 4, BBC Arabic website has the highest readability score with about 20% more than other outlets. This means that the writers tried to use more simple words and avoid complex ones. This made their published articles are simpler to read than articles from other outlets especially CNN Arabic outlet that has the lowest readability score.

5. CONCLUSIONS

In this paper, we presented a new system, called MLAR, for automatic calculation of the Arabic text readability and applied it on the online Arabic news. The proposed system used the Naïve Bayes algorithm in order to build an efficient model for article topic detection with about **95%** accuracy. The readability of the Arabic articles is calculated using OSMAN readability metric that has a good combination of factors that are more related to the Arabic scripts than other

Arabic readability formulas. The proposed system is applied over a collection of some Arabic news articles from four different source outlets. The results show that MLAR system is very helpful in understanding the readability of these articles. In addition, it can be used by the writers themselves to judge the writing style of their articles. In order to make the text more readable, the writers should use more simple words and avoid complex and hard words that are more difficult for the readers.

In future, we are planning to extend MLAR system by adding the semantic relationships between the words in the input Arabic text. Some tools, such as Arabic WordNet Project in [21], have been presented that define the semantic features in the Arabic text and show how words are linked together based on their meaning. This will help greatly in measuring the readability of the Arabic text rather than the current proposed formulas.

6. REFERENCES

[1] Dalecki, L., Lasorsa, D. L., and Lewis, S. C. 2009. The News Readability Problem, Journalism Practice, vol. 3(1), pp. 1-12.

[2] Al-Khalifa, H. S., and Al-Ajlan A. A. 2010. Automatic Readability Measurements of the Arabic Text: An Exploratory Study, the Arabian Journal for Science and Engineering, vol. 35(2C), pp. 103-124.

[3] Compton, D. L., Appleton, A. C., and Hosp, M. K. 2004. Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average to poor decoders, Learning Disabilities Research & Practice, vol. 19, pp. 176-184.

[4] Bailin, A., and Grafstein, A. 2001. The linguistic assumptions underlying readability formulae: A critique, Language and Communication, vol. 21, pp. 285-301.

[5] Barbham, E. G., and Villaume, S. K. 2002. Leveled text: The good news and the bad news, The Reading Teacher, vol. 55, pp. 438-41.

[6] Persampieri, M., Gortmaker, V., Daly III, E. J., Sheridan, S. M., and McCurdy, M. 2006. Promoting parent use of empirically supported reading interventions: Two experimental investigations of child outcomes, Behavioral Interventions, vol. 21, pp. 31-57.

[7] Begeny, J. C., and Greene, D. J. 2014. Can readability formulas be used to successfully gauge difficulty of reading materials, Psychology in the Schools, vol. 51(2), pp. 198-215.

[8] Dale, E., and Chall, J. 1948. A formula for predicting readability: Instructions, Educational Research Bulletin, vol. 27, pp. 37-54.

[9] Flesch, R. 1948. A new readability yardstick, Journal of Applied Psychology, vol. 32, pp. 221-229.

[10] Gunning, R. 1952. The technique of clear writing. New York: McGraw-Hill.

[11] McLaughlin, G. H. 1969. SMOG grading: A new readability formula, Journal of Reading, vo. 22, pp. 639-646.

[12] Spache, G. 1953. A new readability formula for primary grade reading materials, The Elementary School Journal, vol. 53, pp. 410-413.

[13] Al-Heeti, K. 1984. Judgment analysis technique applied to readability prediction of Arabic reading material, Ph.D. Thesis, University of North Colorado.

[14] Al-Tamimi, A., Jaradat, M., Aljarrah, N., and Ghanim, S. 2014. AARI: Automatic Arabic Readability Index, The International Arab Journal of Information Technology, vol. 11(4), pp. 370-378.

[15] El-Haj, M., and Rayson, P. 2016. OSMAN – A Novel Arabic Readability Metric, Proceedings of the Language Resources and Evaluation Conference 2016. European Language Resources Association (ELRA), Slovenia, pp. 250-255.

[16] Mat Daud, N., Hassan, H., and Abdul Aziz, N. 2013. A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty, World Applied Sciences Journal, vol. 21, pp. 168-173.

[17] Al-Thubaity, A. O. 2015. A 700M+ Arabic corpus: KACST Arabic corpus design and construction, Language Resources and Evaluation, vol. 49(3), pp. 721-751.

[18] Saad, M. K., and Ashour, W. 2010. OSAC: Open Source Arabic Corpus, Proceedings of the 6th International Conference on Electrical and Computer Systems (EECS’10), Lefke, North Cyprus, pp. 1-6.

[19] RapidMiner® Data Science Tool: https://rapidminer.com/

[20] The Stanford Natural Language Processing Group, Stanford NLP: http://nlp.stanford.edu/software/

[21] The Arabic WordNet Project: http://globalwordnet.org/arabic-wordnet/