MTReadable: Arabic Readability Corpus for Medical Tests Information

Dimah Alahmdi, Athir Saeed Alghamdi, Neda'a Almuallim, Suaad Alarifi
dalahmadi@kau.edu.sa, aothmanalghamdi0001@stu.kau.edu.sa, nmalmuallim@stu.kau.edu.sa , salarifi@kau.edu.sa
Faculty of Computer and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract
Medical tests are very important part of the health monitoring process. It is performed for various reasons like diagnosing diseases, determining medications effectiveness, etc. Due to that, patients should be able to read and understand the available online tests and results in order to take proper decisions regarding their health condition. In fact, people are varying in their educational level and health backgrounds that make providing such information in an easily readable format by the majority of people considered as a challenge in the health domain since ever. This paper describes the MTReadable corpus which constructed for evaluating the readability of online medical tests. It covered 32 basic periodic check-up tests with over 36k words. These tests information are annotated and labelled based on three readability levels which are easy, neutral and difficult by three non-specialists native Arabic speakers. This paper contributes to enriching the Arabic health research community with an investigation of the level of readability of online medical tests and to be a baseline for further complex health online reports and information.

Key words:
Text mining, Arabic corpus, Readability corpus, Medical Test.

1. Introduction

Health education is an essential part of health domain. People by nature, always look for information about ambiguous questions that come in their minds, especially when it is related to their health. The unreadable doctors' notes and complex information that patients found on the Internet had been usually a barrier preventing them from understanding their condition [1],[2]. One of the most important written medical information is medical tests reports. Nowadays, Medical tests (MT) usually are found online as a part of online medical websites and hospitals. Reports can be accessed by people to understand their health conditions, diagnose diseases which are written by doctors whom set a treatment plan, or monitor it. These reports may include text and image such as X-rays [3]. Medical tests vary considerably on purposes and methods; however, it classified into biochemical tests, pathological tests, and electrophysiological tests [4]. Easily understandable information helps patients to determine their condition, improve it, and make informed decisions regarding their health [5]. Advance and intelligent technology can help to improve patient understanding by for instance eliminating the issue regarding hand-written doctors' notes; although, information itself still complex for some patients, which referred to as health literacy. Health literacy considered a challenge in the health domain ever since. Due to that, many studies conducted over the years to evaluate health-related text readability in different languages using several methods. These methods varied from traditional formulas to annotated corpora and machine learning algorithms all with the aim to predict the information difficulty level in order to enhance its formulas to be readable by any patient. Online generated health reports are crucial to people who use the Internet. Due to the widespread of technology and accessibility using smart devices, patients can read some medical reports and test such as blood test results. Measuring the level of the readability of this medical information will enrich the health literacy [6],[7].

In order to overcome the limitations of current research in Arabic readability corpus, this paper contributions are in twofold:

1. Build an Arabic readability corpus MTReadable for evaluating the readability level of medical tests information.
2. Provide insight for future Arabic studies related to building health corpora by constructing a baseline for Arabic health annotation and defining the associated challenges.

The rest of this paper organized as follows: Section 2 reviews the related work on health and Arabic linguistic corpora. Section 3 describes the methodology followed to construct the corpus. Section 4 describes the MTReadable corpus details. Section 5 discusses the annotation process, used guidelines, and examples of annotated data. After that, section 6 describes the validation of constructed corpus including the used methods and classifiers and its results.

Manuscript received May 5, 2021
Manuscript revised May 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.5.14

Then, Section 7 explores the challenges faced during project completion.

Finally, the paper concludes with Section 8, which highlights some future opportunities for extending this research.

2. Literature Review

This section examines the studies conducted in the field of Arabic linguistics and health text readability assessments in either Arabic or English.

Several studies conducted to handle the evaluation of health-related text readability, one study by [8] discussed the readability of online-accessible health information regarding congestive heart failure. The study focused on 70 out of the first 100 websites at Google search engine. The selected websites evaluated through 6 readability assessment methods. The result determined that most of the selected websites exceeded the recommended readability level of sixth-grade. The paper also highlighted some methods that help to enhance the readability of online health' content. Another study by [9] constructed a corpus to assess the readability level of medicine leaflets written in Arabic by utilizing machine learning algorithms and tools. The corpus annotated the medicine information leaflets with three difficulty levels and it aimed to contribute to the production of leaflets with readable by the majority of consumers. While [10] presented an automated ontology-based annotation methodology. The developed methodology utilizes the developed Arabic OWL ontologies related to food, nutrition, and health domains in addition to linguistic patterns to detect the relationships between labeled entities then link it to the ontology corresponding concepts and properties in order to produce the RDF Despite the significant amount of studies on Arabic corpora, the linguistic-related research field still promising owing to the lack of publicly accessible Arabic corpora [11]. In [12] authors described the built of linguistic corpus to overcome the limitation of existed Quranic Arabic corpora using a semi-automated rule-based technique that analyzes Quran words then morphologically tagging them. Another study in [13], authors worked further to provide a resource for future analysis by describing a multi-level approach for Quranic linguistics annotation include morphological segmentation, part-of metadata for the corresponding Web resources. speech tags, and syntactic analysis.

Related to the Arabic health domain, [14] presented the complexity of the Arabic language nature and explored the problem of poor-structured digital content in Arabic. In addition, this paper compiled different datasets to create a benchmark that can be used in future research in different domains including economics, history, education, health and much more. While the study [15] presented the latest Arabic medical language resources and the challenges associated with addressing these resources. The study also experimented with two strategies to extract medical terms. Where the first strategy used a list of set medical terms, the second was used the equivalent Arabic terms for Latin prefix and suffix. However, the results indicated that the last strategy outperformed.

In fact, Arabic health domain suffers from a lack of well-structured, public readability corpora. This paper describes MTReadable an Arabic annotated readability corpus for evaluating online medical tests' information readability level as an effort to fill that gap.

3. Experiment Set Up

Owing to the lack of studies that examined health-related Arabic online content and the absence of public medical tests' readability corpus, MTReadable is collected and developed to fill this gap. A set of steps were followed in order to meet that goal as follows:

i. Main medical tests' related information has been collected from Arabic reliable sources.

ii. The collected data annotated regarding the readability level.

iii. The collected data were cleaned and prepared, and then the corpus was built.

iv. The corpus has been evaluated by implementing different benchmark algorithms and also corpus testing is applied.

4. Corpus Description

Medical tests are an important part of periodic health screening and essential procedure in tracking treatment effectiveness, diagnose diseases, wherefore understanding such information is a must. Due to drastic differences between people's educational levels, the information which is very clear for some people maybe not understood by others who are older in age or any other factor. Since health is a right for everyone, the health-related information should be provided in a simple and readable by the majority.

At first, the corpus aimed to be constructed from the information available on the Saudi health websites. For example, Ministry of Health [16], information was provided as images, where this paper focus is text data. In addition, King Abdullah Arabic Health Encyclopedia [18] has data however, medical tests it covers few and limited textual medical test data. Accordingly, the corpus data were collected from a reliable Arabic laboratory website "Al-Borg Medical Laboratories" [19]. It is a popular approved medical test center in Saudi Arabia. The corpus is called Medical Test Readability “MTReadable”, it covers the

basic periodic check-up tests information which composed of 32 medical tests Figure 1. Corpus statistic with a total of approximately 2000 sentences and 36 thousand words in Figure 2. This corpus is available upon request from authors in to help research community.

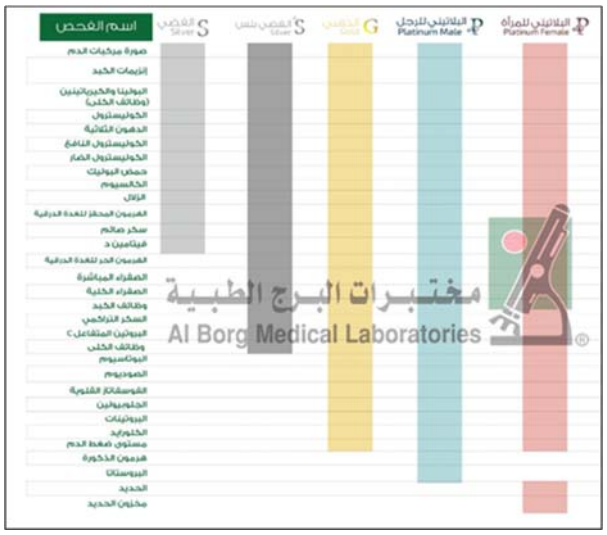

Fig. 1 List of the Main Periodic Check-up tests [19].

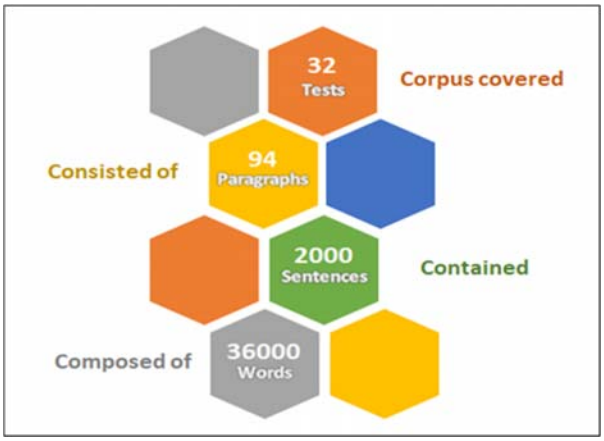

Fig. 2 Corpus General Statistics.

Each test contains a brief description of the test, a detailed description of the sample; how and why the test is done, how the sample collected, and what instructions should be followed before doing the test. Moreover, test results explained in detail including what high, or low level reflects. Also, a set of common questions append with each test explained in Figure 3 and translated and described in Table 1.

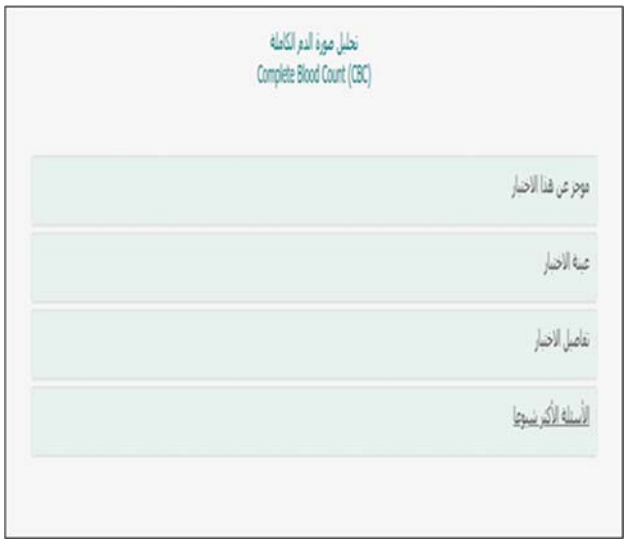

Fig. 3 Figure 3: Test Available Information [19].

Table 1: Sections and Subsections of Test Information.

Sections	Subsections
Test summary	Purpose of the test
	Cases requiring performing the test
	Sample type
	Preparations for test
Test sample	How the test performed and what the doctor is looking for
	How the sample collected
	Required preparations for test to ensure sample quality
Test details	How to benefit from the test
	When is the test required
	What are the test results and what it means
	Other things that should be known about the test
Most common questions	–

The list of main periodic check-up tests gotten from the laboratory in Figure 2, then web searched to collect HTMLs that contains information about them. After that, text extracted from these HTMLs using Python and NLTK library to implement text analysis. For corpus preparation, cleaning the data is required. Spaces and Arabic text diacritics were removed. URLs are also removed only text is kept. Finally, data saved as csv files, some examples are shown below in Figures 4, 5, and 6.

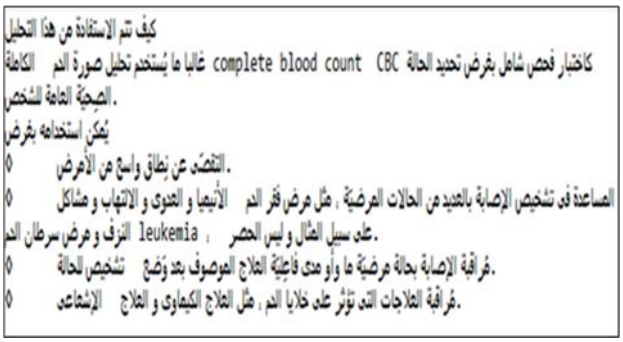

Fig. 4 CBC Test "How to benefit from the test".

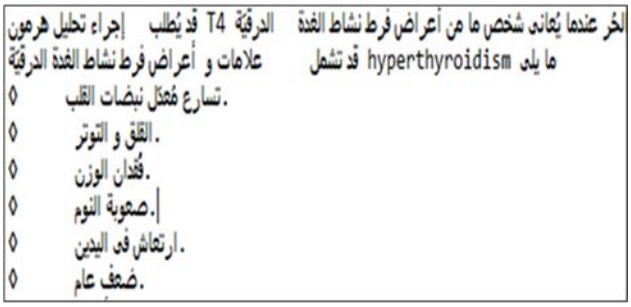

Fig. 5 When required to perform T4 test.

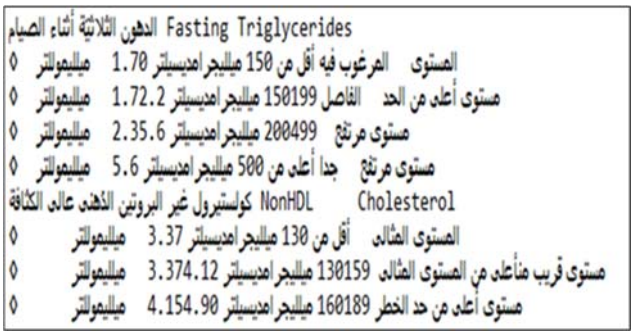

Fig. 6 Triglycerides Levels' indications

5. Annotation

The collected data is organized as paragraphs in order to have description and a clear judgment of information readability and understandability. The annotation process in MTReadable has chosen to be done on the paragraph level. Due to all sections of the test information contain similar information that only differentiates in the extent of the details. This research is focused on the test details section only, because it contains the most informative and valuable information.

5.1 Labels Explanation

In order to label the collected medical test paragraphs. Annotators were provided with a set of guidelines for directing the annotation process aiming to achieve the best results, and it was outlined as follows:

i. Easy: reflects very clear information that understandable by average educated without any further searches or help.
ii. Neutral: reflects readable information which has a few medical or ambiguous terms that not understandable without an additional search.
iii. Difficult: reflects a complex piece of information that has some medical or ambiguous terms and cannot be understood by the average educated person.

If the text contains three or fewer medical terms (difficult words) it is considered as Easy to read. If the number of difficult or ambiguous words between four and six, text considered Neutral. Otherwise, it will be Difficult.

5.2 Annotation Process

The annotation was done by four Arabic native speakers from different educational levels and backgrounds but none of them was a health domain specialist, where the aim is to determine the readability of the text by non-specialists. After providing the annotation guidelines three of the annotators evaluated each test information as "Easy", "Neutral", or "Difficult" to reflect their level of understandability of test information. The fourth annotator can read the other three annotators' judgments before doing the annotation and only resolve the conflict between them. After that, to prove the reliability of the annotation and measure the Inter Annotator Agreement Cohen's kappa was used. The measurements indicated a fair level of agreement between annotators which is 0.4 based on [19]. It thought that the variation of annotators' educational levels was the reason. From Table 2 some annotated examples, where An indicates the annotator.

Table 2: Some annotated examples from MTReadable.

TestName اسم التحليل	Annotated Readability Level مستوى سهولة القراءة			
	A1	A2	A3	A4
CBC تعداد الدم	Neutral عادي	Easy سهل	Easy سهل	Easy سهل
Renal Profile ملف الكبد	Neutral عادي	Neutral عادي	Neutral عادي	Neutral عادي
Lipid Profile ملف الدهون	Easy سهل	Neutral عادي	Easy سهل	Easy سهل
Iron الحديد	Easy سهل	Neutral عادي	Difficult صعب	Neutral عادي

6. Corpus Testing

In order to validate MTReadable corpus, testing by splitting methodology is used. This is done by splitting the corpus data which composed of 24 text files into two sets, first (19 text files) which represent 80% of data considered as training set while the remaining 20% (5 text files) used for testing with no overlap between them validation [20]. In order to validate and benchmark the corpus, three popular classification algorithms in text analysis have been tested, which are Stochastic Gradient Descent (SGD), Naive Bayes, and Logistic Regression [21]. From the annotation process, results showed that the data annotated "Difficult" is zero, the class difficult has almost no instances under it. Therefore, the three classifiers built upon the data with two class labels only which are "Easy" and "Neutral" and the third label is removed. SGD classifier showed superior performance regarding the final evaluation results. We can that SGD achieved the highest accuracy 0.96 in compare with the other two models applied on our dataset. Table 3 below shows the detailed results of the final evaluation of the three classification models include precision, recall and F1-score.

Table 3: Summarization of the final evaluation results.

Labels	Precision	Recall	F1-score	Accuracy
SGD Classifier	0.94	0.83	0.97	**0.96**
Naïve Bayes	0.67	1.00	0.80	0.67
Logistic Regression	0.67	1.00	0.80	0.67

It is noticeable that the annotators couldn't perceive the test information as difficult to read. Basically, they are chosen from different education background secondary school and higher degrees but not in the medical fields. This gives the indication of the type of online users and their ability to recognize the basic medical tests. More medical reports with non-common test information may need more investigation.

7. Research challenges

Through the MTReadable developing and validating of the constructed Arabic corpus, a set of challenges and difficulties were encountered. Some are highlighted here for benefiting future studies.

- *Availability and accessibility of data*. The data explosion generation which we live in today is encouraging to produce massive amounts of data over the Internet. However, these data are not always reliable or accurate. Health domain data especially must be reliable because critical decisions regarding human health might be taken based on it. Hence, an authorized source on online medical data was a very critical decision.
- *Arabic online health content.* Collecting medical tests' information was the most challenging part, the health domain information in Arabic is scarce. Furthermore, Arabic reliable data sources are concerned about clarifying medical tests' information, terms, and the results indications in English. Even that websites which have some information about such tests are written in a difficult language to read and format by the average education person Arabic speaker.
- *Annotation process.* The inter agreement of annotators was just fair, although a guideline was provided, and the process was clearly described before stating. However, the different educational backgrounds of annotators played a role in making stronger agreement.
- *Training and testing.* The accuracy of dataset splits varied with different classifiers, classifiers considered data classes to be two only "Easy" and "Neutral" and ignoring "Difficult" class. We think that might be owing to the rareness of test information annotated as "Difficult". Expanding and enriching of the corpus with more types of medical tests might show a real distribution between classes and improve the results.

8. Conclusion and Future Directions

This paper describes a methodology in constructing MTReadable corpus which aims to evaluate the readability of Arabic medical tests' results information. The corpus composed of approximately 36k words representing the results' indications and main information of the basic periodic check-up tests. This information annotated by non-specialist Arabic native speakers with three labels "Easy", "Neutral", and "Difficult". Afterward, the corpus validated using the splitting method for training and testing the classifier. MTReadable is available for the public research community and we hope it contributes to enriching the Arabic health domain research and inspiring future studies. Some suggested future directions are:

- Extending the corpus by covering more medical tests and variety of medical online report. That will help to enhance the problem of classifying dataset instances with more accurate results.

- The results of classified data can be used by other researchers from the health domain to assess the readability of available online medical data, enhance its structure and writing format to increase the readability level; also it can be used as a guide for writing health information that are readable by the average person.
- The written medical test results' indications and clarifications can be attached by health care providers like hospitals and laboratories with patient test results, that will contribute to educating people and give them a better understanding of their health conditions and online medical reports.

References

[1] Charnock, Deborah. "The DISCERN handbook." Quality criteria for consumer health information on treatment choices. Radcliffe: University of Oxford and The British Library (1998).

[2] Pope, C., S. Ziedland, and N. Mays. "Qualitative research in health care: Analysing qualitative data. 320." BMJ 8.320 (2000): p.7227.

[3] Bustos, Aurelia, et al. "Padchest: A large chest x-ray image dataset with multi-label annotated reports." Medical image analysis 66 (2020):p. 101797.

[4] Sun, Wencheng, et al. "Data processing and text mining technologies on electronic medical records: a review." Journal of healthcare engineering 2018 (2018).

[5] Pinsonneault, Alain, et al. "Integrated health information technology and the quality of patient care: A natural experiment." Journal of Management Information Systems 34.2 (2017): p.457-486.

[6] Daraz, Lubna, et al. "Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet." Journal of general internal medicine 34.9 (2019): 1884-1891.

[7] Al Aqeel, Sinaa, et al. "Readability of written medicine information materials in Arabic language: expert and consumer evaluation." BMC health services research 18.1 (2018): p.1-7.

[8] Kher, Akhil, Sandra Johnson, and Robert Griffith. "Readability assessment of online patient education material on congestive heart failure." Advances in preventive medicine 2017 (2017).

[9] Alotaibi, S., Alyahya, M., Al-Khalifa, H., Alageel, S., & Abanmy, N.. Readability of Arabic medicine information leaflets: a machine learning approach. Procedia Computer Science, 82 (2016), p. 122-126.

[10] Albukhitan, Saeed, Ahmed Alnazer, and Tarek Helmy. "Semantic annotation of arabic web documents using deep learning." Procedia computer science 130 (2018): p.589-596.

[11] Alalyani, Nada, and Souad Larabi Marie-Sainte. "NADA: New Arabic dataset for text classification." International Journal of Advanced Computer Science and Applications 9.9 (2018).

[12] Dukes, Kais, and Nizar Habash. "Morphological Annotation of Quranic Arabic." Lrec. 2010.

[13] Zeroual, Imad, and Abdelhak Lakhouaja. "A new Quranic Corpus rich in morphosyntactical information." International Journal of Speech Technology 19.2 (2016): p.339-346.

[14] Saad, Motaz K., and Wesam M. Ashour. "Osac: Open source arabic corpora." 6th ArchEng Int. Symposiums, EEECS. Vol. 10. 2010.

[15] Samy, Doaa, et al. "Medical Term Extraction in an Arabic Medical Corpus." LREC. 2012.

[16] Health, S. M. o. (2019). Awareness. Retrieved from https://www.moh.gov.sa/Pages/Default.asx

[17] Encyclopedia, King, A, (2019) https://kaahe.org/en-us/Pages/Home/Home.aspx

[18] Laboratories, A. B. M. (2019). Lab Tests Website. Retrieved from https://www.alborg.sa/ar/

[19] Blackman, Nicole J - M., and John J. Koval. "Interval estimation for Cohen's kappa as a measure of agreement." Statistics in medicine 19.5 (2000): p.723-741.

[20] Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.

[21] Salloum, Said A., et al. "A survey of Arabic text mining." Intelligent Natural Language Processing: Trends and Applications. Springer, Cham, 2018. p.417-431.