Readability of Arabic Medicine Information Leaflets: A Machine Learning Approach

Sihaam Alotaibi\textsuperscript{a}, Maha Alyahya\textsuperscript{a}, Hend Al-Khalifa\textsuperscript{a}, Sinaa Alageel\textsuperscript{b} and Nora Abanmy\textsuperscript{b}

\textsuperscript{a}College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia
\textsuperscript{b}College of Pharmacy, King Saud University, Riyadh, Saudi Arabia

Abstract

This paper presents a project that explores the possibility of assessing the readability level of Arabic medicine information leaflets using machine learning techniques. There are a number of popular readability formulas and tools that have been successfully used to assess the readability of health-related information in several languages. However, there is limited work on the readability assessment of health-related information, specifically medicine information leaflets in Arabic. We describe the design of a tool that uses machine learning to assess the readability of medicine information leaflets. We utilize a corpus comprising 1112 medicine information leaflets annotated with three difficulty levels. Based on a study of existing literature, we selected a number of features influencing text difficulty. The tool will help specialized organizations in medicine information leaflets production to produce the leaflets at appropriate level of reading for the majority of leaflets consumers.

1. Introduction

Text readability is described as “the ease of understanding or comprehension because of the writing style” \textsuperscript{1}. Text readability assessment provides a measure of the text appropriateness to target readers. It provides a number of potential benefits in many fields such as education and healthcare. In education, it helps teachers match between texts and the reading abilities of students. In healthcare, it helps to provide health-related information such as medical instructions understood by the average patient.

* Corresponding author. Tel.: +966532221149.
E-mail address: sima1mish@gmail.com
There are a number of popular readability formulas and tools that have been successfully used to assess the readability of health-related information in several languages. To the best of our knowledge, there have been a number of researches on automatic Arabic readability measurements of different types of documents such as curriculum textbooks and web articles, however, no existing research has explored the readability of health-related information, specifically medicine information leaflets in Arabic. Medicine Information leaflets or package inserts are considered an important source of health-related information. Medicine Information leaflets are leaflets including information about medical conditions, doses, side effects that packed with medicines to give the user more information about the medicine.

1.1. Problem definition

In Saudi Arabia, a survey of over 2000 community pharmacy customers found that the number of respondents who read package inserts (PIs), or at least ask someone else to read the PIs exceeds 85%. This indicates that PIs sometimes become the main source of information for some patients for many reasons. First, self-medication is common in Saudi Arabia. Second, the purchase of prescription-only medications (POM) without community pharmacies prescription is common in Saudi Arabia. This means that the patient assumes the bulk of responsibility for medication safety not only for over-the-counter (OTC) medications but also for medications such as antibiotics, nonsteroidal anti-inflammatories, and oral contraceptives. Third, the information provided by physicians or pharmacists that describe dose and frequency, precautions, and adverse effects of their prescribed medications has been shown to be suboptimal.

Challenges in reading and understanding such written medicine information leaflets may represent one cause for the high rates of medication errors, such as poor adherence which attenuates optimum medicine benefit. To eliminate this challenge it is necessary to find ways to improve the readability of these leaflets and one such method is based on the assessment of text readability. The main goal of this project is to develop an Arabic text readability assessment tool for consumers of medicine information leaflets. Therefore, our project aims to answer the following question:

Given some medicine information leaflets that are written in Arabic, can a readability level of the text be measured automatically and accurately, and to what extent is it comparable to human judgment?

2. Literature review

Researchers have tried to use several approaches and features that help to achieve an accurate estimate of the readability level for health-related texts when developing readability assessment tools. Several studies have followed the traditional approach which is based on formulas. The researchers found that readability formulas are not specifically tailored for medical content. However, these formulas revealed to serve as an accepted initial approximation for predicting the consumer's comprehension of health topics. Machine learning techniques such as Support Vector Machine, Decision Trees and Naïve Bayes were used in the development of a model that assesses the readability of web-based health information in a number of studies.

In addition, the readability assessment research that uses computational approach is nearly new for the Arabic language. It has been applied on different types of documents such as curriculum textbooks and web articles. However, there is a lack of computational approaches for the purpose of readability assessment of health-related text in Arabic language.

3. Methodology

In this section, we will describe the tools used in our readability assessment tool, to help to generate the statistical language models and accomplish readability assessment task. Then we will describe the readability assessment tool development phases.
3.1. Used tools

- The SRI Language Modeling Toolkit
  SRILM is a toolkit for building and applying statistical language models (LMs), used in speech recognition and machine translation and other applications. SRILM consists of a set of C++ class libraries implementing language models and a set of executable programs built on top of these libraries to perform standard tasks such as training LMs and testing them on data. SRILM runs on UNIX and Windows platforms. SRILM integrated into our tool in the feature extraction phase to generate statistical language models for each readability level.

- Weka Machine Learning Software
  Weka is a machine learning workbench which provides tools for data pre-processing, classification, regression, clustering, and visualization. It also provides a graphical user interface. Weka toolkit integrated into our tool to generate classification algorithms such as SVM, decision tree, and naïve bayes.

3.2. Development phases

Automatic text classification is the process of classifying text into predefined classes. Recently, the techniques of automatic text classification have been utilized in readability assessment field. The effectiveness of classifying a given text into readability levels depends on factors including the selection of the dataset, the choice of algorithm, and the selection of effective features to be extracted from the dataset. The development of our tool includes the following phases: dataset collection, data preprocessing, features extraction, classifier training, and testing (Fig.1).

![Fig. 1. The Readability Assessment Tool](image)

3.2.1. Dataset collection

Our corpus, acquired from King Abdullah Arabic Health Encyclopedia and the Saudi Food and Drug Agency, contains 1112 Medicine Information leaflets in Arabic, annotated with three difficulty levels: easy, intermediate, and difficult. The annotation process was performed by human annotators. The readability of leaflets will not be assessed for the whole document. Instead we will assess readability of selected sentences that relevant to patients from each leaflet.

3.2.2. Data pre-processing

The preprocessing phase focuses on Arabic natural language processing tools and techniques to preprocess the leaflets. The tools can be used to facilitate the extraction of features for the machine learning algorithm. We used
AraNLP library\textsuperscript{23} to apply some preprocessing steps to the leaflets. It includes a tokenizer, part-of-speech tagger (POS-tagger), normalizer, and a punctuation and diacritic remover. We also applied a normalization to Arabic letters that have different forms to a unified form such as (١,٣,١ to ١). The initial results from preprocessing phase showed an increasing in the accuracy of the lexical features extraction. Moreover, the size of a leaflet dropping a 20\%, which improves the efficiency of features extraction.

3.2.3. Features Selection and Extraction

The main goal of this phase is to identify statistical features that can help in the prediction of readability of medicine information leaflets. Also to represent the output text from the previous step as a feature vector that will be the input of the learning algorithm\textsuperscript{4}. Features and combination of features are evaluated to find the best features that provide metrics that measure readability with high accuracy.

Most of particular features are motivated by prior work (shallow text features, lexical difficulty features and part-of-speech ratio features). Shallow text features utilized by most readability formulas. They concentrate on factors, such as word and sentence length, average word length and average sentence length, based on the hypothesis that longer sentences containing long words are more difficult to read\textsuperscript{2}.

The second feature set is lexical difficulty features. Since vocabulary plays an important role in health related text readability, the most important element taken into account are concerned with lexical features\textsuperscript{2}. For example, the unigram language models which have been shown to be effective in the literatures of readability assessment\textsuperscript{2, 18, 19}. Unigram language models features are expected to capture most of vocabulary’s influence\textsuperscript{2}.

We used the POS-based ratio features inspired by Forsyth\textsuperscript{5}, POS-based ratio features were shown to be the most informative in making predictions. POS-based ratio features are the ratio of the individual POS type in Arabic (nouns, verbs, adjectives, and pronoun) occurrences to the token count in the document.

3.2.4. Machine Learning Algorithm

To investigate the possible contribution of different types of features on the readability prediction of medicine information leaflets, we use machine learning techniques, in particular support vector machines (SVMs), to train a classification model on data with class labels, known as training dataset. Then the SVM classifier will be used on data without class labels, the testing dataset, to produce the most likely readability class label (easy, intermediate, and difficult). We selected the Support Vector Machine (SVM), because of their prior success in automatic readability assessment\textsuperscript{25, 26}. Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and more. SVM formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm produces an optimal hyperplane which categorizes new examples\textsuperscript{27}.

3.2.5. Evaluation

In order to evaluate our automated readability assessment tool, we will use 10-fold cross validation as the evaluation method. The main benefit of cross fold validation is that it maximizes the training set size. It also provides sound test results which are more generalizable. We will measure the performance of our readability assessment tool in term of accuracy rate. The accuracy rate is the percentage of documents for which the readability measure correctly predicted their gold-standard levels.

4. Conclusion

In this paper we described our methodology to develop an automatic text readability assessment tool of medicine information leaflets for the Arabic language. We implemented the first increment in the methodology which is the data preprocessing by utilizing an ANLP library.

For future work, we will complete on the implementation of our methodology increments. We will conduct the performance evaluation of our tool. We will show the contributions of the selected features in the readability of Arabic medicine information leaflets and the major findings will be derived for further improvements.
Acknowledgements

This Project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdul-Aziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (INF 2822).

References