FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

AUTOMATIC ASSESSMENT OF HEALTH INFORMATION READABILITY

Hélder Manuel Mouro Antunes

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Carla Teixeira Lopes

July 22, 2019

© The present thesis has been developed within the project "NORTE-01-0145-FEDER-000016", financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

AUTOMATIC ASSESSMENT OF HEALTH INFORMATION READABILITY

Hélder Manuel Mouro Antunes

Mestrado Integrado em Engenharia Informática e Computação

July 22, 2019

Abstract

The current digital age increases the dissemination of information to a large number of people, with health being one of the most popular topics on the web. Health information can contain words with specific terminology that negatively impact the readability of medical content. This problem gets worse when a reader has low health literacy. Envisioning a search system that can personalize the readability of medical content to the characteristics of its users, in this dissertation, we built machine learning models to assess the readability of health content in the Portuguese and English languages.

As a first step, we evaluated the readability of topics on the web using traditional readability measures. We found that the topic of health is one of the least readable topics according to the metrics used. We also analysed the linguistic differences between the English and Portuguese languages, finding that, in general, Portuguese words have a greater number of syllables. With this, we proposed adaptations to the traditional readability metrics originally created for the English language in order to be applicable to the Portuguese language.

After, we assessed the readability of general texts since a health text has the same properties as any text. We used existing features from the state-of-the-art that provided better results. We created four sets of features: traditional features, word familiarity, lexical richness, and part of speech ratios. The models were built on a dataset used in other researches about the readability of general texts. This dataset has texts with five levels of difficulty corresponding to five grade cycles. The texts were originally in the English language and were translated to the Portuguese language. We achieved accuracy close to 80% in the 5-level annotated corpus for both languages, with a mean absolute error inferior to 0.5. Exclusively for the Portuguese language, we made a regression using 65 school books from different subjects through the 1-12 school level grades. We obtained a mean absolute error of 1.69 years of education.

In the end, we extended the models using health features capable of measuring the specificity of a medical text and the knowledge necessary to comprehend it. We created features based on information retrieval ones like collection frequency (CF), document frequency (DF), and inverse document frequency (IDF). In addition to these features, we applied the Collins-Thompson and Callan's statistical language model and other related features initially created to assess the readability of general texts. In our case, the model measures the log-likelihood of a given health text relative to language models of health news and medico-scientific articles. For that, we built a dataset of health news and medico-scientific articles in English and Portuguese languages. We built binary classification models under a dataset based on Wikipedia. We used health documents of the Simple English Wikipedia and the respective documents in the ordinary Wikipedia. We also translated these documents to the Portuguese language. We obtained a very high accuracy only using the general features (> 90%), reason why the health features do not help so much in the improvement of the performance. Nevertheless, the health features, by themselves, obtain an accuracy closed to 90% in both languages, which can be beneficial to differentiate texts taking into account only the medical terminology difficulty.

Resumo

A era digital atual aumentou a disseminação de informação para um grande número de pessoas, sendo a saúde um dos tópicos mais populares da web. As informações de saúde podem conter palavras com terminologia específica que afetam negativamente a inteligibilidade do conteúdo médico. Esse problema piora quando um leitor tem baixa literacia em saúde. Visando um sistema de busca que possa personalizar a inteligibilidade do conteúdo médico à literacia dos seus utilizadores, nesta dissertação, construímos modelos de aprendizagem automática para avaliar a inteligibilidade de conteúdo de saúde nas línguas portuguesa e inglesa.

Numa primeira fase, nós avaliamos a inteligibilidade de tópicos na web usando medidas traditionais de inteligibilidade. Encontramos que o tópico da saúde é um dos tópicos menos inteligíveis de acordo com as métricas usadas. Também procuramos saber as diferenças linguísticas entre o inglês e português, descobrindo que, em geral, as palavras portuguesas têm um maior número de sílabas. Com isso, propusémos adaptações às métricas tradicionais de inteligibilidade originalmente criadas para o inglês de modo a serem aplicáveis ao Português.

De seguida, avaliamos a inteligibilidade de textos gerais, pois um texto médico tem as mesmas propriedades de qualquer texto. Utilizamos features existentes do estado da arte que reportavam melhores resultados. Criamos quatro conjuntos de features: features tradicionais, familiaridade das palavras, riqueza lexical e rácios de classes gramaticais. Os modelos de classificaçao foram construídos em um conjunto de dados usado em outras pesquisas sobre a inteligibilidade de textos gerais. Este conjunto de dados tem textos com cinco níveis de dificuldade correspondentes a cinco ciclos de estudos. Os textos escritos originalmente em inglês foram traduzidos para o português. Atingimos precisão próxima de 80% no corpus de 5 níveis anotado para os dois idiomas, com erro médio absoluto inferior a 0.5. Exclusivamente para a língua portuguesa, fizemos uma regressão usando 65 livros escolares de diferentes disciplinas do 1^{o} ano até ao 12^{o} ano. Obtivemos um erro absoluto médio de 1.69 anos de educação.

A partir disso, estendemos os modelos utilizando features de saúde capazes de medir a especificidade de um texto médico e o conhecimento necessário para compreendê-lo. Criamos features com base em métricas de information retrieval, como collection frequency (CF), document frequency (DF) e inverse document frequency (IDF). Além desses recursos, aplicamos o modelo de linguagem estatística de Collins-Thompson e Callan e outros recursos relacionados inicialmente criados para avaliar a inteligibilidade de textos gerais. No nosso caso, o modelo mede a probabilidade logarítmica de um dado texto de saúde relativo a modelos de linguagem de notícias de saúde e artigos médico-científicos. Para isso, construímos um conjunto de dados de notícias de saúde e artigos médico-científicos. Nós construímos modelos de classificação binária usando documentos da Wikipedia. Utilizamos documentos de saúde do site Simple English Wikipedia e os respectivos documentos na Wikipédia comum. Também traduzimos esses documentos para o idioma português. Obtivemos uma precisão muito alta apenas usando as features gerais (> 90%), razão pela qual as features de saúde não ajudaram tanto na melhoria do desempenho. No entanto, as features de saúde, por si só, obtêm uma precisão próxima de 90% em ambas os idiomas, o que

pode ser benéfico para diferenciar textos levando em conta apenas a dificuldade da terminologia médica.

Acknowledgements

I would first like to thank my supervisor, Professor Carla Teixeira Lopes, for all the given knowledge, help and guidance throughout all the meetings.

Thanks also to Dr. Dagmara Paiva, Professor Alexandra Pinto and Milaydis Sosa for the ideas about the readability of general and health-related texts and for indicating very useful resources for the realization of this dissertation.

A big thanks to Tiago Devezas and Professor Sérgio Nunes for providing a very useful news dataset that greatly reduced my work.

I would also like to thank the Escola Virtual and Porto Editora for giving us online access to Portuguese school books.

For the Integrated Master in Informatics and Computing Engineering (MIEIC) and in particular for Professor João Pascoal Faria, a big thanks for supporting my participation in two conferences where articles resulting of the work of this dissertation were published.

Finally, thank you to my family and friends for being with me throughout this semester and life.

Hélder Antunes

"I'll leave tomorrow's problems to tomorrow's me."

Saitama

Contents

List of Figures

LIST OF FIGURES

List of Tables

[1] https://github.com/HelderAntunes/readability-corpora#general-content-and-resources-en
[2] https://github.com/HelderAntunes/readability-corpora#general-content-and-resources-pt
[3] https://github.com/HelderAntunes/readability-corpora#health-related-content-and-resources-en
[4] https://github.com/HelderAntunes/readability-corpora#health-related-content-and-resources-pt

Abbreviations

ARI	Automated readability index
AWL	Academic Word List
BMJ	British Medical Journal
CART	Classification and regression tree
CF	Collection frequency
CHV	Consumer Health Vocabulary
CL	Coleman Liau
CRISP-DM	Cross Industry Standard Process for Data Mining
CTTR	Corrected type-token ratio
DF	Document frequency
FK	Flesch Kincaid
GF	Gunning Fog
GLMNET	Lasso and Elastic-Net Regularized Generalized Linear Model
GSL	General Service List
GWL	Google Word List
HD-D	Hypergeometric distribution D
HNL	Health News List
HWL	Health Word List
IDF	Inverse document frequency
IR	Information retrieval
K-NN	K-nearest neighbors algorithm
KEWE	Knowledge-enriched Word Embedding
MAE	Mean absolute error
MAL	Medical Articles List
MATTR	Moving average TTR
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
MSTTR	Mean segmental type-token ratio
MTLD	Measure of textual lexical diversity
MeSH	Medical Subject Heading
NLP	Natural Language Processing
PMC	PubMed Central
POS	Parts-Of-Speech
RMSE	Root-mean-square error
RTTR	Root type-token ratio
SMOG	Simple Measure of Gobbledygook
SVM	Support Vector Machine
TTR	Type-Token Ratio
UMLS	Unified Medical Language System

Chapter 1

Introduction

The current digital age increases the dissemination of information to a large number of people, with health being one of the most popular topics on the web. Among the population with Internet access, not everyone has sufficient literacy to understand overly specialized medical content. With this in mind, we seek to create methods that facilitate the recognition of difficult medical content for the general population. Also, such methods would be helpful in the retrieval of easy-to-read documents by search engines and in the assessment of the effectiveness of text simplification methods.

1.1 Context

Readability is the ease with which a reader can understand a text [Wikb]. The difficulty of a text depends on multiple factors like its contents and legibility. Legibility is the ease with which a reader can recognize individual characters in a text [Wika], and will not be considered in this work.

The content assumes an important factor in the medical-related texts since it must be detailed to objectify the ideas of health professionals and at the same time cannot be difficult to be understandable by general people.

1.2 Motivation and Goals

Nowadays, the Internet allows easy access to information. Over the time, the Internet use has been growing and it is estimated that, in June 2018, 55.1% of the world's population had Internet access, being more accentuated in developed countries where the use rate exceeds 80% or even 90% [Int].

Health is one of the popular topics on the web. A study conducted in 2013 points out that one-third of the United States of America adults use the Internet for self-diagnosis and to understand medical concepts [FD]. The online information passed to the people can be useful only if a person understands it.

While readability measures how easily a text can be read and understood, health literacy measures "the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions" [CPRZ00]. Sørensen et al. [SPR+15] conducted a health literacy survey in eight countries of Europe. The authors used four ordered grades of health literacy: insufficient, problematic, sufficient and excellent. The results show that 12% of participants had insufficient health literacy and 47% showed insufficient or problematic health literacy. These results differed substantially across the countries. Specifically for Portugal, Paiva et al. [PSS+17] assessed the health literacy of a sample of the population using the Newest Vital Sign [WMM+05]. The results show that 72.9% of the Portuguese population has limited health literacy.

Studies that evaluate health literacy tests have shown divergent results about the effectiveness of health literacy. By one hand, results show that health literacy is not directly associated with adherence to medicines or other types of treatment [AMW+13, QPM+13, II13]. By the other hand, some studies conclude that lower health literacy is correlated with worse effectiveness from the treatments and is determinant in health behaviour [OSY+13, SSZ+13].

Therefore, people with all levels of literacy, from low to high, can access medical texts. It is important that the content is completely understood to improve the outcomes of certain treatments, medical prescriptions, or even self-care. Thus, a tool to measure the readability of health content may have several uses. It can be used by:

- government and health care institutions that want to maximize understanding of what they want to communicate with the population;
- the medical professionals that want to maximize the likelihood of the message being passed [KMPL14];
- those who publish on the Web for health consumers;
- computer applications (for instance, search engines that want to personalize the retrieved content to the user's literacy);
- text simplification researchers in the assessment of the performance of medical text simplification tools.

The goal of this dissertation is to create methods to assess health content information readability to English and Portuguese languages. The medical texts readability can be affected by the same issues that affect a general content text. Because of that, we will also create methods to assess general content readability. In order to complete these tasks, we will use machine learning to build predictive models with complex and representative features through the use of natural language processing (NLP) methods.

1.3 Contributions

We extended the state-of-the-art in the health-related content readability, creating new features and methodologies. In the general readability, we used well-known features from the state-of-the-art and applied to the Portuguese language where there is a lack of research, tools, and resources.

Throughout our work, we build datasets and resources that may be useful for future work or for other research. During the presentation of using that resources in the next Chapters, we share the website URL where that resources can be accessed. We used a GitHub repository[1] to share these results:

- Health-related news and medico-scientific articles documents for English and Portuguese languages;
- Medical word lists forming a health vocabulary for English and Portuguese languages;
- Medical word frequency lists calculated through the analysis of health news and medico-scientific articles;
- Paired documents of health-related articles collected from Simple English Wikipedia[2] and ordinary English Wikipedia[3].

At this moment, we have two scientific articles accepted, and we will submit one more paper. The first paper accepted [AL19b] shows that the health content is one of the topics less readable in the web, and the second [AL19a] shows the main differences found in English and Portuguese languages and its impact on the traditional readability formulas.

1.4 Dissertation Structure

Besides the introduction, this dissertation contains four more chapters. Chapter 2 describes the state of the art of readability assessment of general texts and health texts. The problems found in the early chapters and the planned solutions to surpass them will be presented in Chapter 3. Chapters 4 and 5 will show the detailed implementation and the results of our approach to the general readability and health readability, respectively. In the end, Chapter 6 describes the main conclusions and points directions to future work.

[1] https://github.com/HelderAntunes/readability-corpora

[2] https://simple.wikipedia.org/wiki/Main_Page

[3] https://en.wikipedia.org/wiki/Main_Page

Chapter 2

Readability Assessment

In this chapter, we present the current state-of-the-art of readability. We describe the used processes and features to evaluate the readability for general texts. For the health readability, we show the features, datasets and medical resources commonly used in researches.

2.1 Readability assessment of general textual content

Although our focus is on health-related text, this type of content is text and, as so, we will also consider work on the readability assessment of general text. In this section, we will present the existing approaches, their comparison and the current effectiveness of the general readability measures.

Before presenting the approaches and methods to assess readability, we define readability. One definition is given by Richards et al. [Ric02] stating that readability is "how easily written materials can be read and understood. This depends on several factors including the average length of sentences, the number of new words contained, and the grammatical complexity of the language used in a passage". Other authors of classic readability formulas also provided definitions. The author of Simple Measure of Gobbledygook (SMOG) formula, Harry McLaghlin [McL69], defined readability as "the degree to which a given class of people find certain reading matter compelling and comprehensible." An even older definition is provided by Dale and Chall [DC49], stating that readability is “the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers has with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting”.

An important thing to note is the difference between readability and legibility. Legibility is the ease with which a reader can visually recognize individual characters, words, and sentences in a text and depends on typographic factors and design like font, colour or text justification. The readability of a text can, obviously, be influenced by its legibility.

In the next sections, we present two types of approaches to assess the readability of general content: one is more traditional using simpler features, typically based on surface characteristics of text, and the other combines machine learning methods and natural language processing (NLP) techniques.

2.1.1 Traditional approaches

The evaluation of textual readability goes back to the last century with the creation of several formulas of readability. These formulas evaluate the syntactic part with respect to the phrases and the semantic part that refers to the words, generally providing the year of schooling necessary to understand the text. Formula operands are easy to compute. For example, several formulas use the average number of syllables per word to evaluate the lexical part and the mean sentence size to evaluate the sentence difficulty.

One of the most used formulas is the Flesch-Kincaid [Kin75]:

$$RG = 0.39 \times AverageWordsPerSentence + 11.8 \times AverageSyllablesPerWord - 15.59$$

in which RG represents the degree of schooling required to read the text.

Another sub-type of traditional methods include the ‘vocabulary-based’ measures. These methods estimate the semantic difficulty of a word by checking whether the word is in a pre-specified word list. For instance, the revised Revised Dale-Chall formula [CD75] uses a list of 3,000 words familiar to 4th graders and, therefore, contains words more easy to understand. In recent approaches, a word is harder to read if it appears less in a large standard corpus, and easier if it is more frequent. An example of this measure is the Lexile measure [LB04] that uses the Carroll-Davies-Richman text corpus [CR71].

Classic readability measures have limitations. They provide simple formulas that are easy to compute but they are based on surface characteristics of text, ignoring other important aspects such as cohesion, coherence, word ambiguity/specificity and conceptual density (number of ideas in a text). With the current high computational power and the quantity of available data, it is possible to deal with this problem using machine learning and NLP. In the next section, we give an overview of machine learning methods applied to automatic readability assessment.

2.1.2 Machine learning classifiers and NLP methods

The combination of machine learning and NLP enables the use of many features that are not considered in the classic approaches. Because of that, the resultant models give, in general, better prediction accuracy [FM12].

As defined in the survey done by Collins-Thompson [CT15], the approach consists of three steps executed in this order: construction of a gold-standard corpus, the definition of a set of features and the use of a machine learning model. To a more detailed view of these elements and how they relate to each other, see the Figure 2.1.

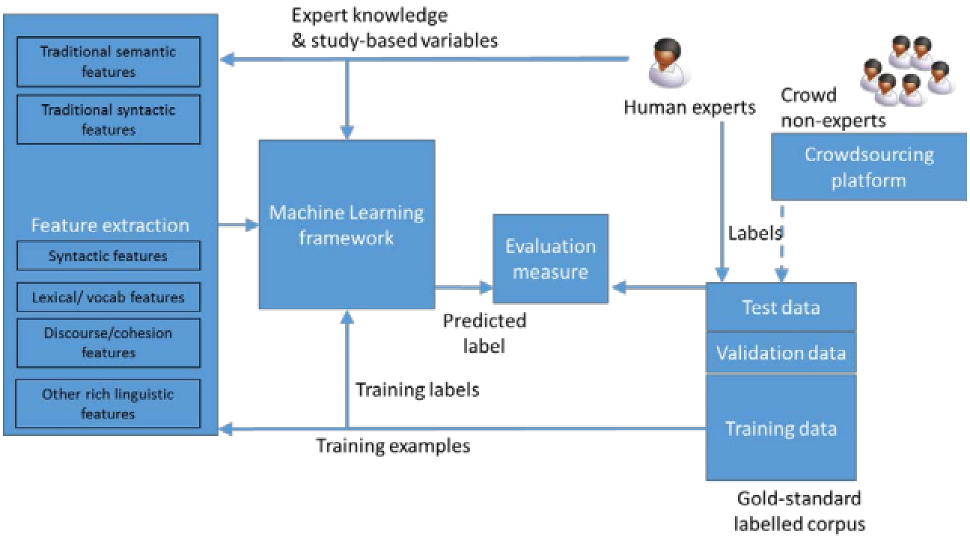


Figure 2.1: Typical computational reading difficulty estimation pipeline [CT15].

2.1.3 Readability features types

One way of categorizing the readability features is also presented in the survey by Collins-Thompson [CT15]. Authors identify the following readability features: legibility, lexico-semantic, morphological, syntax, discourse, higher-level semantics, pragmatic and user knowledge. More information on this is presented in Figure 2.2.

In the next subsections, we detail these features and other categories showing studies and existing tools that use them.

2.1.3.1 Lexico-semantic features

The traditional approaches, in order to estimate the word difficulty, use the average number of syllables/letters per word and, like the 'vocabulary-based' measures, might use a reference list of difficulty/easy words and assess the frequency of the words in that list. Other types of features include the lexical diversity, or type-token ratio (TTR), that is defined as the ratio of different unique terms to the total number of words observed in a text. It is assumed that texts with higher TTR, using a broader vocabulary, are more difficult to understand.

Another lexical feature is provided by statistical language models. Collins-Thompson and Callan' statistical language model [CTC04] consists in assigning to each word a probability of being found in a text of a specified grade. That method uses a set of training data, usually one for each grade level of education. That way, a statistical language model provides a histogram of the probability distribution of a word across all grade levels. For instance, a difficult word would have a histogram with a peak in high grades and a readable word in low grades.

The Word Maturity measure [KL11, LKP11], like the statistical language models, tracks the evolution of words and phrases along the learning stage. But one additional detail is added, the word's usage in multiple contexts giving the reader's degree of knowledge that is expected to

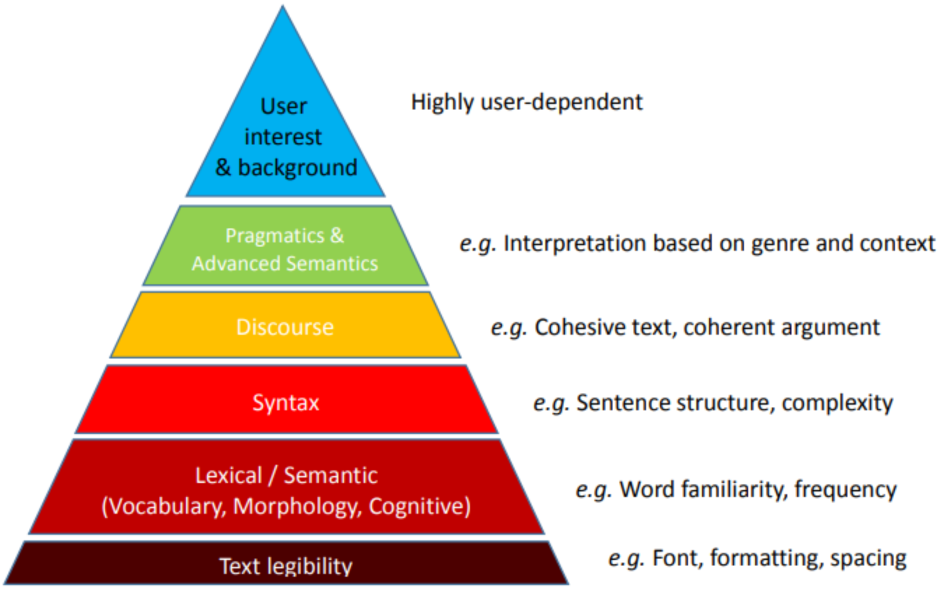

User
interest
& background
Highly user-dependent
Pragmatics &
Advanced Semantics
e.g. Interpretation based on genre and context
Discourse
e.g. Cohesive text, coherent argument
Syntax
e.g. Sentence structure, complexity
Lexical / Semantic
(Vocabulary, Morphology, Cognitive)
e.g. Word familiarity, frequency
Text legibility
e.g. Font, formatting, spacing

Figure 2.2: Types of readability features grouped by Collins-Thompson [CT15].

have. This detail allows a more personalized readability measure. This model uses Latent Semantic Analysis [DDF+90] to extract the topics that characterize the word's context in a particular learning stage or grade level.

Some languages have rich inflectional and a high morphology derivation that changes the meaning of certain words. Different words suffixes or prefixes and other morphological operations have an important role in readability assessment as shown by Hancke et al. [HVM12].

2.1.3.2 Psycholinguistics-based lexical features

Psycholinguistics-based lexical type of features includes word concreteness, average age-of-acquisition (old words should be more difficult to recent generations) and degree of polysemy (one is difficult if it has multiple meanings). The word concreteness feature has been an important feature of readability. Previous works defined word concreteness [PYM68, Ric75] based on perceivability (ability to perceive an object) and abstraction (ability to imagine and understand the concept). It was used by Tanaka et al. [TJKT13] to measure the text comprehensibility. Some studies [CGM08, VM12] have applied cognitive-based lexical features to second-language learners with good results.

2.1.3.3 Syntactic features

Syntactic complexity relates to sentence difficulty and organization. The syntactic difficulty has a high impact on the processing time to comprehend a text [Gib98]. The traditional way of measuring syntactic complexity is through the average sentence length. With the advances of NLP it is possible to perform a deep analysis of texts like parse trees (see an example of a parse tree in the

Figure 2.3). The common syntactic features provided by NLP for readability assessment are the proportion of incomplete parses (for instance, when a sentence does not conform to a given grammar), average parsed tree height, and average number of nouns/verbs/subordinate clauses phrases per sentence.

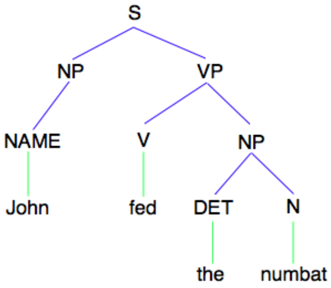


Figure 2.3: Sentence structure presented by a parse tree [Wil].

Pitler and Nenkova [PN08] found that the average number of verb phrases per sentence had the highest correlation with the readability in a news corpus. Other detected featured was the average parse tree height. Many more applications of this type of features can be found in many works thanks to the high number of available of NLP tools [SO05, HCTCE07, KLP+10, TC12, RB15, XKB16, CSD+17, CMB14].

2.1.3.4 Discourse-based features

These features are related to the cohesion and coherence of a text. Cohesion is the dependency between the interpretation of two elements in the same text [HH76]. For instance, one element can only be understood if a previous element was understood. Coherence refers to the logical order of arguments and ideas and the organizational structure of a text. Both properties are significant to the understanding of a text and are ignored in the traditional readability measures. Multiples works use these features, for instance: Pitler and Nenkova study for English texts [PN08], Todirascu et al. [TFG+13] and Dascalu [Das14] for the French language, and Sung et al. [SCC+14] for the Chinese language.

2.1.3.5 Higher-level semantic and pragmatic features

The readability of a text is also dependent on the reader since two different persons with the same educational level can perceive differently the same text. This understanding is not only dependent on the reader education level, but also of the reader domain knowledge and what topics the reader find interesting. This can include text domain, specific idioms, local references, cultural context, and sentiments embedded in the text. Honkela et al. [HIL12] conducted a study that consists of a system that searches content related to the interests of the user by providing encouragement and emotive relevance using higher-level semantic and pragmatic features. In general, this type of features are scarcely used in readability studies and are a topic for future research.

2.1.3.6 Word Embeddings

Word embedding consists of mapping words or phrases to vectors of real numbers. It involves a mathematical embedding from a space with one dimension per word and a very large dimensions for a vocabulary to a continuous vector space.

In 2017, Cha et al. [CGK17] used word embeddings to represent semantic features appropriate for text regression. Word embeddings algorithms hypothesize that the word co-occurrences imply similar meaning or/and context [Har54]. This provides an high-dimensional semantic space where the Euclidean distance between two vectors representing words quantifies their semantic dissimilarity [HAMJ16]. The authors performed a cluster on word embeddings creating a histogram of cluster membership for each word in a text. That histogram, after normalized, is used as a feature in a linear support vector machine (SVM) regression.

Jiang et al. [JGYC18], in 2018, provided the knowledge-enriched word embedding (KEWE), when a vector representing a word has encoded some reading difficulty of the word. This approach expands the work of Cha et al. [CGK17] that assumes that words similar in the semantic aspect have also similar difficult, for instance, words with a similar semantic aspect, such as “man” and “gentleman”, are mapped into close vectors although their reading difficulties are different. Therefore, the authors encode readability features in the word vector, such as acquisition difficulty, usage frequency and structure difficulty (number of syllables, has a suffix, among others).

2.1.4 Models

In general, a computational readability prediction is a function that outputs the school grade level necessary to read an input text or the level of difficulty of the input text. In that sense, the problem can be approached as a classification task (each grade level is a category, ordered or not), regression problem or ranking problem (comparison of readability texts). Regression and classifications are the most used methods [CT15]. However, some studies treat the problem as a ranking problem. For instance, Pitler and Nenkova [PN08] compare the readability of a pair of documents, and Tanaka-Ishii et al. [TITT10] combine the readability pairwise evaluation to order a set of texts by readability. More recently, Cha et al. [CGK17] combines a language modeling by clustering with regression classification model, using word embeddings (see Figure 2.4). The regression model takes as input the histogram of the computed clusters.

As Collins-Thompson’s survey [CT15] points out, the nature of the features is more relevant than the model/learning framework. Obviously, the choice of the learning framework is important, but the main gain of the new approaches in comparison with the traditional measures resides in adding more complex and representative features. For example, in addition to the accuracy, a model can give other details that can be important to decide which model to choose. Adding a confidence level to a prediction can be an interesting output, and only certain models have that capacity, like Bayesian regression or Naive Bayes classifier. Another important application of a readability tool is to point out how a prediction was done. For that, models like decision trees and regression models are more explanatory.

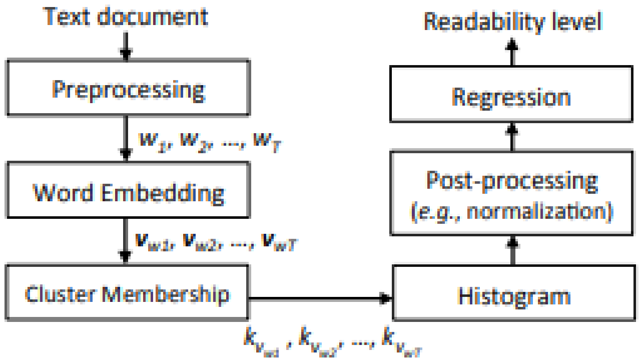


Figure 2.4: System pipeline built by Cha et al. combining clustering model with regression classifiers [CGK17].

2.1.5 Evaluation corpora

An evaluation corpus is a set of passages in which a text is assigned to a school grade or difficulty level. The manual classification is normally made by linguistic experts. It is important to know the process of the expert's classification. Sometimes, existing readability measures are used as support to the manual labelling, generating a performance bias in favour of the readability measures used. Some studies also used texts from school books as datasets, using the school year of the book to label the text contained in that book.

In general, most studies use their own corpora. However, there are some public resources frequently used like: Common Core Appendix B containing 168 docs labeled to US grade levels 2-12, graded articles for elementary students in Weekly Reader Corporation, the WeeBit corpus constructed by Vajjala and Meurers [VM12], the OneStopEnglish corpus built by Vajjala and Lucic [VL18] and several easy/difficult corpora available from Simple English Wikipedia (`simple.wikipedia.org`) and the ordinary Wikipedia (`en.wikipedia.org`).

2.1.6 Evaluation measures

The main used evaluation measure is *rank order correlation* between the predicted text difficulty and the "gold standard" text difficulty label. The advantage of this measure is that only ranks the texts ordering it by metrics comparison, and it is not necessary to normalize the results of a model with the scale of the dataset or the desirable output. Spearman's rho is a very used rank correlation measure that is a nonparametric measure, assessing monotonic relationships (whether linear or not). Another very used measure is the Pearson correlation. The prediction accuracy, measuring the percentage of corrected predictions, is another evaluation measure. This measure ignores the size of the error of an incorrect prediction. The Mean Squared Error (MSE) is a measure of error penalizing large errors, being important to a classification task with high number of classes.

In machine learning models, the cross-validation technique is the most used technique to estimate the performance of a predictive model. Cross-validation tests the model's ability to predict new data that was not used in training phase, in order to detect problems like overfitting or selection bias and to give an insight on the generalization capacity of the model.

2.1.7 Current accuracy of readability measures

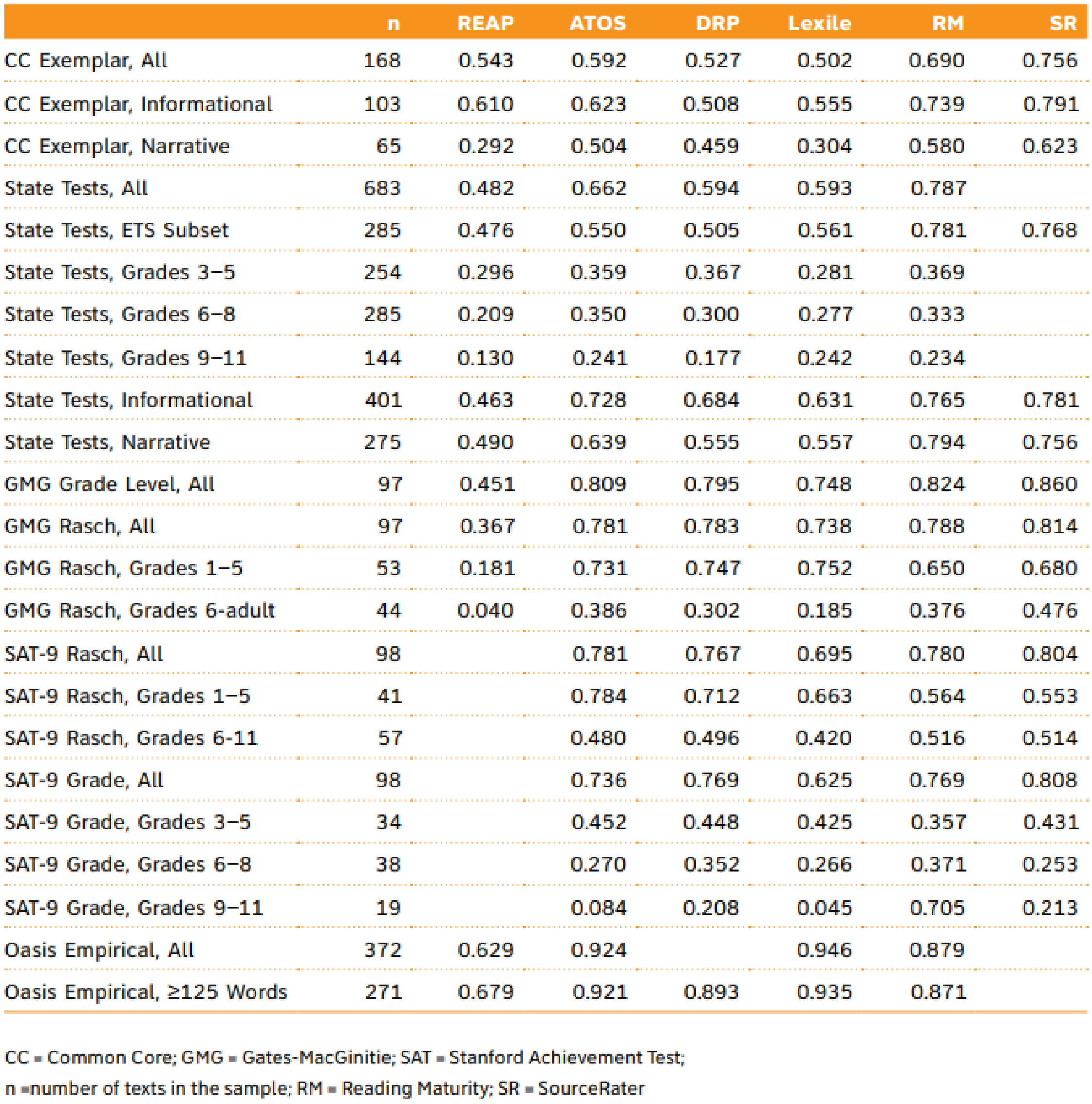
In 2012, Nelson et al. [PL12] assessed six readability measures: Lexile (MetaMetrics), ATOS (Renaissance Learning), Degrees of Reading Power: DRP Analyzer (DRP), REAP (Carnegie Mellon University), SourceRater (SR), and the Pearson Reading Maturity Metric (RM). These metrics were tested in several sets of texts: 1) the set of exemplary texts that were placed into grade levels by education experts and published as Appendix B of the Common Core Standards, 2) a set of standardized state test passages, 3) passages from the Stanford Achievement Test (SAT-9), 4) comprehension passages from the Gates-MacGinitie Reading Test, and 5) passages from the MetaMetrics Oasis platform used for student practice. The authors used the Spearman's rank correlation coefficient (Spearman's rho) as an evaluation measure. The results are shown in Figure 2.5. The performance of the metrics varies across the sets of texts. The REAP metric performed badly in the most sets of texts. The other metrics performed better, reaching often a correlation close to 80% (considering all grades of a set of texts). One important fact was that the readability measures were more accurate in low-grade material than in high-grade material.

2.1.8 Support to the Portuguese language

The work of this dissertation also aims to provide readability tools to the Portuguese language. One of the first systems built was REAP.PT [LMV09], which is a learning support system developed for the Portuguese language. It was adapted from the REAP system originally designed for teaching English as a second language [CTC03]. The REAP.PT bases the calculation of the readability in lexical characteristics, such as the word frequencies used in it. This frequency is captured by a language model based on unigrams. SVM was used to classify the text in a scale of 8 levels, numbered from 5 to 12, corresponding to the level of schooling of the training materials and test used.

Another tool is Coh-Metrix-Port [SA10], an online tool that calculates parameters to measure the cohesion, coherence and difficulty of a text. Specifically developed for Brazilian Portuguese, it was adapted from the Coh-Metrix [GMLC04] system originally developed for the English language.

For evaluation of readability for Portuguese as a second language, the online system LX-CEFR classifies in a scale of five levels of proficiency as defined by QuaREPE (A1, A2, B1, B2 e C1) [MJGP11]. The measurement of the level of difficulty of the texts takes into account four characteristics: Flesch Reading Ease score [Kin75], nouns frequency, average syllables per words, and average words per sentence. Another study realized by Curto et al. [CMB14] shows a system that classifies text readability for European Portuguese, based in the same levels provided by QuaREPE. The classifier makes use of 52 features grouped in 7 types: parts-of-speech (POS),



	n	REAP	ATOS	DRP	Lexile	RM	SR
CC Exemplar, All	168	0.543	0.592	0.527	0.502	0.690	0.756
CC Exemplar, Informational	103	0.610	0.623	0.508	0.555	0.739	0.791
CC Exemplar, Narrative	65	0.292	0.504	0.459	0.304	0.580	0.623
State Tests, All	683	0.482	0.662	0.594	0.593	0.787	
State Tests, ETS Subset	285	0.476	0.550	0.505	0.561	0.781	0.768
State Tests, Grades 3–5	254	0.296	0.359	0.367	0.281	0.369	
State Tests, Grades 6–8	285	0.209	0.350	0.300	0.277	0.333	
State Tests, Grades 9–11	144	0.130	0.241	0.177	0.242	0.234	
State Tests, Informational	401	0.463	0.728	0.684	0.631	0.765	0.781
State Tests, Narrative	275	0.490	0.639	0.555	0.557	0.794	0.756
GMG Grade Level, All	97	0.451	0.809	0.795	0.748	0.824	0.860
GMG Rasch, All	97	0.367	0.781	0.783	0.738	0.788	0.814
GMG Rasch, Grades 1–5	53	0.181	0.731	0.747	0.752	0.650	0.680
GMG Rasch, Grades 6-adult	44	0.040	0.386	0.302	0.185	0.376	0.476
SAT-9 Rasch, All	98		0.781	0.767	0.695	0.780	0.804
SAT-9 Rasch, Grades 1–5	41		0.784	0.712	0.663	0.564	0.553
SAT-9 Rasch, Grades 6-11	57		0.480	0.496	0.420	0.516	0.514
SAT-9 Grade, All	98		0.736	0.769	0.625	0.769	0.808
SAT-9 Grade, Grades 3–5	34		0.452	0.448	0.425	0.357	0.431
SAT-9 Grade, Grades 6–8	38		0.270	0.352	0.266	0.371	0.253
SAT-9 Grade, Grades 9–11	19		0.084	0.208	0.045	0.705	0.213
Oasis Empirical, All	372	0.629	0.924		0.946	0.879	
Oasis Empirical, ≥125 Words	271	0.679	0.921	0.893	0.935	0.871	

CC = Common Core; GMG = Gates-MacGinitie; SAT = Stanford Achievement Test;
n =number of texts in the sample; RM = Reading Maturity; SR = SourceRater

Figure 2.5: Spearman's rho results in study conducted by Nelson et al. [PL12].

syllables, words, chunks and phrases, averages and frequencies, and some extra features using NLP techniques.

There is a lack of research in the Portuguese language as a native language. The system developed by Marujo et al. [LMV09] only uses traditional features and ignores the advanced features that are provided by the current techniques in NLP.

2.2 Readability assessment of health textual content

Medical texts have specific characteristics that can be analysed in more detail. The specific characteristics of the terminology usually employed in medical documents can be used to enlarge the set of features used by classifiers when compared with ones used for general content. For this reason, a general content readability predictive model may be improved if it considers the specificity of health content.

In the next subsections, we present an overview of machine learning methods and the common resources and datasets used to assess the readability of health-related content.

2.2.1 Medical resources

Medical resources have been used to evaluate the semantic aspects of health-related documents, like the extraction of a hierarchy of words to calculate word specificity or identification of concepts related to a word. These features were mainly used by Kauchak et al. [KMPL14] through UMLS and its associated components like Metathesaurus and MeSH. MetaMap [AL10] and Consumer health Vocabulary (CHV) system [ZT06] were used by Palotti et. al [PZH19] to find medical terms. The same researchers used ADAM database [ZTS06] and the medical dictionary OpenMedSpel [e-M] to find the frequency of medical words.

For the English language, there are several reliable medical resources described in the Table 2.1. For non-English languages like the Portuguese language, there do not seem to be many resources.

2.2.2 Health specific features

The readability classifiers of health texts can use the same features of general classifiers, as the lexical, syntax and discourse features continue to play an important role in this type of texts. The additional features that can be considered too mainly assess the semantics aspects of medical vocabulary. In the next subsections, we present the features categories used in the previous works.

2.2.2.1 Medical concept density

The number of medical concepts in a text can influence the text difficulty. Kauchak et al. [KMPL14] ignore words contained in the Dale-Chall List [DC77], a list of frequent and easier words. The concept density was calculated counting the number of the remaining words found in the Metathesaurus of the Unified Medical Language System (UMLS).

Palotti et. al [PZH19] distinguishes consumer medical concepts from expert medical concepts. For counting of consumer medical concepts, they use MetaMap [AL10] to recognize medical words in a text, and use it as entries in the Consumer Health Vocabulary (CHV) [ZT06]. For expert medical concepts, they also used MetaMap, but to count the number of Medical Subject Heading (MeSH - hierarchically-organized terminology for biomedical information) entities.

[1] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/
[2] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/
[3] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/
[4] https://metamap.nlm.nih.gov/
[5] https://www.nlm.nih.gov/mesh/meshhome.html
[6] https://www.openhub.net/p/openmedspel

Resource	**Descritption**
Unified Medical Language System (UMLS)[1]	A system that brings together many health and biomedical vocabularies and standards.
Metathesaurus[2]	The biggest component of UMLS representing a semantic network linking similar names for the same concept.
Consumer Health Vocabulary (CHV)[3]	It connects consumer terms about health to technical terms used by health care professionals.
MetaMap[4]	A system that recognize medical words, discovering Metathesaurus concepts in a given text.
Medical Subject Heading (MeSH)[5]	A resource that provides a hierarchically-organized terminology for indexing and cataloging biomedical information.
ADAM [ZTS06]	A database of abbreviations and their long-forms in the biomedical domain.
OpenMedSpel[6]	A medical dictionary that includes nearly 50,000 medical terms.

Table 2.1: Usual medical resources used in health readability research.

2.2.2.2 Specificity

Kauchak et al. [KMPL14] define specificity as "the technicality of a term in the medical domain". For example, they compare the terms "heart" and "endocardium" that are related to the cardiovascular system. The "heart" term is more accessible and familiar for most people, and is, therefore, less specific.

To measure the specificity of a term, not necessarily health-related, methods like simple word frequency statistics have been applied [CC99]. The study of Kauchak et al. [KMPL14] relies on the depth of a term within a word hierarchy of medical terms. To measure the specificity of a word, they use the hierarchically-organized terminology of MeSH.

2.2.2.3 Ambiguity

According to Kauchak et al. [KMPL14], ambiguity “is the vagueness or uncertainty of the exact meaning of a term”. The results obtained by Rayner et al. [RD86] suggest that the time fixating (related to time understanding) ambiguous words with two equally likely meanings is bigger than the time fixating of ambiguous words with a totally different meaning.

To measure ambiguity, Kauchak et al. [KMPL14] count the number of different concepts that a word is associated with, using the UMLS system already cited. They assume that it is a good heuristic to count the number of multiple meanings of a medical word.

2.2.2.4 Frequency of medical words/acronyms

Miller et al. [MLC+07] use a model where the frequency of occurrence of each word is used as a feature for training a classifier. Later, the data collected was used to train a Naive Bayes classifier. This classifier is a popular method for text categorization, like the prediction of the categories of documents such as spam or legitimate, sports or politics. In this approach, other relevant features to general readability were ignored.

Palotti et al. [PZH19] count the number of words with medical prefix/suffix, the number of medical acronyms (ADAM database [ZTS06]) and the number of words in a medical dictionary OpenMedSpel [e-M].

2.2.3 Datasets

The main works create datasets based on the type of documents that usually are easy or difficult. Kauchak et al. [KMPL14] take texts that are well known more simple, like the Simple English Wikipedia (`http://simple.wikipedia.org/`), and difficulty texts that are more probably in normal English Wikipedia (`http://en.wikipedia.org/`). Miller et al. [MLC+07] take easy documents from web blogs used by people without the medical knowledge and the hard documents from journal articles oriented to medical research.

In general, like the general content readability problem, there is also a properly annotated and organized data for non-English languages.

2.2.4 Current accuracy of health readability measures

Miller et al. [MLC+07] used a Naive Bayes classifier in a corpus annotated with three levels of increasing medical terminology specificity: consumer/patient (collected 50 documents from blog sites), novice health learner (collected from 50 web pages of the City of Hope National Medical Center website), and medical professional (50 journal articles from the Journal of the American Medical Association). The authors tokenized all the words of texts, medical and non-medical words. Summing up the probabilities from all of the tokens, one can obtain numeric estimates representing the likelihood that the document belongs to a given category. With this approach, they obtained an accuracy of 96% using leave-one-out validation.

The study conducted by Kauchak et al. [KMPL14] obtained an accuracy of 84.14% using Random Forest as the learning method. The authors used health-related features on 118,000 simple and difficult sentences from a sentence-aligned corpus. This corpus contained sentences of articles from Simple English Wikipedia and ordinary Wikipedia [CK11]. However, the documents used in this dataset belong to non-medical categories.

Palotti et al. [PZH19] concluded that machine learning predictive models were more suitable to estimate health Web page readability than traditional readability formulas. They tested several measures of readability in texts from the Conference and Labs of the Evaluation Forum eHealth 2015 collection, and found a Pearson correlation of 0.602 using a gradient boosting regressor and a Pearson correlation of 0.438 using the Simple Measure of Gobbledygook Index.

2.3 Summary

The machine learning classifiers combined with more complex and representative features can capture deeper and broader aspects of text than the traditional approaches. This is only possible due to the advances in NLP. The types of readability features can be divided into text legibility, lexical/semantic, syntactic, discourse-based and higher-level semantic, pragmatic features and user interest and background. Text legibility is ignored from readability classifiers, even though it is an important property of a text. The user interest and background are highly user-dependent and need for adaptive readability algorithms, that, at the moment, it is an unexplored field. The other features, provided by NLP algorithms, improve the accuracy of machine learning models in comparison with the simple features like average size sentence/word.

Even so, there is a lack of datasets properly annotated by grade level, and the existing ones that are frequently used and public are mainly for the English language. The lack of data is even rarer for specific domains like health. For the Portuguese language as a second language there are some systems that used advanced features provided by NLP. However, the Portuguese language as a native language doesn't have a system using the same advanced features.

The automatic readability assessment of health content can be done using machine learning and NLP methods and use the relevant features found to treat the general content texts. Apart from that, some medical specific features are added to evaluate the semantic aspects of health concepts included in this type of texts. The works here analysed use features like medical concept density, specificity and ambiguity words, and frequency of words in a corpus. The readability classifiers specialized in health-related texts can be very helpful to identify texts understood by people with low health literacy.

Like in the general readability assessment, the lack of proper and annotated health data is a barrier to research, even more in non-English languages. To the Portuguese language, to the best of our knowledge, there is no readability system to assess health-related content.

Chapter 3

Assessing readability of health-related content in English and Portuguese

In Section 2.1, we presented the traditional and the recent approaches of readability assessment. The features used in the recent approaches through NLP and machine learning proves to be more effective than traditional approaches [FM12]. In Section 2.2, features related to health content readability assessment were presented in the few studies that we found.

We identify some problems in general and health-related content readability assessment that we expose in the next section. We will also present the solution and the methodology to solve the problems identified, as well as the evaluation methods to assess the correct fulfilment of the objectives.

3.1 Problem

We want to contribute with advances in the automatic measurement of the readability of health-related texts since the medical field is very impacting in people's lives. The main goal of our work is to propose new features and combine them with the ones typically used, hoping that helps to increase the accuracy of classifiers models.

A possible classifier model for medical texts should extend features used in the general content texts. We want to use features already used in previous works. For that reason, we will start by creating a general text readability classifier.

Specifically for the Portuguese language, to the best of our knowledge, there is no medical readability tool or research work. Like the health text readability, there is also a lack of support to the Portuguese native language. We detect more advances in Portuguese as a second language [CMB14]. To the Portuguese native language, there is the REAP.PT [LMV09] system that uses only four features making little use of the advances in NLP. These kind of tools are very important to the general population and can be very useful to the simplification of bureaucratic and complicated content provided by government organizations [Gov].

On the other hand, another goal is to contribute to the scientific community through the datasets that we will produce, which is specially important for the Portuguese language.

3.2 Solution

We will use machine learning to build readability classification models, and NLP to generate features to these models. We make use of the more recent and advanced features used until now in previous works, and identify new types of features for health-related content.

We will develop classification models for general information texts and for health-related texts. We will apply it to English and Portuguese languages.

The datasets we will create will be available online to public access and future research.

3.3 Methodology

As shown in the Figure 3.1, the methodology of our work has three phases: initial phase, general readability and health readability.

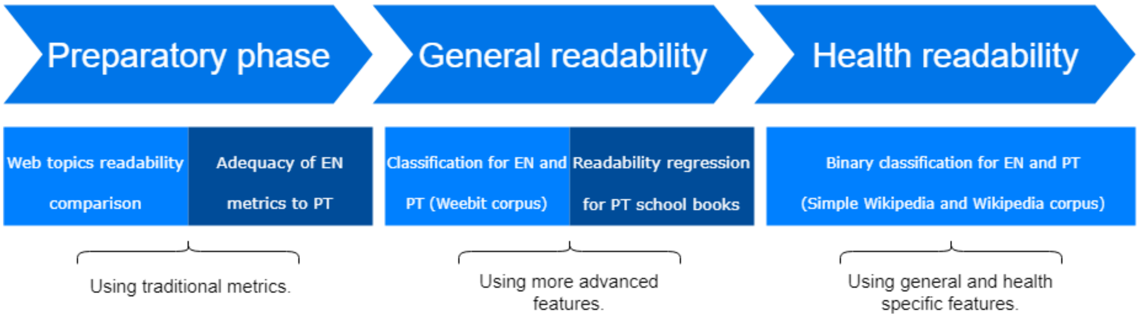

Preparatory phase
General readability
Health readability
Web topics readability comparison
Adequacy of EN metrics to PT
Classification for EN and PT (Weebit corpus)
Readability regression for PT school books
Binary classification for EN and PT (Simple Wikipedia and Wikipedia corpus)
Using traditional metrics.
Using more advanced features.
Using general and health specific features.

Figure 3.1: General methodology.

In the first phase, we assessed and compared the readability of topics on the web and analysed the main differences between the English and Portuguese languages.

By the analysis of the readability through traditional formulas of 278,081 web documents categorized in 20 topics [AL19b], we found that 'Health' topic is the second less readable topic (see the Figure 3.2 for more details). This result shows that medical content has a high average word and sentence size, and a specialized classifier for this topic would be useful for a wide range of people.

In order to analyse the differences between the English and Portuguese languages [AL19a], we applied five traditional formulas - SMOG [McL69], Flesch Kincaid (FK) [Kin75], ARI [SS67], Coleman Liau (CL) [CLL75], and Gunning Fog (GF) [Gun52] - in 10 parallel corpora from the

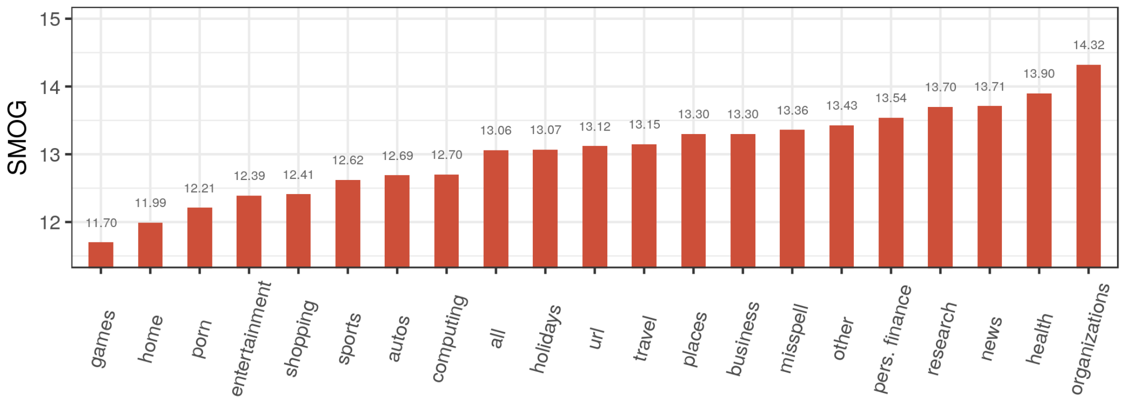


Figure 3.2: Means of SMOG metric ordered by more readable topic in the web pages collected. [AL19b]

OPUS collection[1] of English texts and the corresponding translated Portuguese texts. We verified that the Portuguese language scores lower readability values in comparison with the English language (see Figure 3.3). This difference is more pronounced in the formulas that use the number of syllables per words or number of complex words per sentence (SMOG, FK, and GF). Formulas using the number of characters per word as word difficulty parameter does not differ so much between the languages (ARI and CL). We found out, using texts of school books, that the concept of complex word as a word with 4 or more syllables, instead of 3 or more syllables as originally used in traditional formulas, is more appropriate to the Portuguese language. In the end, for each traditional readability formula, we adapted it to the Portuguese language. The adjusted formulas are shown in Table 3.1. For each school grade, we calculated the mean of the parameters presented in those formulas (WO, SE, etc). Only then we apply a multiple linear regression.

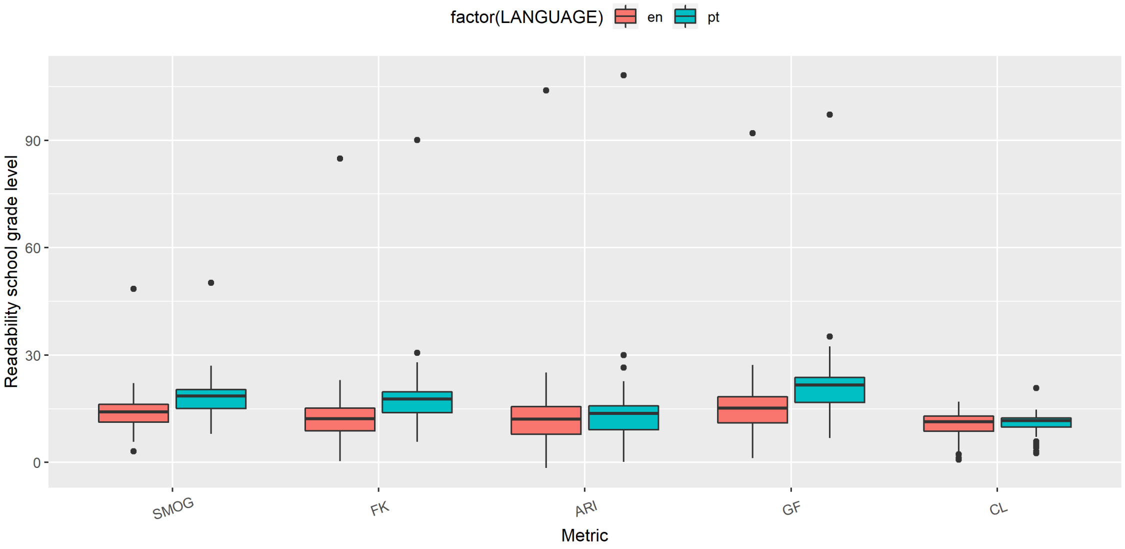


Figure 3.3: Metrics score comparison between languages in all parallel corpora. [AL19a]

[1] http://opus.nlpl.eu/index.php

Metric	Formula	RSE	Error rate
SMOG	$16.830 \times \sqrt{CW \times 30 \div SE} - 23.809$	1.469	0.225
FK	$0.883 \times WO \div SE + 17.347 \times SY \div WO - 41.239$	0.987	0.152
ARI	$6.286 \times CH \div WO + 0.927 \times WO \div SE - 36.551$	1.064	0.164
CL	$5.730 \times CH \div WO - 171.365 \times SE \div WO - 6.662$	1.375	0.212
GF	$0.760 \times WO \div SE + 58.600 \times CW \div WO - 12.166$	1.001	0.154

Table 3.1: Adjusted Portuguese formulas.

CH - characters, CW - complex words, SY - syllables, WO - words, SE - sentences, RSE - residual standard error.

After, we built models for general readability for both languages. The previously created features in general readability were used to built health readability models with a health-specific dataset.

In each Data Mining step of the previous methodology (General and Health readability models), we applied a modified version of CRISP-DM (Cross Industry Standard Process for Data Mining) [SD00]. From the phases of CRISP-DM (Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment), we didn't complete the deployment phase, since this work is intended to be more exploratory about readability features and the models itself. We also ignore the business understanding phase corresponding to the research and study of readability concepts in the current state of the art (Sections 2.1 and 2.2). The modified version of CRISP-DM is presented in the Figure 3.4. We detail the tasks and goals of that version in the next subsections.

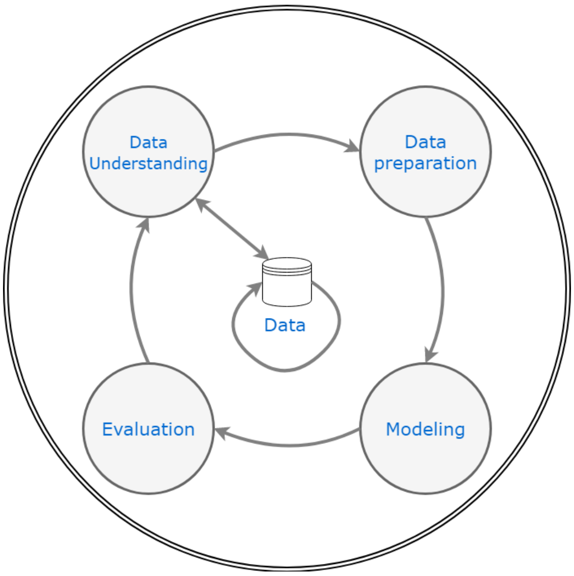


Figure 3.4: Phases of our modified version of CRISP-DM.

3.3.1 Data Understanding

For a better understanding of data, we made an exploratory data analysis using for instance histograms and correlation tables. For that, we used the R language taking advantage of packages like DataExplorer[2] and ggplot2[3].

Along the process, sometimes the data understanding phase shown that the chosen data was not the most appropriate. For instance, the health readability models firstly used health news and medico-scientific articles, but we noted in that the differences between this type of data was very high (average word and sentence length were already differentiating features). Then, we chose a different type of documents, documents of the Simple English Wikipedia and the ordinary Wikipedia, with fewer differences between them.

3.3.2 Data preparation

The datasets used in other research works did not need special preparation, we just transformed it in XML format. On the contrary, the created datasets were collected via web scraping using the Python library BeautifulSoup[4].

The machine learning models don't receive texts as input, and we needed to transform the texts in numbers using suitable features for readability. We used NLP through the Python tools NLTK[5] and Spacy[6].

3.3.3 Modeling

For the modelling phase, we used several machine learning algorithms. Following the recommendations of Jason Brownlee [Bro], we create spot check machine learning algorithms in R. We used 8 models incorporated in the caret package[7] of R language (See Table 3.2).

The hyperparameters are automatically setuped by the train()[8] function of the caret package. The function sets up a grid of tuning parameters, and the best candidate combination of tuning parameters is selected based on optimal resampling statistic.

3.3.4 Evaluation

The evaluation of our models will be made by the cross-validation method (10 folds). For general content, we used the mean absolute error (MAE), since there is a wide target labels to classify (ordered study cycles) and the accuracy. We also use the root-mean-square error (RMSE) in a regression of a set of Portuguese school books. For the health content classifier we used only the accuracy because it is a binary classifier labelling the texts as hard-to-read (health professional)

[2] https://boxuancui.github.io/DataExplorer/

[3] https://ggplot2.tidyverse.org/

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[5] https://www.nltk.org/

[6] https://spacy.io/

[7] https://topepo.github.io/caret/

[8] https://www.rdocumentation.org/packages/caret/versions/4.47/topics/train

[9] The fundamental difference between Random Forests and Bagged CART is that in Random forests, only a subset of features are selected at random out of the total and the best split feature from the subset is used to split each node in a tree, unlike in bagging where all features are considered for splitting a node.

Model	Description
Multi-Layer Perceptron (MLP)	MLPs are fully connected feedforward networks, and probably the most common network architecture in use. Training is usually performed by error backpropagation or a related procedure.
Lasso and Elastic-Net Regularized Generalized Linear Model	Fit a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty.
SVM Radial	Support Vector Machines with Radial Basis Function Kernel.
k-nearest neighbors algorithm (k-NN)	It is a type of lazy learning where the function is only approximated locally and all computation is deferred until classification.
Classification And Regression Tree (CART)	It uses a decision tree, as a predictive model, to go from observations about an item (branches) to conclusions about the item's target value (leaves).
Bagged CART	Bagging (Bootstrap Aggregation) is an ensemble method to improve model accuracy by getting an aggregated value from multiple subsets of a dataset. It uses the CART model as 'bagging' function.
Random Forest	Random forest is machine learning algorithm that fits many classification or regression tree (CART) models to random subsets of the input data and uses the combined result (the forest) for prediction.[9]
Stochastic Gradient Boosting - Generalized Boosted Modeling	Fit a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda. Fits linear, logistic and multinomial, poisson, and Cox regression models.

Table 3.2: Used models incorporated in caret package of R language.

or easy-to-read (consumer/patient). The calculation of these metrics is also provided by the caret package of R language.

3.4 Summary

The main goal of this work is to build a health-related text readability binary classifier and a general text readability classifier for English and Portuguese languages. The general methodology involves creating a classifier to general texts, and then a classifier to health texts with specific medical features. In each Data Mining step, we use a modified version of CRISP-DM to build the final models. In the end, we present a new type of features for medical readability, and fill the lack of resources in general and medical readability for the Portuguese language. We will publish the datasets generated along with the work to future research.

Chapter 4

Predictive models of readability for general textual content

In this Chapter, we describe our experiments regarding the readability classification of general content. We present the used features that evaluate the readability of a general text, and show the results of applying these features to annotated datasets with five difficulty levels for English and Portuguese languages. In the end, just for the Portuguese language, we present the results of a regression applied in Portuguese school books.

4.1 Introduction

The general readability features are also necessary to evaluate the readability of health-related texts. With this in mind, we took advantage of the NLP features to increase the accuracy of traditional approaches and use the current best features found in the state-of-the-art.

In the next sections, we will describe a 5-level classification and a regression applied in a Portuguese school books dataset using the created features.

4.2 5-level Classification

We use a dataset annotated with 5 levels of readability. Each level corresponds to a range of educational years. In the next subsections, we explain in more detail what and how the used dataset, features and the results using these features.

4.2.1 Annotated corpora

We used the WeeBit corpus created by Vajjala et al. [VM12]. The sources of this dataset are the educational newspaper WeeklyReader[1] and the website BBC-Bitesize[2]. The documents of both sources were aggregated in a set of documents labelled with 5 levels (level 2, level 3, level 4, KS3, and GCSE) with the age ranging from 7 to 16. The resume of that dataset is presented in Table 4.1.

Grade level	Age in years	Number of documents
Level 2	7-8	629
Level 3	8-9	801
Level 4	9-10	814
KS3	11-14	644
GCSE	14-16	3500

Table 4.1: WeeBit corpus summary description.

Since the original data is unbalanced, particularly the class GCSE, we randomly selected 808 documents with the GCSE label.

We translated the dataset into Portuguese languages using the googletrans[3] Python library that uses the Google Translate API.

4.2.2 General features

In addition to the traditional features, like the average of sentence/word length, we used NLP to evaluate the word familiarity, the lexical richness and the Part-of-speech (POS) ratios of a given text.

We only consider measures that do not depend on the document's size. For example, features like the number of words or sentences won't be used.

In the next subsections, we present the features of each type of features. We also show Pearson correlation coefficient charts between the features and the school grade (in English and Portuguese datasets) to have an idea of the importance of each feature. All the correlations were statistically significant (p-value < 0.05), what is understandable given the size of the dataset. It is assumed that readability decreases with the school grade.

4.2.2.1 Traditional features

The traditional features considered are shown in the Table 4.2. By the analysis of the figure 4.1, we conclude that the feature *avg_wo_by_se* is the most correlated with the grade level. The second feature more correlated with readability is *cw_gte4* in English and *cw_gt6* in Portuguese. By that, it is possible, again, verify that the Portuguese language have, in general, words with

[1] http://www.weeklyreader.com/

[2] http://www.bbc.co.uk/bitesize

[3] https://pypi.org/project/googletrans/

a greater number of syllables [AL19a]. It is also noted that the feature *avg_sy_by_wo* is more correlated than the feature *avg_ch_by_wo* in both languages. In Portuguese language the feature *avg_ch_by_wo* have a zero correlation, indicating possible outliers.

ID	**Description**
avg_ch_by_wo	Average characters by word.
avg_sy_by_wo	Average syllables by word.
avg_wo_by_se	Average words by sentence.
cw_eqX	number of words with X syllables by total number of words. X in {1, 2, 3, 4, 5, 6, 7}.
cw_gteX	number of words with X or more syllables by total number of words. X in {1, 2, 3, 4, 5, 6, 7}.

Table 4.2: Traditional features description.

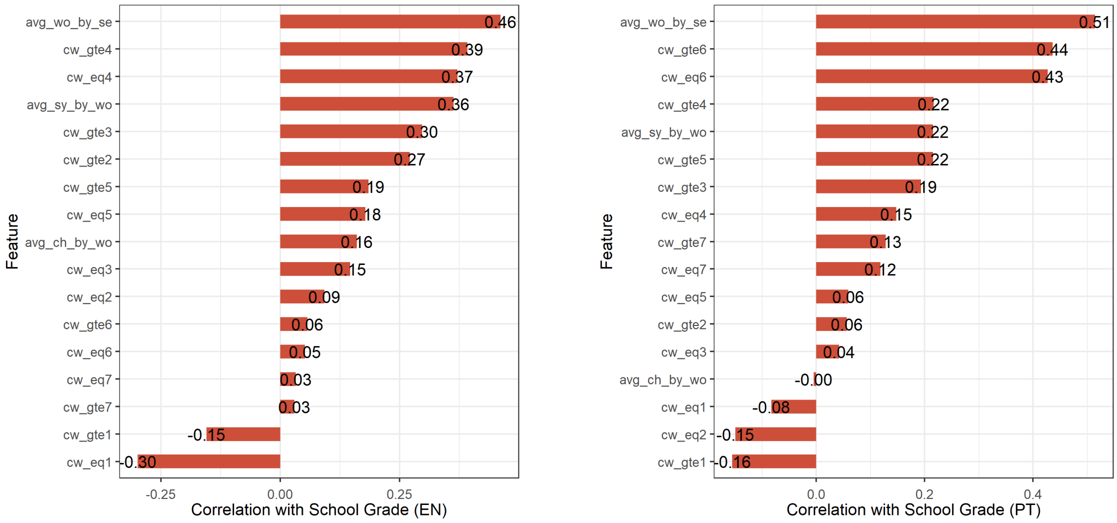


Figure 4.1: Traditional features correlation in the WeeBit corpus.

4.2.2.2 Word familiarity

The word familiarity consists of words more used by people. We used three lists of common words presented in the Table 4.3. These lists have been translated into Portuguese language using Google Translator through the googletrans[4] Python library to apply in Portuguese texts.

The features considered are shown in the Table 4.4. The correlation charts (Figure 4.2) show that the AWL-related features are good features to test low readable text. Its negative value indicates that AWL has academic words that are less readable for most people. The best list of easy words is the GSL in the English language and the first 250 words of the GWL in the Portuguese

[4] https://pypi.org/project/googletrans/
[5] https://www.lextutor.ca/freq/lists_download/
[6] https://github.com/HelderAntunes/readability-corpora#general-content-and-resources-en
[7] https://github.com/HelderAntunes/readability-corpora#general-content-and-resources-pt

Abbr	Name	Description
GSL	General Service List	Roughly 2,000 words representing the most frequent words of English [Wes53]. The words of this list are headwords. For instance 'be' is a headword and possible variations are 'am', 'is' or 'are'. We used a final list[5] with the included variations.
AWL	Academic Word List	It contains 570 headwords with great frequency in a broad range of academic texts [Cox98]. Words present in GSL are not included in this list. Similar to GSL, we included all variations for each headword using the same web resource.
GWL	Google Word List	10,000 most common English words ordered by frequency [Kau]. These words were determined by n-gram frequency analysis of the Google's Trillion Word Corpus [MSA^{+}11].

Table 4.3: Word familiarity lists. We publish these lists in a public repository in the English[6] and Portuguese[7] versions.

language. This shows that the translation of GSL to the Portuguese language, on the contrary of the AWL, did not result in good features.

ID	Description
rare_words_by_words_gsl	Percentage of words not present in GSL.
rare_words_by_words_gsl_heads	Percentage of words not present in GSL (list with only the headwords).
rare_words_by_words_awl	Percentage of words not present in AWL.
rare_words_by_words_awl_heads	Percentage of words not present in AWL (list with only the headwords).
rare_words_by_words_google	Percentage of words not present in GWL.
rare_words_X_by_words_google	Percentage of words not present in the first X words of GWL. X in {250, 500, 1000, 2000, 4000, 8000}.
avg_word_rarity_google	Average word rank in the GWL. Words that are not in GWL have a rank of 10001.

Table 4.4: Word familiarity features description.

4.2.2.3 Lexical richness

We assumed previously that the readability should be independent of text size. This creates some problems in measures of lexical richness that are affected by text size. The type-token ratio (TTR), a traditional measure of lexical richness, is calculated by dividing the number of unique words (types) by all the words (tokens). TTR is affected by text length, with its value decreasing as the text becomes longer.

A study conducted by Torruella et al. [TC13] tests the text length independence in seven measures: TTR (type–token ratio), RTTR (root type-token ratio) [Gui60], CTTR (corrected type-token ratio) [Car64], Mass [Maa72], MSTTR (mean segmental type-token ratio) [Joh44], MTLD (measure of textual lexical diversity) [Mcc05] and HD-D (Hypergeometric distribution D) [MJ07].

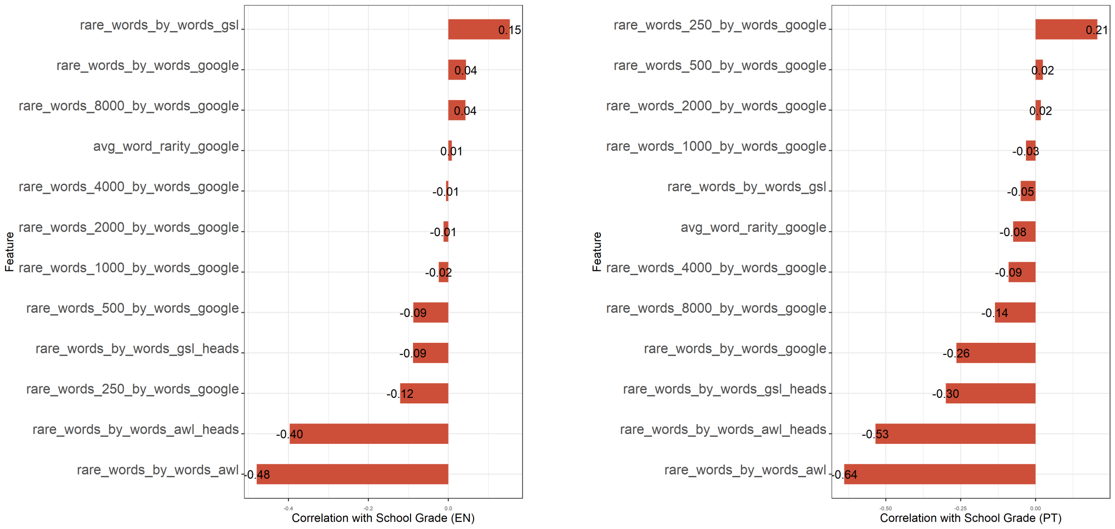


Figure 4.2: Word familiarity features correlation in the WeeBit corpus.

The authors conclude that only the last four are unaffected by text size. In our work we use those four measures and the MATTR (moving average TTR) [CM10] that is also a text size independent measure not considered in the Torruella study.

We use the Python library LexicalRichness[8] to compute these measures. The input parameters of some measures were defined as the default, and a more extensive study would be necessary to define the best input parameters.

The considered lexical richness measures are in Table 4.5. Figure 4.3 shows that these features do not provide too much information about the readability. The best feature found was *hdd*, having a correlation of 0.06 and 0.13 in English and Portuguese languages, respectively.

ID	Description
Maas	$(log(w) - log(t))/(log(w) \times 2)$
msttr	Mean segmental TTR.
mtld	Measure of lexical textual diversity.
hdd	Hypergeometric distribution D.
mattr	Moving average TTR.

Table 4.5: Lexical richness features description.

w - number of words, t - number of unique terms

4.2.2.4 Parts of speech (POS) ratios

We use the Python library Spacy[9] as POS tagger. The POS tags used were: adjectives, adverbs, coordinating conjunctions, determiners, interjections, nouns, numerals, particles, pronouns, proper nouns, punctuations, symbols, verbs, auxiliary verbs, and others (untagged tokens). To use them as features, we divided the number of each type of tag by the number of sentences, being that

[8] https://pypi.org/project/lexicalrichness/

[9] https://spacy.io/

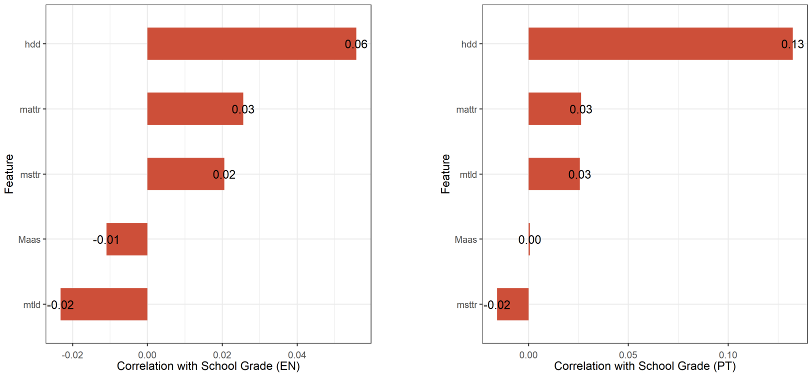


Figure 4.3: Lexical richness features correlation in the WeeBit corpus.

respective features text size independents. We also use the parameter of parse tree height given by the POS tagger tool, calculating the average of sentence parse tree height and use it as a feature.

The final features are presented in the Table 4.6. By the analysis of the Figure 4.4, the feature *avg_parse_tree_heights* have good correlations in the both languages. For the English language, the feature *avg_adpositions_per_sentence* is the most correlated feature, while in the Portuguese language is the *avg_adjectives_per_sentence*. The feature *avg_verbs_per_sentence* also show high correlations in the both languages, corroborating the study of Pitler and Nenkova [PN08].

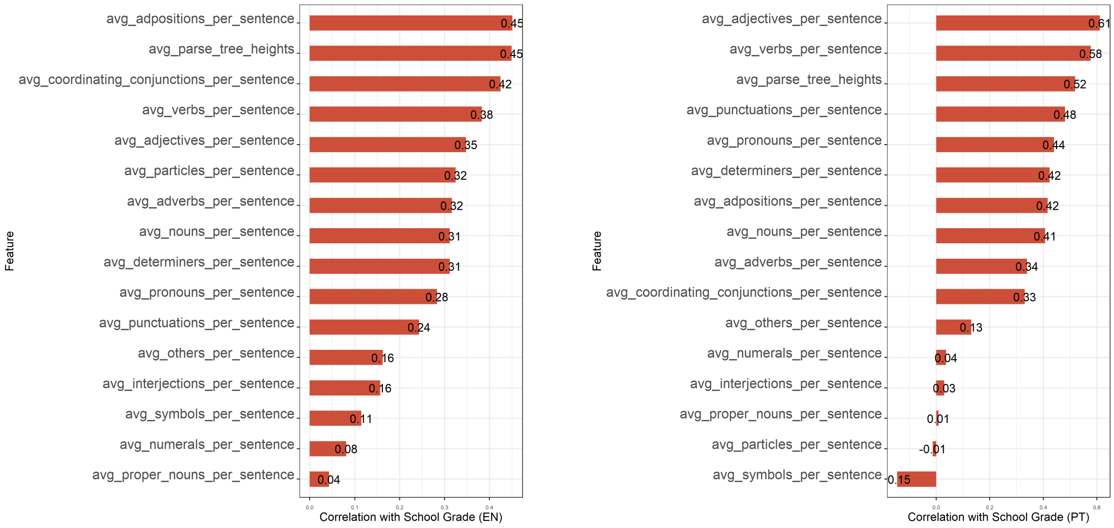


Figure 4.4: POS ratios features correlation in the WeeBit corpus.

4.2.3 Results

The results for the WeeBit corpus in the English language are shown in the Table 4.7. The traditional features perform well in the dataset, getting just behind the POS ratios features. The accuracy of the best model using all the features is 79.3%. This is also the best overall model.

For the Portuguese language, the results are similar like it is shown in Table 4.8. However, the best accuracy, 75.5%, is a bit worse. The main reason for this was the poor performance of

ID	Description
avg_parse_tree_heights	Average sentence parse tree height.
avg_adjectives_per_sentence	Number of adjectives per sentence.
avg_adpositions_per_sentence	Number of adpositions (ex.: in, to, during) per sentence.
avg_adverbs_per_sentence	Number of adverbs (ex.: very, tomorrow, down) per sentence.
avg_coord_conj_per_sentence	Number of coordinating conjunctions (ex.: and, or, but) per sentence.
avg_determiners_per_sentence	Number of determiners (ex.: a, an, the) per sentence.
avg_interjections_per_sentence	Number of interjections (ex.: psst, ouch, bravo, hello) per sentence.
avg_nouns_per_sentence	Number of nouns per sentence.
avg_numerals_per_sentence	Number of numerals (ex.: 2, one, IV) per sentence.
avg_particles_per_sentence	Number of particles (ex.: 's, not) per sentence.
avg_pronouns_per_sentence	Number of pronouns (ex.: I, you, he) per sentence.
avg_proper_nouns_per_sentence	Number of proper nouns (ex.: Mary, John, London) per sentence.
avg_punctuations_per_sentence	Number of punctuations (ex.: '.', '?', '!') per sentence.
avg_symbols_per_sentence	Number of symbols (ex.: $, %, +) per sentence.
avg_verbs_per_sentence	Number of verbs per sentence.
avg_others_per_sentence	Number of others (ex.: sfpsdxma, 23dsd) per sentence.

Table 4.6: POS ratios features description.

Feature set	# Features	Accuracy	MAE	Model
Traditional features	17	66.7%	0.58	GLMNET
Word familiarity	12	55.5%	0.64	SVM
Lexical richness	5	44.9%	0.91	Random Forest
POS ratios	16	75.5%	0.43	Random Forest
All features	50	**79.3%**	**0.40**	Stochastic Gradient Boosting

Table 4.7: Results with the WeeBit corpus in the English language.

the 'POS ratios' feature set. That poor performance can be explained by the worst POS tagger of Spacy library for the Portuguese language. In fact, the estimation of the accuracy of the POS tagger for the Portuguese language is 80.36%[10], while the English POS tagger usually achieves an accuracy of 96.98%[11].

[10] https://spacy.io/models/pt

[11] https://spacy.io/models/en#en_core_web_lg

By the analysis of both Tables, the machine learning models that work better with this dataset were the SVM and Random Forest.

Feature set	**# Features**	**Accuracy**	**MAE**	**Model**
Traditional features	17	66.4%	0.60	Random Forest
Word familiarity	12	50.5%	0.71	SVM
Lexical richness	5	43.8%	0.96	SVM
POS ratios	16	68.2%	0.52	SVM
All features	50	**75.5%**	**0.46**	Random Forest

Table 4.8: Results with the WeeBit corpus in the Portuguese language.

By the analysis of the Figure 4.5, it is possible to make the same conclusions for both languages. The first grade (label 0) and the second grade (label 1) can be confused between them. The same for the second grade and the third grade (label 2). The last two grades (labels 3 and 4) also have more chances of being confused.

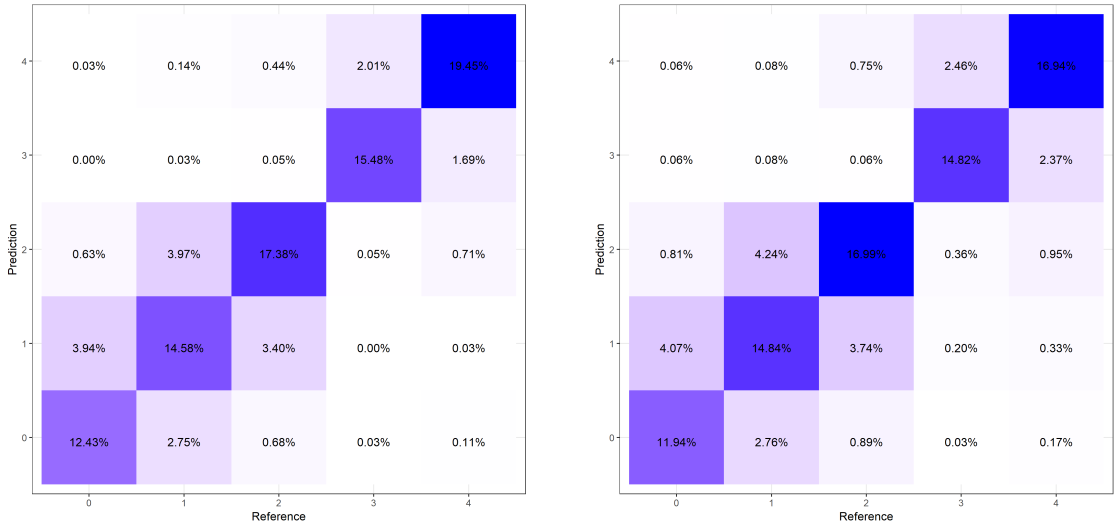


Figure 4.5: Confusion matrix of the best models in the WeeBit corpus for the both languages (left image is for English and the right image is for Portuguese).

4.2.3.1 Comparison of results with previous studies

Vajjala et al. [VM12] obtained 93.3% (split validation, with 80% of data used in train and 20% used in test) in the WeeBit corpus. Our results in the English version show a worse accuracy of 79.3%. The reasons by this difference may be due to us ignoring features that are text size dependent, like Type-Token Ratio (TTR) and others. Since our work is not directly pointed to general readability, we ignore other more advanced features used by Vajjala et al. [VM12], like features that measure the syntactic complexity introduced by Xiaofei Lu [Lu10]. For the Portuguese language, as best of our knowledge, nobody test readability measures and we cannot compare our results.

4.3 Regression on Portuguese school books dataset

We decided to use a set of Portuguese school books used by native students in Portugal.

We extracted the previous features in a set of Portuguese from elementary through high school (from grade 1 to grade 12). The books are from different disciplines, including Portuguese native learning, study of the environment, history, biology, geology, physics and chemistry courses. The books belong to editors Porto Editora[12] and Areal Editores[13] and were collected from the online platform Escola Virtual[14]. A total of 65 books were analyzed. Each page of a book is in the XHTML format, so we parsed it to clean the text. We also divided the texts in excerpts of 60 sentences. Since the higher grades have more text, we balanced the number of excerpts across the grade. In the end, each grade has 120 excerpts of 60 sentences.

Due to a large number of grade levels, we made a regression on the dataset. We applied the same models and used the MAE and RMSE metrics to evaluate the results.

4.3.1 Results

The results of the regression are shown in Table 4.9. Like in the classification task, the 'POS ratios' feature set has the best performance. Again, the traditional feature performs well, accompanied by the 'Word familiarity' feature set. The combination of all the features results in an improved performance, having a mean absolute error of 1.69 (corresponds to an error of 1.69 year/grade). The RMSE metric is always a little bigger than MAE, since it is more sensible to larger errors. SVM was the best model in this dataset, with the Lasso and Elastic-Net Regularized Generalized Linear Model (GLMNET) having a better result with the lexical richness set of features.

Feature set	**# Features**	**MAE**	**RMSE**	**Model**
Traditional features	17	2.32	2.90	SVM
Word familiarity	12	2.37	2.94	SVM
Lexical richness	5	2.67	3.16	GLMNET
POS ratios	16	2.15	2.62	SVM
All features	50	**1.69**	**2.15**	SVM

Table 4.9: Results of the regression applied to the Portuguese school books dataset.

4.4 Summary

For the general text's readability, we used four feature sets. The best features are the 'POS ratio' features, but the traditional features also had a good performance.

We consider the results for the WeeBit corpus good, since its accuracy was almost 80% in English language and was 75.5% and do not consider text size-dependent features. For this dataset,

[12] https://www.portoeditora.pt/

[13] https://www.arealeditores.pt/

[14] https://www.escolavirtual.pt/

the models which fit better are the SVM and Random Forest. Our results in the English version show a worse accuracy in comparison with Vajjala et al. [VM12] that obtained 93.3%. The reasons by this difference may be due to us ignoring features that are text size dependent, like Type-Token Ratio (TTR) and others. Since our work is not directly pointed to general readability, we also ignore other more advanced features. For the Portuguese language, as best of our knowledge, there are no readability measures tested in WeeBit corpus, and we cannot compare our results.

For the Portuguese language, we made a regression in a set of school books, since there is a lack of readability measures using the advanced features of NLP. We obtained a mean absolute error of 1.69, which we consider a good result taking in the account the considered 12 grades levels of the books.

Chapter 5

Predictive models of readability for health textual content

In this Chapter, we describe our experiments regarding the readability classification of health-related content. These features consider the difficulty of health concepts, that is totally ignored by a readability predictive measures of general texts. We also describe the used datasets and the results of applying the health features.

5.1 Introduction

In Section 2.2, we analysed the features and the datasets previously used in other research works. In our approach to the problem, we created new type of features. In the next section, we present the constructed binary classifiers including the used corpora, the created health features and the respective results.

5.2 Binary Classification

We created a dataset of health texts with 2 levels of readability. The simple readability level is directed to people with lower health literacy, while the other level is directed to people with higher health literacy. In the next subsections, we explain how we created the dataset, the features and the results of applying those features.

5.2.1 Annotated corpora

We used the Simple English Wikipedia[1] to collect simple documents. Simple English Wikipedia has the goal to be accessible for everyone, including children and adults who are learning English.

[1] https://simple.wikipedia.org/wiki/Main_Page

The articles are expected to be more readable since there are guidelines about writing simple texts[2] like the use basic English terms and the use of simple sentence structure.

For the construction of the dataset, we selected health articles in the respective webpage category[3]. From that page, we did a 2-level depth search, collecting articles of subcategories of 'Health' like 'Nutrition' and 'Disability'. The corresponding normal English page URL was obtained by replacement of 'simple' by 'en' in the original simple webpage URL. For Portuguese dataset construction, we chose to translate the ordinary and Simple English Wikipedia.

In total, we collected 3014 Wikipedia articles in simple and normal versions for English and Portugues languages. These articles are presented in a public GitHub repository[4].

5.2.2 Health features

Like the features for general texts, we created features that are independent of text length. For instance, a feature like the number of health-related words cannot be used because larger texts are more prone to have more health-related words. Instead, that feature can be transformed in the number of health-related words by number total of words.

For health features construction, we used three health word lists (Health Word List (HWL), Health News List (HNL) and Medical Articles List (MAL) described in the Table 5.1.

Abbr	Name	Description
HWL	Health Word List	It represents health vocabulary. This list doesn't provide any information about word frequencies. For the English language, we web scraped a medical glossary[5] with 27480 terms. For the Portuguese language, we web scraped the Dicionário Médico website[6] with 9712 medical terms.
HNL	Health News List	It contains the health words found in easy-to-read health-related content (health news). For each health word of HWL, there is the respective collection frequency (CF) and document frequency (DF, number of documents in which the word appears).
MAL	Medical Articles List	It contains the health words found in hard to read health-related content (medical articles). For each health word of HWL, there is the respective collection frequency (CF) and document frequency (DF).

Table 5.1: Health word lists. We publish these lists in a public repository in the English[7] and Portuguese[8] versions.

[2] https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

[3] https://simple.wikipedia.org/wiki/Category:Health

[4] https://github.com/HelderAntunes/readability-corpora

[5] https://www.online-medical-dictionary.org/glossary.html

[6] https://www.dicionáriomédico.com

[7] https://github.com/HelderAntunes/readability-corpora#health-related-content-and-resources-en

[8] https://github.com/HelderAntunes/readability-corpora#health-related-content-and-resources-pt

The health news were collected through web scraping using the Python library Beautiful Soup[9]. The health news websites used for the English language were Medical News Today[10] (1479 documents) and ScienceDaily in health category[11] (2100 documents), totalizing 3579 documents. For the Portuguese language, we collected 192 health news from DN Life[12], 581 from Notícias Magazine[13], 215 from Men's Health[14] and 275 from Women's Health[15], totalizing 1263 documents.

We collected English health scientific articles from the British Medical Journal (BMJ)[16] from 2009 to 2019, totalizing 1905 documents. BMJ is a scientific journal that publishes in all areas of health, guaranteeing the coverage of most medical concepts. We used web scraping in the website articles of BMJ found in PubMed Central (PMC)[17], ignoring tables, captions, titles and the references. For the Portuguese language, we collected 1094 articles from Acta Médica Portuguesa[18]. This scientific journal also publishes in all fields of medicine. The articles were in the PDF format, so we transformed them into the HTML format, using the Java library Pdf2Dom[19] to facilitate the parsing. The parsing of HTML documents was accomplished by filtering the text elements of a certain font size representing the main content. The useless text like title, references, captions, tables have a different size of the main content and was ignored.

The health news and the medico-scientific articles in both languages are also shared in a public repository[20].

Next, we present the health features based on health words proportion and rank, information retrieval (IR), and the Collins-Thompson and Callan's statistical model.

5.2.2.1 Health words - proportion

We use the MAL and HNL to have a representation of the distribution of word in hard-to-read and easy-to-read health texts. The features of this category are shown in the Table 5.2. According to the Figure 5.1, these features do not provide much information about the readability. The only notable point is that the number of health words belonging to medical articles is higher than the number of health words belonging to health news as the level of Wikipedia increases. Nevertheless, the feature showing that fact, *articles_words_minus_news_words_by_health_words*, has a low correlation value (0.01 and 0.05 in English and Portuguese, respectively).

[9] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[10] https://www.medicalnewstoday.com/
[11] https://www.sciencedaily.com/news/top/health/
[12] https://life.dn.pt/categoria/saude/
[13] https://www.noticiasmagazine.pt/categoria/estilos/saude/
[14] https://menshealth.pt/categoria/saude/
[15] https://www.womenshealth.pt/categoria/saude/
[16] https://www.bmj.com/
[17] https://www.ncbi.nlm.nih.gov/pmc/
[18] https://www.actamedicaportuguesa.com/
[19] https://github.com/radkovo/Pdf2Dom
[20] https://github.com/HelderAntunes/readability-corpora

ID	Description
num_health_words_by_words	Number of health words in HWL divided by number of words.
num_health_articles_words_by_words	Number of health words in MAL divided by number of words.
num_health_news_words_by_words	Number of health words in HNL divided by number of words.
articles_words_minus_news_words _by_health_words	num_health_articles_words_by_words - num_health_news_words_by_words

Table 5.2: Health words proportion features description.

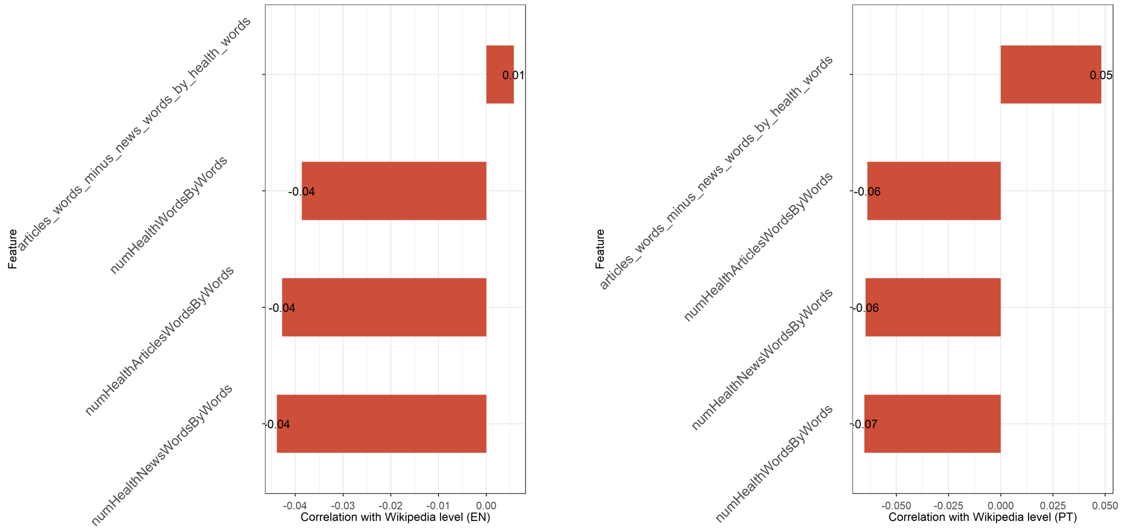


Figure 5.1: Health words proportion features correlation in the Wikipedia dataset.

5.2.2.2 Health words - Collection frequency (CF)

During the preparation of HNL and MAL, we counted how many times each health word appeared in the respective collections (see all features in Table 5.3). By the analysis of the Figure 5.2, it is possible to observe that as the Wikipedia level increases the degree of similarity with the health news taking into account the value of CF decreases ($r(avg\_word\_CF\_news) < 0$ in both languages) and with the medical articles increases ($r(avg\_word\_CF\_articles) > 0$ in both languages). The maximum value of CF having in account the articles and news are very correlated with Wikipedia level. The reason behind this can be the fact that documents of ordinary Wikipedia are larger, and the probability of finding a very common health word (high CF) is higher.

5.2.2.3 Health words - Document frequency (DF)

We also count the number of documents of each health word appeared in health news and medical articles (Table 5.4 for more details). The correlations of these type of features are shown in the Figure 5.3. The same conclusions of the features using CF can be applied.

ID	Description
avg_word_CF_news	Average value of health words CF in HNL.
avg_word_CF_articles	Average value of health words CF in MAL.
max_word_CF_news	Maximum value of health words CF in HNL.
max_word_CF_articles	Maximum value of health words CF in MAL.
min_word_CF_news	Minimum value of health words CF in HNL.
min_word_CF_articles	Minimum value of health words CF in MAL.
max_minus_min_word_CF_news	max_word_CF_news - min_word_CF_news
max_minus_min_word_CF_articles	max_word_CF_articles - min_word_CF_articles

Table 5.3: Health words CF features description.

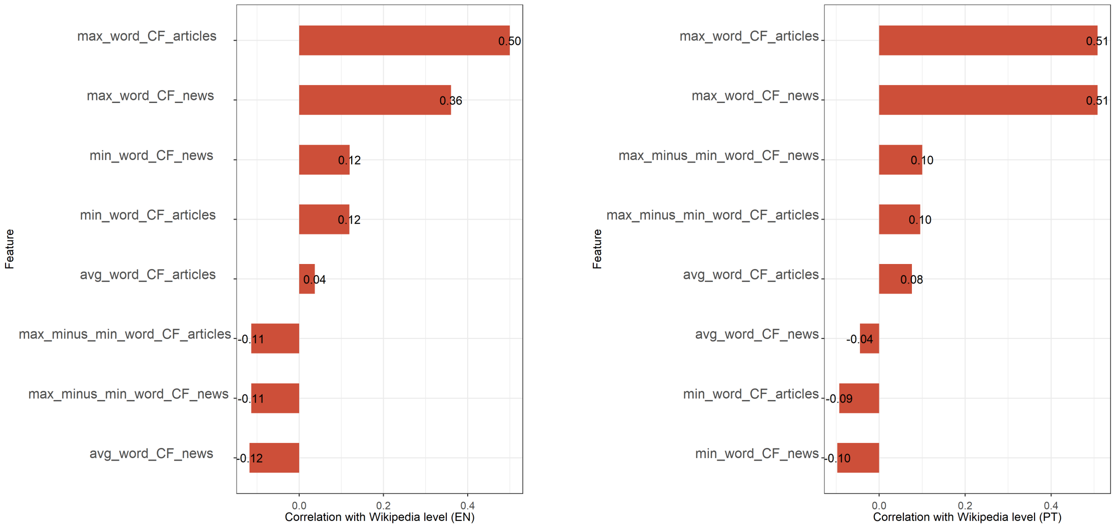


Figure 5.2: Health words CF features correlation in the Wikipedia dataset.

ID	Description
avg_word_DF_news	Average value of health words DF in HNL.
avg_word_DF_articles	Average value of health words DF in MAL.
max_word_DF_news	Maximum value of health words DF in HNL.
max_word_DF_articles	Maximum value of health words DF in MAL.
min_word_DF_news	Minimum value of health words DF in HNL.
min_word_DF_articles	Minimum value of health words DF in MAL.
max_minus_min_word_DF_news	max_word_DF_news - min_word_DF_news
max_minus_min_word_DF_articles	max_word_DF_articles - min_word_DF_articles

Table 5.4: Health words DF features description.

5.2.2.4 Health words - Inverse document frequency (IDF)

The inverse document frequency (IDF) measures how much information a word provides. A word provides little information if it's a common word across all documents, and provides more information if it's a rare word across the documents. A word with a high IDF is a very informative and specific word. The formula to calculate it is:

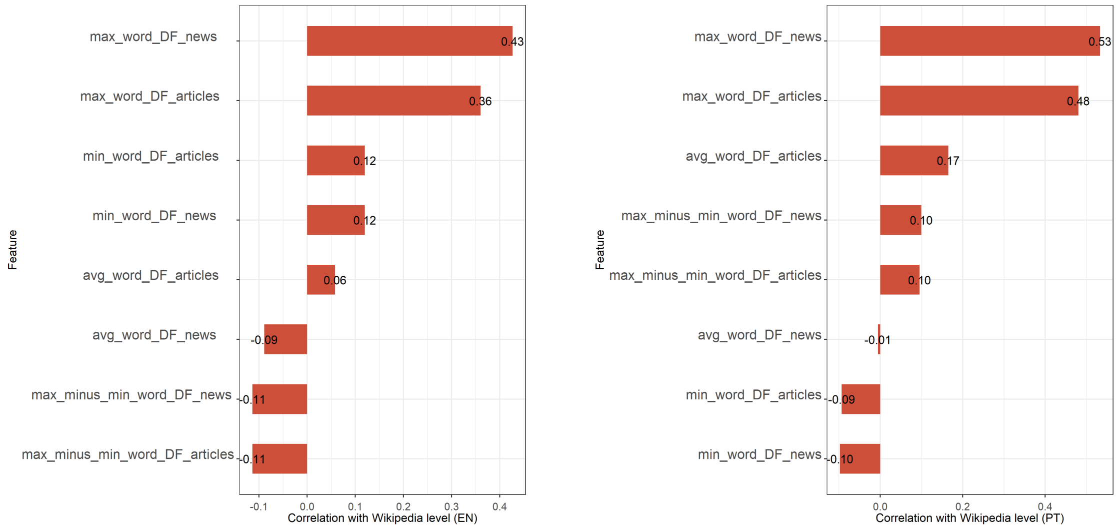


Figure 5.3: Health words DF features correlation in the Wikipedia dataset.

$$idf(t,D) = \log(\frac{N}{|\{d \in D : t \in d\}|})$$

with:

- t: the term or word under analysis
- D: the collection of documents (Health news or medical articles)
- N: total number of documents in a collection (Number of health news or number of medical articles)
- $|\{d \in D : t \in d\}|$: number of documents where the term t appears.

To avoid a division-by-zero, we smoothed the previous formula adjusting the denominator (adding 1 to the denominator is a common adjust), resulting in:

$$idf(t,D) = \log(\frac{N}{|\{d \in D : t \in d\}| + 1})$$

The final features of this category are shown in the Table 5.5. The Figure 5.4 shows that the average specificity of the health words increases as the Wikipedia level increases taking into account news or articles (r(*avg_word_IDF_news*), r(*avg_word_IDF_articles*) > 0 in both languages). The correlations, in the both languages, of the features *max_word_IDF_news* and *max_word_IDF_articles* are the greatest, showing that as the Wikipedia level increases it is more probable to find very difficult health terms. The correlations of the features *min_word_IDF_news* and *min_word_IDF_articles* are greater than zero in English language, showing that the more easier word of a medical text have more chances to be less readable as the Wikipedia level increases. For the Portuguese language, that features correlations are slightly less than zero (r = -0.08) and does not provide much information.

ID	Description
avg_word_IDF_news	Average value of health words IDF in HNL.
avg_word_IDF_articles	Average value of health words IDF in MAL.
max_word_IDF_news	Maximum value of health words IDF in HNL.
max_word_IDF_articles	Maximum value of health words IDF in MAL.
min_word_IDF_news	Minimum value of health words IDF in HNL.
min_word_IDF_articles	Minimum value of health words IDF in MAL.
max_minus_min_word_IDF_news	max_word_IDF_news - min_word_IDF_news
max_minus_min_word_IDF_articles	max_word_IDF_articles - min_word_IDF_articles

Table 5.5: Health words IDF features description.

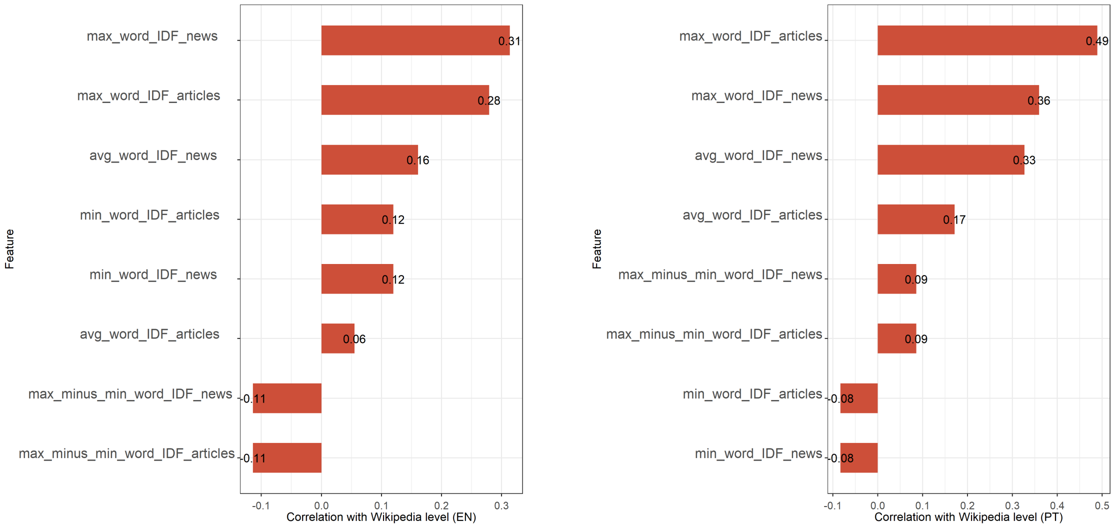


Figure 5.4: Health words IDF features correlation in the Wikipedia dataset.

5.2.2.5 Health words - Ranking

We ordered the lists HNL and MAL by decreasing order of document frequency. A word positioned in the top of the rank will have a high DF, meaning that appears in a large percentage of documents. On the contrary, a word in the lowest positions will have a low DF, and appears in a little percentage of documents, being more informative and specific.

Features of this category are presented in the Table 5.6. The Figure 5.5 presents the correlations of these features. The features *max_word_rank_news* and *max_word_rank_articles* have an high correlation in both languages. This means that the health word with the highest rank (the less readable word) has a higher value of rank as the Wikipedia level increases. The features *max_word_rank_news* and *max_word_rank_articles* have a correlation greater than zero in both languages, meaning that the average ranking of a health word (the average readability) increases as the Wikipedia level increases.

ID	Description
avg_word_rank_news	Average value of health words rank in HNL.
avg_word_rank_articles	Average value of health words rank in MAL.
max_word_rank_news	Maximum value of health words rank in HNL.
max_word_rank_articles	Maximum value of health words rank in MAL.
min_word_rank_news	Minimum value of health words rank in HNL.
min_word_rank_articles	Minimum value of health words rank in MAL.

Table 5.6: Health words ranking features description.

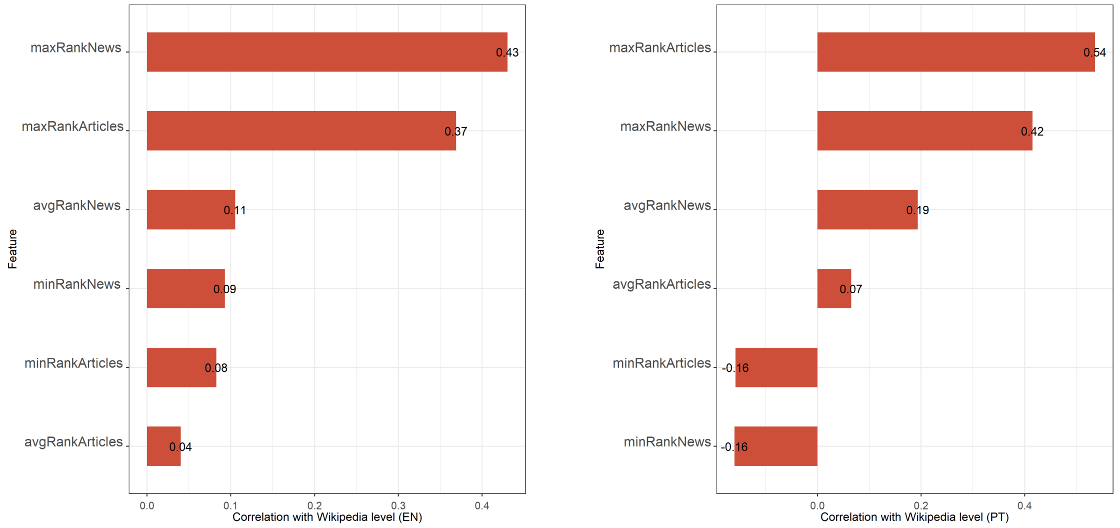


Figure 5.5: Health words ranking features correlation in the Wikipedia dataset.

5.2.2.6 Collins-Thompson and Callan's statistical language model

The Collins-Thompson and Callan's approach was originally used to predict the 12 American grade levels of a text [CTC04]. The model consists of a language model, more specifically, a smoothed unigram model, since it assumes that the probability of a token is independent of the surrounding tokens, given the grade language model. The authors created 12 language models corresponding to the 12 American grade levels. In each model, a given word has different probabilities of appearing (see Figure 5.6 to more details).

To predict the grade level, the authors calculated the log-likelihood of a sample text for each grade level, through the formula:

$$L(G_i \mid T) = \sum_{w \in T} \log(P(w \mid G_i))$$

where:

- G_i is the grade level (from 1 to 12)
- T is the sample text
- w is a word in the text T

- $P(w \mid G_i)$ is the probability of the word w appear in texts of grade level G_i

In the end, it is selected the grade of the language model having the maximum likelihood (see the Figure 5.7).

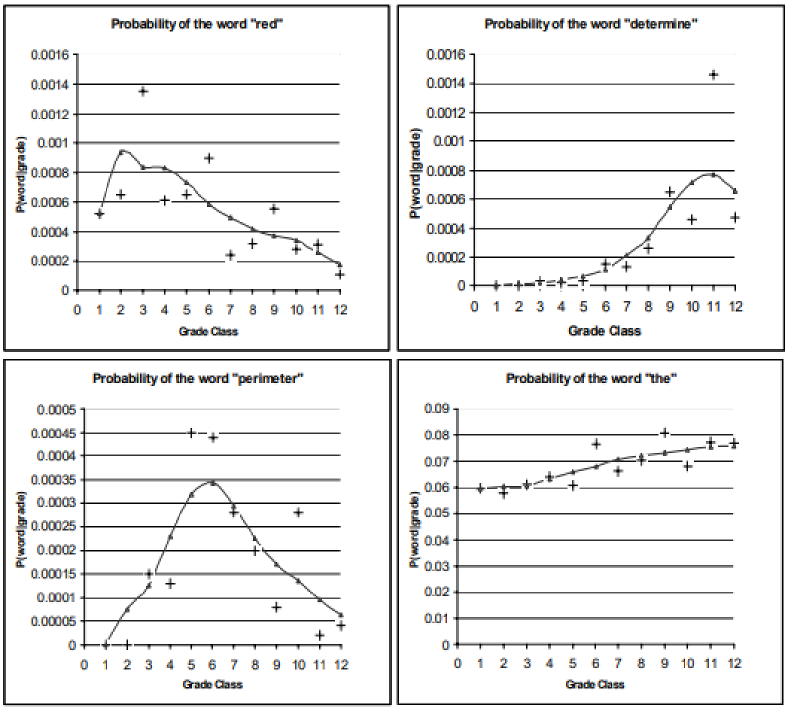


Figure 5.6: Probability distribution of the words ‘red’, ‘determine’, ‘the’, and ‘perimeter’ across 1-12 grades [CTC04].

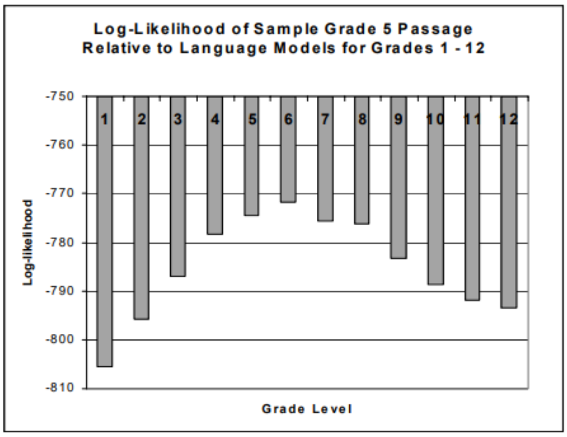


Figure 5.7: The log-likelihood of a grade 5 passage relative to the language models for grades 1 to 12. The maximum log-likelihood in this example is achieved for the grade 6 language model. [CTC04].

We apply the same approach to the health texts. Our initial base language models correspond to easy-to-read texts (represented by health news) and hard-to-read texts (represented by medical articles). The probability of a given word appearing in these languages models was calculated in two ways, using CF or DF:

$$(1) \quad P_{CF}(w \mid G_i) = \frac{CF(w)}{N} \quad , \quad (2) \quad P_{DF}(w \mid G_i) = \frac{DF(w)}{D}$$

,

where:

- G_i is the collection or language model. It can be the health news or medical articles.
- $CF(w)$ is the collection frequency of the word w.
- $DF(w)$ is the document frequency of the word w.
- N is the total number of health words in the collection G_i.
- D is the total number of documents in the collection G_i.

We calculated the log-likelihood for health news and medical articles using the two ways of probability calculation, resulting in the features shown in the Table 5.7. The Figure 5.8 shows the correlation of these features. The best-correlated feature, in both languages, is *abs_diff_log_likelihood_DF*. The feature *abs_diff_log_likelihood_CF* is clearly affected by outliers or does not give much information about the difficulty of the text. The Collins-Thompson and Callan's statistical language model using P_{DF} applied to health news and medical articles is the second best-correlated feature in both languages, showing again that the measure of the likelihood is preferable using the value of P_{DF}. The first 4 features in Table 5.7 are not much correlated and have the same value in each language. The reason by this is that the existence of one health word that does not belong to MAL or HNL or has few frequencies on those lists can decrease the value of that features too much (averages or means are very sensitive to extreme values).

5.2.3 Results

The results for the Wikipedia dataset in the English language are presented in Table 5.8. The general features perform very well achieving an accuracy of 91.8%. For that reason, the health features do not improve much the overall accuracy, reaching a value of 92.2%. The good results achieved by the general features are explained by the existing guidelines in the Simple English Wikipedia. An existing guideline is to write a small sentence, having a strong impact on the traditional feature 'words by sentences'. These guidelines also have a positive impact on other general features.

Specifically, in the health features, the statistical language model performs much better than all the others (88.7%), and for that reason, the accuracy of all health features is equal to the accuracy of a model using only features of the statistical language model. The features of the 'Health words

ID	Formula
log_likelihood_news_CF_by_health_words	$[\sum_{w \in T} \log(P_{CF}(w \mid HNL))] / \mid T \mid$
log_likelihood_articles_CF_by_health_words	$[\sum_{w \in T} \log(P_{CF}(w \mid MAL))] / \mid T \mid$
log_likelihood_news_DF_by_health_words	$[\sum_{w \in T} \log(P_{DF}(w \mid HNL))] / \mid T \mid$
log_likelihood_articles_DF_by_health_words	$[\sum_{w \in T} \log(P_{DF}(w \mid MAL))] / \mid T \mid$
abs_diff_log_likelihood_CF	$\sum_{w \in T} [\log(P_{CF}(w \mid MAL)) - \log(P_{CF}(w \mid HNL))]$
rel_diff_log_likelihood_CF	$\frac{\sum_{w \in T} [\log(P_{CF}(w \mid MAL)) - \log(P_{CF}(w \mid HNL))]}{\max(abs(\sum_{w \in T} \log(P_{CF}(w \mid MAL))), abs(\sum_{w \in T} \log(P_{CF}(w \mid HNL))))}$
abs_diff_log_likelihood_DF	$\sum_{w \in T} [\log(P_{DF}(w \mid MAL)) - \log(P_{DF}(w \mid HNL))]$
rel_diff_log_likelihood_DF	$\frac{\sum_{w \in T} [\log(P_{DF}(w \mid MAL)) - \log(P_{DF}(w \mid HNL))]}{\max(abs(\sum_{w \in T} \log(P_{DF}(w \mid MAL))), abs(\sum_{w \in T} \log(P_{DF}(w \mid HNL))))}$
thompson_callan_model_CF	$\begin{cases} 1 & \sum_{w \in T} \log(P_{CF}(w \mid MAL)) \geq \sum_{w \in T} \log(P_{CF}(w \mid HNL)) \\ 0 & otherwise \end{cases}$
thompson_callan_model_DF	$\begin{cases} 1 & \sum_{w \in T} \log(P_{DF}(w \mid MAL)) \geq \sum_{w \in T} \log(P_{DF}(w \mid HNL)) \\ 0 & otherwise \end{cases}$

Table 5.7: Collins-Thompson and Callan's model related features description.

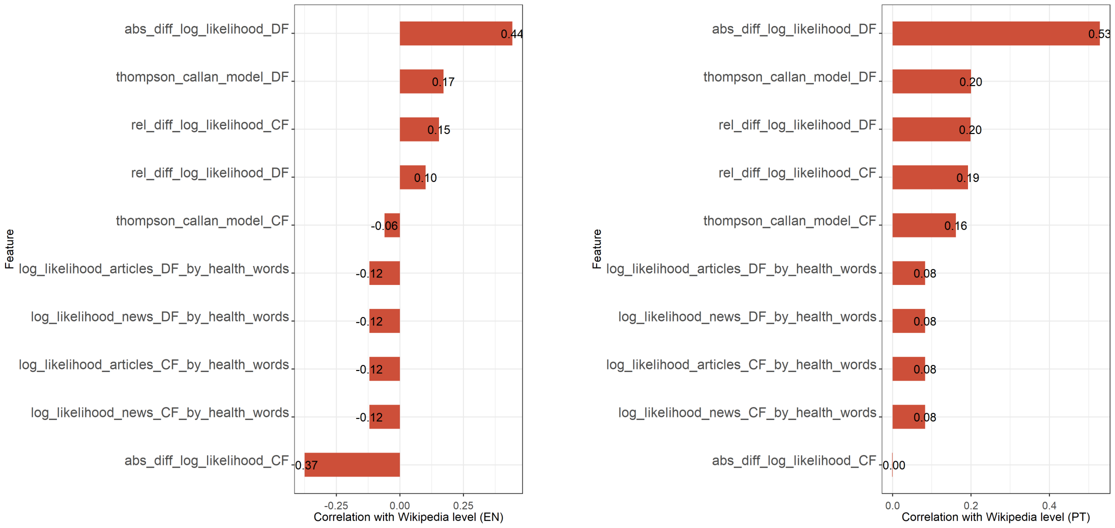


Figure 5.8: Collins-Thompson and Callan's statistical language model related features correlation in the Wikipedia dataset.

- proportion' category are the worst performer in the dataset. This particular set is better used to identify the subject of a text, and not so much to differentiate the medical specificity. The other type of features has good performances (a little more than 80%). The reason behind this is that those feature sets use the same concept of medical word specificity by accessing and comparing the frequency of the health words.

Feature set	**# Features**	**Accuracy**	**Model with highest accuracy**
General features	50	91.8%	Random Forest
Health words - proportion	4	72.4%	Stochastic Gradient Boosting
Health words - CF	8	85.9%	Random Forest
Health words - DF	8	86.2%	Stochastic Gradient Boosting
Health words - IDF	8	83.3%	Random Forest
Health words - Ranking	6	81.4%	Stochastic Gradient Boosting
Collins-Thompson and Callan's model	10	88.7%	Stochastic Gradient Boosting
All Health features	44	88.7%	Stochastic Gradient Boosting
All features	94	**92.2%**	Random Forest

Table 5.8: Results with the Wikipedia dataset in the English language.

The results for the Portuguese language are very similar to the English language results (see Table 5.9 for more details). Again, health features do not improve accuracy, but it performs well in the same.

In the Wikipedia dataset, the model that adjusts better is the Stochastic Gradient Boosting. We don't have the sufficient knowledge to explain why, but we emphasize that sometimes other algorithms perform close to that algorithm, having an accuracy 0.1% worse.

Feature set	**# Features**	**Accuracy**	**Model with highest accuracy**
General features	50	**94.7%**	Stochastic Gradient Boosting
Health words - proportion	4	77.7%	Stochastic Gradient Boosting
Health words - CF	8	86.3%	Stochastic Gradient Boosting
Health words - DF	8	86.3%	Stochastic Gradient Boosting
Health words - IDF	8	85.2%	Stochastic Gradient Boosting
Health words - Ranking	6	85.2%	Stochastic Gradient Boosting
Collins-Thompson and Callan's model	10	88.3%	Stochastic Gradient Boosting
All Health features	44	89.6%	Stochastic Gradient Boosting
All features	94	**94.7%**	Stochastic Gradient Boosting

Table 5.9: Results with the Wikipedia dataset in the Portuguese language.

By the analysis of the Figure 5.9, in the English language it is a little more easy to identify easy-to-read Wikipedia documents, and in the Portuguese language the different type of documents have similar precision.

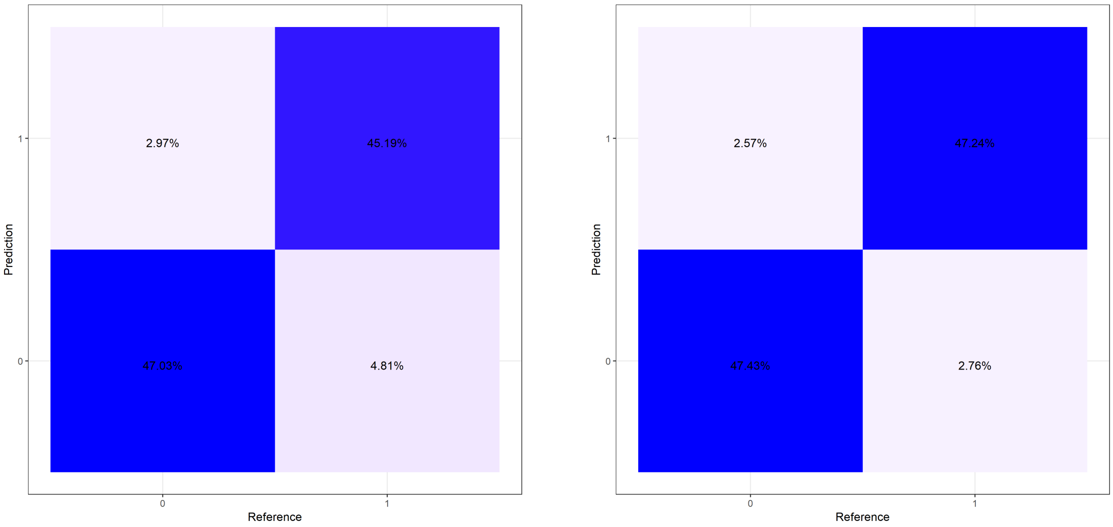


Figure 5.9: Confusion matrix of the models with the highest accuracy (using all features) in the Wikipedia dataset for the both languages (Left matrix is for English and the right image is for Portuguese). '0' label is for easy-to-read documents and '1' label is for hard-to-read documents.

5.2.3.1 Comparison of results with previous studies

Other studies do not use the same datasets used here. The most similar to that used herein was used by Kauchak et al. [KMPL14], using the dataset created by Coster et. al [CK11]. This dataset has the particularity of having sentence-aligned phrases from Simple English Wikipedia and the ordinary Wikipedia. Also, note that this dataset contains medical and non-medical articles. The authors achieved an accuracy of 84.14% using general features and health features with the 10-folds cross-validation (same used in this dissertation). Using only health features, they achieve 78.31% accuracy with specificity (see section 2.2.2.2 for more details) and 79.74% with ambiguity (see section 2.2.2.3 for more details). Although our results are better (92.2% using all features and 88.7% using only health features in English dataset), it should be noted that the dataset used in our study includes entire document texts rather than aligned sentences. Since one sentence is less representative than the whole document, a dataset using aligned sentences is more difficult to classify.

Miller et. al [MLC$^+$07] used 150 entire documents of three levels of increasing medical terminology specificity (consumer/patient, novice health learner, medical professional). They used a Naive Bayes classifier as machine learning model. This model is often used in text classification. In their approach, the probabilities from all of the tokens representing the likelihood that the document belongs to a given category are used to classify a document. They obtained an accuracy of 96%. Unlike us, they consider all tokens, not just tokens related to health terms, which may explain their better performance. Terms commonly found in articles, such as 'study', 'analysis', and 'discussion', can be impacting on the final result, without specifically measuring the semantic difficulty of medical concepts. In our approach of health features creation, we only analysed the medical terms and we consider the 88.7% and 89.6% of accuracy, in English and Portuguese respectively, good results in comparison with the Miller et. al [MLC+07] approach.

5.3 Summary

We used six feature sets: the proportion of health words, health words CF/DF/IDF/Ranking and a statistical language model. The features were built using health news, representing easy-to-read texts, and scientific medical articles, representing the hard-to-read texts.

We created a dataset collecting documents from Simple English Wikipedia and the ordinary Wikipedia, forming a two-level dataset for binary classification.

The results show that the health features do not improve the accuracy of the general features, but its performance is good in both languages reaching close to the 90% of accuracy. Even though health features do not help differentiate the documents can be extremely useful to give an idea of the health knowledge necessary to comprehend a health-related text.

The results obtained here are better than the results of Kauchak et al. [KMPL14] (we obtained 92.2% of accuracy and they obtained 84.14% of accuracy using general and health features). However, although the datasets have the same source (Simple English Wikipedia and ordinary Wikipedia), they are significantly different, since their dataset is formed by sentences-aligned and our dataset includes entire documents. A test should be conducted in the same dataset to make a fair comparison.

Chapter 6

Conclusions and Future Work

This chapter exposes the main conclusions of our work. We present the scientific contributions and the possible improvements and experiments to be carried out in the future.

The aim of this dissertation was the development of machine learning predictive models to assess the readability of health-related texts for English and Portuguese languages. The general methodology consisted of starting by analyse the readability of several topics (health included) using traditional readability measures and inquire about the linguistic differences between the English and Portuguese languages. After, we treat health-related texts as general texts evaluating the readability with features found in the state-of-the-art. In the end, we extended the features to calculate the specificity of health text and the medical knowledge necessary to comprehend it.

We evaluated the readability of topics on the web using traditional readability metrics. We found that the topic of health is the second less readable topic according to the metrics used, showing the importance of readability predictive models to this subject. We also analysed the linguistic differences between the English and Portuguese languages, finding that, in general, Portuguese words have a greater number of syllables. With this, we proposed adaptations to the traditional readability metrics originally created for the English language in order to be applicable to the Portuguese language using multiple linear regression on simple features.

In order to access the readability of general texts, we chose the WeeBit corpus, a dataset used in several research works about text readability. The dataset consists of excerpts of texts annotated in 5 levels of difficulty corresponding to grade levels between the ages 7-16. We translated the dataset to the Portuguese language. After that, we applied to the texts known features of the state-of-the-art grouping them in: traditional features, word familiarity, lexical diversity and parts of speech ratios. We were careful to not create features that were dependent on the size of the text, excluding as a feature, for instance, the number of sentences in a text. The traditional features refers to the known metrics like the number of syllables by words or the number of words by sentences. We calculated the word familiarity using lists of common words in the English language (common English words found in Google's Trillion Word Corpus and the General Service List), and a list of common academic words. The lexical diversity of a text was calculated using several metrics. We excluded the metrics that are dependent on the text size, like Type-Token ratio (TTR), one of the most known measure of lexical diversity. In parts of speech ratios, we calculated several ratios according to the grammatical function of each word, like nouns, verbs or adjectives. The models built under these features performed 79.3% of accuracy in the English WeeBit data and 75.5% in Portuguese. We find this results satisfactory, since the mean absolute error was very low in both languages (0.40 in English and 0.46 in Portuguese) and we didn't consider text-size dependent features. There are better results in previous researches (93.3% Vajjala et al. [VM12]) that consider features dependent on text size and other features that we do not consider since the general readability is not the main focus of this dissertation. In addition to these classification models, we made a regression in a dataset of excerpts of Portuguese school books using the previous features. The regression obtained a mean absolute error of 1.69, which we find a good given the high number of grade levels (from 1 to 12 grade level). This particular regression fills the current gap in assessing the readability of native Portuguese language, using advances NLP features.

In addition to features to assess the general readability, we built specific features to measure the health readability of health texts. These features were built based on two created lists of the frequency of medical words in two collections of words collected from health news and medical scientific articles. In each list, we calculated, for each medical word, the number of times that occurred in the respective collection and the number of documents in which the word appears. With these calculations, we were capable of generating features based on informational retrieval metrics like (collection frequency, document frequency, and inverse document frequency) and health words ranking. Besides that, we extended the Collins-Thompson and Callan's statistical language model, originally applied to the 12 American school grade levels. In our case, we used this model to calculate the log-likelihood of a text relative to health news and medical scientific articles. To create the dataset which servers as input to predictive models, we used medical documents from Simple English Wikipedia, representing the simple health texts, and the corresponding documents in the normal Wikipedia, representing the less readable health texts. Since these documents are originally in the English language, we translated them to the Portuguese language. With all that, we built binary classifiers using general readability features and health features. In both languages, the health features didn't improve the performance of the general features in differentiating the documents. Nevertheless, the health features, by itself, had a good accuracy (88.7% in English and 89.7% in Portuguese). The general features obtained a very high accuracy (91.8% in English and 94.7% in Portuguese), explaining the fact that the health features did not improve the performance of general features. The results obtained here are better than the results of Kauchak et al. [KMPL14] (we obtained 92.2% of accuracy and they obtained 84.14% of accuracy using general and health features). However, the datasets are significantly different, since their dataset is formed by sentences-aligned and our dataset includes entire documents. A test should be conducted in the same dataset to make fair comparisons. With the work developed, we extended the current features used in the state-of-the-art in the field of health readability, using features related with information retrieval and applying the Collins-Thompson and Callan's statistical language model originally applied to the calculation of the readability of general texts. Also, this study, as our knowledge, does the first contribution in this field to the Portuguese language.

6.1 Future Work

With the work developed in this dissertation, there are many directions to improve and extend our research.

The scope of this dissertation does not include the deploy of a system, like a website. That system could be very useful to health professionals that write medical texts, and even to general people to assess the health difficulty of the posts in blogs and other documents. Such a system that implements the assessment of the health readability could make suggestions for readability improvement. These suggestions would be based on text simplification.

In general readability, in particular, to the Portuguese language, it would be interesting to apply the Collins-Thompson and Callan's statistical language model and testing on the translated dataset of WeeBit corpus. In health readability, we used the Wikipedia-based dataset to create the models of health readability. Other option would be a collection of other health news and medical scientific papers, different of the lists used to calculate the measures of frequency for each medical word. Another idea would be the construction of a dataset of health-related sentences properly annotated instead of using whole documents, and analyse the impact of text length in health-related content.

The approach followed in this dissertation was applied to the health, but it could be applied to other areas. Two interesting topics would be finances and laws, that easily it can have specific terms which impedes a better understanding of people with low literacy in these subjects.

References

[AL10] Alan R. Aronson and François Michel Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 2010.

[AL19a] Hélder Antunes and Carla Teixeira Lopes. Analyzing the adequacy of readability indicators to a non-english language. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*. Springer International Publishing, September 2019.

[AL19b] Hélder Antunes and Carla Teixeira Lopes. Readability of web content - an analysis by topic. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, June 2019.

[AMW$^+$13] Fatima Al Sayah, Sumit R. Majumdar, Beverly Williams, Sandy Robertson, and Jeffrey A. Johnson. Health Literacy and Health Outcomes in Diabetes: A Systematic Review, 2013.

[Bro] Jason Brownlee. Spot check machine learning algorithms in r (algorithms to try on your next project). https://machinelearningmastery.com/spot-check-machine-learning-algorithms-in-r/. (Accessed on 06/15/2019).

[Car64] John B. Carroll. *Language And Thought*. Prentice-Hall, 1964.

[CC99] Sharon A Caraballo and Eugene Charniak. Determining the specificity of nouns from text. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[CD75] J.S. Chall and E Dale. *Readability Revisited: The New Dale-Chall Readability Formula*. Research Branch report. Cambridge, MA: Brookline Books, 1975.

[CGK17] Miriam Cha, Youngjune Gwon, and H T Kung. Language Modeling by Clustering with Word Embeddings for Text Readability Assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2003–2006, New York, NY, USA, 2017. ACM.

[CGM08] Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 2008.

[CK11] William Coster and David Kauchak. Simple English Wikipedia: A New Text Simplification Task William. *Research Journal of Agricultural Sciences*, 2011.

[CLL75] Meri Coleman and T L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284, 04 1975.

[CM10] Michael A. Covington and Joe D. McFall. Cutting the gordian knot: The moving-average type-token ratio (MATTR), 2010.

[CMB14] Pedro Curto, Nuno Mamede, and Jorge Baptista. Automatic readability classifier for European Portuguese. In *INFORUM 2014 – Simpósio de Informática*, pages 309–324, 2014.

[Cox98] A J Coxhead. The Academic Word List. *Occasional Publication Number 18*, 1998.

[CPRZ00] S C Ratzan, Ruth Parker, C R Selden, and Marcia Zorn. *National Library of Medicine Current Bibliographies in Medicine: Health Literacy*. Bethesda, MD: National Institutes of Health, jan 2000.

[CR71] Davies P. Carroll, J. B. and B. Richman. *The American heritage word frequency book*. Boston : Houghton Mifflin, 1971.

[CSD^{+}17] Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, 2017.

[CT15] Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135, jan 2015.

[CTC03] Kevyn Collins-Thompson and Jamie Callan. Information retrieval for language tutoring: An overview of the REAP project. In *In SIGIR '03: Proceedings of the 26thAnnual International ACM SIGIR Conference onResearch and Development in Information Retrieval*, pages 544–545, jan 2003.

[CTC04] Kevyn Collins-Thompson and James P Callan. A Language Modeling Approach to Predicting Reading Difficulty. *Proceddings of the Annual Conference of the North American Chapter of the Association of Computational Linguistics (NAACL/HLT)*, 2004.

[Das14] Mihai Dascalu. *ReaderBench (2) - Individual Assessment through Reading Strategies and Textual Complexity*, pages 161–188. Springer International Publishing, Cham, 2014.

[DC49] E. Dale and J. S. Chall. The Concept of Readability. *Elementary English*, 1949.

[DC77] E Dale and J S Chall. *A Formula for Predicting Readability*. Bureau of Educational Research, 1977.

[DDF^{+}90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990.

[e-M] Openmedspel (en-us) | apache openoffice extensions. https://extensions.openoffice.org/en/project/openmedspel-en-us. (Accessed on 02/06/2019).

[FD] Susannah Fox and Maeve Duggan. Health Online 2013 | Pew Research Center. http://www.pewinternet.org/2013/01/15/health-online-2013.

[FM12] Thomas Franc and Eleni Miltsakaki. Do NLP and machine learning improve traditional readability formulas ? In *NAACL-HLT (Workshop - PITR)*, pages 49–57, 2012.

[Gib98] Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 1998.

[GMLC04] Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of text on cohesion and language. In *Behavior Research Methods, Instruments, and Computers*, 2004.

[Gov] Programa simplificar | simplificar. http://historico.simplificar.gov.pt/programa. (Accessed on 02/07/2019).

[Gui60] P.L. Guiraud. *Problèmes et méthodes de la statistique linguistique*. Paris: Presses universitaires de France, 1960.

[Gun52] Robert Gunning. *The technique of clear writing*. McGraw-Hill New York, 1952.

[HAMJ16] Tatsunori B Hashimoto, David Alvarez-Melis, and Tommi S Jaakkola. Word Embeddings as Metric Recovery in Semantic Spaces. *Transactions of the Association for Computational Linguistics*, 2016.

[Har54] Zellig S. Harris. Distributional Structure. *WORD*, 1954.

[HCTCE07] Michael J Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Proceedings of NAACL HLT*, 2007.

[HH76] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. London: Longman, 1976.

[HIL12] Timo Honkela, Zaur Izzatdust, and Krista Lagus. Text mining for wellbeing: Selecting stories using semantic and pragmatic features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[HVM12] Julia Hancke, Sowmya Vajjala, and Detmar Meurers. Readability Classification for German using lexical, syntactic, and morphological features. In *COLING*, 2012.

[II13] Racquel Richardson Ingram and L. Louise Ivanov. Examining the Association of Health Literacy and Health Behaviors in African American Older Adults: Does Health Literacy Affect Adherence to Antihypertensive Regimens. *Journal of Gerontological Nursing*, 2013.

[Int] World internet users statistics and 2018 world population stats. https://www.internetworldstats.com/stats.htm. (Accessed on 01/30/2019).

[JGYC18] Zhiwei Jiang, Qing Gu, Yafeng Yin, and Daoxu Chen. Enriching Word Embeddings with Domain Knowledge for Readability Assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378. Association for Computational Linguistics, 2018.

[Joh44] Wendell Johnson. I. A program of research. *Psychological Monographs*, 56(2):1–15, 1944.

[Kau] Josh Kaufman. first20hours/google-10000-english: This repo contains a list of the 10,000 most common english words in order of frequency, as determined by n-gram frequency analysis of the google's trillion word corpus. https://github.com/first20hours/google-10000-english. (Accessed on 06/04/2019).

[Kin75] J.P. Kincaid. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis, 1975.

[KL11] Kirill Kireyev and Thomas K Landauer. Word Maturity: Computational Modeling of Word Knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[KLP$^+$10] Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. Learning to Predict Readability using Diverse Linguistic Features. *Computational Linguistics*, 2010.

[KMPL14] David Kauchak, Obay Mouradi, Christopher Pentoney, and Gondy Leroy. Text simplification tools: Using machine learning to discover features that identify difficult text. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2616–2625, 2014.

[LB04] Colleen Lennon and Hal Burdick. The lexile framework as an approach for reading measurement and success. *Electronic publication on www. lexile. com*, 2004.

[LKP11] Thomas K. Landauer, Kirill Kireyev, and Charles Panaccione. Word maturity: A new metric for word knowledge. *Scientific Studies of Reading*, 2011.

[LMV09] N. Mamede I. Trancoso J. Pino M. Eskenazi J. Baptista L. Marujo, J. Lopes and C. Viana. Porting REAP to European Portuguese. *In ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2009)*, 2009.

[Lu10] Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 2010.

[Maa72] H. D. Maas. Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 1972.

[Mcc05] Philip M Mccarthy. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 2005.

[McL69] Harry G McLaughlin. {SMOG} grading - a new readability formula. *Journal of Reading*, pages 639–646, 1969.

[MJ07] Philip M. McCarthy and Scott Jarvis. vocd: A theoretical and empirical evaluation. *Language Testing*, 2007.

[MJGP11] F. Sousa M. J. Grosso, A. Soares and J. Pascoal. *QuaREPE - Quadro de Referência para o Ensino de Português no Estrangeiro. Documento Orientador*. Lisboa: Ministério da Educação e Ciência/Direção Geral de Inovação e Desenvolvimento Curricular, 2011.

[MLC$^+$07] Trudi Miller, Gondy Leroy, Samir Chatterjee, Fan Jie, and Brian Thoms. A classifier to evaluate language specificity of medical documents. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2007.

[MSA$^+$11] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[OSY$^+$13] Theodore A. Omachi, Urmimala Sarkar, Edward H. Yelin, Paul D. Blanc, and Patricia P. Katz. Lower health literacy is associated with poorer health status and outcomes in chronic obstructive pulmonary disease. *Journal of General Internal Medicine*, 2013.

[PL12] J. Nelson, C. Perfetti, and M. Liben. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical Report submitted to the Gates Foundation. https://achievethecore.org/content/upload/nelson_perfetti_liben_measures_of_text_difficulty_research_ela.pdf, 2012. (Accessed on 01/21/2019).

[PN08] E Pitler and A Nenkova. Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

[PSS$^+$17] Dagmara Paiva, Susana Silva, Milton Severo, Pedro Moura-Ferreira, Nuno Lunet, and Ana Azevedo. *Limited Health Literacy in Portugal Assessed with the Newest Vital Sign*, volume 30. Acta Med Port, dec 2017.

[PYM68] Allan Paivio, John C. Yuille, and Stephen A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 1968.

[PZH19] Joao Palotti, Guido Zuccon, and Allan Hanbury. Consumer Health Search on the Web: Study of Web Page Understandability and Its Integration in Ranking Algorithms. *J Med Internet Res*, 21(1):e10986, 2019.

[QPM$^+$13] Patricia Quinlan, Kwanza O. Price, Steven K. Magid, Stephen Lyman, Lisa A. Mandl, and Patricia W. Stone. The Relationship Among Health Literacy, Health Knowledge, and Adherence to Treatment in Patients with Rheumatoid Arthritis. *HSS Journal*, 2013.

[RB15] A R Razon and J A Barnden. A new approach to automated text readability classification based on concept indexing with integrated part-of-speech n-gram features. In *International Conference Recent Advances in Natural Language Processing, RANLP*, 2015.

[RD86] Keith Rayner and Susan A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 1986.

[Ric75] John T E Richardson. Imagery, Concreteness, and Lexical Complexity. *Quarterly Journal of Experimental Psychology*, 27(2):211–223, 1975.

[Ric02] Jack C Richards 1943-. *Longman dictionary of language teaching and applied linguistics*. Third edition / Jack C. Richards and Richard Schmidt ; with Heidi Kendricks and Youngkyu Kim. London ; New York : Longman, 2002., 2002.

[SA10] Carolina Evaristo Scarton and Sandra Maria Aluísio. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural : adaptando as métricas do Coh-Metrix para o Português. *Linguamatica*, 2:45–62, 2010.

[SCC$^+$14] Yao Ting Sung, Ju Ling Chen, Ji Her Cha, Hou Chiang Tseng, Tao Hsing Chang, and Kuo En Chang. Constructing and validating readability models. *The method of integrating multilevel linguistic features with machine learning*, 2014.

[SD00] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz Daimlerchrysler, Colin Shearer, and Rüdiger Wirth Daimlerchrysler. Crisp-dm 1.0. https://www.the-modeling-agency.com/crisp-dm.pdf, 2000. (Accessed on 06/14/2019).

[SO05] Sarah E. Schwarm and Mari 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.

[SPR$^+$15] Kristine Sørensen, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, Barbara Kondilis, Demosthenes Agrafiotis, Ellen Uiters, Maria Falcon, Monika Mensing, Kancho Tchamov, Stephan Van Den Broucke, and Helmut Brand. Health literacy in Europe: Comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health*, 2015.

[SS67] Eric A. Smith and R. Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14, 1967.

[SSZ$^+$13] Xinying Sun, Yuhui Shi, Qingqi Zeng, Yanling Wang, Weijing Du, Nanfang Wei, Ruiqian Xie, and Chun Chang. Determinants of health literacy and health behavior regarding infectious respiratory diseases: A pathway model. *BMC Public Health*, 2013.

[TC12] François Thomas and Fairon Cédrick. An " AI readability " formula for French as a foreign language. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,*, pages 466–477, 2012.

[TC13] Joan Torruella and Ramon Capsada. Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. *Procedia - Social and Behavioral Sciences*, 2013.

[TFG$^+$13] Amali Todirascu, Thomas François, Nuria Gala, Cédrick Fairon, Anne-Laure Ligozat, and Delphine Bernhard. Coherence and Cohesion for the Assessment of Text Readability. In *10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, 2013.

[TITT10] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. Sorting texts by readability. *Computational Linguistics*, 2010.

[TJKT13] Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, 2013.

[VL18] Sowmya Vajjala and Ivana Lucic. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *English Conference Papers, Posters and Proceedings.*, 2018.

[VM12] Sowmya Vajjala and Detmar Meurers. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *In Proceedings of the seventh workshop on building educational applications using NLP*, 2012.

[Wes53] M. West. *A General Service List of English words*. London, UK: Longman, Green & Co, 1953.

[Wika] Legibility - wikipedia. https://en.wikipedia.org/wiki/Legibility. (Accessed on 06/21/2019).

[Wikb] Readability - wikipedia. https://en.wikipedia.org/wiki/Readability. (Accessed on 06/21/2019).

[Wil] Bill Wilson. Grammars and parsing. http://www.cse.unsw.edu.au/~billw/cs9414/notes/nlp/grampars/grampars.html. (Accessed on 06/24/2019).

[WMM+05] Barry D Weiss, Mary Z Mays, William Martz, Kelley Merriam Castro, Darren A DeWalt, Michael P Pignone, Joy Mockbee, and Frank A Hale. Quick assessment of literacy in primary care: the newest vital sign. *Annals of family medicine*, 3(6):514–522, nov 2005.

[XKB16] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016.

[ZT06] Qing T. Zeng and Tony Tse. Exploring and developing consumer health vocabularies, 2006.

[ZTS06] Wei Zhou, Vetle I. Torvik, and Neil R. Smalheiser. ADAM: Another database of abbreviations in MEDLINE. *Bioinformatics*, 2006.