Title

ABSTRACT

In this paper we investigate methods to estimate the understandability of health Web pages, and use these to improve the retrieval of information for people seeking health advice on the Web. Understandability plays a key role in ensuring that people accessing health information are capable of gaining insights that can assist them with their health concerns and choices. The access to unclear or misleading information has been shown to negatively impact on the health decisions of the general public.

Our investigation considers methods to automatically estimate the understandability of health information in Web pages, provides a thorough evaluation of these methods using human assessments as well as an analysis of pre-processing factors affecting understandability estimations, as associated pitfalls. Furthermore, lessons learnt for estimating Web page understandability are applied to the construction of retrieval methods that pay specific attention to retrieving information understandable by the general public.

We find that machine learning techniques are more suitable to estimate health Web page understandability than traditional readability formulas devised by the linguistic community and often used as guidelines and benchmarking by health information providers on the Web. Learning to rank effectively exploit these estimates to provide the general public with more understandable search results. These results are important for specialised search services tailored to support the general public in seeking health advice on the Web.

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
. 2017. Title. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Search engines are concerned with retrieving relevant information to support a user's information seeking task. Commonly, signals about the topicality or aboutness of a piece of information to a query is used to estimate relevance, with other relevance dimensions like understandability, trustworthiness, etc. [] being relegated to a secondary position, or completely neglected. While this may be a minor problem for many information seeking tasks, there are some specific tasks in which dimensions other than topicality have an important role in the information seeking and decision making process. The seeking of health information and advice on the Web by the general public is one such task.

A key problem when searching the Web for health information is that this can be too technical, unclear, unreliable, generally misleading, and can lead to unfounded escalations and poor decisions []. Where correct information exists, it can be hard to find amongst the noise, spam, and irrelevant information. In high-stakes search tasks such as this, access to poor information can lead to poor decisions which ultimately can have a significant impact on our health and well-being []. In this work we are specifically interested in the understandability of health information retrieved by search engines.

The use of general purpose Web search engines like Google and Bing for seeking health advice has been largely questioned and criticised [], despite the commendable efforts these players have put in providing increasingly better health information to their users, e.g., the Google Health Cards []. Ad-hoc solutions to support the general public in searching and accessing health information on the Web have been implemented, typically supported by governments initiatives or medical practitioners associations, e.g., HealthDirect.gov.au, HealthOnTheNet.org. These solutions aim to provide *better* health information to the general public. For example, Health On The Net's mission statement is "to guide Internet users to reliable understandable accessible and trustworthy sources of medical and health information". But, do the solutions these providers currently employ actually provide this type of information to the health-seeking general public? As an illustrative example, we analysed the top 10 search results retrieved by Health On The Net and Health Direct[1] in answers to 300 search queries used within the CLEF 2016 Ehealth evaluation task (see Section **??**). Table/Figure **??** reports the distribution of understandability scores for these search results (note, we did not assess the topical relevance of these results). Understandability scores were computed with the most effective estimation methods from Section **??**. For comparison, we report also the distribution of understandability scores of the "optimal" search results for those queries, as found from the CLEF 2016 Ehealth evaluation task (relevant results that have the highest understandability scores). The results clearly indicate that, despite these solutions are expressively aimed at supporting the general public in accessing understandable health information, they often fail to do so.

In this paper ...

==== ARRIVED TO HERE =====

An existing concern of communicators is knowing if their message is well understood by their target public. Writing for a group of readers other than one's own is difficult. This fact motived research over the past century on the development of readability formulas which can, through a single number, provide hints on the difficulty of a text. Readability formulas were then widely adopted by different groups of the society, showing their effectiveness when increasing the amount learnt by recruits in the army [10] or the readership of newspaper [17].

Nevertheless, with advent and popularity of Web, in which anyone can write about anything, content has been largely generated without proper concern with its understandability. Again: writing for a group of readers other than one's own is difficult and in some domains this might be dangerous. The medical domain is one of such []. The problem, as reported by X, is that health consumers

WOODSTOCK'97, July 1997, El Paso, Texas USA
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Results retrieved on XXX.

might put themselves at risk if they misunderstand the content of what they read on the Web []. Typically health consumers, such as patients and their next-of-kin, start their searches through commercial search engines []. Thus, researchers in various areas of medicine have assessed the understandability of the material retrieved by popular search engines, often finding that they are harder to understand than they should be (e.g., [1, 6, 7, 12, 16, 20]). Mostly, this kind of research also relies on the output of readability formulas to deem if a Web page is easy or hard to understand.

However, as recently discussed in Palotti et al.[15], the automated use of readability formulas firstly requires the content extraction from Web documents. Palotti identified that the decision of appending a period at the end of each element in a list or table extracted from the HTML might result either in a single very long or many very short sentences, which drastically affects the interpretation of readability formula results. The same readability formula may yield results that vary from *suitable even for kids* to *understandable only if you are an experts* with only adding or removing a single period, a 'minor' implementation decision when cleaning HTML pages. One limitation of Palotti's work is not evaluating if either of these interpretations is correct.

In this paper, we propose a vast investigation on the preprocessing of Web documents and the correlation of their understandability estimation with human assessments. For that, we take advantage of the understandability assessments made in recent Information Retrieval campaigns in CLEF eHealth ([14, 24]).

During the CLEF eHealth 2015 and 2016 campaigns [14, 24], organizers explored Internet communities, such as Reddit AskDocs[2] - in which people seek medical advice free of charge, to generate health consumer searches. Participants in CLEF eHealth campaigns were then instruct to search a considerable portion of the Web[3] for documents that could be topically relevant for each consumer query. In addition to topical relevance, medical students and health professionals serving as assessors were instructed to assess how easy to understand and how much they would trust the information contained in each assessed document. In this work, we extensively use the understandability assessments collected in these campaigns.

We correlate the human assessments with the output of various readability formulas applied with different preprocessing settings. Our first intent is to complete Palotti's work, providing guidelines for the automatic use of these formulas in the Web documents in the medical domain. We take advantage of the framework built to further investigate other understandability estimators besides the readability formulas. Finally, we use the lessons learnt in our correlation analysis to demonstrate how is possible to retrieve more understandable medical content.

add roadmap?

2 RELATED WORK

Readability formulas? CLEF work? CIKM paper? White/MSR papers?....???

[2]https://www.reddit.com/r/AskDocs/

[3]In 2016 campaing ClueWeb-12 B was used: http://lemurproject.org/clueweb12/

3 WHICH READABILITY FORMULA TO USE

Palotti, Zuccon and Hanbury [15] compared six different HTML cleaning methods and their impact in the use of readability formulas. They evaluate three methods to extract the content of a Web page from its HTML source: BeautifulSoap 4, which just naively removed HTML tags, Boilerpipe[4] and Justext[5], two approaches to eliminate boilerplate text together with HTML tags. Henceforth these methods are referred as *Naive*, *Boilerpipe - Boi* and *Justext - Jst*, respectively. The authors noticed that the text in HTML tags often missed a correct punctuation mark. For example, the text extract from titles, menus, tables or lists could be interpreted as many short sentences of few very long sentences, depending only whether a period is forced at the end of sentences. These two preprocessing options are henceforth called *ForcePeriod - FP* and *DoNotForcePeriod - DNFP*.

Their experiments found that the use of *Naive* preprocessing was associated with larger variances in the understandability score estimated by readability formulas, while the use of both *Boi* and *Jst* were more stable. Coleman-Liau Index (CLI) was the most stable metric among all tested in Palotti's work[15].

In our experiments, we use the Pearson, Kendall Tau and Spearman's Rank correlation to compare the understandability label assigned to each document by human assessors and by readability formulas. Pearson correlation is used to calculate the strength of two linearly related variables, while the Kendall and Spearman are rank correlation, i.e., they act on the rank of the variables instead of their values. We opt to report on all three correlation coefficients as all three are equally used in the literature, and thus we allow other researchers to compare our results across.

We show in Figures 1 and 2 the correlation scores of each traditional readability metric with the human assessments made in CLEF 2015 and 2016, respectively. We observe that the *Naive* preprocessing also results in the lowest correlation, no matter which correlation measure or readability formula is used. Also, when the *Naive* preprocessing is used, the variant *DoNotForcePeriod* yields higher correlations than the variant *ForcePeriod*, but when using a higher quality HTML cleaner, such as Justext or Boilerplate, the results indicate that the use of *ForcePeriod* should be preferred.

Among all the readability formulas and preprocessing methods, SMOG with *ForcePeriod* preprocessing and Dale-Chall Index with *DoNotForcePeriod* are the best ones respectively for 2015 and 2016. Although there is no single best readability measure or best preprocessing strategy in all scenarios, CLI and FRE with *Justext* are stable options, with correlation coefficients as high as the best ones in both campaigns. Thus, we confirm Palotti's advice for the use of CLI, as it has shown once more to be the most robust measure to variances due to use of *ForcePeriod* or *DoNotForcePerid*.

4 (MORE) UNDERSTANDABILITY ESTIMATORS

The correlation of readability formulas as shown in Figures 1 and 2 is not strong, with no correlation coefficient being higher than 0.5. Our next intent is comparing the correlation coefficient of

[4]Add ref.

[5]Add ref.

the traditional readability formulas with other methods for understandability estimation, including an evaluation of other humans performing the same task. For that, we devise and group several methods into semantically related groups which will be following presented. We summarize all methods in Table 2.

Traditional Readability Formulas: This group contains all the traditional readability formulas listed in Section 2 (or maybe in Table **??**).

Raw Components of Readability Formulas: This group comprises the building blocks that make up the traditional readability measures. Some examples include the average number of characters per word or the average number of syllables in a sentence[6].

General Medical Vocabularies: This group includes methods such as the number of words with a medical prefix or suffix, i.e. beginning or ending with Latin or Greek particles (e.g., amni, angi, algia, arteri), acronyms [7] or medical vocabularies such as the International Statistical Classification of Diseases and Related Health Problems (ICD), Drugbank and the OpenMedSpel dictionary[8]. Methods listed here were matched with documents using a simple keywords matching.

Consumer Vocabulary Features: the Consumer Health Vocabulary (CHV) is a prominent medical vocabulary dedicated to mapping consumer (layperson) vocabulary to technical terms. It attributes a score for each of its concepts with respect to their difficulty, with lower/higher scores for harder/easier concepts. We used MetaMap once again to map the content of Web documents, as done in Chapters **??** and **??**. We further use MetaMap options to also filter only concepts identified as symptoms or diseases, using the same definitions from Section **??**.

Expert Vocabulary Features: The hierarchy of Medical Subject Headers (MeSH) was previously used in the literature to identify hard concepts, assuming that a concept that is deep in the hierarchy is harder than a shallow one [21]. As done with CHV, we used MetaMap to map the content of Web documents to MeSH and explore symptoms and disease concepts separately.

Natural Language: This group comprises commonly used metrics in the natural language processing field: the ratio of part-of-speech (POS) classes, the number of entities in a text, the sentiment polarity and the ratio of words found in English vocabularies. The Python package NLTK 3.2[9] was employed for sentiment analysis and POS tagging. The GNU Aspell[10] dictionary was used as a standard English vocabulary and a stopword list was built by merging the stopword lists of the Indri[11] and Terrier[12] toolkits.

HTML Features: The aim of this group is to represent a web page by its HTML content. We hypothesize that a Web page rich of images or with its content well summarized in tables can potentially ease hard subjects such as medicine. We identify a large number of HTML tags in this group with the Python library BeautifulSoap v4.4[13].

Word Frequency Features: Common and known words are usually frequent words, while unknown and obscure words are rare. This idea is implemented in readability formulas such as the Dale-Chall index which uses a list of common words and counts the number of words that fall outside this list [5]. In this work we model word frequency in a straightforward manner: we sort the frequency of all words in a corpus and normalize the ranking of word frequency such that values close to 100 are attributed to common words and values close to 0 to rare words. We explore three different corpora in this work:

- Medical Reddit: Reddit is an Internet forum with a sizable user community which is responsible for generating content. Any user can start a discussion receiving replies from any other user. This discussion forum is intensively used for health purposes, for example in the Reddit community AskDocs licensed nursers and doctors (subject to user identity verification) advise help seekers free of charge. We selected six of such communities (medical, AskDocs, AskDoctorSmeeee, Health, WomensHealth, Mens_Health) and downloaded all user interactions using the Python Reddit API Wrapper (PRAW[14]), v5.1. In total 43,018 discussions were collected.
- PubMed Central: PubMed Central (PMC) is an online digital database of freely available full-text biomedical literature playing a similar role to physicians as the ACM Digital Library does to computer scientists. We used in this work the same collection crafted for the TREC Clinical Decision Support Track 2014 and 2015 (TREC-CDS)[15] consisting of 733,138 articles.
- Medical English Wikipedia: we filtered articles from a Wikipedia dump[16] (May 1st 2017), that contained an Infobox[17] in which at least one of the following words appeared as a property: ICD10, ICD9, DiseasesDB, MeSH, MeSHID, MeshName, MeshNumber, GeneReviewsName, Orphanet, eMedicine, MedlinePlus, drug_name, Drugs.com, DailyMedID, LOINC. Figure 3 illustrates a Wikipedia page that is marked as medical because of its Infobox entries. This idea was successfully implemented in Soldaini et al. [19] and our filtering process resulted in a collection of 11,942 articles. Note that this procedure highly favors precision over recall.

A summary of the statistics of these three collections is reported in Table 1. In order to calculate word frequency, we removed words that occur less than 5 times in a corpus. Finally, we ignore out of vocabulary (OV) words in our calculations, unless it is explicitly stated.

Machine Learning on Text - Regressors and Classifiers: In a recent survey, Kevin Collins-Thompson reports that the future of understandability estimation relies on Machine Learning [4]. A challenge in using Machine Learning in this task is defining the background corpora used as training set. A possible setup for our work could have used CLEF 2015 assessments to learn a model for CLEF 2016 and vice-versa, but instead, we opt for a more reusable

[6]Words were divided into syllables using the Python package Pyphen http://pyphen.org/
[7]The acronym list was obtained from the ADAM database [23]
[8]http://extensions.openoffice.org/en/project/openmedspel-en-us
[9]http://www.nltk.org/
[10]http://www.aspell.net/
[11]http://www.lemurproject.org/indri/
[12]http://www.terrier.org/
[13]https://www.crummy.com/software/BeautifulSoup/
[14]https://praw.readthedocs.io/
[15]http://www.trec-cds.org/
[16]https://dumps.wikimedia.org/enwiki/
[17]A Wikipedia infobox is a template containing structured information that appear on the right of Wikipedia pages to summarize key aspects of concepts

Table 1: General statistics for the auxiliar collections used in this work

Statistic	Medical Wikipedia	Medical Reddit	PubMed Central
Number of Docs.	11,868	43,019	733,191
Number of Words	10,655,572	11,978,447	144,024,976
Number of Unique Words	467,650	317,106	2,933,167
Avg. Words per Doc.	898.90 ± 1351.76	278.45 ± 359.70	227.22 ± 270.44
Avg. Char per Doc.	5107.81 ± 7618.57	1258.44 ± 1659.96	1309.11 ± 1447.31
Avg. Char per Word	5.68 ± 3.75	4.52 ± 3.52	5.76 ± 3.51

solution for the medical/health domain. We employed the three datasets explained in Section **??** and assume different labels according to the average difficulty of documents in these collections:

- Medical Reddit (label 1): Documents in this collection are expected to be written in a colloquial style, and thus the easiest to understand. All the conversations are in fact explicitly directed to assist inexpert health consumers;
- Medical English Wikipedia (label 2): Documents in this collection are expected to be less formal than scientific articles, but more formal than a Web forum like Reddit;
- PubMed Central (label 3): Documents from this collection are expected to be written in a highly formal style, as the target audience of these documents are physicians, nursers and researchers in the biomedical domain.

Models were trained on a Latent Semantic Analysis (LSA) empirically set to have ten dimension based on word counts in documents in these three collections. We model two different tasks a classification one and a regression task. Different labels for the regression could be employed, for example, a label 5 to PubMed Central documents to emphasize that these documents are explicitly made for expert users. We did not explore the effects of different labels in this work, it is left as future work.

5 TOP MEASURES FROM EACH GROUP

We correlated each individual understandability estimator listed in Table 2 with the human assessments collected in CLEF eHealth 2015 and 2016 campaigns. We report in Table 3 the best metric for each group according to Pearson, Spearman or Kendall correlation. For some groups, such as the readability formula group, the highest correlated metric was the same for different correlation measure: SMOG Index in CLEF eHealth 2015 and Dale-Chall Index in 2016. We highlight the top score value of each correlation measure in each group. Note that there is no single case in which three different metrics were the top correlated for each different correlation measure.

hypotesis that kendall tau and spearman always point to the same winner

Interestingly, Table 3 shows that the polysyllable words, best formula component metric for CLEF 2015 data, is the main metric for the SMOG formula, the best readability formula for CLEF 2015. Likewise, the number of difficult words, best formula component metric for CLEF 2016, is the main metric for Dale-Chall index, the best readability formula for CLEF 2016.

The top correlation for MeSH group, number of MeSH concepts, reaches much lower correlation than the top correlation metric for the CHV group, the scores of CHV concepts. The dominating metrics for the Natural Language group are the number of pronouns, the number of stopwords and the number of out of vocabulary words; all these are consistently more correlated than metrics in the MeSH and CHV group. In turn, the top correlations for the HTML group, counts of P tags and list tags, were the weakest. P tags are used to create paragraphs in a Web page, being roughly a proxy for text lengthiness. Top estimators for the word frequency group are based on the Medical Reddit and PubMed counts, with correlations as high as the readability formulas. Finally, the group with the highest correlated estimators are the regressors and classifiers, with top estimators being the Neural Network regressor and the multinomial Naive Bayes.

this section misses some sort of conclusion or at least a link to the next section

6 WHICH PREPROCESSING APPROACH TO PREFER

We further investigate the preprocessing steps with the groups of features introduced in Table 2. For that, we present in Figures 4 and 5 the box plot of different correlation metrics divided by preprocessing alternative for CLEF eHealth 2015 and 2016. For instance, the very first box plot in the upper part of these figures shows the absolute Pearson's rank correlation of different readability metrics when using a combination of Naive and ForcePeriod as preprocessing steps. Boxes extend from the lower to upper quartile values of the data, with a line at the median. Whiskers extend from the box to show the range of the data. Flier points are those past the end of the whiskers, usually interpreted as outlier values.

We also include in Figures 4 and 5 boxes for the summary of the 3 preprocessing procedures to remove HTML, the use of HTML features, which is done without any preprocessing and the comparison with other human assessors. For CLEF eHealth 2015, we used as human assessments the additional assessments made by unpaid medical students and health consumers (see [13]), while for CLEF eHealth 2016 data, we randomly selected 100 pages that were assessed by another assessor. add at least another person doing assessments. The correlations with human assessments provide important insights on how hard and subjective understandability assessments are.

Figure 4 shows the correlations for CLEF eHealth 2015 assessments. The choice of preprocessing method had the highest impact on the traditional readability formula group, with the Naive preprocessing clearly underperforming the other preprocessing methods. The choice of the Naive method was also the worst with the raw readability formula components and word frequency estimators, but, interestingly, it was a good choice, if not the best one, for all other groups. The highest correlations were archived by the regressors and classifiers, independently of the preprocessing method used.

Similarly to Figure 4, Figure 5 reports the findings for CLEF eHealth 2016. This time, though, the Naive preprocessing method was clearly underperforming for most of the groups analysed, including regressors and classifiers.

In order to further understand our experiments, we compared the median of each pair of preprocessing strategy showed in Figures 4 and 5 and present the results in Table 4. For instance, the entry *FP* < *DNFP* counts the number of times the median value for ForcePeriod was superior to DoNotForcePeriod when comparisons with the same HTML cleaning method was used, e.g. Naive ForcePeriod versus Naive DoNotForcePeriod. From all comparisons, the ones

Table 2: Metrics used as understandability proxies; ⋆: raw values are used. ⋄: values normalised by number of words in a documents are used. †: values normalised by number of sentences in a document are used.

Group	Metric
Traditional Readability Formulas	Automated Readability Index (ARI) [18]
	Coleman-Liau Index (CLI) [3]
	Dale Chall Index (DCI) [5]
	Flesch-Kincaid Grade Level (FKGL) [9]
	Flesch Reading Ease (FRE) [9]
	Gunning Fog Index (GFI) [8]
	Lasbarhetsindex (LIX) [2]
	Simple Measure of Gobbledygook (SMOG) [11]
Raw Components of Readability Measures	# of Characters ⋆⋄†
	# of Words ⋆†
	# of Sentences ⋆⋄
	# of Difficult Words (Dale Chall list [5]) ⋆⋄†
	# of Words Longer than 4 chars ⋆⋄†
	# of Words Longer than 6 chars ⋆⋄†
	# of Words Longer than 10 chars ⋆⋄†
	# of Words Longer than 13 chars ⋆⋄†
	# of Number of Syllables ⋆⋄†
	# of Polysyllable Words (>3 Syllables) ⋆⋄†
Medical Vocabularies	# of Words with Medical Prefix ⋆⋄†
	# of Words with Medical Suffix ⋆⋄†
	# of Acronyms ⋆⋄†
	# of ICD Concepts ⋆⋄†
	# of Drugbank ⋆⋄†
	# of Words in medical dict. (OpenMedSpel) ⋆⋄†
Consumer Health Vocabulary (CHV) [22] Features	CHV Mean Score for all Concepts ⋆⋄†
	# of CHV Concepts ⋆⋄†
	CHV Mean Score for Symptom Concepts ⋆⋄†
	# of CHV Symptom Concepts ⋆⋄†
	CHV Mean Score for Disease Concepts ⋆⋄†
	# of CHV Disease Concepts ⋆⋄†
Medical Subject Headers (MeSH)	# of MeSH Concepts ⋆⋄†
	Average Tree of MeSH Concepts ⋆⋄†
	# of MeSH Symptom Concepts ⋆⋄†
	Average Tree of MeSH Symptom Concepts ⋆⋄†
	# of MeSH Disease Concepts ⋆⋄†
	Average Tree of MeSH Disease Concepts ⋆⋄†
Natural Language	Positive Words ⋆⋄†
	Negative Words ⋆⋄†
	Neutral Words ⋆⋄†
	# of verbs ⋆⋄†
	# of nouns ⋆⋄†
	# of pronouns ⋆⋄†
	# of adjectives ⋆⋄†
	# of adverbs ⋆⋄†
	# of adpositions ⋆⋄†
	# of conjunctions ⋆⋄†
	# of determiners ⋆⋄†
	# of cardinal numbers ⋆⋄†
	# of particles or other function words ⋆⋄†
	# of other POS (foreign words, typos) ⋆⋄†
	# of punctuation ⋆⋄†
	Height of part-of-speech parser tree ⋆⋄†
	# of Entities ⋆⋄†
	# of Stopwords ⋆⋄†
	# of words not found in Aspell Eng. dict. ⋆⋄†

Group	Metric
HTML Features	# of Abbr tags
	# of A tags
	# of Blockquote tags
	# of Bold tags
	# of Cite tags
	# of Div tags
	# of Forms tags
	# of H1 tags
	# of H2 tags
	# of H3 tags
	# of H4 tags
	# of H5 tags
	# of H6 tags
	# of Hs (any H above)
	# of Img tags
	# of Input tags
	# of Link tags
	# of DL tags
	# of UL tags
	# of OL tags
	# of List (DL + UL + OL)
	# of Q tags
	# of Scripts tags
	# of Spans tags
	# of Table tags
	# of P tags
Word Frequency	25th percentil English Wikipedia
	50th percentil English Wikipedia
	75th percentil English Wikipedia
	Mean Rank English Wikipedia
	Mean Rank English Wikipedia - Includes OV
	25th percentil Medical Reddit
	50th percentil Medical Reddit
	75th percentil Medical Reddit
	Mean Rank Medical Reddit
	Mean Rank Medical Reddit ncludelude OV
	25th percentil Pubmed
	50th percentil Pubmed
	75th percentil Pubmed
	Mean Rank Pubmed
	Mean Rank Pubmed - Includes OV
	25th p. Wikipedia+Reddit+Pubmed
	50th p. Wikipedia+Reddit+Pubmed
	75th p. Wikipedia+Reddit+Pubmed
	Mean R. Wiki.+Reddit+Pubmed
	Mean R. Wiki.+Reddit+Pubmed - w. OV
Regressor	Linear Regressor
	Gradient Boosting Regressor
	Multi-layer Perceptron Regressor
	Random Forest Regressor
	Support Vector Machine Regressor
Classifier	Logistic Regression
	Gradient Boosting Classifier
	Multinomial Naive Bayes
	Multi-layer Perceptron Classifier
	Random Forest Classifier
	Support Vector Machine Classifier

Table 3: Metrics with highest correlation per group. In bold are the metric that archived the highest correlation for a correlation measure.

Dataset	Group	Metric	Preprocessing	Pearson	Spearman	KendallTau
CLEF 2015	Readability Formula	SMOG Index	Justext NFP	**0.438**	**0.388**	**0.286**
	Formula Component	Avg. Number of Polysyl. Words per Word	Justext FP	**0.429**	0.364	0.268
		Avg. N. of Polysyl. Words per Sentence	Justext NFP	0.192	**0.388**	**0.286**
	Medical Vocabulary	Avg. N. Medical Prefixes per Word	Naive FP	**0.314**	0.312	0.229
		Number of Medical Prefixes	Naive FP	0.131	**0.368**	**0.272**
	CHV	CHV Mean Score for all Concepts	Naive FP	**0.371**	**0.314**	**0.228**
	MeSH	Number of MeSH Concepts	Naive FP	**0.227**	**0.249**	**0.178**
	Natural Language	N. of words not found in Aspell Dict.	Justext NFP	**0.351**	0.276	0.203
		Number of Pronouns per Word	Naive FP	0.271	**0.441**	**0.325**
	HTML	Number of P Tags	None	**0.219**	**0.196**	**0.142**
	Word Frequency	Mean Rank Medical Reddit - Includes OV	Justext NFP	**0.435**	0.277	0.197
		25th percentil Pubmed	Justext NFP	0.330	**0.347**	**0.256**
	Regressors	Neural Network Regressor	Boilerp. NFP	**0.602**	0.394	0.287
		Neural Network Regressor	Justext FP	0.565	**0.438**	**0.324**
	Classifiers	Multinomial Naive Bayes	Naive FP	**0.573**	**0.477**	**0.416**
CLEF 2016	Readability Formula	Dale Chall Index	Justext FP	**0.439**	0.381	0.264
		Dale Chall Index	Boilerp. FP	0.437	**0.382**	**0.264**
	Formula Component	Avg. Difficult Words Per Word	Boilerp. FP	**0.431**	**0.379**	**0.262**
	Medical Vocabulary	Avg. Prefixes per Sentence	Justext FP	**0.263**	0.242	0.164
		ICD Concepts Per Sentence	Justext NFP	0.014	**0.253**	**0.172**
	CHV	CHV Mean Score for all Concepts	Justext FP	**0.329**	0.313	0.216
		CHV Mean Score for all Concepts	Boilerp. FP	0.329	**0.325**	**0.224**
	MeSH	Number of MeSH Concepts	Boiperp. NFP	**0.201**	0.166	0.113
		Number of MeSH Disease Concepts	Boiperp. NFP	0.179	**0.192**	**0.132**
	Natural Language	Avg. Stopword Per Word	Boiperp. FP	**0.344**	0.312	0.213
		Number of Pronouns	Boiperp. FP	0.341	**0.364**	**0.252**
	HTML	Number of Lists	None	**0.114**	0.021	0.015
		Number of P Tags		0.110	**0.123**	**0.084**
	Word Frequency	Mean Rank Medical Reddit	Boiperp. NFP	**0.387**	0.312	0.214
		50th percentil Medical Reddit	Justext NFP	0.351	**0.315**	**0.216**
	Regressors	Neural Network Regressor	Justext NFP	**0.454**	**0.373**	0.258
		Random Forest Regressor	Boiperp. NFP	0.389	0.355	**0.264**
	Classifiers	Multinomial Naive Bayes	Justext FP	**0.461**	**0.391**	**0.318**

that were statistically significant according to a t-test are shown inside parentheses.

The upper part of Table 4 shows results for the comparisons between ForcePeriod (FP) and DoNotForcePeriod (DNFP). Although the interpretation of readability formulas is drastically affected by this choice of preprocessing, as learning in Chapter **??**, the correlation results are not. The number of times FP reached a higher correlation than DNFP is roughly the same that DNFP was higher than FP. The bottom part of Table 4 shows the comparisons made for Naive, Justext and Boilerpipe. Results for CLEF 2015 contrast with 2016, while Naive was sightly better than Boilerpipe and Justext in 2015, it was the worst in almost all 2016 comparisons. Also, the comparisons between Justext and Boilerpipe are exactly the opposite from 2015 to 2016.

Table 4: Exhaustive Comparison summary using the data from Figures 1.2 and 1.3. Numbers inside parentheses represent the number of tests that yielded p < 0.05 in a two-tailed t-test

Comparison	CLEF 2015				CLEF 2016			
	Pearson	Spearman	Kendall Tau	Total	Pearson	Spearman	Kendall Tau	Total
FP > DNFP	8 (0)	11 (4)	11 (3)	30 (7)	16 (10)	10 (3)	11 (4)	37 (17)
FP < DNFP	16 (5)	12 (5)	12 (6)	40 (16)	8 (0)	12 (2)	11 (2)	31 (4)
FP == DNFP	0	1	1	2	0	2	2	4
Naive > Jst	11 (7)	9 (6)	9 (5)	29 (18)	1 (0)	0 (0)	0 (0)	1 (0)
Naive < Jst	6 (4)	8 (4)	8 (4)	22 (12)	16 (12)	17 (13)	17 (13)	50 (38)
Naive == Jst	0	0	0	0	0	0	0	0
Naive > Boi	12 (7)	10 (6)	10 (5)	32 (18)	0 (0)	0 (0)	0 (0)	0 (0)
Naive < Boi	5 (4)	7 (3)	7 (3)	19 (10)	16 (12)	17 (13)	17 (13)	51 (39)
Naive == Boi	0	0	0	0	0	0	0	0
Jst > Boi	10 (7)	16 (9)	14 (8)	40 (24)	9 (4)	9 (4)	4 (2)	17 (8)
Jst < Boi	7 (2)	1 (0)	3 (1)	11 (3)	8 (2)	8 (2)	13 (2)	34 (6)
Boil == Jst	0	0	0	0	0	0	0	0

7 EXPERIMENTING WITH UNDERSTANDABILITY

The data use made our analysis here was also used in the Information Retrieval branch of CLEF eHealth. We focus our attention to the CLEF eHealth 2016 campaign leaving 2015 experiments offline[18].

We start by the defining the evaluation measure that we will use here. In CLEF eHealth campaign, organizers used a modification of RBP which ties together the relevance of a document with any other relevance dimension, in this case in particular, with understandability. Mathematically, it consists in adding an understandability factor to the RBP formula, as shown:

add formula here

The drawback of such evaluation metric is that we cannot separately evaluate each dimension. We propose, instead, to separately evaluate a ranking list with respect to its topical relevance and its understandability:

- P@10r: a document is topically relevant if assessed as somewhat relevant or highly relevant. This metric counts the number of relevant documents in the top 10 documents of a ranked list.
- P@10u: a document is relevant for this metric if the understandability score is smaller than a threshold U. Like P@10r, we count the number of relevant documents in the first 10 docs of a ranked list. We use $U = 40$ in our experiments.

I decided to use this threshold based on the data. I will need to add a figure

During the campaign, organizers opt to use shallow pools and focus on highly ranked documents, using P@10r as one official metric for topical relevance. It makes our choice of metric natural. Likewise it is traditionally done with F measure, we combine P@10r and P@10u with an harmonic mean: $F_ru = 2 \times \frac{P@10r \times P@10u}{P@10r + P@10u}$

8 CONCLUSION

There is an abundance of factors that affect how readability is perceived by users. In this chapter we devised and studied a large number of readability estimators, ranging from traditional readability formulas extensively used in the past 50 years to state-of-the-art machine learning algorithms. We grouped them into semantically related groups in order to visualize their correlation with human assessments collected during CLEF eHealth campaigns in 2015 and 2016.

Complementary to our previous chapter, we evaluated how preprocessing steps impact the readability estimation in traditional readability formulas and in other modern estimators. We empirically learnt the importance of preprocessing steps when applying readability formulas, as the highest correlations happen when other than the Naive method is used. For the most modern estimators, such as the ones based on machine learning methods, the correlation is less sensible to the preprocessing steps.

We also studied the correlation of each individual readability formula to the human assessment to provide insights on which formula should be preferred. Our analysis concluded that the Simple Measure of Gobbledygook (SMOG) and Dale-Chall Index (DCI) were the most correlated metrics for the two datasets studied and, together with Coleman-Liau Index (CLI) and the Flesch Reading Ease (FRE) are the most stable metrics across datasets, and therefore, should be preferred.

Finally, this chapter serves as a basis for the following chapters of this work, as the learning to rank methods will largely use the estimators devised and analysed here.

[18] Link to experiments will be available upon acceptance of this manuscript

REFERENCES

[1] Samuel R Atcherson, Ashley E DeLaune, Kristie Hadden, Richard I Zraick, Rebecca J Kelly-Campbell, and Carlos P Minaya. 2014. A Computer-Based Readability Analysis of Consumer Materials on the American Speech-Language-Hearing Association Website. *Contemporary Issues in Communication Science & Disorders* 41 (2014).

[2] C. H. BjÃűrnsson. 1983. Readability of Newspapers in 11 Languages. *Reading Research Quarterly* 18, 4 (1983), 480–497. http://www.jstor.org/stable/747382

[3] Meri Coleman and T. L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology* (1975).

Table 5: Reports on the experiments with 4 base runs: the top 3 runs of CLEF eHealth 2017 and a plain baseline run. The second (indices 5-8) and third (indices 9-12) parts shows results of a RegressorTree based on top features from Table X. The forth part of this table shows results when reranking top 20 results based on Dale-Chall Index. Finally, we combine selected runs with reciprocal rank fusion in the last part of this table. Results shows that regression on top 15 improves understandability to the detriment of topical relevance. Regression on top 20 it is even more agreessive. Dale Chall presents the same logic, but understandability gains are smaller compared to regressor results. Finally, the combination with rrf yields the best F_ru scores. I still need to compute if we get statistically significant improvements or not.

Index	Rerank	Run	P_r@10	P_u@10	F_ru	RBP	uRBP	Unj@10	P_r@10	P_u@10	F_ru*
1	No Rerank	BM25 Q.E.	31.43	49.70	38.51	29.05	18.65	0.02	31.70	50.83	39.05
2		GUIR	29.67	50.97	37.50	28.15	18.12	0.01	30.20	51.80	38.16
3		ECNU	29.33	49.27	36.77	27.70	17.55	0.01	29.50	49.87	37.07
4		Plain BM25	26.47	46.33	33.60	25.28	16.05	0.06	27.63	48.93	35.32
5	Regress Top 15	Based on Run 1	27.87	49.40	35.63	26.38	17.17	0.01	28.10	49.60	35.88
6		Based on Run 2	27.27	48.57	34.93	25.95	19.31	0.01	27.30	48.60	34.96
7		Based on Run 3	26.60	50.10	34.75	25.98	18.68	0.00	26.67	50.10	34.81
8		Based on Run 4	23.73	50.87	32.37	22.86	15.68	0.09	24.97	52.67	33.87
9	Regress. Top 20	Based on Run 1	27.70	**52.13**	34–29	36.18	26.10	20.06	0.02	28.17	52.53	36.67
10		Based on Run 2	26.70	**53.10**	35.53	25.44	19.62	0.01	26.83	53.33	35.70
11		Based on Run 3	26.17	**52.67**	34.96	24.88	19.21	0.02	26.43	53.17	35.31
12		Based on Run 4	23.30	**51.43**	32.07	22.53	18.01	0.12	25.10	55.20	34.51
13	Dale-Chall top 20	Based on Run 1	29.63	49.93	37.19	26.45	17.78	0.02	29.87	50.37	37.50
14		Based on Run 2	28.37	51.10	36.48	25.91	16.96	0.01	28.60	51.40	36.75
15		Based on Run 3	27.70	50.80	35–27	35.85	24.54	16.49	0.02	28.03	51.47	36.30
16		Based on Run 4	24.13	49.97	32.55	21.28	14.60	0.10	26.17	53.17	35.07
17	Reciprocal Rank Fusion	Runs 1 and 9	30.70	52.37	**38.71**	27.93	20.20	0.02	31.10	53.30	39.28
18		Runs 2 and 10	30.67	51.90	**38.55**	28.34	20.11	0.01	30.97	52.53	38.96
19		Runs 3 and 11	30.40	51.37	**38.20**	27.82	19.72	0.01	30.67	51.90	38.55
20		Runs 4 and 12	28.67	50.67	**36.62**	26.33	18.68	0.03	29.37	51.87	37.50
21	LTR XBoost	BM25 Only IR Feat & IR label	23.43	43.97	30.57	22.43	14.34	0.12	26.57	50.90	34.91
22		BM25 & Unders. Feat.	29.33	41.00	34.20	28.14	16.91	0.17	34.37	51.10	**41.10**

[4] Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165, 2 (2014), 97–135.

[5] Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability: Instructions. *Educational Research Bulletin* 27, 2 (1948), 37–54. http://www.jstor.org/stable/1473669

[6] PR Fitzsimmons, BD Michael, JL Hulley, and GO Scott. 2010. A readability assessment of online Parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh* 40, 4 (2010), 292–296.

[7] Mark A Graber, Cathy M Roller, and Betsy Kaeble. 1999. Readability levels of patient education material on the World Wide Web. *Journal of Family Practice* 48, 1 (1999), 58–59.

[8] Robert Gunning. 1952. *The Technique of Clear Writing.* McGraw-Hill.

[9] J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. *Derivation of New Readability Formulas for Navy Enlisted Personnel.* National Technical Information Service.

[10] G. R. Klare, J. E. Mabry, and L. M. Gustafson. 1955. The relationship of style difficulty to immediate retention and to acceptability of technical material. In *Journal of Education Psychology*. 287–295.

[11] G. Harry McLaughlin. 1969. SMOG Grading - a New Readability Formula. *Journal of Reading* (1969).

[12] Andrew Meillier and Shyam Patel. 2017. Readability of Healthcare Literature for Gastroparesis and Evaluation of Medical Terminology in Reading Difficulty. *Gastroenterology Research* 10, 1 (2017), 1–5.

[13] Joao Palotti, Guido Zuccon, Johannes Bernhardt, Allan Hanbury, and Lorraine Goeuriot. 2016. Assessors Agreement: A Case Study across Assessor Type, Payment Levels, Query Variations and Relevance Dimensions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF'16 Proceedings*. Springer International Publishing.

[14] João Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth J. F. Jones, Mihai Lupu, and Pavel Pecina. 2015. ShARe/CLEF eHealth Evaluation Lab 2015, Task 2: User-centred Health Information Retrieval. In *Working Notes for CLEF 2015 Conference, Toulouse, France, September 8-11, 2015.*

[15] João Rafael de Moura Palotti, Guido Zuccon, and Allan Hanbury. 2015. The Influence of Pre-processing on the Estimation of Readability of Web Documents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1763–1766.

[16] Chirag R Patel, Deepa V Cherla, Saurin Sanghvi, Soly Baredes, and Jean Anderson Eloy. 2013. Readability assessment of online thyroid surgery patient education materials. *Head & neck* 35, 10 (2013), 1421–1425.

[17] Bernard Lionel Perry. 1954. *Newspaper readability: A pilot study.* Ph.D. Dissertation. Boston University, USA.

[18] E. A. Smith and R. J. Senter. 1967. *Automated Readability Index.* Aerospace Medical Research Laboratories.

[19] Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. *Retrieving Medical Literature for Clinical Decision Support*. Springer International Publishing, Cham, 538–549.

[20] R Constance Wiener and Regina Wiener-Pla. 2014. Literacy, pregnancy and potential oral health changes: The internet and readability levels. *Maternal and child health journal* 18, 3 (2014), 657–662.

[21] Xin Yan, Raymond Y.K. Lau, Dawei Song, Xue Li, and Jian Ma. 2011. Toward a semantic granularity model for domain-specific information retrieval. *ACM Transactions on Information Systems* 29, 3, Article 15 (July 2011), 46 pages. https://doi.org/10.1145/1993036.1993039

[22] Qing Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association* 13, 1 (2006), 24–29.

[23] W. Zhou, V. Torvik, and N. Smalheiser. 2006. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics* 22, 22 (2006), 2813–2818. /brokenurl# http://publication.wilsonwong.me/load.php?id=233282083

[24] Guido Zuccon, Joao Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Mueller, Julie Budaher, and Anthony Deacon. 2016. The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In *CLEF 2016-Conference and Labs of the Evaluation Forum*, Vol. 1609. 15–27.

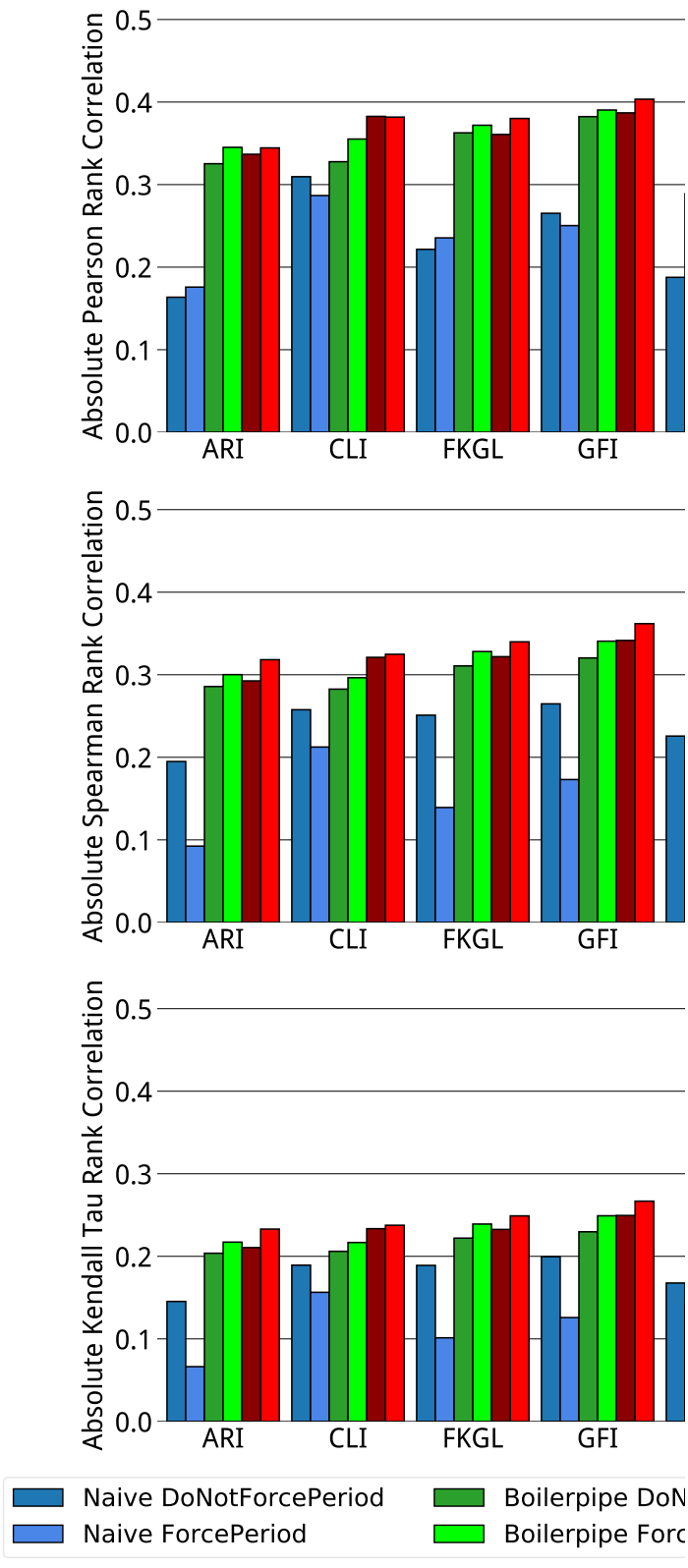


Figure 1: Correlation of different readability measures and the understandability scores collected in CLEF eHealth 2015

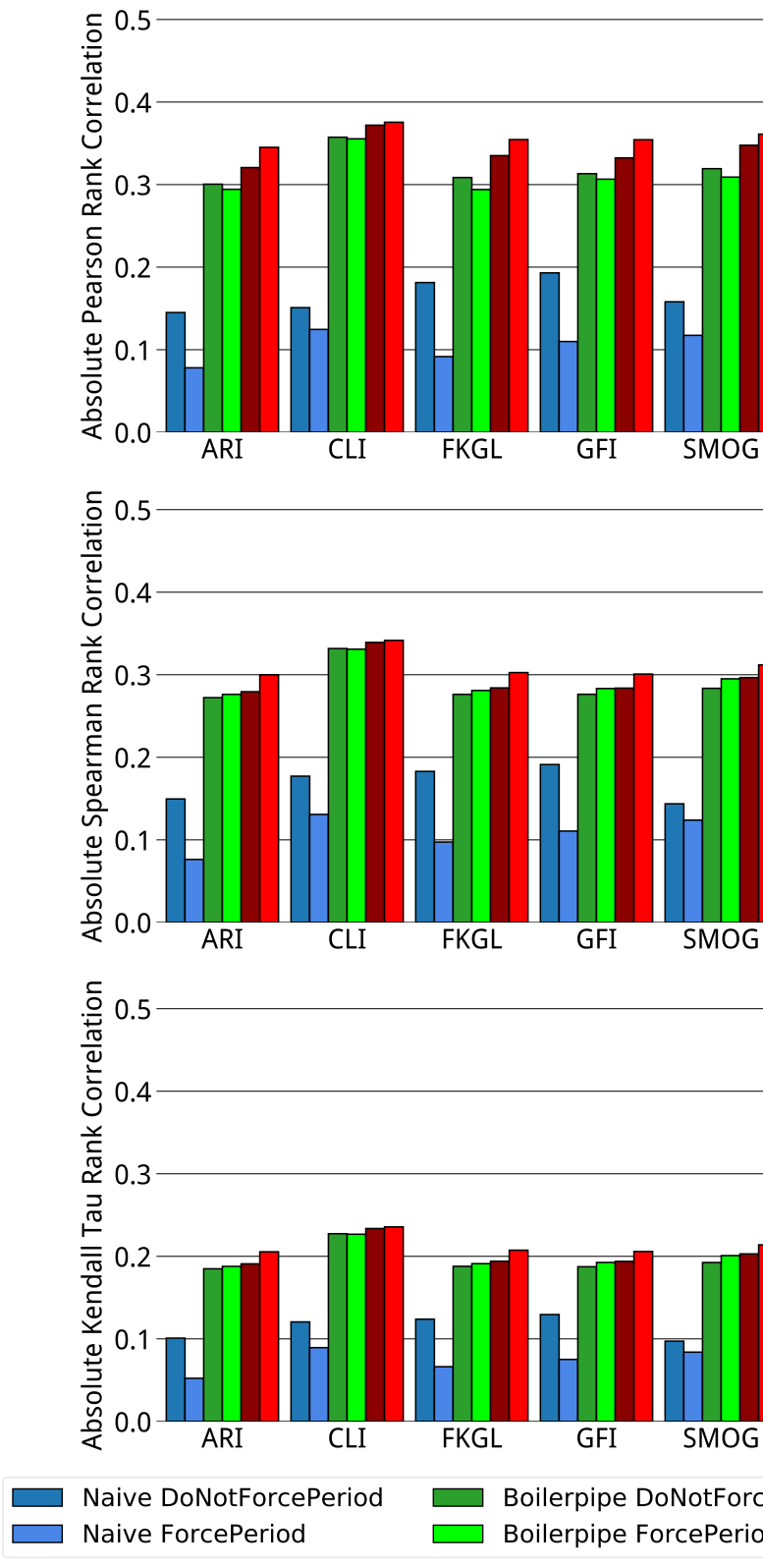


Figure 2: Correlation of different readability measures and the understandability scores collected in CLEF eHealth 2016.

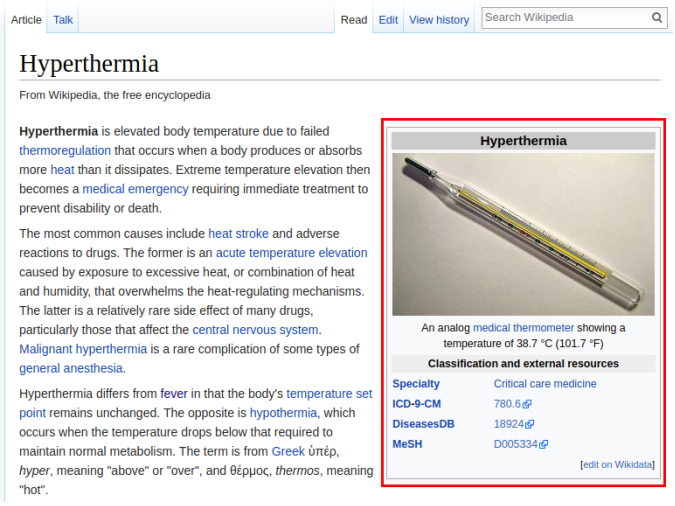
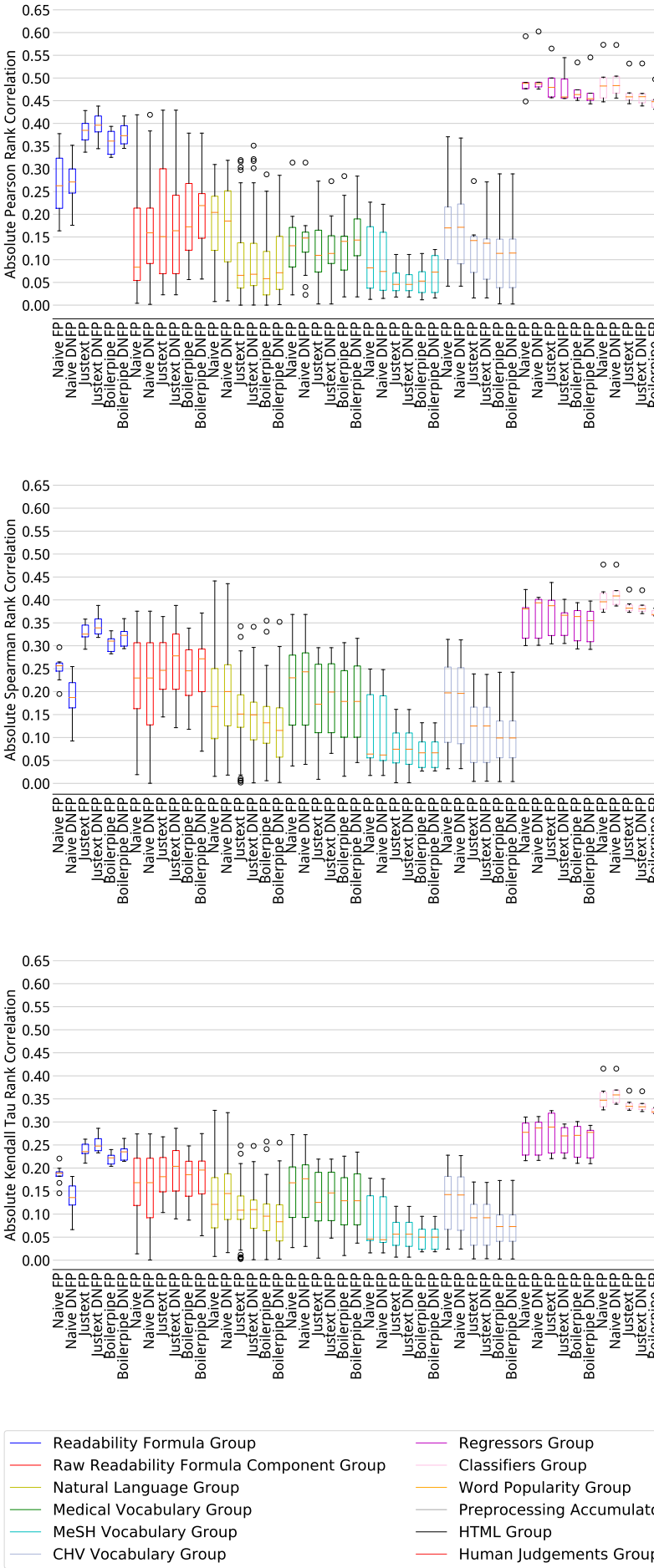
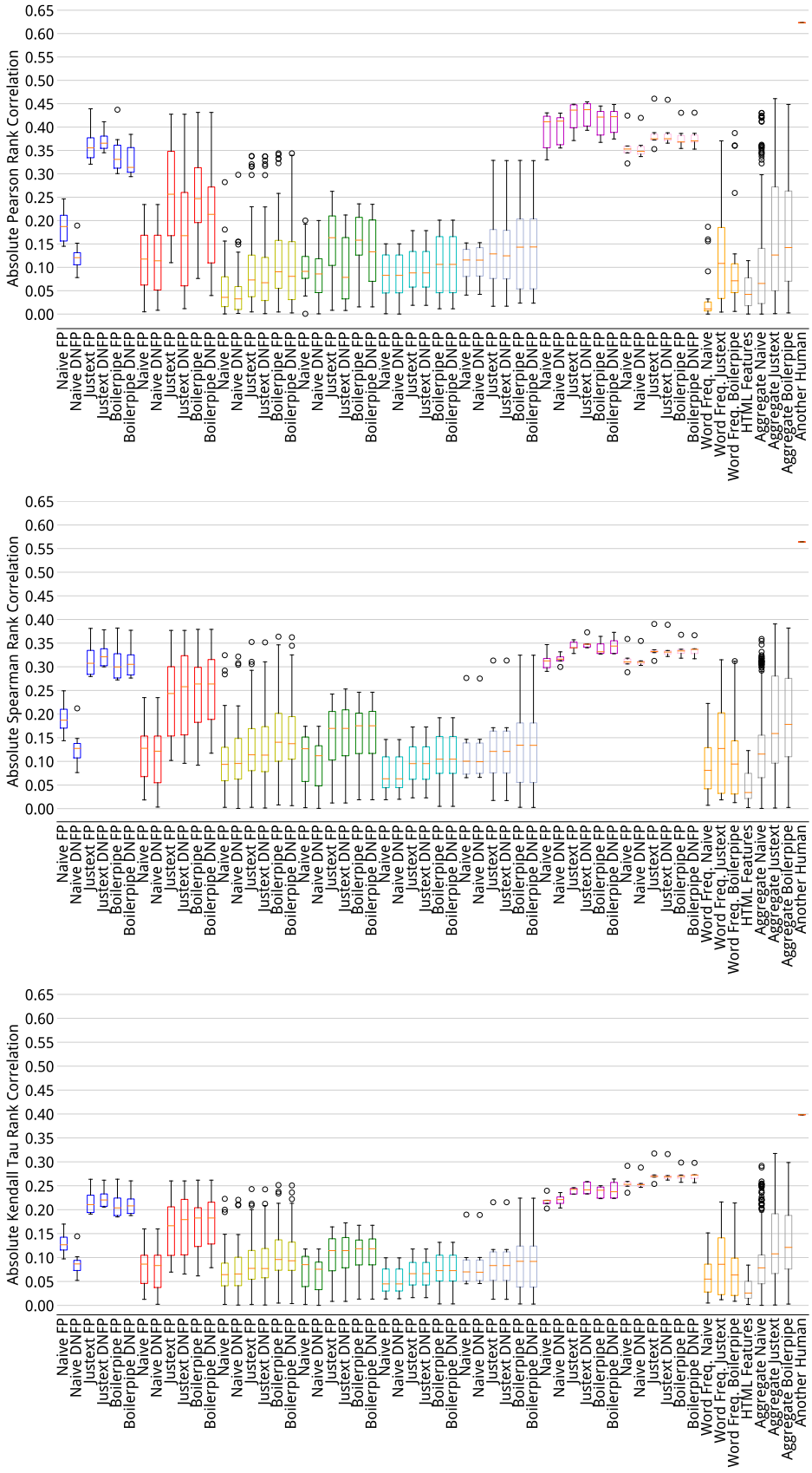


Figure 3: Wikipedia page on hyperthermia. A rectangular red box identify the Infobox on the right hand side containing entries for Specialty, ICD-9-CM, DiseasesDB and MeSH.




Absolute Pearson Rank Correlation
Absolute Spearman Rank Correlation
Absolute Kendall Tau Rank Correlation
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Naive FP
Naive DNFP
Justext FP
Justext DNFP
Boilerpipe FP
Boilerpipe DNFP
Word Freq. Naive
Word Freq. Justext
Word Freq. Boilerpipe
HTML Features
Aggregate Naive
Aggregate Justext
Aggregate Boilerpipe
Another Human
Readability Formula Group
Raw Readability Formula Component Group
Natural Language Group
Medical Vocabulary Group
MeSH Vocabulary Group
CHV Vocabulary Group
Regressors Group
Classifiers Group
Word Popularity Group
Preprocessing Accumulator Group
HTML Group
Human Judgements Group