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, 14 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 are 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 [21]. 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 [20, 21]. In this work we are specifically interested in the understandability of health information retrieved by search engines, and in improving search results to favour information understandable by the general public.

The use of general purpose Web search engines like Google and Bing for seeking health advice has been largely questioned and criticised [1, 9, 10, 14, 17, 22], 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 services 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 we propose and investigate methods for the estimation of understandability for health information in Web pages. In doing so, we also study the influence of HTML processing methods on these estimations, and their pitfalls. Then, we investigate how understandability estimations can be integrated into retrieval methods to enhance the quality

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.

of the retrieved health information, with particular attention to its understandability by the general public.

2 RELATED WORK

There is a large body of literature that has examined the understandability of Web health content when the information seeker is a member of the general public. For example ... A common finding of these studies is that, in general health content available on Web pages are often hard to understand by the general public; this includes content that is retrieved in top rank positions by current commercial search engines.

Previous Linguistics and Information Retrieval research has attempted to devise computational methods for the automatic estimation of text readability and understandability [2], and for the inclusion of these within search methods. Computational approaches to understandability estimations include (1) readability formulas, which exploit word surface characteristics of the text, (2) matching with specialised dictionaries or terminologies, often compiled with information about understandability difficulty, (3) machine learning approaches.

Measures such as Coleman-Liau Index (CLI), Dale-Chall Index, XXX (FRE), XXX (SMOG) [] belong to the first category. These measures generally rely on ... While these measures have been widely used in studies investigating the understandability of health content retrieved by search engines, e.g. [2], Palotti et al. found that these measures are heavily affected by the methods used to extract text from the HTML source [16]. They were able to identify specific settings of a HTML preprocessing pipeline that provided consistent estimates. We shall revisit this work in more details in Section 3, as we further investigate this problem by comparing the effect of HTML preprocessing on text understandability estimations in light of explicit human assessments.

The use of the Consumer Health Vocabulary (CHV) or of terminologies like the MEdical Subject Headings (MESH) belong to the second category. TODO: here speak about CHV, but also about Dawei Song's use of UMLS/Mesh ontology level.

Use of Machine Learning for understandability: anything in health specifically?

3 WHICH READABILITY FORMULA TO USE

Palotti, Zuccon and Hanbury [16] 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[3] and Justext[4], 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[16].

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 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 **??**).

[2]TODO: difference between readability and understandability
[3]Add ref.
[4]Add ref.

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[5].

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[6] or medical vocabularies such as the International Statistical Classification of Diseases and Related Health Problems (ICD), Drugbank and the OpenMedSpel dictionary[7]. 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 NLM MetaMap[8] to map contents occurring in Web documents to the CHV vocabulary. We further use MetaMap options to also filter only concepts identified as symptoms or diseases. Similar approach is commonly used in the literature [] maybe cite something that is not ours.

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 [23]. As done with CHV, we used MetaMap to map the content of Web documents to MeSH, exploring symptoms and disease concepts separately.

Natural Language: This group comprises commonly used metrics in the natural language processing field, such as 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 [8]. 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[14] is an Internet forum with a sizable user community which is responsible for generating its 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[15] 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 library PRAW[16]. In total 43,018 discussions were collected.
- PubMed Central: PubMed Central[17] 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)[18] consisting of 733,138 articles.
- Medical English Wikipedia: we filtered articles from a Wikipedia dump[19] (May 1st 2017), that contained an Infobox[20] 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. In case we need space, I suggest we drop this figure

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, unless it is explicitly stated otherwise, we ignore out of vocabulary (OV) words in our calculations,

Machine Learning on Text - Regressors and Classifiers: In a recent survey, Kevin Collins-Thompson reports

[5] Words were divided into syllables using the Python package Pyphen http://pyphen.org/
[6] The acronym list was obtained from the ADAM database [25]
[7] http://extensions.openoffice.org/en/project/openmedspel-en-us
[8] https://metamap.nlm.nih.gov/
[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://www.reddit.com/
[15] https://www.reddit.com/r/AskDocs/
[16] https://praw.readthedocs.io/
[17] https://www.ncbi.nlm.nih.gov/pmc/
[18] http://www.trec-cds.org/
[19] https://dumps.wikimedia.org/enwiki/
[20] 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

that the future of understandability estimation relies on Machine Learning [6]. 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 solution for the medical/health domain. We employed the three datasets described in Table 1 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 here 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 6 to PubMed Central documents would emphasize that these documents are explicitly made for expert users, being 3 times harder than Wikipedia ones. 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

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 [15]), 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.

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) [5]
	Dale Chall Index (DCI) [8]
	Flesch-Kincaid Grade Level (FKGL) [12]
	Flesch Reading Ease (FRE) [12]
	Gunning Fog Index (GFI) [11]
	Lasbarhetsindex (LIX) [3]
	Simple Measure of Gobbledygook (SMOG) [13]
Raw Components of Readability Measures	# of Characters ⋆♢†
	# of Words ⋆†
	# of Sentences ⋆♢
	# of Difficult Words (Dale Chall list [8]) ⋆♢†
	# 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) [24] 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

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

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**

Naive DoNotForcePeriod. From all comparisons, the ones 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.

7 EXPERIMENTING WITH UNDERSTANDABILITY

Our experiments are based on the data collected during the Information Retrieval branch of CLEF eHealth. To the best of our knowledge that was the only venue in which topical and understandability assessments were collected. We show here experiments with the CLEF eHealth 2016 campaign leaving 2015 experiments offline[21].

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

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

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 document relevance any other relevance dimension, in this case in particular, understandability [26]. Mathematically, it consists in adding an understandability factor to the RBP formula: $uRBP(\rho) = (1-\rho)\sum_{k=1}^{K} \rho^{K-1} r(d@K)u(d@k)$, with r(d@k) representing the gain in retrieving a topically relevant documents at rank k and u(d@k) the gains coming from the understandability factor of a document at rank k.

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_r@10: a document is topically relevant if assessed as somewhat relevant or highly relevant. This metric counts the number of topically relevant documents in the top 10 documents of a ranked list.
- P_u@10: a document is relevant for this metric if the understandability score is smaller than a threshold U. Like P_r@10, we count the number of relevant documents in the first 10 documents 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 to sup-

During the campaign, organizers opt to use shallow pools and focus on highly ranked documents, using P_r@10 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_r@10 and P_u@10 with an harmonic mean: $F_{ru} = 2 \times \frac{P_r@10 \times P_u@10}{P_r@10 + P_u@10}$.

When reranking runs, we risk to bring to the top of the rankings documents that were not assessed when the task took place. For this reason, we inform the average percentage of documents that were not assessed in the top 10 documents of each run (Unj@10) and we inform P_r@10$*$, P_u@10$*$ and $F_{ru}*$, modified measures that excluding all unassessed documents for the ranking list before evaluation. These last three metrics showcase the potential of a run if it had been included in the pool set.

We start our experiments by showing at the top of Table 5 (indices 1-4) the performance of the top 3 systems in CLEF eHealth 2016 together with a straightforward BM25 baseline run made with Terrier toolkit. Our further experiments will use not only these runs as a comparison base, but modify these runs when necessary.

The first batch of experiments, indices 5-8, consists of reranking the top 20 documents based solely on the Dale-Chall index, as it was the metrics with highest correlation with human judgements for this collection. The topical relevance of the reranked resulting lists is immediately hurt in benefit of understandability. Note how P_r@10 and P_r@10$*$ decrease while P_u@10 and in special P_u@10$*$ increase. The percentage of unassessed documents reach 10% and, unfortunately, only looking at RBP and uRBP we cannot see the understandability benefits brought by the rerank method.

The next batch of experiments, indices 9-19, swaps the Dale-Chall Index by a Linear Regression (LR) learned using the top features of Table 3[22]. We experiment with reranking the top 15 (more more conservative method), 20 and 50 (a more heterodox method). We clearly see the trade-off between understandability and relevance when we increase from top 15 to top 50: while the P_r@10 falls 5 points, P_u@10 increases 10. Nevertheless, $F_{ru}*$ shows that more conservative methods provide the best trade-offs. When comparing the use of Dale-Chall and Linear Regression reranking the same number of documents (indices 13-16), we notice that LR results for P_u@10$*$ are always superior to Dale-Chall, meaning it does a better job when bringing the easier documents to the top of the rank, but that not necessarily means the final rank is better, as we can see comparing the results for $F_{ru}*$.

[22]Experiments with various machine learning methods were performed as well and will be available in a website upon paper acceptance. The results of other methods are similar

Our third batch of experiments, indices 20-31, aims to combine the good ranking for topicality coming from the original rank list with the best rankings for understandability, coming from the LR reranking. For that, we use the Reciprocal Rank Fusion (RRF) method, which combines runs on the rank of documents instead of scores [7]. Reranking with LR and combining runs with RRF allows to obtain the best trade-offs so far, with F_{ru} of the best method in CLEF being improved from 39.05 to 39.68, with comes from a very small decrease in P_u@10∗ (from 31.70 to 31.60) and a 4.9% increase in P_u@10∗ (from 50.83 to 53.30).

Our final batch of experiments, indices 32-34, are based on learning-to-rank. We decided to use the pairwise implementation of the gradient boosting approach from the XGBoost framework [4] due to its recent good results in many machine learning tasks . Evaluating other learning to rank methods and frameworks are left as future work. We evaluated three different settings based on the plain BM25 run shown in index 4, each one with a different combination of feature set and relevance labels:

actually, I should use XGBoost method instead of LR

- Index 32: we used as relevance labels only the topical labels and as feature set the score of different information retrieval models (BM25, Dirichlet LM and PL2). This represents the typical use of learning to rank aiming to optimize only the topical relevance results. Unfortunately that was not the case and not even the P_r@10∗ was improved.
- Index 33: we used as relevance labels the linear combination of topical relevance and understandability. Additionally to the IR features, we used the 10 features correspondent to the best features according to Kendall tau in Table 3. Results show that this resulted in higher F_{ur}∗ when compared to its correspondent approaches (indices 4, 8, 12, 16, 19, 23, 27 and 31) due to a high increase in P_r@10∗, which reveals understandability features improving topicability.
- Index 34: On top of our previous run, we included here all features shown in Table 2. Note that despite the very high number of unassessed documents, this run shows improvements even in P_r@10 obtaining the same results as the third best run in the campaign (ECNU). In its turn, the P_r∗ which ignores the unassessed documents is 8.4% higher than the best system, while F_{ru}∗ is 5.2% higher.

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.

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] Shirley Ann Becker. 2004. A study of web usability for older adults seeking online health resources. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11, 4 (2004), 387–406.

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

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

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

[6] 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.

[7] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[8] 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

[9] 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.

[10] 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.

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

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

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

Table 5: Different batchs of experiments are shown below, all based in the runs listed in rows indices 1-4. The last row shows our best result based on learning to rank which modifies a simple BM25 baseline (row 4) and reaches results that are higher than the best system in CLEF eHealth 2016.

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. (Best Run)	31.43	49.70	38.51	29.05	18.65	0.02	31.70	50.83	39.05
2		GUIR (Runner Up Run)	29.67	50.97	37.50	28.11	18.12	0.01	30.20	51.80	38.16
3		ECNU (Thrid Place)	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	33.69	25.28	16.05	0.06	27.63	48.93	35.32
5	Dale-Chall Top 20	Based on Run 1	28.43	49.57	36.14	24.64	16.78	0.10	30.63	54.57	39.24
6		Based on Run 2	25.90	52.80	34.75	22.19	15.36	0.06	27.27	55.90	36.65
7		Based on Run 3	26.87	53.07	35.67	23.01	15.70	0.05	27.83	54.90	36.94
8		Based on Run 4	24.60	50.43	33.07	21.58	14.83	0.09	26.30	53.00	35.16
9	Linear Regre. Top 15	Based on Run 1	30.17	50.50	37.77	26.10	17.92	0.07	31.50	53.13	39.55
10		Based on Run 2	29.23	53.03	37.69	25.16	17.23	0.03	29.73	54.13	38.38
11		Based on Run 3	28.23	52.57	36.74	24.21	16.53	0.02	28.83	53.30	37.42
12		Based on Run 4	26.00	49.53	34.10	22.75	15.67	0.07	26.83	51.00	35.16
13	Linear Regre. Top 20	Based on Run 1	27.57	49.87	35.51	24.00	16.71	0.11	30.03	54.93	38.83
14		Based on Run 2	26.50	54.00	35.55	22.45	15.58	0.06	28.13	57.03	37.68
15		Based on Run 3	26.33	53.50	35.29	22.67	15.72	0.05	27.50	55.77	36.84
16		Based on Run 4	24.40	50.37	32.87	20.95	14.68	0.10	26.00	53.20	34.93
17	Linear Regre. Top 50	Based on Run 1	18.83	43.40	26.27	16.32	11.89	0.29	26.33	59.93	36.59
18		Based on Run 2	18.57	45.60	26.39	16.18	11.83	0.26	24.83	60.23	35.17
19		Based on Run 3	18.17	43.83	25.69	15.56	11.26	0.28	24.20	59.13	34.34
19		Based on Run 4	16.33	43.57	23.76	14.13	10.30	0.29	22.77	58.20	32.73
20	RRF Top 15 (LR & Orig.)	Based on Run 1	30.43	50.40	37.95	28.46	19.09	0.04	31.37	52.47	39.26
21		Based on Run 2	29.43	52.93	37.83	27.17	18.15	0.02	30.00	53.80	38.52
22		Based on Run 3	28.70	51.87	36.95	26.78	17.82	0.02	29.10	52.63	37.48
23		Based on Run 4	26.17	48.90	34.09	24.84	16.56	0.06	27.27	51.27	35.60
24	RRF Top 20 (LR & Orig.)	Based on Run 1	30.30	50.23	37.80	27.97	18.92	0.06	31.60	53.30	39.68
25		Based on Run 2	28.73	53.83	37.47	26.54	17.89	0.03	29.43	55.23	38.40
26		Based on Run 3	28.57	52.73	37.06	26.21	17.57	0.03	29.30	54.07	38.00
27		Based on Run 4	25.67	50.30	33.99	24.53	16.59	0.07	27.00	53.07	35.79
28	RRF Top 50 (LR & Orig.)	Based on Run 1	27.63	50.77	35.79	25.58	17.71	0.12	30.50	56.63	39.65
29		Based on Run 2	26.00	53.40	34.97	23.62	16.46	0.09	28.37	58.10	38.12
30		Based on Run 3	25.47	52.40	34.28	23.52	16.22	0.09	27.80	57.07	37.39
31		Based on Run 4	24.10	50.47	32.62	22.10	15.27	0.12	26.33	56.03	35.83
32	LTR XGBoost	Run 4 Only IR Feat & IR label	23.43	43.97	30.57	22.43	14.34	0.12	26.57	50.90	34.91
33		Run 4 + Few ReadFeat	23.37	39.47	29.35	21.21	13.15	0.21	29.50	50.73	37.31
34		Run 4 + All ReadFeat	29.33	41.00	34.20	28.14	16.91	0.17	34.37	51.10	41.10

[14] 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.

[15] 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.

[16] João 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.

[17] 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.

[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, 538–549.

[20] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*. ACM, New York, NY, USA, 3–12. https://doi.org/10.1145/2484028.2484053

[21] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Transactions on Information Systems* 27, 4, Article 23 (Nov. 2009), 37 pages. https://doi.org/10.1145/1629096.1629101

[22] 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.

[23] 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

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

[25] 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

[26] 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.

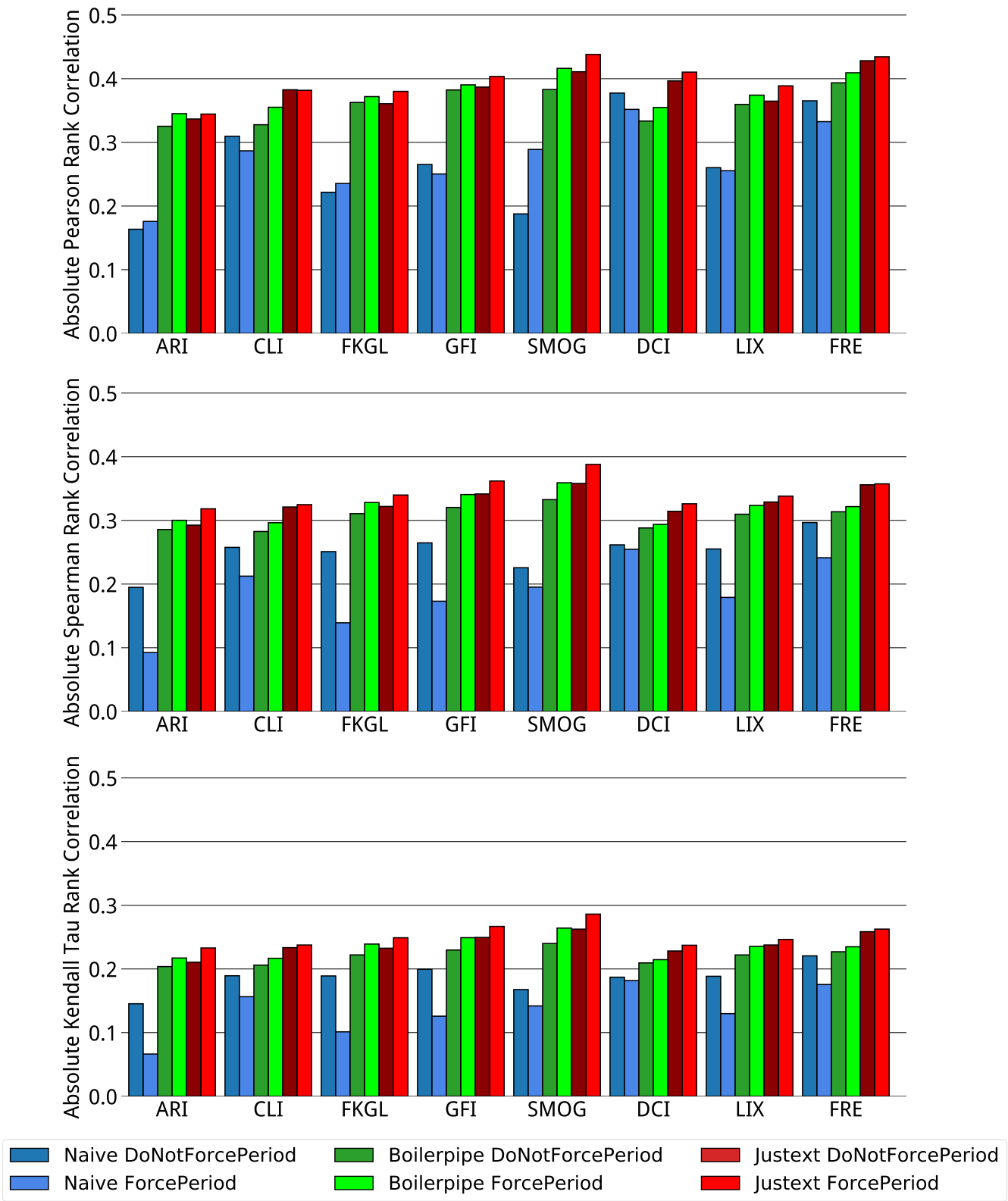


Figure 1: Correlation of different readability measures and the understandability scores collected in CLEF eHealth 2015

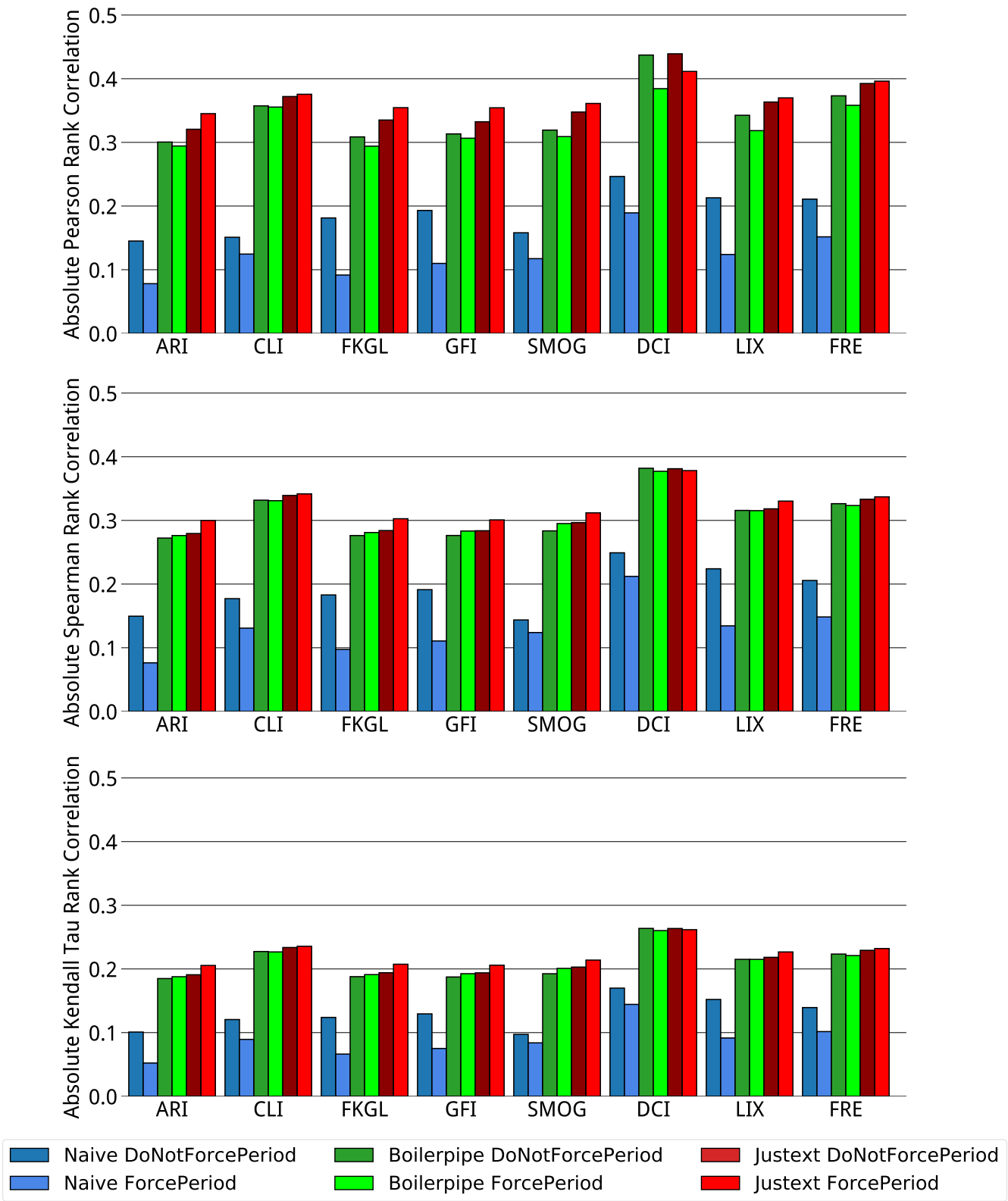


Figure 2: Correlation of different readability measures and the understandability scores collected in CLEF eHealth 2016.

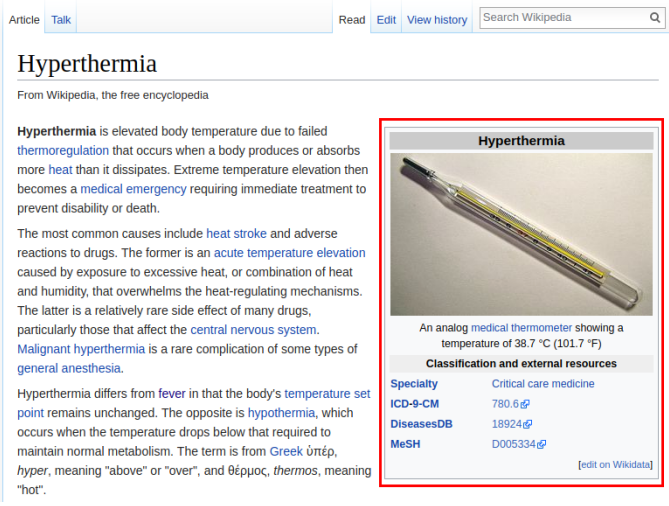
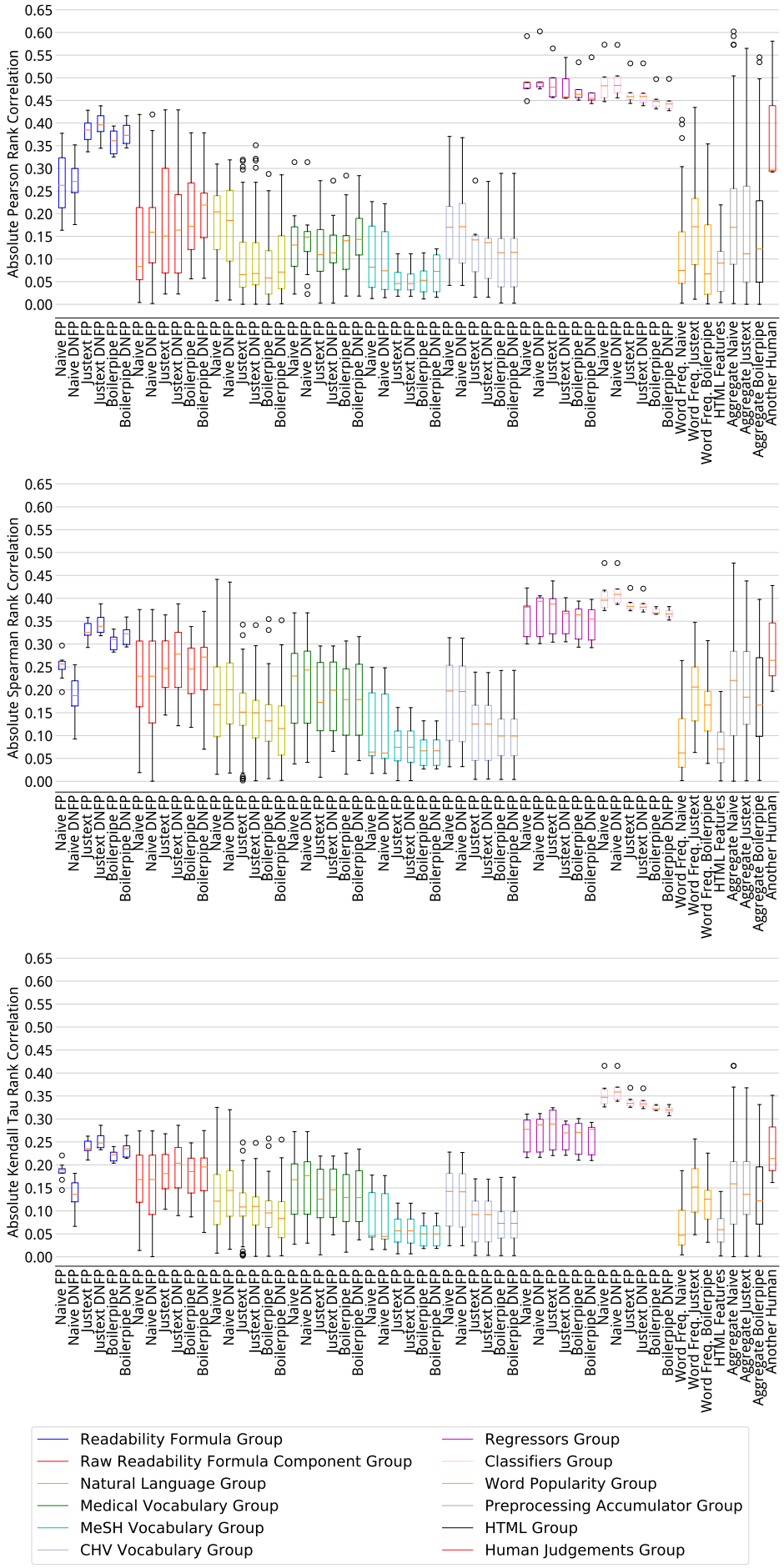
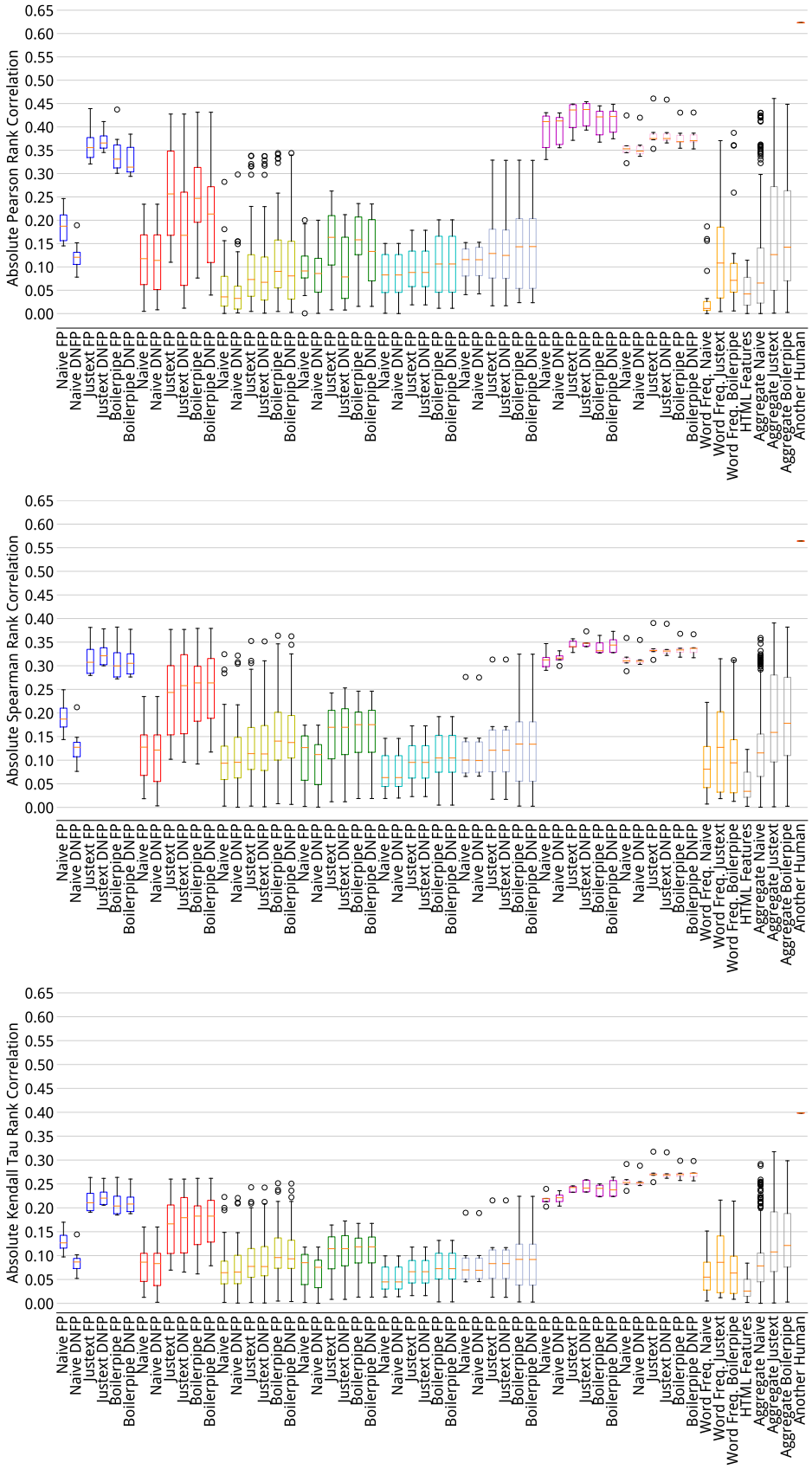


Figure 3: Wikipedia page on hyperthermia. A rectangular red box identify the Infobox on the right hand side containing entries for Specialty, ICD-9-CM, DiseasesDB and MeSH.


Absolute Pearson Rank Correlation
Absolute Spearman Rank Correlation
Absolute Kendall Tau Rank Correlation
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
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


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