**Type of Paper:**

Original Article (Appendix)

**Title:**

Appendix to: A Study of Web Page Understandability for Consumer Health Search

**Authors:**

Joao Palotti1,3, MSc; Guido Zuccon2, PhD; Dr. Allan Hanbury3,4

**Affiliations:**

1 Qatar Computing Research Institute, Doha, Qatar

2 Queensland University of Technology, Brisbane, Australia

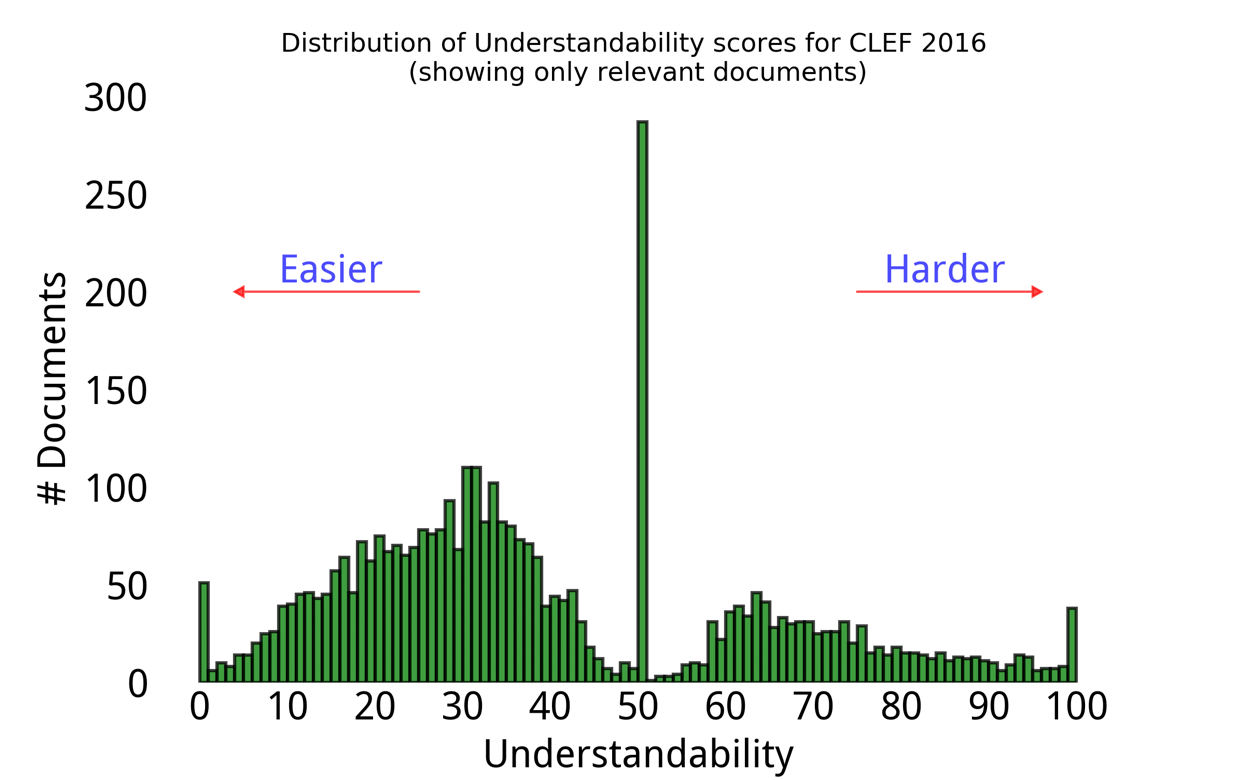
3 TU Wien, Vienna, Austria

4 Complexity Science Hub, Vienna, Austria

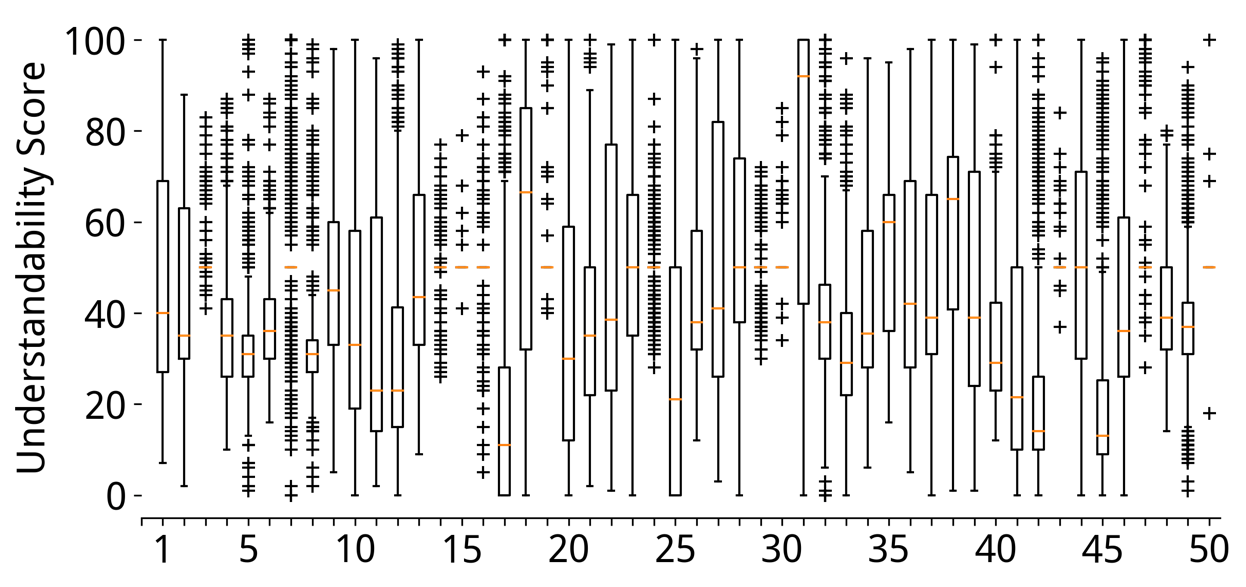
**Corresponding Author:**

Dr. Allan Hanbury   
TU Wien, Favoritenstrasse 9-11/194-04, 1040 Vienna, Austria. Phone: +43 1 58801 188310. Email: allan.hanbury@tuwien.ac.at

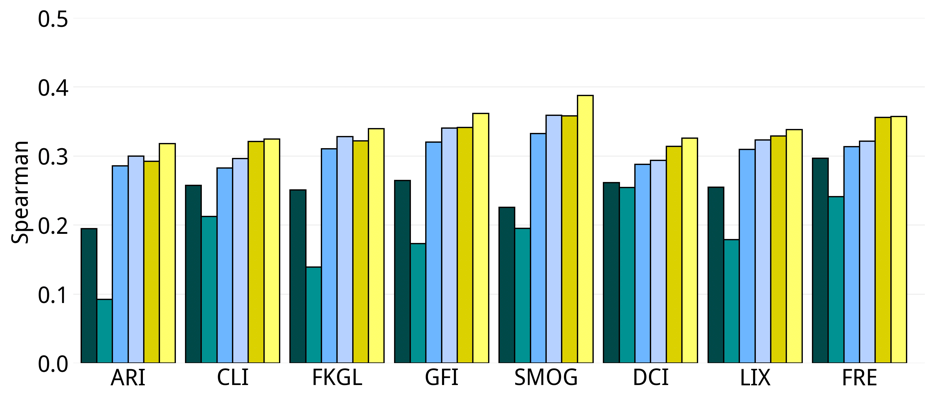
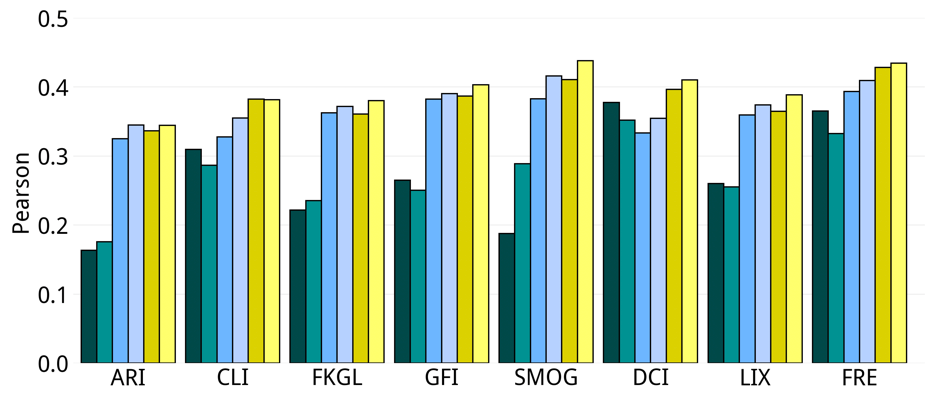
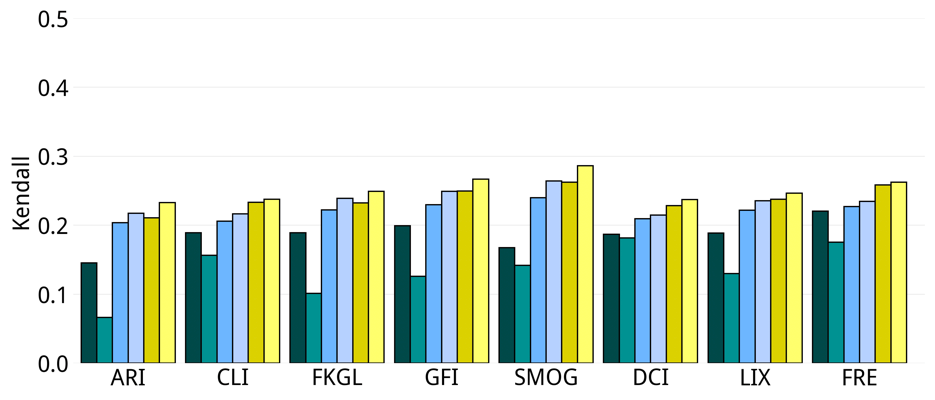
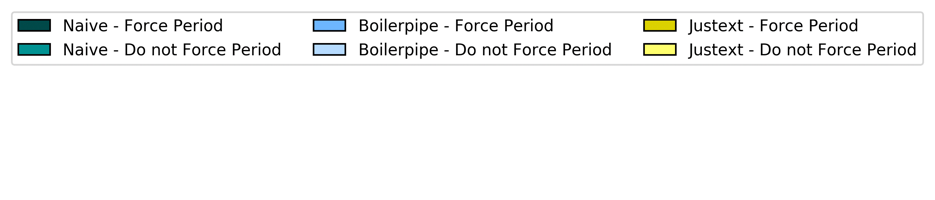
# Appendix



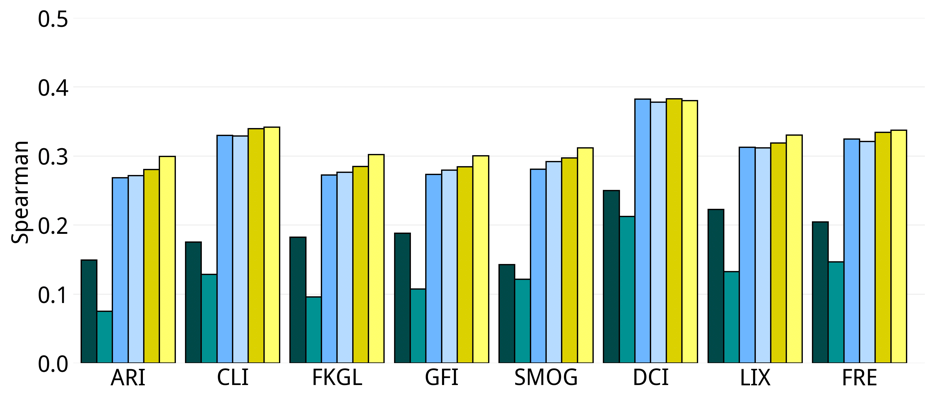
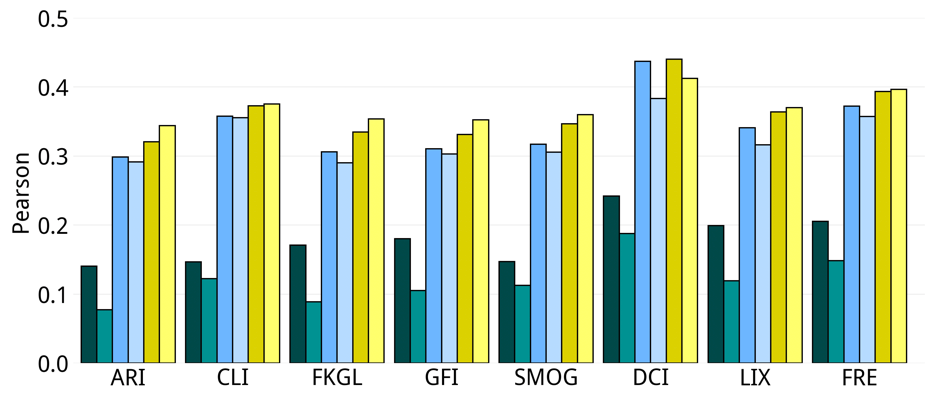
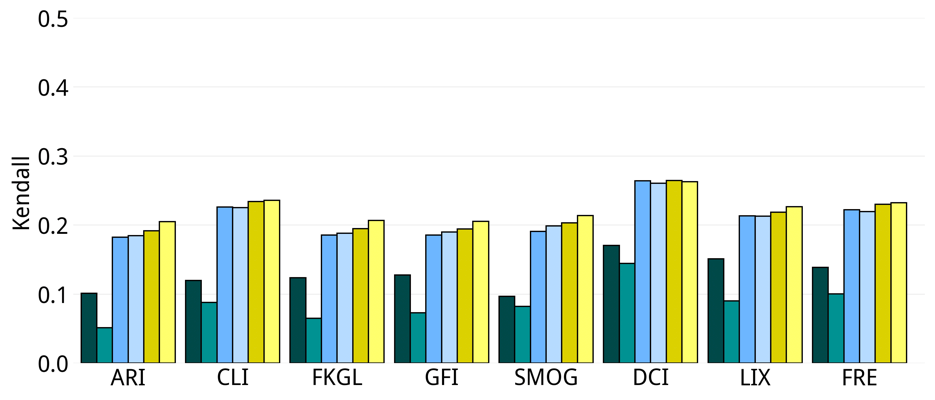
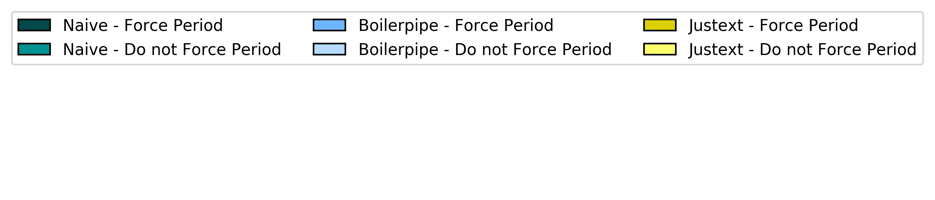
***Figure 1.*** *Understandability label distribution for CLEF eHealth 2016.*



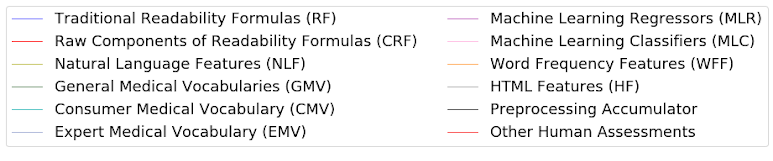
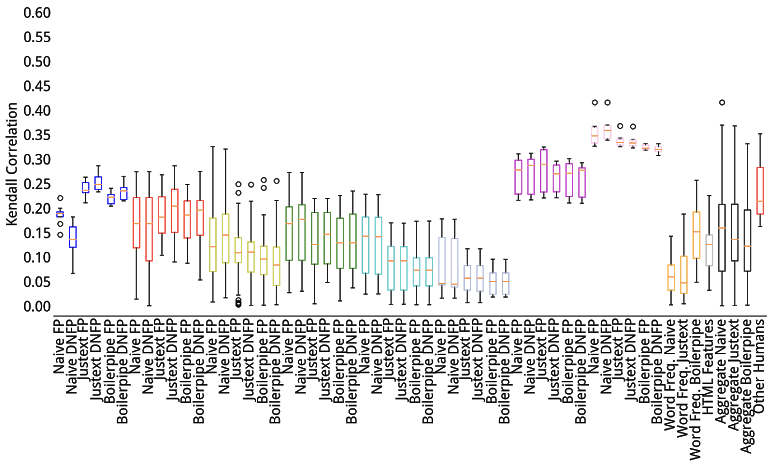
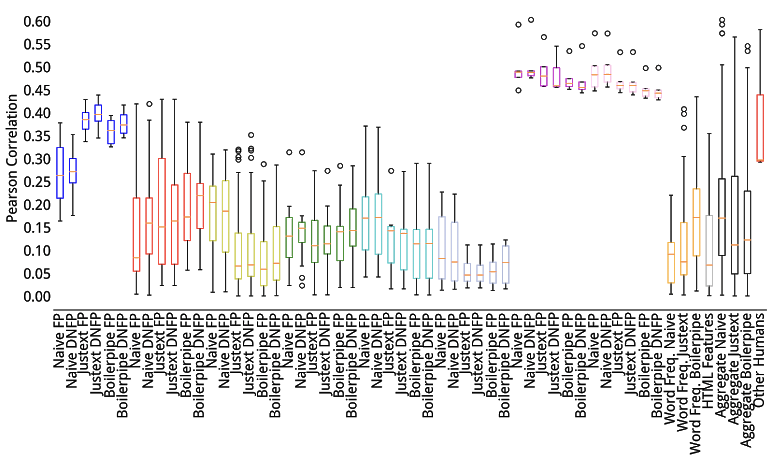
***Figure 2.*** *Understandability distribution broken per topic for CLEF eHealth 2016.*

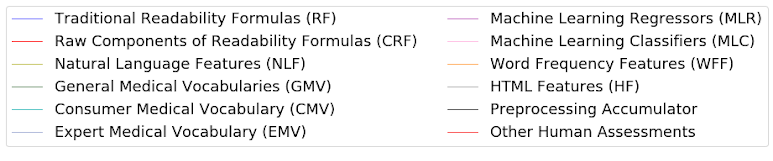
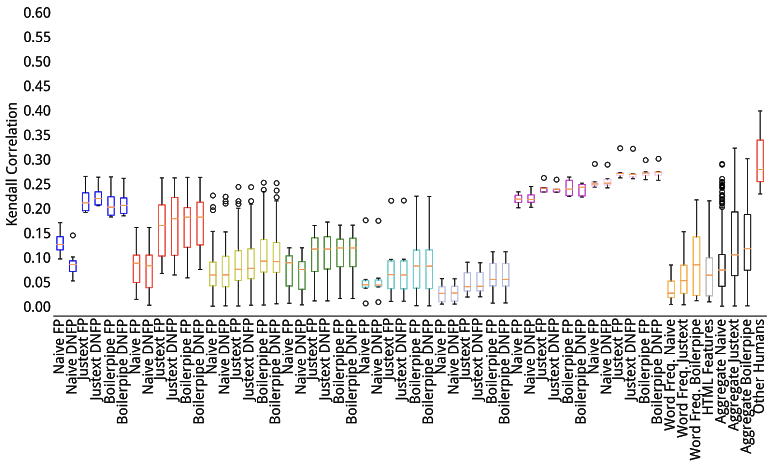
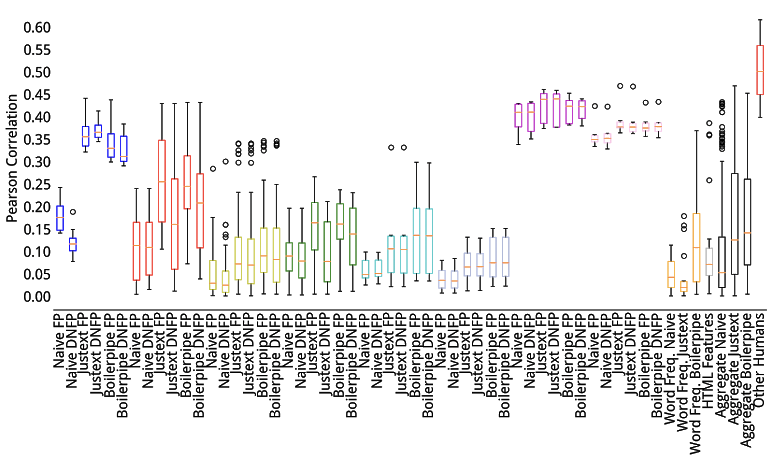
***Figure 3.*** *In Palotti et al. [1], authors investigates the influence of HTML preprocessing when readability formulas are used to estimate Webpage understandability. They found that readability formulas are heavily affected by the methods used to extract text from the HTML source, but they did not measure how correlated each method was with a human ground truth. We further extended Palotti et al.’s work to understand the influence of HTML preprocessing on automatic understandability methods and establish best practices. We show the correlation of each preprocessing combination with the ground truth assessments for CLEF 2015.*

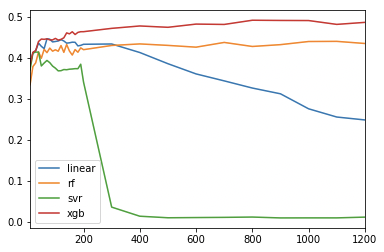
***Figure 4.*** *In Palotti et al. [1], authors investigates the influence of HTML preprocessing when readability formulas are used to estimate Webpage understandability. They found that readability formulas are heavily affected by the methods used to extract text from the HTML source, but they did not measure how correlated each method was with a human ground truth. We further extended Palotti et al.’s work to understand the influence of HTML preprocessing on automatic understandability methods and establish best practices. We show the correlation of each preprocessing combination with the ground truth assessments for CLEF 2016.*



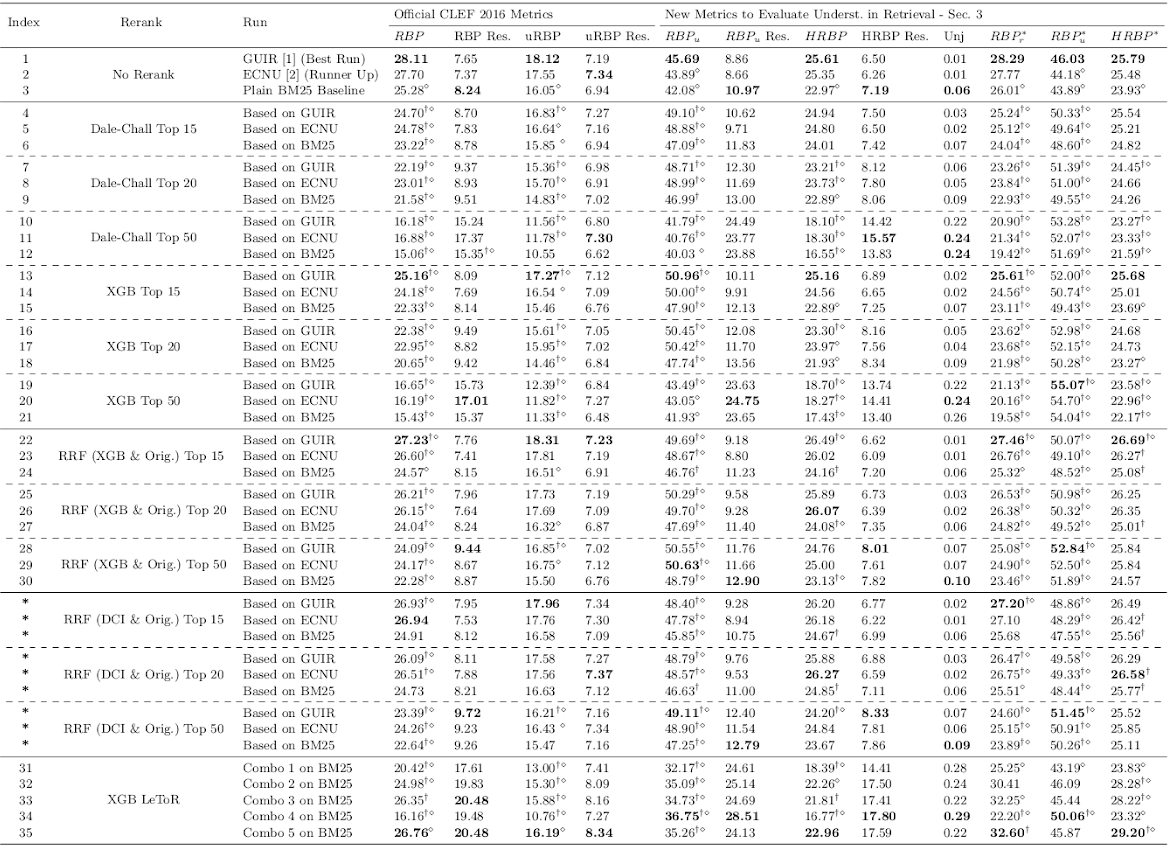
***Figure 5.*** *Correlations between understandability estimators and human assessments for CLEF 2015.*



***Figure 6.*** *Correlations between understandability estimators and human assessments for CLEF 2015.*



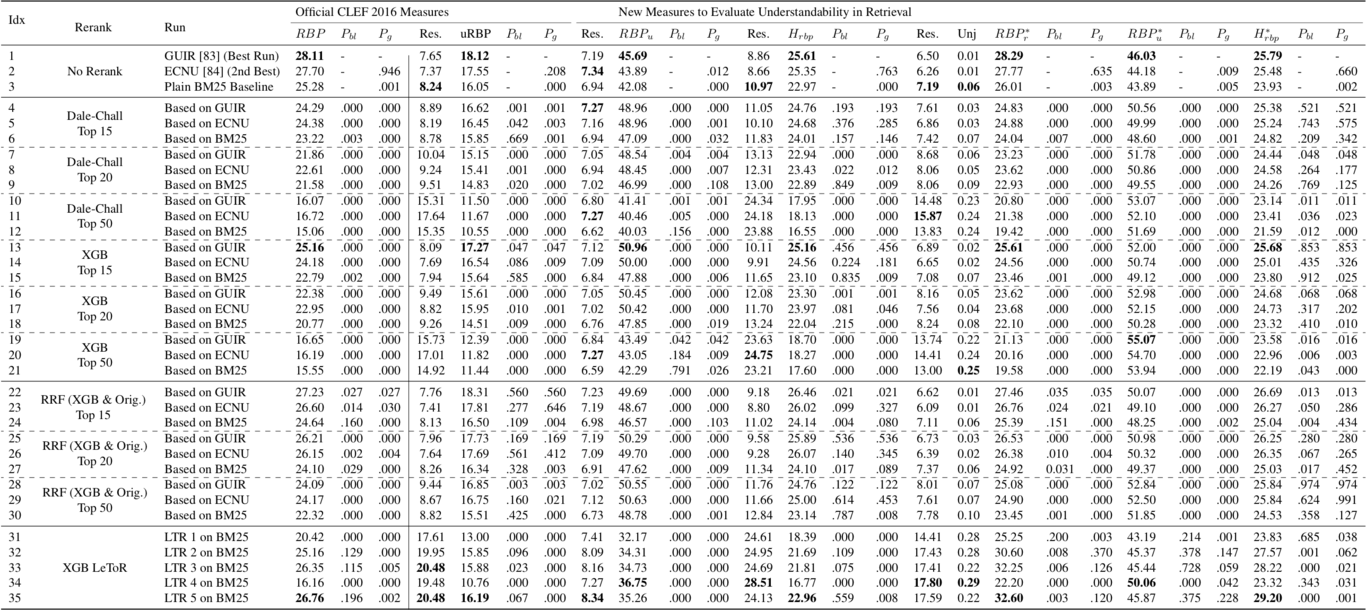
***Figure 7.*** *To study the impact of feature sets, we varied the number of features each regressor algorithm was trained with. For that, we selected features with Python’s f\_regression algorithm from Sklearn package. The eXtreme Gradient Boosting (XGB) and Random Forest Regressor were marginally influenced by the selection of features, therefore we used XGB with all features devised in our experiments.*



***Table 1.*** *We expanded retrieval results including the experiments with Reciprocal Rank Fusion (RRF) of Dale-Chall Index Runs (indices 4-12) and the original runs (indices 1-3). These results confirm the superiority of using XGB rather than DCI in terms of understandability of the results retrieved.*



**Table 2.** Results obtained by integrating understandability estimations within retrieval methods on CLEF 2015. Baseline runs are reported at table indices 1-3 (the index column is labelled Index). Re-ranking experiments are reported at indices 4-21. Fusion experiments are reported at indices 22-30. Learning to rank experiments are reported at indices 31-35. All measures were calculated up to rank n = 10. P-values are rounded to the third decimal digit and reported with respect to the ECNU run () and the BM25 baseline (). The highest result of each set of experiments is reported in bold face.



**Table 3.** Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1-3 (the index column is labelled Index). Re-ranking experiments are reported at indices 4-21. Fusion experiments are reported at indices 22-30. Learning to rank experiments are reported at indices 31-35. All measures were calculated up to rank n = 10. P-values are rounded to the third decimal digit and reported with respect to the GUIR run () and the BM25 baseline (). The highest result of each set of experiments is reported in bold face.

# References

[1] Palotti J, Zuccon G, Hanbury A. 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. New York, NY, USA: ACM; 2015. p. 1763–1766. doi:10.1145/2806416.2806613.