1.4 External Validation of Method Assumptions and Quality Score

Before using our method in a field application, we empirically test and confirm that the highest-ranking websites in the search engine indeed score highest in terms of our developed quality score components. For this task, we used around 8,500 relevant keywords and about 1.42 million ranked websites from 4 main industry sectors and 36 specific industries: details on the distribution of these keywords across industries and the number of scraped website content are reported in Table W1.4.1, columns (1) and (2). Using Wilcoxon rank sum group comparison tests, Table W1.4.2 illustrates that the poorer the search engine ranking, the lower the quality scores compared to the top 10 ranked content tends to be for all quality score components with the exception for content uniqueness (s_d).

Recall that the content uniqueness score s_d compares a given piece of content to the top 10 search results. The observed pattern in Table W1.4.2 suggests that a given result from outside of the top 10 set is more unique compared to the top 10 set than a result from the top 10 set is unique compared to other top 10 results. As noted in the manuscript, this may arise because the top 10 ranked websites consistently reflect similar topics. A contributor to the similarity among the top 10 ranked websites are the topics and keywords discussed in the content, as reflected by the higher scores for (s_a and s_k, respectively) compared to lower ranked websites. In assessing uniqueness relative to the content of the top 10 ranked websites, as these sites tend to cover similar topics and keywords compared to lower ranked sites, it is not surprising to find that the uniqueness score among the top 10 ranked websites is lower than the uniqueness score associated with lower ranked websites. Thus, we can ascertain that fine-tuning on the top 10 ranked websites' content will produce the most optimal content, and supporting the use of our quality

score as a measure of content optimality. We aggregate the search engine results into groups (i.e., top 10, search engine ranks 11-20, search engine ranks 21-99, search engine ranks 100-200) to summarize the results.

Table W1.4.1: Empirical Setup for Validating Method and Quality Score Assumptions

Industry Sector	**Industry**	**(1) Number of Keywords**	**(2) Number of Scraped Rankings & Websites**	**(3) Number of Selected Keywords**	**(4) Number of Generated Texts**
I.	Coal Mining	100	14,678	5	5,000
	Forestry	501	87,537	9	9,000
	Grazing	100	18,021	10	10,000
	Hunting	100	17,621	7	7,000
	Fishing	500	77,210	10	10,000
	Quarrying	176	18,448	8	8,000
II.	Automobile production	270	42,303	10	10,000
	Textile production	150	26,960	9	9,000
	Chemical engineering	230	43,288	8	8,000
	Aerospace production	250	57,149	10	10,000
	Energy utilities	150	29,767	10	10,000
	Breweries & bottlers	150	30,691	9	9,000
	Construction	150	21,757	7	7,000
	Ship building	70	14,058	9	9,000
	Jewelries	245	45,097	9	9,000
III.	Retailing	150	27,717	9	9,000
	Transportation	450	60,222	9	9,000
	Restaurants	230	32,539	9	9,000
	Clerical service	300	49,188	9	9,000
	Mass media	300	39,784	9	9,000
	Tourism	300	41,174	10	10,000
	Insurance	150	27,581	10	10,000
	Banking	270	44,007	9	9,000
	Healthcare	150	30,478	10	10,000
	Law	230	43,717	9	9,000
	IT service	324	50,670	19	19,000
	Art & galleries	150	27,167	9	9,000
	Cafes	230	35,382	9	9,000
	Grocery stores	500	80,814	10	10,000
	Media agencies	150	29,180	10	10,000
IV.	Government	300	50,074	9	9,000
	University	349	54,775	11	11,000
	Culture	300	57,704	9	9,000
	Libraries	100	15,715	9	9,000
	Research	100	9,938	10	10,000
	Education	278	62,518	10	10,000

Table W1.4.2: External Validation of Method Assumptions Statistics

Industry Sector	Ranks of Content Compared to Top 10	Topic (s_a)[1]	Keywords (s_k)[1]	Uniqueness (s_d)[1]	Readability similarity (s_r)[1]	Naturality similarity (s_n)[1]
I.	Top 10	.27 (.16)	.23 (.23)	.93 (.22)	.74 (.57)	.75 (.50)
	11 - 20	.23 (.17)**	.18 (.23)**	.96 (.11)**	.70 (.62)**	.58 (.58)**
	21 - 99	.18 (.15)**	.13 (.20)**	.96 (.09)**	.65 (.55)**	.58 (.58)**
	100 - 200	.15 (.15)**	.09 (.20)**	.97 (.09)**	.70 (.62)**	.67 (.58)**
II.	Top 10	.31 (.17)	.26 (.22)	.95 (.15)	.70 (.57)	.67 (.50)
	11 – 20	.25 (.16)**	.20 (.22)**	.97 (.09)**	.62 (.62)**	.58 (.50)**
	21 - 99	.22 (.17)**	.16 (21)**	.97 (.08)**	.59 (.59)**	.58 (.50)**
	100 - 200	.17 (.15)**	.11 (.21)**	.97 (.07)**	.57 (.57)**	.50 (.50)**
III.	Top 10	.35 (.22)	.31 (.30)	.94 (.17)	.72 (.60)	.75 (.50)
	11 - 20	.29 (.21)**	.25 (.29)**	.96 (.10)**	.70 (.60)**	.67 (.58)**
	21 - 99	.23 (.20)**	.17 (.26)**	.97 (.08)**	.64 (.60)**	.58 (.58)**
	100 - 200	.18 (.17)**	.10 (.22)**	.98 (.06)**	.57 (.62)**	.50 (.58)**
IV.	Top 10	.31 (.20)	.26 (.27)	.95 (.10)	.72 (.60)	.62 (.58)
	11 - 20	.27 (.20)**	.21 (.25)**	.97 (.08)**	.68 (.57)**	.58 (.58)**
	21 - 99	.22 (.19)**	.14 (.21)**	.97 (.07)**	.62 (.60)**	.57 (.58)**
	100 - 200	.16 (.16)**	.07 (.18)**	.97 (.06)**	.62 (.59)**	.42 (.67)**

[1]Reported numbers are group medians and IQRs in parentheses. Statistical significance codes come from Wilcoxon rank sum 2-group comparison tests between top 10 ranked websites and the content with specific rankings as stated in column 2; statistical significance codes (one-tailed): *0.05 level, **0.01 level; assumptions (e.g., non-normality of data) for all Wilcoxon rank-sum 2-group comparison tests are confirmed.

The results of Table W1.4.2 are consistent for smaller sets of search engine rankings (which correspond to a given page of search engine results) for the single exemplary industry sector III (Table W1.4.3) and for specific industries (e.g., tourism) (Table W1.4.4). The observed patterns are consistent across industries and the tests for significance are insensitive to industry

sector aggregations and/or data-groupings; more detailed industry-specific results are available from the authors upon request.

Table W1.4.3: External Validation of Method Assumptions Statistics for Industry Sector III

Industry Sector	Ranks of Content Compared to Top 10	Topic (s_a)[1]	Keywords (s_k)[1]	Uniqueness (s_d)[1]	Readability similarity (s_r)[1]	Naturality similarity (s_n)[1]
III.	Top 10	**.35 (.22)**	**.31 (.30)**	**.94 (.17)**	**.72 (.60)**	**.75 (.50)**
	11 - 20	.29 (.21)**	.25 (.29)**	.96 (.10)**	.70 (.60)**	.67 (.58)**
	21 - 30	.27 (.20)**	.22 (.27)**	.96 (.09)**	.66 (.60)**	.58 (.50)**
	31 - 40	.25 (.20)**	.20 (.27)**	.97 (.09)**	.66 (.62)**	.58 (.58)**
	41 - 50	.24 (.20)**	.19 (.26)**	.97 (.08)**	.63 (.59)**	.58 (.58)**
	51 - 60	.23 (.19)**	.17 (.25)**	.97 (.08)**	.64 (.57)**	.58 (.58)**
	61 - 70	.23 (.19)**	.16 (.25)**	.97 (.08)**	.63 (.62)**	.58 (.67)**
	71 - 80	.22 (.19)**	.15 (.25)**	.97 (.08)**	.66 (.62)**	.58 (.67)**
	81 - 90	.21 (.18)**	.15 (.24)**	.97 (.08)**	.64 (.64)**	.50 (.58)**
	91 - 100	.19 (.18)**	.12 (.23)**	.97 (.08)**	.62 (.62)**	.50 (.58)**
	101 - 110	.19 (.19)**	.12 (.24)**	.97 (.07)**	.61 (.62)**	.50 (.58)**
	111 - 120	.19 (.18)**	.11 (.23)**	.98 (.07)**	.62 (.59)**	.50 (.58)**
	121 - 130	.20 (.18)**	.12 (.24)**	.98 (.07)**	.57 (.55)**	.50 (.58)**
	131 - 140	.18 (.19)**	.11 (.25)**	.97 (.08)**	.60 (.55)**	.50 (.58)**
	141 - 150	.17 (.16)**	.10 (.20)**	.97 (.07)**	.62 (.59)**	.50 (.58)**
	151 - 160	.16 (.16)**	.09 (.20)**	.97 (.07)**	.53 (.57)**	.42 (.67)**
	161 - 170	.16 (.16)**	.09 (.21)**	.97 (.08)**	.57 (.62)**	.50 (.58)**
	171 - 180	.16 (.15)**	.08 (.19)**	.97 (.07)**	.55 (.56)**	.50 (.58)**
	181 - 190	.16 (.15)**	.08 (.19)**	.97 (.07)**	.61 (.62)**	.58 (.58)**
	191 - 200	.15 (.14)**	.07 (.17)**	.97 (.07)**	.49 (.55)**	.50 (.58)**

[1]Reported numbers are group medians and IQRs in parentheses. Statistical significance codes come from Wilcoxon rank sum 2-group comparison tests between top 10 ranked websites and the content with specific rankings as stated in column 2; statistical significance codes (one-tailed): *0.05 level, **0.01 level; assumptions (e.g., non-normality of data) for all Wilcoxon rank-sum 2-group comparison tests are confirmed.

Table W1.4.4: External Validation of Method Assumptions for the Tourism Sector

Industry	Ranks of Content Compared to Top 10	Topic (s_a)[1]	Keywords (s_k)[1]	Uniqueness (s_d)[1]	Readability similarity (s_r)[1]	Naturality similarity (s_n)[1]
Tourism	Top 10	**.37 (.32)**	**.36 (.40)**	**.88 (.27)**	**.63 (.60)**	**.58 (.58)**
	11 - 20	.32 (.21)**	.30 (.29)*	.95 (.18)**	.60 (.53)	.50 (.63)*
	21 - 30	.30 (.22)**	.26 (.28)**	.95 (.09)**	.48 (.54)**	.42 (.58)**
	31 - 40	.28 (.32)**	.26 (.26)**	.95 (.11)**	.42 (.48)**	.42 (.50)**
	41 - 50	.28 (.20)**	.26 (.27)**	.96 (.08)**	.49 (.53)**	.41 (.50)**
	51 - 60	.25 (.16)**	.21 (.24)**	.97 (.08)**	.53 (.51)**	.42 (.58)**
	61 - 70	.25 (.19)**	.23 (.27)**	.95 (.11)**	.43 (.52)**	.42 (.50)**
	71 - 80	.23 (.18)**	.18 (.27)**	.97 (.08)**	.43 (53)**	.33 (.50)**
	81 - 90	.20 (.21)**	.14 (.25)**	.97 (.07)**	.51 (.55)**	.33 (.50)**
	91 - 100	.23 (.21)**	.17 (.28)**	.96 (.09)**	.43 (.59)**	.33 (.48)**
	101 - 110	.19 (.19)**	.15 (.26)**	.97 (.10)**	.43 (.57)**	.33 (.58)**
	111 - 120	.16 (.17)**	.10 (.22)**	.97 (.08)**	.34 (.52)**	.25 (.35)**
	121 - 130	.16 (.18)**	.07 (.22)**	.96 (.11)**	.32 (.51)**	.33 (.46)**
	131 - 140	.14 (.15)**	.08 (.20)**	.97 (.07)**	.36 (.49)**	.42 (.46)**
	141 - 150	.17 (.19)**	.13 (.23)**	.97 (.10)**	.33 (.55)**	.29 (.56)**
	151 - 160	.12 (.10)**	.06 (.14)**	.97 (.07)**	.46 (.40)**	.25 (.25)**
	161 - 170	.10 (.11)**	.04 (.10)**	.96 (.09)**	.33 (.45)**	.38 (.50)**
	171 - 180	.16 (.12)**	.10 (.16)**	.97 (.11)**	.51 (.49)**	.33 (.48)**
	181 - 190	.12 (.19)**	.04 (.18)**	.97 (.03)**	.46 (.62)**	.25 (.50)**
	191 - 200	.13 (.11)**	.10 (.14)**	.98 (.07)**	.59 (.48)	.50 (.58)

[1]Reported numbers are group medians and IQRs in parentheses. Statistical significance codes come from Wilcoxon rank sum 2-group comparison tests between top 10 ranked websites and the content with specific rankings as stated in column 2; statistical significance codes (one-tailed): *0.05 level, **0.01 level; assumptions (e.g., non-normality of data) for all Wilcoxon rank-sum 2-group comparison tests are confirmed.

Appendix References

Baayen RH, Shafaei-Bajestan E (2019) Analyzing linguistic data: A practical introduction to statistics. Package 'languageR'. Version 1.5.0. *CRAN*. Accessed May 20, 2019, https://cran.r-project.org/web/packages/languageR/languageR.pdf

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A, (2018) "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software*. 3(30). https://doi.org/10.21105/joss.00774

Berger J, Sherman G, Ungar L (2020b) TextAnalyzer. Accessed November 11, 2020, http://textanalyzer.org

Bronnenberg BJ, Kim JB, Mela CF (2016) Zooming in on choice: How do consumers search for cameras online? *Marketing Science*. 35(5):693-712.

Danaher PJ, Mullarkey GW, Essegaier S (2006) Factors affecting website visit duration: A cross-domain analysis. *Journal of Marketing Research*. 43(2):182-194.

Edelman B, Zhenyu L (2016) Design of search engine services: Channel interdependence in search engine results. *Journal of Marketing Research*. 53(6):881-900.

Flanigan, AJ, Metzger, MJ (2007) The role of site features, user attribtues, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*. 9(2):319-342. https://doi.org/10.1177/1461444807075015

Jerath K, Ma L, Park YH (2014) Consumer click behavior at a search engine: The role of keyword popularity. *Journal of Marketing Research*. 51(4):480-486.

Kamoen N, Holleman B, Bergh H (2013) Positive, negative, and bipolar questions: The effect of question polarity on ratings of text readability. *Survey Research Methods*. 7(3):181-189.

Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*. 37(6):930-952.

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao ELT, Palma MA (2020) Basic robust statistics. Package 'robustbase'. Version 0.93-6. *CRAN*. Accessed May 20, 2020, https://cran.r-project.org/web/packages/robustbase/robustbase.pdf

Pennebaker JW, Booth RJ, Boyd RL, Francis ME (2015) Linguistic inquiry and word count: LIWC2015. Austin, TX: Pennebaker Conglomerates. Accessed November 1, 2020, www.LIWC.net.

Pitler E, Nenkova A (2008) Revisiting Readability: A unified framework for predicting text quality. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 186-195.

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. OpenAI.

Roberts C (2010) Correlations among variables in message and messenger credibility scales. *American Behavioral Scientist*. 54(1):43-56.

Rocklage MD, Rucker DD, Nordgren LF (2018) Persuasion, emotion and language: the intent to persuade transforms language via emotionality. *Psychological Science*. 29(5):749-760.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomze AN, Kaiser L, Polosukhin I (2017) Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. 1-15.