## 1.6 External Validation of Method Performance

In this section, we assess the generalizability of our proposed method across keywords and industries using our quality score measure. For this purpose, we randomly choose 338 keywords from the approximately 8,500 keywords used previously (typically 9 or 10 keywords for each of the 36 industries) and generated 338,000 pieces of content (1,000 for each single keyword), of which the method automatically selected the best scoring 338 texts (1 for each keyword). Descriptives are in Table W1.4.1, columns (3) and (4).

Table W1.6 reports the difference in medians between the machine generated content and the top 10 ranked websites for all five quality score components in bold, with Wilcoxon rank sum group comparison tests as a statistical difference indicator. We find that the raw machine outperforms the top 10 ranked content for most quality score components in all four industry sectors (Table W1.6). For example, our method outperforms the top 10 ranked websites in terms of topic consistency ($s_a$) by ~9% in industry sector I (+.09**), scoring at 34% in topic consistency. The uniqueness of the generated content ($s_d$), is the only quality indicator that shows a slightly lower value in comparison to the top 10 ranked websites (e.g., -.03** (-3%) in industry sector III), though being at a high value in absolute terms (e.g., ~87% in industry sector III).

Table W1.6: Machine vs. Top 10 Quality Score (All Industry Sectors)

| Industry Sector | Statistics | Topic ($s_a$)[1] | Keywords ($s_k$)[1] | Uniqueness ($s_d$)[1] | Readability similarity ($s_r$)[1] | Naturality similarity ($s_n$)[1] |
|---|---|---|---|---|---|---|
| I. | **Raw Machine vs. Top 10**[1] | **+.09**** | **+.14**** | **+.03*** | **+.31**** | **+.25**** |
| | Raw Machine Median[2] | .34 | .34 | .88 | .91 | .83 |
| II. | **Raw Machine vs. Top 10**[1] | **+.08**** | **+.13**** | **-.02** | **+.22**** | **+.24*** |
| | Raw Machine Median[2] | .40 | .40 | .88 | .83 | .83 |
| III. | **Raw Machine vs. Top 10**[1] | **+.10**** | **+.14**** | **-.03**** | **+.22**** | **+.07** |
| | Raw Machine Median[2] | .43 | .44 | .87 | .83 | .67 |
| IV. | **Raw Machine vs. Top 10**[1] | **+.11**** | **+.15**** | **-.04*** | **+.31**** | **+.23**** |
| | Raw Machine Median[2] | .40 | .40 | .88 | .91 | .83 |

[1] Difference in quality score component median value: raw machine generated content vs. real top 10 ranked websites; p-value from Wilcoxon rank sum 2-group comparison tests between machine generated content and top 10 ranked websites; statistical significance codes (one-tailed): *0.05 level, **0.01 level;
[2] Median quality score component value for raw machine generated content; n=338;

# 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.