Marketing Science Institute Working Paper Series 2021

Report No. 21-107

# Supporting Content Marketing with Natural Language Generation

Martin Reisenbichler, Thomas Reutterer, David Schweidel, and Daniel Dan

## *INTRODUCTION*

Efforts toward decision support and automation have a long history in marketing science. They encompass diverse areas including recommendation systems (e.g., Ansari et al. 2000), programmatic online advertising (Choi et al. 2020), dynamic website design (e.g., Hauser et al. 2009), and dynamic pricing (e.g., Natter et al. 2008). More recently, marketing automation has incorporated text analysis for applications such as examining market structure and performing competitive analysis (Lee and Bradlow, 2011, Netzer et al. 2012). Yet, to date, the literature has not examined the extent to which automation may support the development of marketing content.

Researchers have suggested that automation offers an advantage when data for training is readily available, and inputs and outputs are well defined (e.g., Bucklin et al. 1998, Brynjolfsson and Mitchell 2017). We assert that digital text (e.g., Berger et al. 2020a) coupled with natural language generation (NLG) and a customized fine-tuning process offer the potential for automation to support content marketing (Heaven 2020). With 70% of marketers investing in content marketing and nearly a quarter of marketers planning to increase their expenditures[1], applying automation to content marketing could reduce production costs and increase the rate at which new content is produced. To illustrate the potential for NLG to support content marketing, we apply it to the context of drafting content for search engine optimization (SEO).

SEO is essential to achieve high organic page rankings in search engines to increase traffic and in turn revenue. It is a multibillion-dollar business, a major activity of firm's digital marketing efforts and on par with search engine advertising (SEA) in terms of spending (e.g., Liu and Toubia 2018, Berman and Katona 2013). Research has found that organic search listings offer benefits compared to SEA (Nagpal and Petersen 2019), including lower costs and increased

[1] https://www.hubspot.com/state-of-marketing

trustworthiness among consumers (Purcell et al. 2012). Due to the competition for higher rankings in organic search results (e.g., Bar-Ilan 2006, Luh et al 2015), firms invest heavily in content creation and SEO, typically relying on SEO experts to create content which is both costly and time consuming. Given the frequent updates to search engine algorithms, content creators often rely on heuristics (Sheffield 2020), resulting in uncertainty in the outcome of SEO investments (Berman and Katona 2013).

As content is a primary ranking factor in search engines (e.g., Google 2020, Liu and Toubia 2018), research has analyzed the drivers of rankings. Early research included manual analyses (e.g., Danaher et al. 2006) and the identification of factors such as title length and page length (e.g., Zhu and Wu 2011, Salminen et al. 2019). Recent research on website content has sought to identify optimal word distributions using methods such as TF-IDF (term frequency – inverse document frequency), latent semantic analysis (LSA; e.g., Luh et al. 2016), and latent Dirichlet allocation (LDA; e.g., Liu and Toubia 2018). While LDA has been used extensively within the marketing literature, it ignores the context in which words appear. Word embeddings (e.g., Timoshenko and Hauser 2019) have emerged as a means of recognizing the context in which words appear, representing text as a multi-dimensional vector. The incorporation of word embeddings into a machine learning framework enables us to not only analyze existing website content to capture the context in which SEO keywords appear, but also to generate new content.

Recent years have seen significant advances in machine-generated content. Deep learning methods such as Long-Short Term Memory (LSTM), convolutional, recurrent, and recursive neural networks (Marchenko et al. 2020) have been used as the building blocks for text generation. Large scale pre-trained transformer language models like GPT-2 (e.g., Radford et al. 2019) and GPT-3 (Brown et al. 2020) have been introduced for NLG tasks and have proven

superior to previous methods due to their novel attention mechanism constructs (Vaswani et al. 2017).

We propose a semi-automated content generation method for developing SEO content by combining NLG with fine-tuning and a human editor to refine the content.[2] The human refinement ensures that published content does not fall into the "uncanny valley" (Mori et al. 2012) in which consumers may adversely react (e.g., Luo et al. 2019, Longoni and Cian 2020). Comparing the resulting content to human-created content, we find that the content is similar across a number of linguistic dimensions. In two field studies, we demonstrate that our semi-automated content ranks higher in search engines. In addition to its superior performance, our approach reduces the content production time and hence the associated labor costs by more than 90% compared to the traditional SEO content production.

## *A SEMI-AUTOMATED CONTENT DEVELOPMENT ALGORITHM*

In developing new content for SEO manually, a keyword (i.e., a search query) is initially selected. Next, research is conducted on textual features of the top-ranking competing websites (Sheffield 2020, Luh et al. 2016). Finally, content is created that resembles that of the top-ranked websites. We depict this typical workflow of contemporary content marketing practice in Figure 1. (Tables and Figures follow References throughout.)

In Figure 2, we illustrate our proposed method for semi-automated content generation that mimics this production process.

[2] We make use of the open source GPT-2 model. Our approach could be generalized to accommodate advances in NLG such as GPT-3 when public access is made available.

Once a keyword (e.g., "IT service management") has been specified, the top-*T* ranked search engine results are captured and the content from those links is scraped. The content of these pages ($top\_txt_{1,\ldots,T}$) is then used as an input to a machine learning model which combines the pre-trained GPT-2 345M model (Radford et al. 2018, Radford et al. 2019) with the top search engine ranked content to fine-tune GPT-2 for our SEO application. We then derive a quality score for each piece of content generated during the fine-tuning process, with the top-scoring content being provided to a human editor for revision.

Pre-trained NLG models such as GPT-2 are broadly applicable and not tailored to a particular context. For example, the GPT-2 implementation that we use in our empirical application is trained on a corpus of 8 million English text documents to predict the next term that occurs in a sequence.[3] Should one use the pre-trained GPT-2 model to generate text based on a search keyword, it would not necessarily resemble the text that typically occurs on a website. To leverage the pre-trained GPT-2 model and its semantic and syntactic language knowledge, we use the pre-trained model parameters $\Theta$ as initial values and apply the GPT-2 model to the text from the top $T$=10 search engine results[4]. That is, we fine-tune the model on the text that we wish to mimic, starting with a general linguistic structure of the English language. As the model is trained, $\Theta$ is updated. This process merges application-specific content with the pre-trained language model and is essential to ensure that the produced content incorporates the keyword, and industry- and domain-specific language structures (to reflect sub-keywords, industry specific terms, topics, etc.) that appear in the top ranked search results. As we increasingly fine-tune the

[3] We use the open source GPT-2 model available at https://github.com/minimaxir/gpt-2-simple. We provide a brief overview of the GPT-2 model in the Web Appendix (Figure W1). For simplification, we speak of "words," while GPT-2 is using BPE (Byte Pair Encoding) and tokens (i.e., learned pieces of words).

[4] We opt for a value of $T$=10 since many search engines by default display the top 10 organic search results for a given query on the first result page and thus are instantly visible to a user.

GPT-2 model, we generate content throughout the process at regular checkpoints ($Chp_1,\ldots,Chp_x$). Throughout this process, the model may eventually learn the language structure of just the search results, generating content that effectively reproduces the text of the search results on which it was trained. As we will discuss, this can adversely affect search engine rankings because the generated content would be deemed as too similar to existing content.

For our content generation method to work "on the fly," a number of features have been incorporated at each stage of our algorithm for it to work automatically and reliably for any specified keyword. We summarize the most essential high-level features in Table 1. More details are available from the authors upon request.

Each GPT-2 generated piece of content, $gen\_txt_n$, is scored based on its anticipated SEO performance, as measured through a quality score $qs_g$ that is based on five key criteria (e.g., Google 2020, Sheffield 2020): the overall topic treated in the content ($s_a$), keyword integration ($s_k$), content uniqueness ($s_d$), text naturality ($s_n$) and readability ($s_r$).[5]

$$qs_g = s_a * s_k * s_d * s_n * s_r, with\ 0 \leq qs_g \leq 1 \tag{1}$$

The content topic ($s_a$) is assessed using the mean cosine similarity between the word distributions (after stop words have been removed) of a generated piece of content (where $F_{gen}$ denotes the term frequency vector of *gen_txt*) and each of the top $T$ search results (where $F_{top}$ is the word frequency vector for *top_txt; w* denotes the vector components).

$$s_a = \frac{1}{T}\sum_{t=1}^{T}\frac{F_{gen}\cdot F_{top}}{||F_{gen}||\,||F_{top}||} = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{w=1}^{W}F_{genw}F_{topw}}{\sqrt{\sum_{w=1}^{W}F_{genw}^2}\sqrt{\sum_{w=1}^{W}F_{topw}^2}} \tag{2}$$

[5] Without loss of generalizability, the quality score could be adapted to incorporate other linguistic components.

Keyword integration ($s_k$) is measured in a similar fashion. However, rather than using all words in *gen_txt* and *top_txt* to derive the word distributions $F_{gen}$ and $F_{top}$, we use only the 10 most frequently occurring words in *gen_txt* and *top_txt*.

To measure content uniqueness ($s_d$), we calculate the number of duplicated *n*-grams of size *k*+1 in *gen_txt* compared to *n*-grams in *gen_txt* ∪ $top\_txt_{1,\ldots,T}$, where *k* is the length of the keyword, which we denote $n_{Tg}$. To obtain $n_{Tg}$, we sum over the counts of the number of duplicates of each *n*-gram type. Letting $n_{ag}$ be the number of all possible *n*-grams in *gen_txt*, we measure uniqueness as the fraction of unduplicated (i.e., unique) *n*-grams, where $n_{dg}$ is the number of duplicated *n*-grams both due to repetitions within *gen_txt* and between *gen_txt* and $top\_txt_{1,\ldots,T}$.[6]

$$s_d = \left(1 - \frac{n_{dg}}{n_{ag}}\right), with\ 0 \leq s_d \leq 1 \tag{3}$$

Text naturality ($s_n$) assesses the similarity of the generated text to the top search results on 12 linguistic measures of naturalness (Baayen and Shafaei-Bajestan 2019). For each dimension, we perform a non-parametric one-sample Wilcoxon signed rank test between the naturalness score obtained by *gen_txt* and the distribution of scores of $top\_txt_{1,\ldots,T}$. $s_n$ is the proportion of non-significant tests, with higher scores suggesting that text naturalness is consistent with the top-ranking search results. We follow a similar procedure for score readability ($s_r$), using 47 measures of readability (Benoit et al. 2020).

The set of measures $\{s_a, s_k, s_n, s_r\}$ ensures that the generated content is similar to the top-ranked search results. The content uniqueness component ($s_d$) requires our dynamic fine-tuning process, as content that is deemed too similar to the current top-ranked search results will be penalized by search engine algorithms. Intuitively, one would expect that the measures $\{s_a, s_k, s_n, s_r\}$ improve with more fine-tuning while $s_d$ diminishes as the content becomes more

[6] Additional details on measuring content uniqueness appear in the Web Appendix.

similar to the top-ranked search results. This is illustrated in Figure W2b of the Web Appendix, which shows the extent of fine-tuning that produces the content with the highest quality scores.

Based on an extensive validation study, we confirm that the top-ranked search engine websites indeed score highest on our quality score components by analyzing ~1.42 million ranked websites corresponding to ~8,500 keywords from the 4 main industry sectors and 36 specific industries (see Web Appendix A2). For our experiments, we fine-tune our model for 200 training steps for each keyword, generating 100 pieces of content at each $20^{th}$ step, which resulted in 1,000 generated texts per keyword which we rank based on their quality scores. In the Web Appendix A3, we report details on hyper-parameter value selection as well as the results of a supplemental analysis that assesses at which training steps the highest scoring content is generated and ensure that fine-tuning for 200 training steps is sufficient. In addition, we confirm that our method outperforms the real top 10 ranked content in terms of the quality score by randomly selecting 338 keywords (~9 to 10 for each industry) of the above ~8,500 keywords and generating content for these. Details are reported in Web Appendix A4.

After ordering and selecting the best content based on the quality score ($sel\_txt_{1,\ldots,N}$), our method outputs an ordered list of content for final selection and revision of a desired single piece of content by a human. To demonstrate the role that fine-tuning and human editing plays in the SEO content generation algorithm, in Table 2 we present the texts derived at three critical steps for an illustrative keyword from our empirical application and their associated quality score components. First, we provide the text from the (basic) pre-trained GPT-2 model. Second, we show the text that arises from the fine-tuning process. Lastly, we provide the text from the fine-tuning after minimal edits have been made by a human editor.

While in this particular example the pre-trained GPT-2 model yields text that scores high on uniqueness, readability and naturalness ($s_d$, $s_r$ and $s_n$), it fails to yield a word distribution that is consistent with the top-ranked search results, measured by $s_a$ and $s_k$. The word distribution is informed by the top-ranked search results during the fine-tuning process and is reflected by the increased $s_a$ and $s_k$ scores after fine-tuning. While the off-the-shelf pre-trained GPT-2 model produces readable content, as the quality scores indicate, it is not suited for SEO purposes. Rather, fine-tuning is necessary to tailor the content to a given application, be it SEO content, legal briefs or social media content. We include further examples of machine-generated content in Table W5 of the Web Appendix.

## *APPLICATION IN THE IT SERVICE INDUSTRY*

### *Experimental Setup and Robustness Checks*

To test the empirical performance of our semi-automated content generation machine, we collaborate with a mid-sized international commercial company in the IT service industry. Four experimental groups produced content for the company's website. The groups consist of (1) 19 novices (untrained marketing students who received a written stimulus that broadly stated the task), (2) 19 quasi-experts (marketing students who were trained in class and received a written instruction and a clear direction of how to do it), (3) 5 SEO experts (professionals with at least two months experience in the SEO industry who received the novices' stimulus[7]), and (4) the semi-automated SEO content writing machine with revisions made by a company employee who was instructed to keep content changes to a minimum.

[7] Survey instructions are reported in Table W6 in the Web Appendix B1.

Groups (1) – (3) produced content via an online survey. The incentive for Groups 1 and 2 (the student groups) was 15 € per produced content and credit for a marketing course. The incentive for Group 3 (SEO experts) was 40 € per produced content. All groups produced content for different, randomly assigned industry-specific keywords (e.g., "IT procurement" or "IT service maintenance"), resulting in 19 pieces of content per experimental group in total, except for the SEO expert group (due to time and cost considerations) that produced nine pieces of content for randomly selected keywords.

The company selected the pool of keywords used in our experiment based on its usual procedure (i.e., grounded on the monthly search volume, competition, fit to the firm and keyword-strategy). Content production took place within the same week and in the same geographic location so that all texts had the same state of search engine results as a basis, which we controlled for via daily crawls. To control for content length across the groups, we provided participants with a guideline on text length in terms of number of words. Based on a Kruskal Wallis group comparison, the human content writing groups did not differ in their education ($\chi^2(3)=.60$, $\eta^2=.01$, $p=.745$) or writing skills (for which the SEO experts scored a bit higher; $\chi^2(3)=5.89$, $\eta^2=.12$, $p=.053$), and the time invested conducting research on the target keyword / topic ($\chi^2(3)=.28$, $\eta^2=.00$, $p=.868$) and writing ($\chi^2(3)=3.76$, $\eta^2=.08$, $p=.153$). Descriptive statistics on the content length and changes are provided in Table 3.

*Search Engine Rankings Performance*

Each piece of content is published on its own page at day 0 on the company website, with all pages being composed of the exact same elements and structure, and each URL consisting of the keyword and a random alphanumeric suffix. To compare the search engine performance of

the semi-automated content to human-generated content, we track the top 300 search engine rankings (i.e., 30 pages of the results) for 215 days after the texts were released.

Figure 3 depicts the number of generated pieces of content per group that made it into the search engine ranking (grey bars) and into the top 10 listings (black bars). As shown in Table 4, in stark contrast to all human groups ($\chi^2(3)=576.91$, $\eta^2=.67$, $p<.000$), almost all semi-automated content ranks in the search engine with high stability over the observation period. Moreover, in contrast to the human groups ($\chi^2(3)=630.51$, $\eta^2=.73$, $p<.000$), the semi-automated approach produces more content that appears on the first page of search results (a top 10 ranking) during the observation period, after which the reduction in visibility adversely affects performance (Baye et al. 2016, Bar-Ilan 2006, Luh et al. 2016).[8]

In Table 5, we compare the quality score components from each experimental group and the top 10 ranking search results. The topic ($s_a$), keyword ($s_k$), and readability ($s_r$) scores are higher for the raw and semi-automated content compared to the remaining experimental groups and the top 10 ranked websites as well as the lowest ranked search results, while human created content scores higher in uniqueness.

*Consumer Content Perceptions*

The capability of our semi-automated procedure to generate content that produces longer-lasting search engine rankings as compared to human-written text is important from an SEO perspective. In addition to search engine rankings, the content must also appeal to the human readers. In particular, possible unnatural patterns and related issues with artificial content should

[8] Post-hoc comparisons between experimental groups are presented in Table W7 of the Web Appendix B2.

be avoided (e.g., Radford et al. 2019), as they may contribute to adverse perceptions among consumers.

To examine the differences in consumer perceptions between the semi-automated and human content, we collect data from English speaking MTurk participants in the United States (n=588).[9] We randomly assigned one piece of content to each participant, yielding an even distribution of participants across experimental conditions. Following a short introduction and instructions on reading the content, participants rated the content on scales for readability (Pitler and Nenkova 2008), understandability (Kamoen et al. 2013), credibility (Roberts 2010), attitude toward the content (Kamoen et al. 2013), content naturality, consumers' willingness to further inform themselves on the service, and willingness to buy the service. We provide further details on survey participant's instructions (Table W8), used scale items and pairwise correlations (Tables W9 and W10) in the Web Appendix B3.

Table 6 shows the perceptions of content by experimental group. Our results illustrate that the semi-automated content is generally perceived no differently than human-generated content.

To further probe the similarity in content from the different experimental conditions, we conduct analyses using LIWC (Pennebaker et al. 2015), the evaluative lexicon (Rocklage et al. 2018), and the text analyzer (Berger et al. 2020b) software packages that apply various lexica, analyses and scales to assess linguistic properties along psychological dimensions including concreteness, familiarity, and emotionality. The analysis reveals that differences between the semi-automated and human content are minor along most dimensions. We observe differences in

[9] An ideal experimental scenario would enable us to compare performance of the semi-automated and fully automated content. As this research was conducted in the field in collaboration with a corporate partner, we were unable to conduct such an experiment on their website.

the use of concrete language, with SEO experts exhibiting the highest level and novices the lowest. We also observe differences in language that evokes certainty, with the novice and quasi-expert groups using such language more than the semi-automated content and SEO experts. The full results are reported in Table W11 of the Web Appendix B3.

*Website Engagement*

Having compared performance in terms of consumer perceptions and linguistic content, we next examine the impact of using semi-automated content on firm performance in terms of consumers' engagement with the website (e.g., Bronnenberg et al 2016, Jerath et al 2014, Edelman and Zhenyu 2016). We collect website traffic data for 215 days after the experimental content was posted. During this time, the content received 146 page views from 71 unique website visits arising from organic search results. Consistent with prior research, a series of $\chi^2$ tests (e.g., Ghose et al. 2019, Azzopardi et al. 2018) reveal that semi-automated content performs better than human-generated content on the basis of the number of page views ($\chi^2(3)=141.01$, $p<.000$), page views from unique website visits ($\chi^2(3)=75.65$, $p<.000$), and the number of sessions started on the website through the SEO content (44, $\chi^2(3)=65.15$, $p<.000$). These results are consistent with the higher search engine rankings and the consumer search behavior that typically favors clicking on few, top ranked pages (Azzopardi et al. 2018). The semi-automated content also results in longer visits per visited page ($\chi^2(3)=232.15$, $p<.000$), suggesting better content performance (Danaher et al. 2006). Based on a short survey for website visitors, we also derive three proxies for expected performance: absolute buying affinity, relative buying affinity and expected sales (see footer of Table 7 for more information). These metrics suggest that the semi-automated content offers superior performance to content crafted by SEO experts, beyond

simply generating more page views due to its higher ranking, indicating a substantial positive financial impact on the company in the future. These results are presented in Table 7.[10]

*Reducing Production Costs*

We collected responses from all experiment participants on the amount of time needed for content production, as well as the company's time records, which we report in Table 8. The semi-automated approach outperforms all other experimental groups, enabling a single employee to significantly increase her annualized output. In general, we see more labor time investment in groups that are more skilled. Assuming the average annual salary (~45,000 €) and work hours (~1,567h) from publicly available labor statistics for the country in which the IT service provider is based, the cost associated with producing a single unit of content decreases from the company's current cost of 272.81 € to 15.79 € using the semi-automated procedure. Over the five-year period between 2015 and 2019, the company manually produced 439 units of content at a total cost of 119,765 €. If our semi-automated method were available, our proposed method would have resulted in a cost of 6,933 €, resulting in a savings of 112,832 € (~94%).

## *APPLICATION IN THE EDUCATION SECTOR*

We conduct a second field study in collaboration with a large, internationally recognized public business school. In this study, an employee of the organization revised 6 pieces of machine-generated content, each targeted at an industry-specific keyword (e.g., "master program in marketing") and replaced the existing human-generated content that targeted the same keyword.

[10] Table 7 reports traffic arising from organic search. Table W12 of the Web Appendix B4 shows statistics for traffic arising from direct links.

The median amount of content changed by the employee is 198 words (22.21%, median length-revised-machine=870 words, IQR=73) with a competitive investment in time (median reviser-time-investment=1.75 hours, IQR=.25, min=1.30, max=3.13).

After observing the rankings of the pre-existing human-generated content for 30 days, an employee replaced them with the semi-automated content. Similar to the IT service application, the semi-automated content outperforms human-generated content in search engine rankings. Figure 4 depicts the number of pages that made it into the ranking (grey bars) and the portion that made it into the top 10 search results (black bars), clearly demonstrating the improvement in search engine performance. Tracking rankings for 208 days, the semi-automated content outperforms the previous content based on the number of pages that are ranked ($\chi^2(1)=100.05$, $\eta^2=.49$, $p<.000$) and that appear in the top 10 results ($\chi^2(1)=101.28$, $\eta^2=.49$, $p<.000$).[11]

## *DISCUSSION*

In recent years, we have witnessed considerable advances in NLP and NLG methods. In this research, we demonstrate the potential for NLG methods to support content marketing. Coupling machine learning and NLG with a content editor, we propose a semi-automated approach for SEO content generation that is not only similar to human-generated content along a number of linguistic dimensions, but outperforms manual content creation in search engine ranking, visitor engagement and production efficiency.

Though a human editor only needs to make minor changes, her role is still essential. Though the semi-automated approach is designed to mimic the content that performs well in

[11] Additional details of this field study are presented in the Web Appendix C.

search engines, this does not take factors such as brand personality (e.g., Aaker 1997) or voice (e.g., Schmitt and Zhang 1998, Carnevale et al. 2017) into account. Future research may extend our methodology by considering two textual inputs: top ranking websites to inform the substantive content and a brand's own content to inform its tone. The latter could be accomplished by adjusting the fine-tuning of the transformer language model by focusing on texts consistent with some pre-defined target brand positioning statements. The semi-automated method could also be extended by modifying the quality score to incorporate a measure of brand fit (Robinson et al. 2015) or estimating quality component weights dynamically to account for changes to search algorithms. The current research could also be generalized to generate content for the multiple communication channels that brands employ such as blogs and social media.

As consumers become more accustomed to interacting with machine-generated content (e.g., Luo et al. 2019), it will be important to monitor how consumers react to such interactions. Additional research is needed to understand how consumers will react to machine-generated content throughout the customer journey (Puntoni et al. 2021). It is possible that consumers may react favorably to the automation of certain types of content such as blog posts but less favorably to other types of content such as social media posts. Consumer reactions may also differ based on the industry of the firm making use of machine-generated content (Longoni and Cian 2020).

As automation is applied to an increasing number of marketing tasks, there are broader implications that should be considered. The ability to reduce the costs associated with content marketing suggests that pricing can be reduced or output increased. Examining a single firm in isolation, our results demonstrate the potential to increase the return on marketing investment. Given the choice between manually created content, semi-automated content and eventually fully automated content, future research will need to investigate the impact on the competitive

equilibrium in terms of how firms will position themselves and the content they will choose to employ.

The availability of semi-automated content will also have workforce implications. As with other forms of marketing automation, the demand for labor to perform some tasks will diminish (Brynjolfsson and Mitchell 2017). While there will be less demand on generating an initial draft of content, there may be increased demand for those who can effectively edit content to compensate for automated content's shortcomings. More nuanced and differentiating style may become an increasingly important component of a brand's voice. We may also observe increased demand for those who can assess the negative consequences associated with using content that is ill-suited for its purpose (Wilson et al. 2017), such as after search engines modify the weights associated with different content characteristics.

References


Aaker JL (1997) Dimensions of brand personality. *Journal of Marketing Research*. 34(3):347-356.

Ansari A, Essegaier S, Kohli R (2000) Internet Recommendation Systems. *Journal of Marketing Research*. 37(3):363-75.

Azzopardi L, Thomas P, Craswell N (2018) Measuring the utility of search engine result pages: an information foraging based measure. *SIGIR '18: The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*. 605-614.

Bar-Ilan J, Mat-Hassan M, Levene M (2006) Methods for comparing rankings of search engine results. *Computer Networks*. 50(10):1448-1463.

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

Baye MR, De Los Santos B, Wildenbeest MR (2016) Search engine optimization: What drives organic traffic to retail sites? *Journal of Economics & Management Strategy*. 25(1):6-31.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A, Lua JW, Kuha J, Lowe W, Müller C, Young L, Soroka S, Fellows I (2020) Quantitative analysis of textual data. Package 'quanteda'. Version 2.1.1. *CRAN*. Accessed May 20, 2019, https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

Berger J, Humphreys A, Ludwig A, Moe WW, Netzer O, Schweidel DA (2020a) Uniting the tribes: Using text for marketing insight. *Journal of Marketing*. 84(1):1-25.

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

Berman R, Katona Z (2013) The role of search engine optimization in search marketing. *Marketing Science*. 32(4):644-651.

Blei DM (2012) Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM*. 55(4):77-84.

Blei DM, Lafferty JD (2009) Topic models. In: Srivastava A, Sahami M (eds). Chapman and Hall/CRC. Data mining and knowledge discovery series. (Taylor and Francis Group, LLC, New York).

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

Brynjolfsson E, Mitchell T (2017) What can machine learning do? Workforce implications. *Science*. 358(6370):1530-1534.

Bucklin R, Lehmann D, Little J (1998) From decision support to decision automation: A 2020 vision. *Marketing Letters*. 9(3):235-246.

Budzianowski P, Vulic I (2019) Hello, it's GPT-2 - How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. *arXiv*. Working paper. arXiv:1907.05774v2

Carnevale M, Luna D, Lerman D (2017) Brand linguistics: a theory-driven framework for the study of language in branding. *International Journal of Research in Marketing*. 34(2):572-591.

Choi H, Mela CF, Santiago RB, Leary A (2020) Online Display Advertising Markets: A Literature Review and Future Directions. *Information Systems Research*. 31(2):556-575.

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.

Evans MP (2007) Analysing Google rankings through search engine optimization data. *Internet Research*. 17(1):21-37.

Ghose A, Ipeirotis PG, Li B (2019) Modeling consumer footprints on search engines: An interplay with social media. *Management Science*. 65(3):1363-1385.

Google (2020). Optimize your content. In: *Search engine optimization (SEO) starter guide*. Accessed September 27, 2020, https://support.google.com/webmasters/answer/7451184?hl=en

Hauser JR, Urban GL, Liberali GG, Braun M (2009) Website morphing. *Marketing Science*. 28(2):202-223.

Heaven WD (2020) OpenAI's new language generator GPT-3 is shockingly good – and completely mindless. *MIT Technology Review*. Accessed July 20, 2020, https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/.

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.

Kahn A, Baharudin B, Hong LL, Kahn K (2010) A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*. 1(1):4-20.

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.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *Journal of Marketing Research*. 48(5):881-894.

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

Longoni C, Cian L (2020) Artificial intelligence in utilitarian vs. hedonic contexts: the 'word-of-machine' effect. *Journal of Marketing*. Forthcoming.

Luh CJ, Yang SA, Huang TLD (2016) Estimating Google's search engine ranking function from a search engine optimization perspective. *Online Information Review*. 40(2):239-255.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*. 38(6):937-947.

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

Marchenko OO, Radyvonenko OS, Ignatova TS, Titarchuk PV, Zhelezniakov DV (2020) Improving text generation through introducing coherence metrics. *Cybernetics and Systems Analysis*. 56(1):13-21.

Mori M, MacDorman KF, Kageki N (2012) The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*. 19(2):98-100.

Nagpal M, Petersen A (2019) Keyword selection strategies in search engine optimization: How relevant is relevance? *Marketing Science Institute Working Paper Series 2019*. Report No. 19-113.

Natter M, Reutterer T, Mild A, Taudes A (2007) Practice prize report—an assortmentwide decision-support system for dynamic pricing and promotion planning in DIY retailing. *Marketing Science*. 26(4):576-583.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science*. 31(3):521-543.

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.

Purcell K, Brenner J, Rainie L (2012) Search engine use 2012. *Pew Internet and American Life Project.* Accessed September 27, 2020, https://www.pewresearch.org/internet/2012/03/09/search-engine-use-2012/

Puntoni S, Reczek RW, Giesler M, Botti S (2021) Consumers and artificial intelligence: an experiential perspective. *Journal of Marketing*. Forthcoming.

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

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. OpenAI.

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

Robinson AB, Tuli KR, Kohli AK (2015) Does brand licensing increase a licensor's shareholder value? *Management Science*. 61(6):1436-1455.

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.

Salminen J, Corporan J, Marttila R, Salenius T, Jansen BJ (2019) Using machine learning to predict ranking of webpages in the gift industry: Factors for search engine optimization. *ICIST 2019: Proceedings of the 9th International Conference on Information Systems and Technologies*. 6:1-8.

Schmitt BH, Zhang S (1998) Language structure and categorization: a study of classifiers in consumer cognition, judgment, and choice. *Journal of Consumer Research*. 25(2):108-122.

Sheffield JP (2020). Search engine optimization and business communication instruction: Interviews with experts. *Business and Professional Communication Quarterly*. 83(2):153-183.

Tang J, Meng Z, Nguyen XL, Mei Q, Zhang M (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of the 31st International Conference on Machine Learning, Beijing, China*. JMLR: W&CP. 32:1-9.

Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Science*. 38(1):1-20.

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.

Wilson HJ, Daugherty PR, Morini-Bianzino N (2017) The jobs that artificial intelligence will create. *MIT Sloan Management Review*. Accessed March 23, 2020, https://sloanreview.mit.edu/article/will-ai-create-as-many-jobs-as-it-eliminates/.

Figure 1: Prototypical Manual SEO Content Production Workflow


Figure 2: Overall Method Concept and Procedure

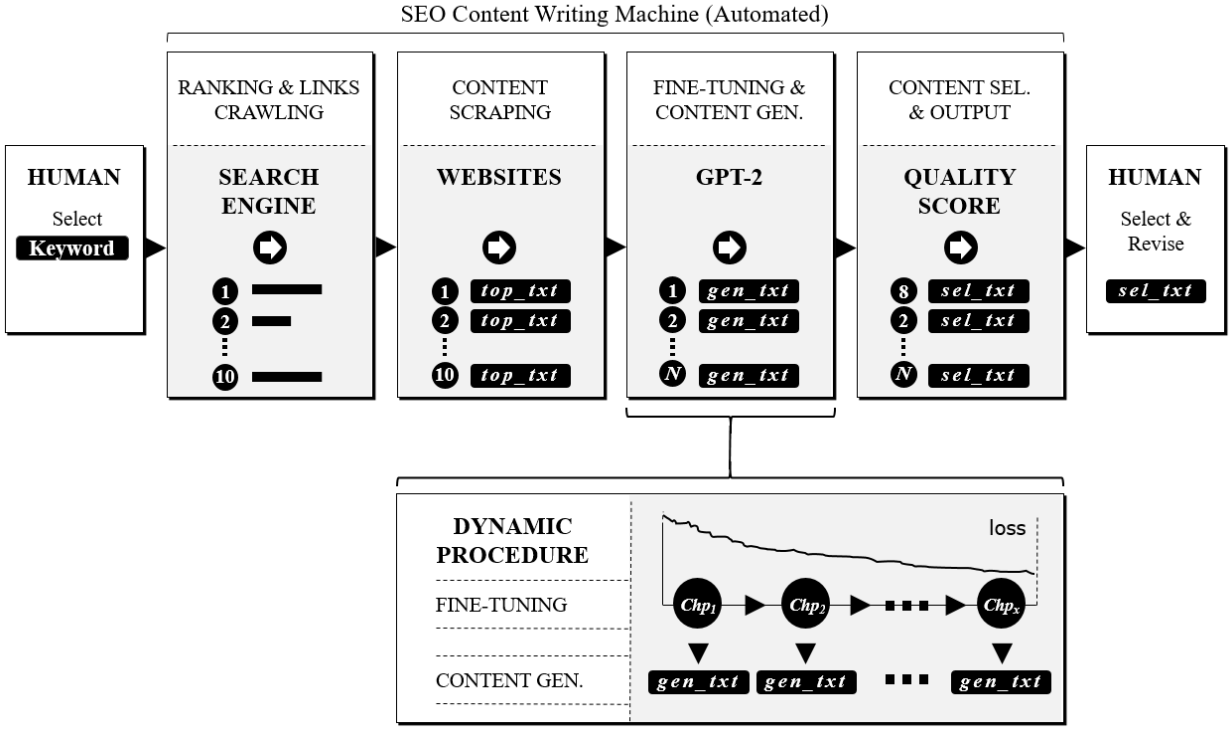


Table 1: Developed and Implemented Software Features

| Method Step | Description of Software Features |
|---|---|
| **Ranking & links crawling** | Provides live updates on job status (i.e., job start and stop times, update on current actions), simulates human browsing behavior (i.e., irregularly timed actions), identifies organic links and corrects for duplicate entries, malspecified link abbreviations and prefixes. |
| **Content scraping** | Provides live updates on job status (i.e., job start and stop times, update on current actions), simulates a local user client, incorporates enhanced security protocols, recognizes main website content and separates scraped content from code, scripts, HTML commands, server messages, etc.; it incoporates multi-redundancy for automation and standardized output for machine learning. |
| **Fine-tuning & content generation** | Dynamic GPT-2 model fine-tuning/training, implementing a fallback model in the event of corrupted text generation, saving intermittent fine-tuned states into checkpoints and outputs a strcutured list of generated content. |
| **Content selection & output** | Quality score is automatically calculated for each piece of generated content, automated redundancies (errors are detected and handled); an intuitively annotated output in the form of an ordered list of suggested generated and selected content is provided for human revision. |

Table 2: Example of Generated Content (Keyword: "IT maintenance support")

| Stage | Quality Score | Generated Content |
|---|---|---|
| **GPT-2 basic** | $s_a$ = .15<br>$s_k$ = .11<br>$s_d$ = .79<br>$s_n$ = .25<br>$s_r$ = .72<br>$qs_g$ = .002 | IT maintenance support, support staff, and support personnel, including:<br>• the Department's Technical Assistance Services (TAS) office;<br>• the Department's Information Technology Services Management Office (ITSMO);<br>• the Department's Computer-Aided Design and Engineering office; and<br>• the Department's Human Resources office.<br><br>The Department of Defense (DOD) will continue to provide support to the Federal Government as it prepares for the next major upgrade to the Department's information technology capabilities. This upgrade will include improvements to information security, the handling of sensitive and classified information, and the provision of training to employees. […] |
| **GPT-2 fine-tuned** | $s_a$ = .51<br>$s_k$ = .57<br>$s_d$ = .78<br>$s_n$ = .17<br>$s_r$ = .77<br>$qs_g$ = .030 | **It maintenance support**<br><br>We offer comprehensive services to make sure that your IT infrastructure runs efficiently and securely. Our services include system, service and maintenance support. Our goal is to be a reliable IT department with a good attitude.<br><br>Our suggestions and support services are flexible and allow you to select certain components or outsource the whole of your IT infrastructure. Our IT maintenance and support services are more in-line with your product strategy and require a higher initial investment. We can arrange for you to have a specialist IT support specialist onboard who can devote more time to your IT infrastructure maintenance needs.<br><br>**System, service and maintenance maintenance maintenance**<br><br>System maintenance is maintenance of software applications and hardware that directly affect operations such as core processes, web content management systems and so on. […] |
| **GPT-2 fine-tuned & revised** | $s_a$ = .52<br>$s_k$ = .61<br>$s_d$ = .81<br>$s_n$ = .17<br>$s_r$ = .70<br>$qs_g$ = .029 | **IT maintenance support**<br><br>We offer comprehensive services to make sure that your IT infrastructure runs efficiently and securely. Our services include system, service and maintenance support. Our goal is to be a reliable IT department, providing you with a good attitude.<br><br>Our suggestions and support services are flexible and allow you to select certain components or outsource the whole of your IT infrastructure. Our IT maintenance and support services are more in-line with your product strategy and require a lower initial investment. We can arrange for you to have an IT support specialist onboard who can devote more time to your IT infrastructure maintenance needs. As business decisions are also influenced by the level of support provided by IT maintenance and support, IT maintenance and support should be considered as a third part of business strategy.<br><br>**System, service and maintenance**<br><br>System maintenance is maintenance of software applications and hardware that directly affect operations such as core processes, web content management systems and so on. […] |

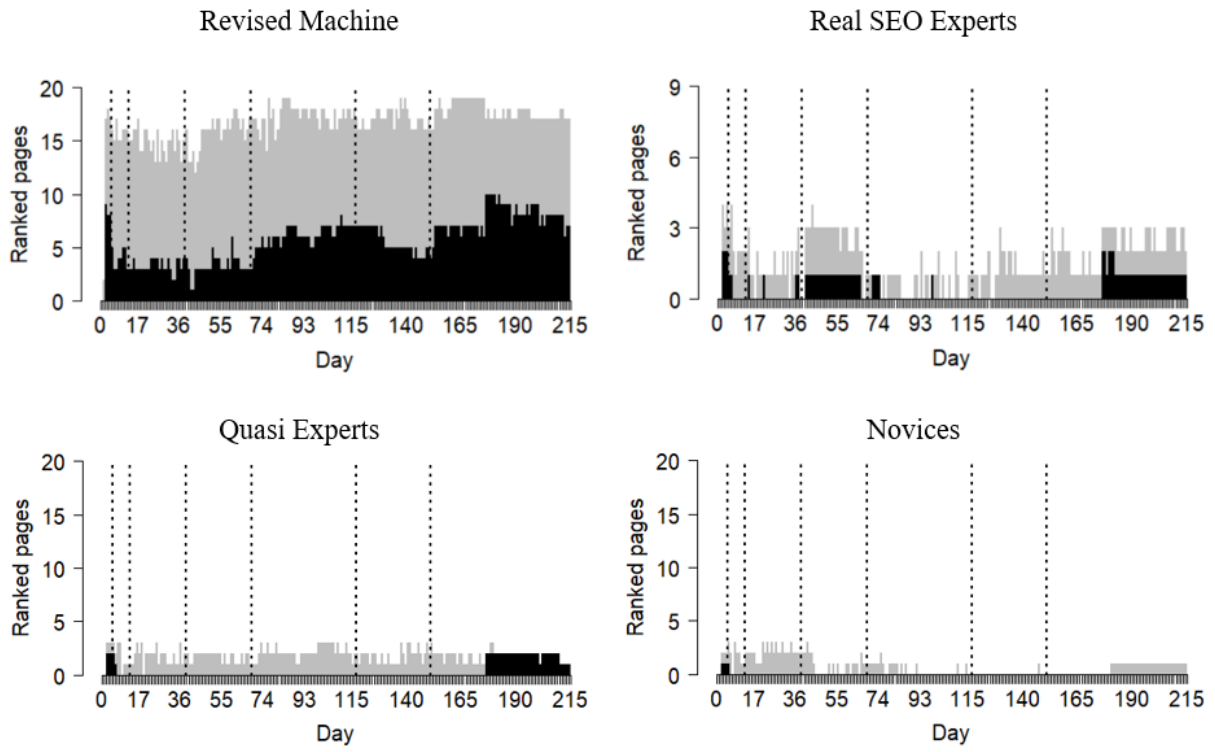
Human revision in our field experiment reported below: = human reviser corrected parts of content; _ = shifted position of content part within generated content by human reviser

Table 3: Descriptives for Content Production

| Dimension | Groups | Descriptives | | | |
|---|---|---|---|---|---|
| | | Median | (IQR) | Min | Max |
| **Produced content length (in words)** | Revised machine | 807 | (67) | 632 | 899 |
| | Real SEO Experts | 729 | (84) | 578 | 771 |
| | Quasi Experts | 694 | (69.5) | 498 | 749 |
| | Novices | 711 | (48.5) | 377 | 966 |
| **Content change (raw vs. revised)**[1] | Change in % | 9.04 | (3.77) | 3.31 | 21.45 |
| | Change in words | 74.00 | (36.50) | 27.00 | 154.00 |

[1]This includes every possible change between the raw machine and revised machine output like added words, deleted words, and words with at least one changed letter (including changed letter capitalization).

Figure 3: Number of Pages in Ranking and in the Top 10 Search Results per Day




⋮ Search engine core update release; Total number of pages in search engine rankings; Number of pages ranked in top 10 search results;

Table 4: Search Engine Rankings Performance Comparison (IT Service Sector)

| Dimension | Group | Descriptives | | | | | Kruskal-Wallis[2] | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $n_p$[1] | Median | (IQR) | Min | Max | $\chi^2$ | $\eta^2$ | *df* | *p* |
| **Pages in ranking / day** | **Revised Machine** | 19 | **17.00** | **(2.00)** | **12** | **19** | 576.91 | .67 | 3 | <.000** |
| | Real SEO Experts | 9 | 1.00 | (1.00) | 0 | 4 | | | | |
| | Quasi Experts | 19 | 2.00 | (1.00) | 0 | 3 | | | | |
| | Novices | 19 | .00 | (.00) | 0 | 3 | | | | |
| **Pages in top 10 / day** | **Revised Machine** | 19 | **6.00** | **(3.00)** | **1** | **10** | 630.51 | .73 | 3 | <.000** |
| | Real SEO Experts | 9 | .00 | (1.00) | 0 | 2 | | | | |
| | Quasi Experts | 19 | .00 | (.00) | 0 | 2 | | | | |
| | Novices | 19 | .00 | (.00) | 0 | 1 | | | | |
| **Mean rankings / day** | **Revised Machine** | 19 | **43.58** | **(33.58)** | **16.63** | **132.63** | 637.46 | .74 | 3 | <.000** |
| | Real SEO Experts | 9 | 268.11 | (33.58) | 171.22 | 301 | | | | |
| | Quasi Experts | 19 | 271.21 | (15.84) | 254.47 | 301 | | | | |
| | Novices | 19 | 301.00 | (14.42) | 256.47 | 301 | | | | |

[1]$n_p$=number of pages per experimental group. n=215 (days) for each group; [2]Statistical significance codes: *0.05 level, **0.01 level; Post hoc group comparison tests are in the appendix. Compared numbers are daily aggregate numbers. For mean rankings / day: we coded non-ranking pages with the value 301 (i.e., 1 place lower than the max observable ranking).

Table 5: Quality Score Components Group Comparisons to Top 10 Ranked Websites

| Quality Score Component | Group | Descriptives | | | | Wilcoxon rank sum[1] | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Median | (IQR) | Min | Max | W | z | *r* | *p* |
| **Topic ($s_a$)** | **Revised machine** | **.40** | **(.13)** | **.35** | **.68** | **296** | **3.36** | **.20** | **<.000**** |
| | Raw Machine | .45 | (.13) | .33 | .61 | 299 | 3.59 | .21 | <.000** |
| | Real SEO Experts | .39 | (.04) | .29 | .49 | 136 | 1.86 | .11 | .031* |
| | Quasi Experts | .36 | (.09) | .10 | .61 | 222 | 1.19 | .06 | .117 |
| | Novices | .29 | (.08) | .20 | .56 | 150 | -.89 | -.05 | .815 |
| | Worst 10 | .19 | (.07) | .21 | .54 | 344 | -5.39 | -.31 | <.000** |
| **Keywords ($s_k$)** | **Revised machine** | **.44** | **(.18)** | **.32** | **.74** | **293** | **3.27** | **.19** | **<.000**** |
| | Raw Machine | .48 | (.17) | .31 | .62 | 297 | 3.53 | .21 | <.000** |
| | Real SEO Experts | .39 | (.09) | .16 | .51 | 127 | 1.44 | .09 | .075 |
| | Quasi Experts | .38 | (.13) | .01 | .70 | 219 | 1.10 | .06 | .135 |
| | Novices | .31 | (.22) | .05 | .61 | 148 | -.95 | -.06 | .830 |
| | Worst 10 | .16 | (.11) | .14 | .61 | 335 | -4.97 | -.29 | <.000** |
| **Uniqueness ($s_d$)** | **Revised Machine** | **.90** | **(.06)** | **.81** | **1.00** | **145** | **-1.02** | **-.06** | **.153** |
| | Raw Machine | .84 | (.12) | .52 | .94 | 60 | -3.66 | -.21 | <.000** |
| | Real SEO Experts | .98 | (.04) | .08 | 1.00 | 156 | 2,82 | .17 | .997 |
| | Quasi Experts | .99 | (.03) | .87 | 1.00 | 303 | 3.59 | .20 | .999 |
| | Novices | .98 | (.07) | .79 | 1.00 | 275 | 2.77 | .16 | .997 |
| | Worst 10 | .95 | (.04) | .89 | .98 | 108 | 2.11 | .12 | .017* |
| **Readability ($s_r$)** | **Revised Machine** | **.87** | **(.17)** | **.47** | **1.00** | **340** | **4.64** | **.27** | **<.000**** |
| | Raw Machine | .96 | (.09) | .70 | 1.00 | 359 | 5.21 | .30 | <.000** |
| | Real SEO Experts | .60 | (.46) | .21 | 1.00 | 109 | .61 | .04 | .271 |
| | Quasi Experts | .53 | (.39) | .11 | 1.00 | 155 | -.75 | -.04 | .776 |
| | Novices | .57 | (.43) | .85 | .96 | 185 | .12 | .00 | .454 |
| | Worst 10 | .47 | (.10) | .26 | .68 | 312.5 | -3.84 | -.22 | <.000** |
| **Naturality ($s_n$)** | **Revised Machine** | **.67** | **(.38)** | **.17** | **1.00** | **216** | **1.02** | **.06** | **.153** |
| | Raw Machine | .92 | (.38) | .42 | 1.00 | 305.5 | 3.66 | .21 | <.000** |
| | Real SEO Experts | .83 | (.25) | .33 | 1.00 | 165 | 3.19 | .19 | <.000** |
| | Quasi Experts | .75 | (.38) | .17 | .83 | 228.5 | 1.39 | .08 | .082 |
| | Novices | .58 | (.38) | .00 | 1.00 | 199 | .53 | .03 | .299 |
| | Worst 10 | .35 | (.15) | .18 | .53 | 359 | -5.19 | -.30 | <.000** |

[1]One-tailed tests, direction for each test according to the idiosyncrasies of the field; statistical significance codes: *0.05 level, **0.01 level;

Table 6: Consumer Content Perception

| Dimension | Descriptives (Mean, SD) | | | | Kruskal Wallis[1] | | | |
|---|---|---|---|---|---|---|---|---|
| | Revised Machine | Real SEO Experts | Quasi Experts | Novices | $\chi^2$ | $\eta^2$ | *df* | *p* |
| Readability | 3.81 (1.01) | 4.06 (.82) | 3.87 (.99) | 3.87 (1.05) | 2.85 | .01 | 3 | .414 |
| Understandability | 3.34 (.95) | 3.54 (.89) | 3.51 (.96) | 3.49 (.99) | 3.45 | .01 | 3 | .327 |
| Credibility | 3.88 (.77) | 3.99 (.69) | 3.89 (.79) | 3.96 (.83) | 2.11 | .00 | 3 | .549 |
| Attitude Toward the Content | 3.05 (1.04) | 3.32 (.86) | 3.21 (.93) | 3.35 (.96) | 7.48 | .01 | 3 | .058 |
| Content Naturality | 3.23 (1.11) | 3.49 (1.04) | 3.47 (1.10) | 3.43 (1.15) | 4.79 | .01 | 3 | .187 |
| Willingness to Further Inform | 48.95 (30.49) | 55.12 (30.39) | 50.15 (29.58) | 56.94 (29.85) | 8.39 | .02 | 3 | .038* |
| Willingness to Buy | 45.92 (30.87) | 52.46 (29.15) | 48.36 (30.30) | 53.26 (30.18) | 5.53 | .01 | 3 | .137 |

[1]Statistical significance codes: *0.05 level, **0.01 level; n=551;

Table 7: Consumer Behavior (Organic Search Source Only)

| Dimension | Descriptives (∑) | | | | One-Sample Chi-Squared[1] | | |
|---|---|---|---|---|---|---|---|
| | Revised Machine | Real SEO Experts | Quasi Experts | Novices | $\chi^2$ | *df* | *p* |
| No. of Pages with Pageviews | 14 | 2 | 4 | 7 | 12.26 | 3 | .006** |
| No. of Pages with Pageview in % | 73.68 | 22.22 | 21.05 | 36.84 | 47.08 | 3 | <.000** |
| Pageviews | 98 | 10 | 14 | 24 | 141.01 | 3 | <.000** |
| Unique Pageviews | 49 | 3 | 7 | 12 | 75.65 | 3 | <.000** |
| Entrances | 44 | 3 | 7 | 11 | 65.15 | 3 | <.000** |
| Exit Rate (means) | .40 | .13 | .46 | .41 | - | - | - |
| Bounce Rate | .00 | .00 | .00 | .00 | - | - | - |
| Avg. Usage Duration (Abs., sums) | 3532 | 77 | 784 | 442 | 7561.00 | 3 | <.000** |
| Avg. Usage Duration (Rel.)[2] | 252 | 39 | 196 | 63 | 232.15 | 3 | <.000** |
| Returning Visitors (Abs.) | 49 | 7 | 7 | 12 | 65.96 | 3 | <.000** |
| Returning Visitors (Rel.)[2] | 3.50 | 3.50 | 1.75 | 1.71 | - | - | - |
| Buying Affinity (Abs.)[3] | 2340 | 140 | 341 | 506 | 3729.80 | 3 | <.000** |
| Buying Affinity (Rel.)[2,4] | 167 | 70 | 85 | 72 | 64.93 | 3 | <.000** |
| Exp. Sales (for U.P.*100)[5] | 98 | 6 | 14 | 24 | 151.30 | 3 | <.000** |

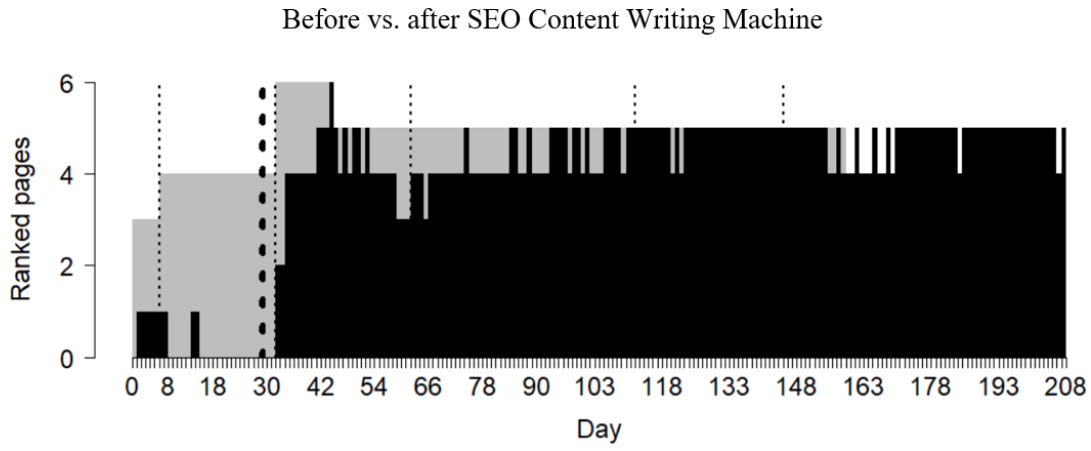
[1]Statistical significance codes: *0.05 level, **0.01 level;
[2](Rel.) = the absolute value (Abs.) divided by No_of_Pages_with_Pageviews
[3]Buying Affinity (Abs.) = Unique_Pageviews*Willingness_to_Buy (survey measured);
[4]Buying Affinity (Rel.) = Buying_Affnity (Abs.)/No._of_Pages_with_Pageviews;
[5]Exp. Sales (for U.P.*100) = (Unique_Pageviews/100*Expected_Sales_Rate)*100, where the expected sales rate is 2% (obtained from past company reports);

Table 8: Labor Time, Cost and Savings for Content Production

| Category | Factor | Company (Real) | Revised Machine | Real SEO Experts | Quasi Experts | Novices |
|---|---|---|---|---|---|---|
| Human labor time for content production | Median (hours) | 9.50 | **.55** | 4.10 | 2.58 | 3.60 |
| | IQR (hours) | 3.69 | **.23** | 1.80 | 2.58 | 3.55 |
| | Min (hours) | 4.50 | **.28** | 1.00 | .90 | .80 |
| | Max (hours) | 21.50 | **1.20** | 7.00 | 28.80 | 12.00 |
| Production output & cost per year[1] | Produced content units | 164.95 | **2,849.09** | 382.20 | 607.36 | 435.28 |
| | Production level (%) | 100.00 | **1,627.27** | 131.71 | 268.22 | 163.89 |
| | Cost per content unit (€) | 272.81 | **15.79** | 117.74 | 74.09 | 103.38 |
| | Cost for 164.95 units (€) | 45,000 | **2,605** | 19,421 | 12,221 | 17,052 |
| | Cost for 2,849.09 units (€) | 777,272 | **45,000** | 335,454 | 211,090 | 294,545 |
| Possible real financial impact (2015 to 2019)[1] | Produced content units | 439 | **439** | 439 | 439 | 439 |
| | Cost (€) | 119,765 | **6,933** | 51,688 | 32,525 | 45,384 |
| | Possible savings (€) | | **112,832** | 68,077 | 87,239 | 74,380 |

[1]Assumed employees' labor time & salary: 39 hours per week, 1,567 hours per year; 45,000 € per year;

Figure 4: Number of Pages in Ranking and in the Top 10 per Day (Education Sector)




⋮ Day at which human made content was replaced by revised machine content; ⋮ Search engine core update; Total number of pages in ranking; Number of pages ranked in top 10;

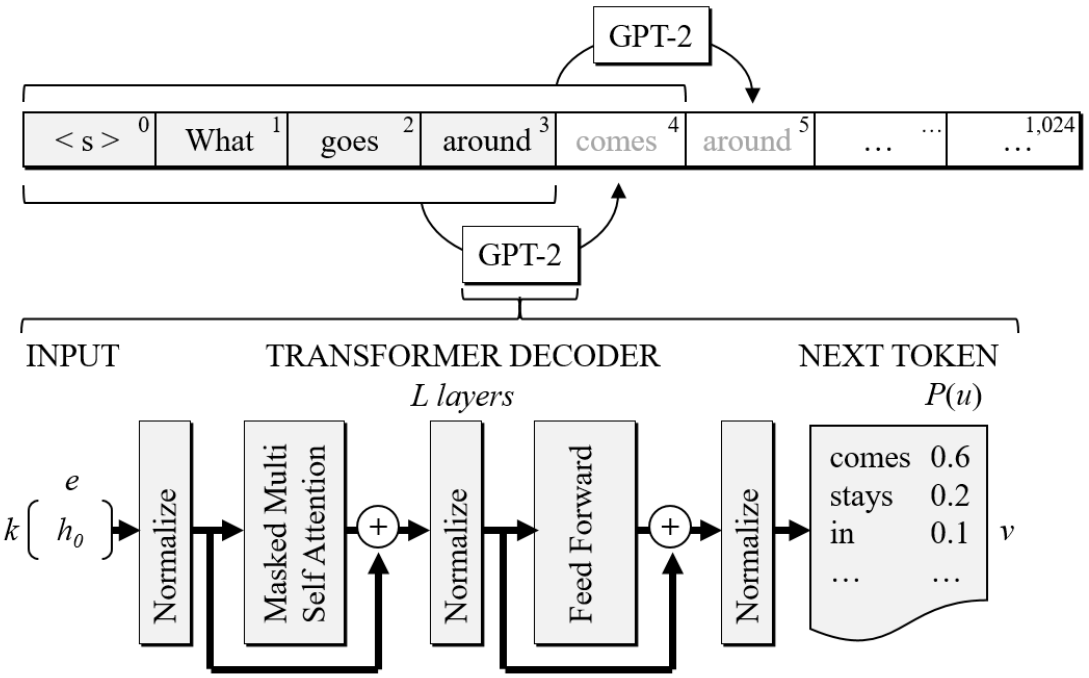
Web Appendix

## Web Appendix A: Technical Modeling and Validation Notes

### A1: GPT-2 Model Description

To get a sense of transformer-based NLG models, we briefly illustrate the mechanics of the popular GPT-2 model. Given a sequence of tokens with context window size $k$, $U=(u_{-k},\ldots,u_{-1})$, the objective of the autoregressive model GPT-2 is to accurately "predict" the next likely word (Figure W1) by sampling from a probability distribution over its entire learned vocabulary (consisting of 50,257 tokens) conditional on the given word sequence and on a pre-trained neural network with parameters $\Theta$. Model pre-training tries to maximize the likelihood in equation (W1) for an unsupervised corpus of words ($\mathcal{U}$) (Radford et al. 2018).

Figure W1: The GPT-2 Model[12]




[12] Visualization derived from Radford et al. 2018, and adapted to depict the updated GPT-2 architecture.

$$L_1(\mathcal{U}) = \sum_i logP(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta) \quad \text{(W1)}$$

$$h_0 = UW_e + W_p \quad \text{(W2)}$$

$$h_l = transformer\_block_{(h_{l-1})} \forall i \in [1, L] \quad \text{(W3)}$$

$$P(u) = softmax(h_L W_e^T) \quad \text{(W4)}$$

In essence, GPT-2 relies on word and given context meaning information to generate its output distribution over its vocabulary. More specifically, the data input consists of a matrix $h_0$ (W2), where the given word sequence $U$, word meaning information in terms of word embeddings $W_e$, and sequential word position information in terms of position embeddings $W_p$ are combined. As illustrated in Figure W1, information from $h_0$ is extracted, transformed, added and normalized multiple times (to ease processing), and projected into the embedding space $e$ by $L$ layers of decoder transformer blocks (W3). This information includes the extent of putting attention on given sequence words using multi-headed self attention (Vaswani et al. 2017), and high dimensional hidden language states on how to shift the focus in the embedding space $e$ to recreate natural word sequences from position wise feed forward neural networks. The output of the final block $h_L$ projects all this information into the embedding space and is multiplied with GPT-2's original (unconditional) transposed word embeddings matrix $W_e^T$ to assess which word from the GPT-2 vocabulary best matches the information contained in $h_L$ (W4). The multiplication of $h_L$ and $W_e^T$ can be thought of as a similarity or matching between the embedding space distribution of the output of $h_L$ (containing meaning, position, attention, and hidden language states information) and the unconditional embedding space distribution of each respective vocabulary word. More similarity of a vocabulary word in terms of its embedding to $h_L$ will result in a higher probability in GPT-2's output distribution. GPT-2 then obtains a

probability distribution over its vocabulary $P(u)$ (4) and can sample the upcoming word in the sequence from the most likely words in $P(u)$.

Using the above procedure, GPT-2 learned and stored word probabilities for given word sequences represented in its 345 million parameters (including its vocabulary, embeddings, attention weight matrices, and $\Theta$) using 8 million English text documents with a broad topical variety. Neural network parameters $\Theta$ were first initialized and then trained on batches of 512 sequences. The loss function refers to the language modeling cross entropy loss, where 1 is assigned to the word that appears next ($u_i$) in the training sequence (e.g., "comes" in Figure W1), and 0 to all other words in GPT-2's vocabulary, and compare the log transformed GPT-2 softmaxed output probability value $P_u$ for that respective word to appear next. A loss of 0 means the GPT-2 prediction was in perfect accordance with the actual next word (i.e., 1), the higher the deviation of the GPT-2 prediction (e.g., 1-0.6 = 0.4 for "comes" in Figure W1) to the actual word, the more the loss value increases. During training, GPT-2 performs this process on batches and minibatches of several sequences before updating $\Theta$.

**A2: 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 all 4 main industry sectors and 36 specific industries (details are reported in Table W1). Using Wilcoxon rank sum group comparison tests, Table W2 illustrates that the worse the search engine ranking, the higher the difference to the top 10 ranked content

tends to be for all quality score components, except for content uniqueness ($s_d$). Note that the latter is lower for the top 10 ranked websites since these consistently reflect similar topics as opposed to lower ranked websites. Thus, we can ascertain that fine-tuning on the top 10 ranked websites' content will produce the most optimal content, and approve our quality score as a measure of content optimality.

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

| Industry Sector | Industry | Number of Keywords | Number of Scraped Rankings & Websites | Number of Selected Keywords | 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 W2: 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 ($s_r$)[1] | Naturality ($s_n$)[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;

### A3: Validation of Method Fine-Tuning Process

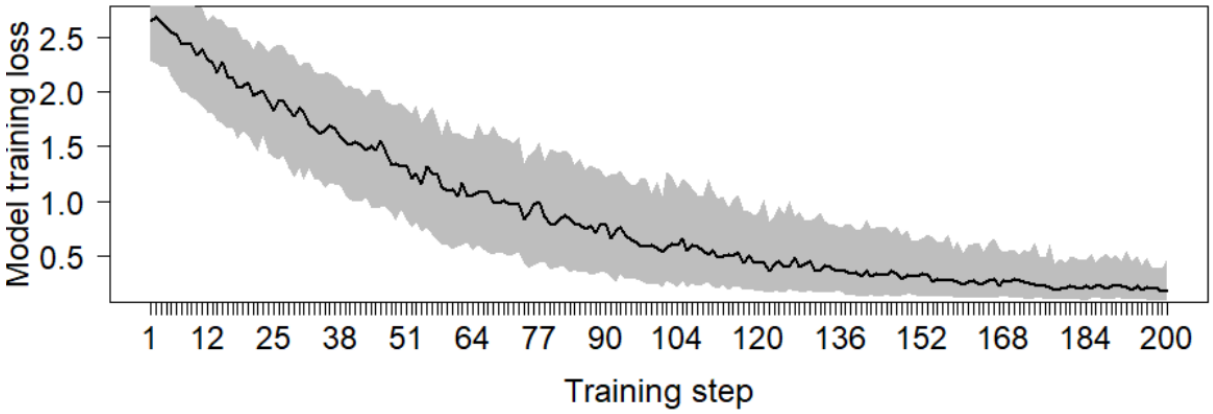
For our experiments, we fine-tune our model for 200 training steps for each keyword, generating 100 pieces of content at each 20$^{th}$ training step which resulted in 1,000 generated texts per focal keyword, of which our method then selected the best scoring pieces of content using the above proposed quality score. Similar to the approach taken by Liu and Toubia (2018), based on prior literature and on several test runs, we set the hyper-parameters top_$k$ = 40, and temperature

= 0.7 (which effectively regulates the randomness in GPT-2's sampling process and output content). Next, we show that fine-tuning for 200 training steps is sufficient and examine factors that determine at which training step our proposed method selects the most optimal content.

Figure W2a illustrates the increasing capability of the model to accurately predict words given prior word sequences over the 200 model training steps using the median (black line) and IQR (grey area) of the Loss measure (Radford et al. 2018) over all keyword trainings for the experiments presented in the main text. While model fit is consistently improving, Figure W2b shows that based on the quality score, the most optimal content commonly comes from mid training steps (between 60 and 160), while an extremely low and an extremely high amount of training steps entail a lower probability to produce the most optimal content. Thus, using 200 training steps for fine-tuning is sufficient.

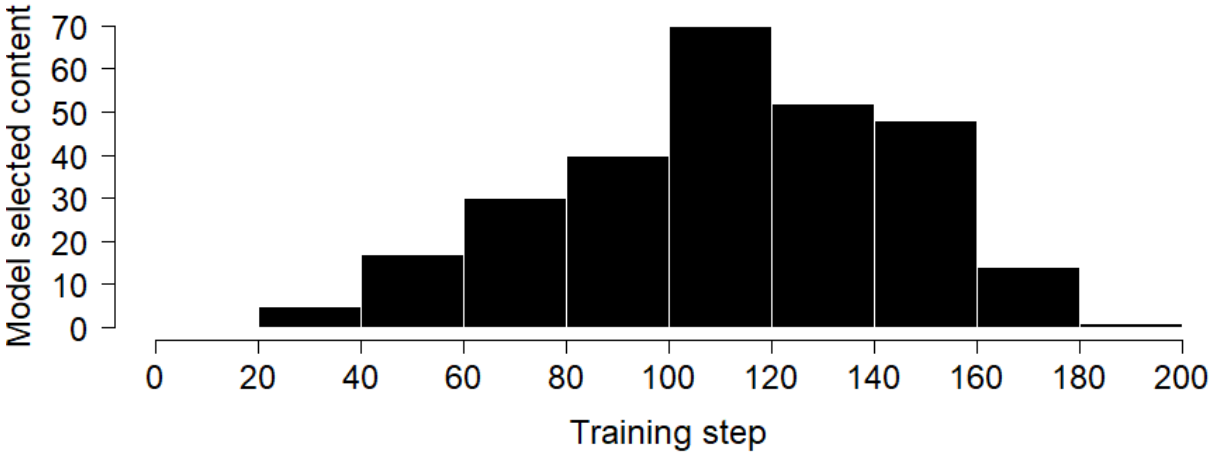
Using a robust regression (robust against violations of classic data assumptions of regression, see Maechler et al. 2020) for the training steps generating the "best" texts with highest overall quality scores on the quality score components, we observe in Table W3 that the content uniqueness of the top 10 ranked websites is the most important determinant for at which training step the most optimal content is generated. That means, if the top 10 ranked websites are more unique, i.e., many top ranked websites that we fine-tune on, do not extensively copy phrases from each other, our method selects content from a later training phase ($B$=117.88, $t$=4.47, p<.000) since the risk to pick up the repetitive language patterns is lower. Interestingly, the regression model explains just ~11% of the variance in the data (Adj.$R^2$=.1084), meaning that the probabilistic fine-tuning and text generation processes of the GPT-2 model has a considerable impact on at which training step the most optimal content is generated.

Figure W2a: Median & IQR Model Fit (Over all Trainings for Keywords)




— Median of model training loss for all model trainings; ■ IQR of model training loss for all model trainings

Figure W2b: Quantity of Model Selected Most Optimal Content vs. Training Step




■ Quantity of mean top model selected content (for each keyword, we extracted the top scoring generated content and calculated the mean training step from which these came from)

Table W3: Quality Score Factors Determining the Training Step for Optimal Content Selection

| **Robust Regression**[1] | | | | |
|---|---|---|---|---|
| Independent Variables | *B* | Std. Error | *t* | *p* |
| Intercept | 54.79 | 26.56 | 2.06 | .039* |
| Topic ($s_a$) + Keywords ($s_k$) of Top 10 | 4.82 | 6.92 | 0.69 | .486 |
| Uniqueness ($s_d$) of Top 10 | 117.88 | 26.37 | 4.47 | <.000** |
| Readability ($s_r$) + Naturality ($s_n$) of Top 10 | -31.07 | 9.59 | -3.24 | .001** |
| Adjusted $R^2$ of regression model: .1084 | | | | |

[1]Dependent variable: Model training step at which most optimal content was selected based on quality score; statistical significance codes: *0.05 level, **0.01 level; because of strong pairwise correlations, we combined $s_a$ and $s_k$ as well as $s_r$ and $s_n$ into one variable by adding them up.

## A4: 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.

Table W4 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 W4). 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 W4: Machine vs. Top 10 Quality Score (All Industry Sectors)

| Industry Sector | Statistics | Topic ($s_a$) | Keywords ($s_k$) | Uniqueness ($s_d$) | Readability ($s_r$) | Naturality ($s_n$) |
|---|---|---|---|---|---|---|
| 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;

To demonstrate the versatility of our approach, Table W5 presents four abbreviated examples of machine-generated content for keywords from varied industries ("buy wedding ring near me," "best e bike insurance," "aerospace component manufacturer," and "state library bookshop") that have not yet been revised by a human.

Table W5: Examples of Generated Content (Raw, Unrevised Machine Output[1])

**Buy wedding ring near me**

Search for your perfect wedding band online today.

**Find Your [brand name] Wedding Band**

Here at [brand name], we are constantly adding to our extensive selection of wedding bands to help you capture the perfect moment. Below, you can find a short summary of some of the more popular styles of wedding bands.

**The Joy of Enchantment**

This style of wedding band is all about personalization. Traditionally, diamond engagement rings come with a beautiful diamond and sigil engraved setting. […]

**Best e bike insurance**

**Best value electric bike insurance**

We offer a range of products to suit every budget and circumstance.
Our products can provide additional protection for theft, damage, accident, negligence, fraud and more.
Our customer service reps are here to tailor a program to your needs.

**Bicycle Roadside Assistance**

If you find yourself with a disabled bike well come pick you up. […]

**Aerospace component manufacturer**

With a revenue of $8 billion, [brand name] is the third-largest aerospace company in the world. It provides products for both global aerospace and defense, manufacturing nacelle systems, avionics systems, control surfaces, electrical systems, mechanical and aerospace systems, and flight simulator systems.

The company also provides aftermarket services, like draping and electrical systems, in addition to product development. [brand name] specializes in fabricating complex components and sub-assemblies for military and commercial uses. Examples of products it creates are for use in communication systems, missiles and aviation defense, and in the aerospace industry generally. […]

**State library bookshop**

- Readings
- Accessibility
- Contact Us
- Our Story

[brand name] was founded in 2006, and our mission is to enrich our community by offering a wide range of unique and innovative titles and gifts. Our shop is the first of its kind in Australia to be dedicated to learning and sharing our Australian heritage.

At the bookstore, you'll find a wide range of titles to choose from, including a wide range of Australian titles. From the award-winning noir mystery to the whimsical romance, we've got something for everyone.
[…]

[1]Generated real brand and product names are replaced by the tag [brand name], and headlines are printed in bold to ease reading.

## Appendix B: Supplemental Information for the IT Service Sector Application

### B1: Participants' Survey Instructions

Table W6 reports the stimulus for the content writing groups in our IT service industry experiment.

Table W6: Participants' Survey Instructions for Content Writing

| Content Writing Group | Instructions[1] |
|---|---|
| Novices | [Short introduction stating the goal of this study, strict anonymization, the incentive and a contact person for questions.]<br><br>Imagine you are a marketing employee in an IT service company.<br><br>Your manager approaches you to **write a Google search engine optimized (SEO) text for a single site on the website of your IT company**, that elaborates on a specific service. You should **write the text in a way that it ranks well in Google**. That means, it should preferably appear on page 1 in the Google search results.<br><br>• The text should be written for the keyword / search term / topic: "**IT maintenance**" (i.e., for IT maintenance provided as a service by your company to firms).<br>• It should be written for ranking well in **Google in [Country blinded], set to English language** (please use the link below).<br>• For ranked **example sites see**: https://www.google.com/search?num=100&hl=en&q=it+maintenance<br>• It should be **original, unique content**, invented by you (i.e., NO copies).<br>• It should be written **in English language**.<br>• It should contain **around 700 to 800 words** (ca. 2 A4 pages).<br><br>**Your text:** (Please write your text in the following text field.) |

Quasi Experts

[Short introduction stating the goal of this study, strict anonymization, the incentive and a contact person for questions.]

Imagine you are a marketing employee in an IT service company.

Your manager approaches you to **write a Google search engine optimized (SEO) text for a single site on the website of your IT company**, that elaborates on a specific service. You should **write the text in a way that it ranks well in Google**. That means, it should preferably appear on page 1 in the Google search results.

- The text should be written for the keyword / search term / topic: "**IT maintenance**" (i.e., for IT maintenance provided as a service by your company to firms).
- It should be written for ranking well in **Google in [Country blinded], set to English language** (please use the link below).
- For ranked **example sites see**: https://www.google.com/search?num=100&hl=en&q=it+maintenance
- It should be **original, unique content**, invented by you (i.e., NO copies).
- It should be written **in English language**.
- It should contain **around 700 to 800 words** (ca. 2 A4 pages).

**How to write a SEO optimized text?**

- **Integrate the main keyword** ("IT maintenance") **or parts of it most often compared to the other words in your text**.
- **Write about subtopics / content** that you can find on the top ranked websites for the main keyword.
- **Align the word distribution of your text** with the word distribution of the top ranked websites for the main keyword (i.e., **put the right words with the right frequencies into your text**).
- For the **word distribution analyses use**: https://wordcounter.net/ (Please be aware that the tool doesn't count common stopwords like "it".)
- **Prevent keyword stuffing** (i.e., don't integrate keywords overly often and in an unnatural way into your text).
- Try to give your text a **good readability and structure**.

**Your text:** (Please write your text in the following text field.)

---

Real SEO Experts

[Short introduction stating the goal of this study, strict anonymization, the incentive and a contact person for questions.]

Imagine you are a marketing employee in an IT service company.

Your manager approaches you to **write a Google search engine optimized (SEO) text for a single site on the website of your IT company**, that elaborates on a specific service. You should **write the text in a way that it ranks well in Google**. That means, it should preferably appear on page 1 in the Google search results.

- The text should be written for the keyword / search term / topic: "**IT maintenance**" (i.e., for IT maintenance provided as a service by your company to firms).
- It should be written for ranking well in **Google in [Country blinded], set to English language** (please use the link below).
- For ranked **example sites see**: https://www.google.com/search?num=100&hl=en&q=it+maintenance
- It should be **original, unique content**, invented by you (i.e., NO copies).
- It should be written **in English language**.
- It should contain **around 700 to 800 words** (ca. 2 A4 pages).

**Your text:** (Please write your text in the following text field.)

---

[1]Keywords and links were adapted in each survey.

**B2: Post Hoc Test for Achieved Search Engine Rankings**

Table W7 reports results of a Kruskal Nemenyi post hoc test to compare the differences between each pair of experimental groups. Except for the comparison between SEO experts and quasi experts in terms of the pages in the ranking and the pages ranked in the top 10 search results, we find that the search engine performances are statistically different at the 0.05 level (Table W7).

Table W7: Post Hoc Tests: Search Engine Rankings Performance Comparison (IT Service Sector)

| | | **Kruskal Nemenyi Post Hoc Test (*p*)** | | |
|---|---|---|---|---|
| **Dimension** | **Group** | Real SEO Experts | Quasi Experts | Novices |
| **Pages in ranking / day** | Revised Machine | <.000** | <.000** | <.000** |
| | Real SEO Experts | | .160 | <.000** |
| | Quasi Experts | | | <.000** |
| **Pages in top 10 / day** | Revised Machine | <.000** | <.000** | <.000** |
| | Real SEO Experts | | .302 | <.000** |
| | Quasi Experts | | | .017* |
| **Mean rankings / day** | Revised Machine | <.000** | <.000** | <.000** |
| | Real SEO Experts | | .000** | <.000** |
| | Quasi Experts | | | <.000** |

[1]Statistical significance codes: *0.05 level, **0.01 level, chi-square approximated;

**B3: Consumer Content Perceptions in the IT Industry Application**

Table W8 reports the instructions for the survey participants.

Table W8: Participants' Survey Instructions

| **Survey Instructions** |
| --- |
| **Dear study participant**<br><br>Thank you for participating in our study on SEO & text writing. Your input is vital for us. In the following, besides answering some demographic questions, **we will ask you to read and assess 1 text**.<br><br>**It will take you 5 minutes at most to finish the survey.**<br><br>Please **read all questions and the text mindfully and completely**, and **answer all questions as honestly and spontaneously as possible**. Follow your intuition, there are **no right or wrong answers**.<br><br>All information that you provide to us will be **strictly treated as anonymous**.<br>Thank you for your kind support.<br><br>Sincerely,<br>[…]<br><br>[New survey page]<br><br>**Imagine, you are looking for an IT service for your company, and you come across a website with the text below. Please take a look at it.**<br><br>[Randomized piece of content]<br><br>[Questions to assess content] |

To assure data quality in our survey based content consumer perception experiment, we implemented honeypots (for antispam), attention and honesty checks (i.e., reverse coded items and same questions worded a bit differently), and excluded all surveys with a completion time lower than 1.50 minutes, leaving us with 551 surveys for our analyses. We performed scale reliability checks using Cronbach's Alpha including deleting offset items. Using a series of

Kruskal Wallis tests, we assured that participants' properties did not differ substantially between the experimental conditions in terms of the time to finish the survey ($\chi^2(3)=3.38$, $\eta^2=.01$, $p=.337$), the participants' gender ($\chi^2(3)=2.00$, $\eta^2=.00$, $p=.572$), the highest completed level of education ($\chi^2(3)=3.08$, $\eta^2=.01$, $p=.380$), age ($\chi^2(3)=.25$, $\eta^2=.00$, $p=.969$), and English reading proficiency ($\chi^2(3)=.41$, $\eta^2=.00$, $p=.939$).

Table W9 reports operationalizations, literature sources, and scale reliability metrics for the content user perception experiment that we conducted using a survey.

Table W9: Operationalizations & Measures of Main Variables for Survey

| Variable | Items | Source | Scale Reliability[1] |
|---|---|---|---|
| Readability | Bipolar 5-point scale with following items:<br>"Please indicate whether you perceive the text above as …<br>● poorly written – well written<br>● poorly readable – well readable<br>● not fitting together well – fitting together well<br>● not understandable – understandable<br>● not interesting – interesting" | Pitler and Nenkova 2008 | .91 |
| Understandability | Bipolar 5-point scale with following items:<br>"Please indicate whether you perceive the text above as …<br>● complicated – simple<br>● unclear – clear<br>● chaotic – orderly<br>● illogically arranged – logically arranged<br>● wordy – concise<br>● difficult – easy" | Kamoen et al. 2013 | .88 |
| Credibility | Bipolar 5-point scale with following items:<br>"Please indicate whether the text above is …<br>● unbelievable – believable<br>● inaccurate – accurate<br>● not trustworthy – trustworthy<br>● biased – not biased<br>● incomplete – complete" | Roberts 2010, Flanigan and Metzger 2000 | .87 |
| Attitude toward the content | Bipolar 5-point scale with following items:<br>"Please indicate whether you feel that the text above is …<br>● distant – appealing<br>● reluctant – inviting<br>● boring – fascinating<br>● impersonal – personal<br>● monotonous – varied<br>● interesting – uninteresting" | Kamoen et al. 2013 | .89 |

[1]Cronbach's Alpha with optimized number of items

In addition to the scales employed in W9, we measure content naturality using two items. On bipolar five-point scales, we ask respondents to indicate whether they believe that the content feels artificial vs. feels natural, and machine made vs. human made. We also ask two question to assess future intent. To gauge willingness to further inform, we use a slider from 0 to 100 and ask respondents to indicate how they agree with the statement: "I want to further inform myself about the company providing the service." To measure willingness to buy, we use a slider from 0 to 100 and ask respondents to indicate how much they agree with the statement: "I am willing to buy the described service."

Table W10 reports pairwise correlations between user perception variables using Kendall's tau b, illustrating high correlations between these items.

Table W10: Consumer Content Perception: Dimensions' Intercorrelations

| **Dimension** | **Kendall's tau b ($\tau_b$)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Readability | Understandability | Credibility | Attitude Toward the Content | Content Naturality | Willingness to Further Inform | Willingness to Buy |
| Readability | 1.00** | .59** | .57** | .50** | .52** | .41** | .42** |
| Understandability | | 1.00** | .43** | .58** | .57** | .44** | .46** |
| Credibility | | | 1.00** | .40** | .44** | .33** | .37** |
| Attitude Toward the Content | | | | 1.00** | .58** | .52** | .53** |
| Content Naturality | | | | | 1.00** | .44** | .49** |
| Willingness to Further Inform | | | | | | 1.00** | .69** |
| Willingness to Buy | | | | | | | 1.00** |

[1]Statistical significance codes: *0.05 level, **0.01 level, one-tailed; n=551;

Table W11 illustrates computational analyses using LIWC (Pennebaker et al. 2015), the evaluative lexicon (Rocklage et al. 2018), and the text analyzer (Berger et al. 2020b) software packages that apply various lexica, analyses and scales to assess linguistic properties along psychological dimensions including concreteness, familiarity, and emotionality. The analysis reveals that differences between the semi-automated and human content are minor along most dimensions.

Table W11: Consumer Content Perception (Computational Analysis)

| **Dimension** | **Descriptives (Mean, SD)**[1] | | | | **Kruskal Wallis**[2] | | | |
|---|---|---|---|---|---|---|---|---|
| | Revised Machine | Real SEO Experts | Quasi Experts | Novices | $\chi^2$ | $\eta^2$ | *df* | *p* |
| Concreteness | 323.10 (7.45) | 326.00 (5.37) | 321.30 (7.48) | 318.60 (4.28) | 9.67 | .15 | 3 | .021* |
| Familiarity | 574.14 (7.95) | 578.14 (12.73) | 579.22 (9.33) | 581.47 (9.14) | 7.14 | .11 | 3 | .067 |
| Emotionality | 3.28 (.66) | 3.33 (.38) | 3.47 (.55) | 3.53 (.47) | 3.07 | .05 | 3 | .380 |
| Emotional Valence | 6.15 (.89) | 6.23 (.86) | 6.45 (.77) | 6.69 (.72) | 3.70 | .06 | 3 | .296 |
| Negations | .004 (.005) | .005 (.003) | .006 (.003) | .007 (.006) | 3.28 | .05 | 3 | .351 |
| Interrogatives | .011 (.006) | .009 (.004) | .013 (.006) | .013 (.008) | 2.31 | .04 | 3 | .509 |
| Causation | .028 (.009) | .030 (.013) | .032 (.015) | .026 (.009) | 2.07 | .03 | 3 | .558 |
| Certainty | .011 (.005) | .013 (.005) | .021 (.009) | .019 (.009) | 16.24 | .25 | 3 | .001** |
| Tentativeness | .022 (.010) | .027 (.014) | .022 (.010) | .022 (.009) | 1.44 | .02 | 3 | .697 |
| Differentiation | .020 (.009) | .026 (.014) | .021 (.009) | .021 (.011) | 1.25 | .02 | 3 | .740 |
| Focus on future | .009 (.006) | .013 (.006) | .011 (.006) | .015 (.007) | 8.54 | .13 | 3 | .036* |

[1]Dimension scales: for concreteness, familiarity scale range: 100 (abstract, unfamiliar) to 700 (concrete, familiar), emotionality scale range: 0 (no emotion) to 9 (high emotion), emotional valence scale range: 0 (highly negative) to 9 (highly positive); other dimensions like negations, interrogatives, etc., represent percentages of total words in the text;
[2]Statistical significance codes: *0.05 level, **0.01 level; n=66;

**B4: Website Engagement Arising from Direct Links (IT Service Application)**

Table W12 reports statistics for the user behavior for visitors coming from direct links (e.g., links in emails, on other webpages, etc.) to the focal experimental pages on the website.

Table W12: Consumer Behavior (Direct Links Source Only)

| Dimension | Descriptives (∑) | | | | One-Sample Chi-Squared[1] | | |
|---|---|---|---|---|---|---|---|
| | Revised Machine | Real SEO Experts | Quasi Experts | Novices | $\chi^2$ | *df* | *p* |
| No. of Pages with Pageviews | 19 | 8 | 19 | 19 | 5.58 | 3 | .134 |
| No. of Pages with Pageview in % | 100 | 89 | 100 | 100 | .93 | 3 | .817 |
| Pageviews | 441 | 83 | 176 | 427 | 344.78 | 3 | <.000** |
| Unique Pageviews | 216 | 42 | 88 | 212 | 166.79 | 3 | <.000** |
| Entrances | 195 | 34 | 73 | 196 | 168.07 | 3 | <.000** |
| Exit Rate (means) | .40 | .39 | .41 | .42 | - | - | - |
| Bounce Rate | .01 | .03 | .00 | .01 | - | - | - |
| Avg. Usage Duration (Abs., sums) | 216 | 8 | 250 | 40 | 348.41 | 3 | <.000** |
| Avg. Usage Duration (Rel.)[2] | 11 | 1 | 13 | 2 | 16.93 | 3 | <.000** |
| Returning Visitors (Abs.) | 225 | 41 | 88 | 215 | 178.10 | 3 | <.000** |
| Returning Visitors (Rel.)[2] | 11.84 | 5.12 | 4.63 | 11.32 | - | - | - |
| Buying Affinity (Abs.)[3] | 7899 | 1957 | 3949 | 10714 | 7556.50 | 3 | <.000** |
| Buying Affinity (Rel.)[2,4] | 416 | 245 | 208 | 564 | 226.65 | 3 | <.000** |
| Exp. Sales (for U.P.*100)[5] | 432 | 84 | 176 | 424 | 333.58 | 3 | <.000** |

[1]Statistical significance codes: *0.05 level, **0.01 level;
[2](Rel.) = the absolute value (Abs.) divided by No_of_Pages_with_Pageviews
[3]Buying Affinity (Abs.) = Unique_Pageviews*Willingness_to_Buy (survey measured);
[4]Buying Affinity (Rel.) = Buying_Affnity (Abs.)/No._of_Pages_with_Pageviews;
[5]Exp. Sales (for U.P.*100) = (Unique_Pageviews/100*Expected_Sales_Rate)*100, where the expected sales rate is 2% (obtained from past company reports);

**Appendix C: Supplemental Information for the Education Sector Application**

Table W13 reports group comparison tests for the search engine ranking performance of the experimental groups "revised machine" and "human" using Kruskal-Wallis tests. Precisely, Table W13 shows that the revised machine outperforms the human content generating group in terms of the number of pages that got into the search engine ranking, the pages ranked in the top 10, and in terms of mean ranking.

Table W13: Search Engine Rankings Performance Comparison (Education Sector)

| Dimension | Group | Descriptives | | | | | Kruskal-Wallis[2] | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $n_p$[1] | Median | (IQR) | Min | Max | $\chi^2$ | $\eta^2$ | *df* | *p* |
| **Pages in ranking / day** | **Revised Machine** | 6 | **5.00** | **(.00)** | **4.00** | **6.00** | 100.95 | .49 | 1 | <.000** |
| | Human | | 4.00 | (.00) | 3.00 | 4.00 | | | | |
| **Pages in top 10 / day** | **Revised Machine** | 6 | **5.00** | **(1.00)** | **3.00** | **6.00** | 101.28 | .49 | 1 | <.000** |
| | Human | | .00 | (1.00) | .00 | 1.00 | | | | |
| **Mean rankings / day** | **Revised Machine** | 6 | **52.67** | **(1.83)** | **4.17** | **101.50** | 67.49 | .33 | 1 | <.000** |
| | Human | | 117.20 | (4.13) | 112.50 | 162.30 | | | | |

[1]$n_p$=number of pages per experimental group. n=208 (days); [2]Statistical significance codes: *0.05 level, **0.01 level; Compared numbers are daily aggregate numbers. For mean rankings / day: we coded non-ranking pages with the value 301 (i.e., 1 place lower than the max observable ranking).

Table W14 reports statistics for the experimental groups (e.g., Revised machine, Humans, etc.) and the lowest ranked results on the 5 quality score components. The results are consistent with our findings from the IT service industry experiment.

Table W14: Quality Score Components Group Comparisons to Top 10 Ranked Websites (Education Sector)

| Quality Score Component | Group | Descriptives | | | | Wilcoxon rank sum[1] | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Median (IQR) | | Min | Max | W | z | *r* | *p* |
| **Topic ($s_a$)** | Revised machine | .48 | (.06) | .43 | .52 | 24 | 2.41 | .14 | .008** |
| | Raw machine | .50 | (.08) | .40 | .55 | 34 | 2.62 | .16 | .004** |
| | Humans | .35 | (.10) | .27 | .46 | 16 | .38 | .02 | .650 |
| | Worst 10 | .17 | (.02) | .12 | .22 | 00 | -3.07 | -.89 | .001** |
| **Keywords ($s_k$)** | Revised machine | .50 | (.04) | .45 | .57 | 24 | 2.41 | .14 | .008** |
| | Raw machine | .54 | (.06) | .41 | .60 | 35 | 2.85 | .17 | .002** |
| | Humans | .38 | (.09) | .29 | .52 | 21 | .39 | .02 | .349 |
| | Worst 10 | .11 | (.04) | .06 | .18 | 00 | -3.07 | -.89 | .001** |
| **Uniqueness ($s_d$)** | Revised Machine | .94 | (.07) | .92 | 1.00 | 18 | 1.26 | .07 | .896 |
| | Raw machine | .90 | (.06) | .83 | .99 | 17 | .08 | .00 | .468 |
| | Humans | .12 | (.01) | .07 | .22 | 00 | -3.07 | -.18 | .001** |
| | Worst 10 | .93 | (.02) | .91 | .96 | 16 | .39 | .11 | .650 |
| **Readability ($s_r$)** | Revised Machine | .77 | (.19) | .64 | 1.00 | 23 | 2.15 | .13 | .015* |
| | Raw machine | .93 | (.10) | .77 | 1.00 | 36 | 2.81 | .17 | .002** |
| | Humans | .77 | (.29) | .34 | .98 | 24 | .85 | .05 | .197 |
| | Worst 10 | .52 | (.08) | .45 | .61 | 5 | -2.04 | .59 | .021* |
| **Naturality ($s_n$)** | Revised Machine | 83 | (.00) | .83 | .92 | 25 | 2.59 | .15 | .004** |
| | Raw machine | .83 | (.23) | .50 | 1.00 | 30 | 1.84 | .11 | .032* |
| | Humans | .38 | (.46) | .08 | 1.00 | 12 | -1.01 | -.06 | .845 |
| | Worst 10 | .42 | (.21) | .28 | .88 | 11 | -1.04 | -.30 | .194 |

[1] One-tailed tests, direction for each test according to the idiosyncrasies of the field; statistical significance codes: *0.05 level, **0.01 level;