informs
http://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Vol. 41, No. 3, May–June 2022, pp. 441–452
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Frontiers: Supporting Content Marketing with Natural Language Generation

Martin Reisenbichler,[a,]* Thomas Reutterer,[a] David A. Schweidel,[b] Daniel Dan[c]

[a] Department of Marketing , Vienna University of Economics and Business, Vienna A-1020, Austria; [b] Goizueta Business School, Marketing Area, Emory University, Atlanta, Georgia 30322; [c] School of Applied Data Science, Modul University, Vienna, Vienna A-1190, Austria
*Corresponding author

Contact: martin.reisenbichler@wu.ac.at (MR); thomas.reutterer@wu.ac.at, https://orcid.org/0000-0003-1276-8239 (TR); dschweidel@emory.edu, https://orcid.org/0000-0003-2665-3272 (DAS); daniel.dan@modul.ac.at (DD)

Received: June 30, 2021
Revised: September 5, 2021
Accepted: September 15, 2021
Published Online in Articles in Advance: February 25, 2022

https://doi.org/10.1287/mksc.2022.1354

Copyright: © 2022 INFORMS

Abstract. Advances in natural language generation (NLG) have facilitated technologies such as digital voice assistants and chatbots. In this research, we demonstrate how NLG can support content marketing by using it to draft content for the landing page of a website in search engine optimization (SEO). Traditional SEO projects rely on hand-crafted content that is both time consuming and costly to produce. To address the costs associated with producing SEO content, we propose a semiautomated methodology using state-of-the-art NLG and demonstrate that the content-writing machine can create unique, human-like SEO content. As part of our research, we demonstrate that although the machine-generated content is designed to perform well in search engines, the role of the human editor remains essential. Comparing the resulting content with human refinement to traditional human-written SEO texts, we find that the revised, machine-generated texts are virtually indistinguishable from those created by SEO experts along a number of human perceptual dimensions. We conduct field experiments in two industries to demonstrate our approach and show that the resulting SEO content outperforms that created by human writers (including SEO experts) in search engine rankings. Additionally, we illustrate how our approach can substantially reduce the production costs associated with content marketing, increasing their return on investment.

History: K. Sudhir served as the senior editor and Olivier Toubia served as associate editor for this article. This paper was accepted through the Marketing Science: Frontiers review process.
Supplemental Material: Data and the web appendices are available at https://doi.org/10.1287/mksc.2022.1354.

Keywords: SEO • content marketing • natural language generation • transfer learning

Introduction

Natural language generation (NLG) has seen several applications, designed to serve both consumers and commercial users. Digital voice assistants and chatbots use NLG to respond to user inputs. Email and text messages employ NLG to suggest the words that likely follow text that has been entered. NLG is being used for applications such as drafting emails from bullet points, crafting short product descriptions, and summarizing website content for social media posts. Despite the deployment of such technologies, there has been no research on the efficacy of NLG to support content marketing. Moreover, although marketing investments in artificial intelligence (which encompasses NLG) are increasing, most fail to see a return on their investment (Ascarza et al. 2021).

With 70% of marketers investing in content marketing and nearly a quarter of marketers planning to increase their expenditures,[1] NLG can reduce the costs associated with creating marketing content and increase the rate at which new content is produced. Given the availability of digital text (e.g., Berger et al. 2020b), we assert that NLG can support the production of domain-specific marketing content (Heaven 2020). To illustrate this, we apply it to the context of drafting content for search engine optimization (SEO).

SEO is essential to achieve high organic search engine rankings to increase traffic, and consequently revenue. SEO is a major component of firms' digital marketing efforts, on par with search engine advertising (SEA) in terms of spending (e.g., Berman and Katona 2013, Liu and Toubia 2018). Due to the competition for higher rankings in organic search results (e.g., Bar-Ilan et al. 2006, Luh et al. 2016), firms invest heavily in search engine optimized content, 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 factor in search engine rankings (e.g., Liu and Toubia 2018, Google 2020), SEO research has explored its ranking drivers. Early research included manual analyses (e.g., Danaher et al. 2006) and the identification of factors such as title and page length (e.g., Salminen et al. 2019). Recent work has sought to identify optimal word distributions using methods such as term frequency–inverse document frequency (TF-IDF), latent semantic analysis (LSA; e.g., Luh et al. 2016), and latent Dirichlet allocation (LDA; e.g., Liu and Toubia 2018). 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 multidimensional vector. The incorporation of embeddings combined with a quality measure aligned with critical search engine ranking factors into a machine learning framework enables us to analyze existing website content to capture the context in which keywords appear, and to generate new content.

A Semiautomated Content Development Algorithm

We propose a human in the loop, semiautomated content-generation method for developing SEO 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 to the content (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, our semiautomated content outperforms human-created content in search engines. Moreover, our approach reduces the content-production time and hence the associated labor costs by more than 91% compared with the traditional SEO content production.

In developing SEO website landing page content manually, a main keyword (i.e., a search query) is initially selected. Next, research is conducted on textual features of the top-ranking websites (Luh et al. 2016, Sheffield 2020). 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.

In Figure 2, we illustrate our proposed method for semiautomated content generation that mimics this process.

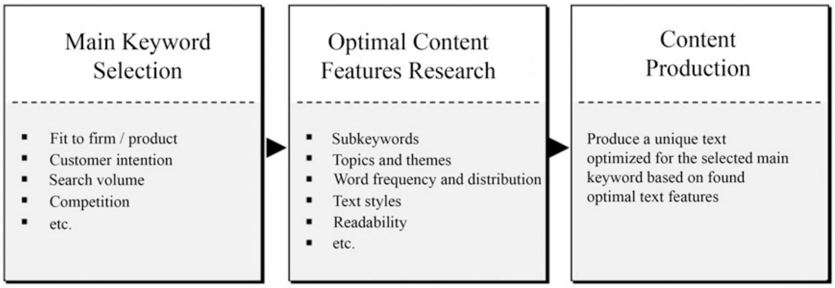
Once a keyword (e.g., "IT service management") has been specified by a human, the top T search engine results for that keyword are captured and the content from those websites is scraped. The content of these pages ($top\_txt_{1,\ldots,}\ top\_txt_T$) is then used to update the pretrained GPT-2 345M model (Radford et al. 2018, Radford et al. 2019), similar to a Bayesian updating of parameters when new data become available, to generate new content that resembles optimal text structures of the top T ranked search engine results. For each piece of generated content, we derive a quality score and provide the top-scoring content to a human editor for revision to ensure the accuracy, company fit, and other aspects such as brand tone or voice of the content.

Pretrained Language Models and Fine-Tuning

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) are the building blocks for text generation. Large-scale pretrained transformer language models such as 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).

To provide the intuition underlying the transformer-based GPT-2 NLG model, we offer a brief overview of the mechanics. Whereas word embeddings have been used in the marketing literature to represent text as vectors (Timoshenko and Hauser 2019), GPT-2 uses byte pair encoding (BPE) and tokens (i.e., learned and encoded pieces of words). To facilitate our exposition of how the transformer-based language model operates, we describe the model in terms of words. Using a large corpus of digital text collected from various online sources, the pretrained language model GPT-2 makes use of word embeddings and position embeddings in which word meaning information and

Figure 1. Prototypical Manual SEO Content-Production Workflow



sequential language patterns are represented. Given a sequence of words, $U = (u_{-k}, \ldots, u_{-1})$, the autoregressive model predicts the likely next word by sampling from a probability distribution over its entire learned vocabulary (consisting of 50,257 words) conditional on the given word sequence and on a pretrained neural network with parameters Θ (Figure 3).

GPT-2 relies on word and given context meaning information to generate its output distribution over its vocabulary. The input matrix h_0 combines a given word sequence, meaning in terms of word embeddings, and sequential word position information in terms of position embeddings. GPT-2 then extracts, transforms, adds, and normalizes information from h_0 into the embedding space e by using L layers of decoder transformer blocks (Figure 3). This information includes the extent of attention put on given sequence words using multiheaded self-attention (Vaswani et al. 2017), and high-dimensional hidden language states that shift the focus in the embedding space e to recreate natural word sequences from position-wise feed-forward neural networks. The information contained in the final block's output (h_L) is used to sample the likely next word from GPT-2's vocabulary. In sum, for a given word sequence, GPT-2 outputs a probability vector (P_u), which provides the likelihood of a given word occurring next. GPT-2 learns and stores word probabilities for given word sequences represented in its 345 million parameters (including its word and position embeddings, attention weight matrices, and Θ) using eight million English text documents with a broad topical variety. We refer the reader to the web Appendix 1.1 for a more detailed technical explanation of GPT-2.

Pretrained NLG models such as GPT-2 are broadly applicable and not tailored to a particular context. Should one use the pretrained GPT-2 model to generate text based on a search keyword, it would not necessarily resemble the text typically found on a website. To leverage the pretrained GPT-2 model and its semantic and syntactic language knowledge, we use the pretrained model parameters as initial values and apply the GPT-2 model to the text from the top $T = 10$ search engine results, a process referred to as fine-tuning.[4]

Fine-tuning enables domain-specific applications of pretrained language models. GPT-2 can be fine-tuned using song lyrics, enabling the generation of new potential lyrics. Fine-tuning on books such as the *Harry Potter* series or Shakespeare's works allow for the creation of new content in the style of the original author. In essence, we use the general linguistic structure of the English language as captured by the parameters Θ of the pretrained GPT-2 model as a starting point. As we present in more detail in web Appendix 1.5, the fine-tuning process employed in our empirical applications then updates these parameters to reflect the content and language patterns of the

Figure 2. Overall Method Concept and Procedure

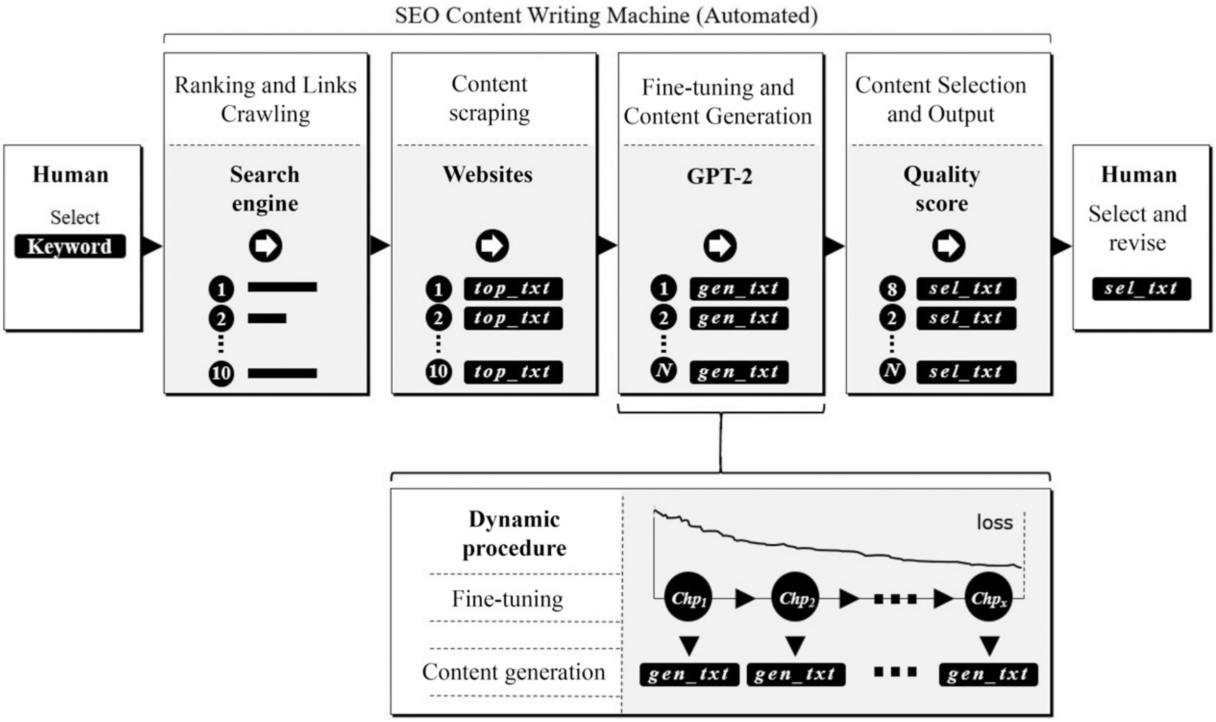
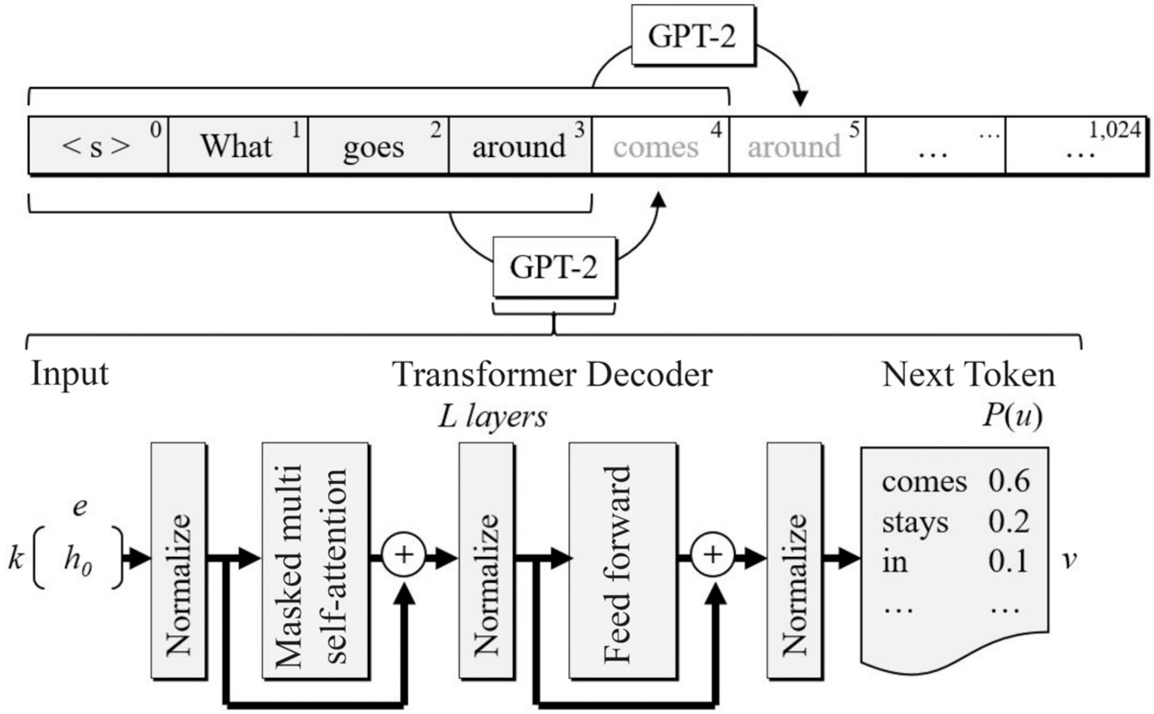


Figure 3. The GPT-2 Transformer Model[3]



text from the top 10 search results. This results in a generative language model that is capable of producing content for a particular keyword in the linguistic style of the top-ranking search results. In addition, to ascertain topical focus, the GPT-2 generated website texts use the focal keyword as the main headline and seed sequence (i.e., as a given word sequence to generate content from) for text generation.

This process merges application-specific content with the pretrained language model and is essential to ensure that the produced content incorporates the keyword, and industry- and domain-specific language structures (to reflect subkeywords, industry-specific terms, topics, etc.) that appear in the top-ranked search results. The fine-tuning process is less resource intensive than estimating a new, pretrained language model. As we increasingly fine-tune the GPT-2 model, we generate content throughout the process at regular checkpoints ($Chp_1, \ldots, Chp_x$). At each checkpoint, we use the fine-tuned model to generate the text that follows the search query text. Web Appendix 1.2 provides a description of the software features developed to automate the fine-tuning process.

Content Quality Score

A piece of content (generated at a checkpoint), gen_txt_n, is scored for its anticipated SEO performance, which we measure by constructing a quality score qs_g. This is based on five key criteria derived from industry practice and guidelines (e.g., Google 2020, Sheffield 2020) that focus on core content-related aspects of the search engine and are highly stable over time: the overall topic treated in the content (s_a), keyword integration (s_k), content uniqueness (s_d), naturality similarity (s_n), and readability similarity (s_r):

$$qs_g = s_a \cdot s_k \cdot s_d \cdot s_n \cdot s_r, \text{ with } \quad 0 \le qs_g \le 1. \qquad (1)$$

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 $FGen$ denotes the term frequency vector of gen_txt) and each of the top T search results (where $FTop_t$ is the word frequency vector corresponding to top_txt_t; w indexes the vector components):

$$\begin{aligned} s_a &= \frac{1}{T}\sum_{t=1}^{T} \frac{FGen \cdot FTop_t}{\|FGen\| \, \|FTop_t\|} \\ &= \frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{w=1}^{W} FGen_w FTop_{tw}}{\sqrt{\sum_{w=1}^{W} FGen_w^2}\sqrt{\sum_{w=1}^{W} FTop_{tw}^2}}. \end{aligned} \qquad (2)$$

Keyword integration (s_k) is measured in a similar fashion. However, in constructing this component, we use only the 10 most frequently occurring words in gen_txt and $top\_txt_{1,\ldots,T}$.

To measure content uniqueness (s_d), we calculate the number of duplicated n-grams of size $kw + 1$ in gen_txt compared with n-grams in $gen\_txt \cup top\_txt_{1,\ldots,T}$, where kw is the length of the main keyword. 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 $top\_txt_{1,\dots,T}$:

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

Naturality similarity (s_n) assesses the similarity of the generated text to the top search results on 12 linguistic measures of naturalness, which include measures for assessing the lexical richness and composition of a text. For each dimension, we perform a nonparametric one-sample Wilcoxon signed rank test between the naturalness score obtained by gen_txt and the distribution of scores from $\{top\_txt_1, \dots, top\_txt_T\}$. Higher scores of s_n, which is the proportion of nonsignificant tests, suggest that the naturalness of the generated text is consistent with the top search results. We follow a similar procedure to measure readability similarity (s_r), using 46 measures of readability. We provide additional details of how we measure content uniqueness, naturality and readability similarity in web Appendix 1.3.

To confirm that top-ranked search engine websites score highest on our quality score components, we conduct an extensive validation study by analyzing nearly 1.5 million ranked websites corresponding to approximately 8,500 keywords across four business sectors and 36 specific industries (see web Appendix 1.4). The measures $\{s_a, s_k, s_n, s_r\}$ ensure that the generated content is similar to the top-ranked search results, with a high quality score meaning that the generated content indeed resembles the top 10 ranked content. Consider the components s_a and s_k. The top-ranked content elaborates on isolated subtopics and aspects of the keyword for which it has been optimized. Our approach not only creates content that mimics heuristics such as the keyword density, but also the context in which the keywords appear and the broader linguistic structure of the top performing search results, enabling the content to correspond to a search engine user's query, and consequently perform better in search engine rankings. Similarly, a higher score for s_n and s_r for the machine-generated content means that it reflects the naturality and readability patterns found in the top 10 ranked content.

The uniqueness component (s_d) requires our dynamic fine-tuning process, as content deemed too similar to the current top-ranked search results will be penalized by search engine algorithms. A higher s_d score means that the machine content is more unique compared with the top 10 ranked content. 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 similar to the top-ranked search results, leading to the most optimal content selection from intermediate fine-tuning steps (for a more detailed discussion, see web Appendix 1.5). To ensure that our method outperforms the real top 10 ranked content in terms of the quality score, we generated content for more than 300 randomly selected keywords from the aforementioned 8,500 keywords and compared the resulting quality score to the top-ranking content for each keyword. We refer readers to web Appendix 1.6 for details of this validation and to web Appendix 1.8 for a robustness check that uses an alternative formulation of the quality score based on the same five components.

For our experiments, we fine-tune our model for 200 training steps for each keyword, generating 100 pieces of content at each 20th step, which resulted in 1,000 generated texts per keyword. After ordering and selecting the best content based on the quality score ($sel\_txt_{1,\dots,N}$), 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 roles that fine-tuning and human editing play in the SEO content-generation algorithm, we present fragments of landing page texts in Table 1 derived throughout the production process for an illustrative keyword from our empirical application and the associated quality score components. First, we provide the text from the pretrained GPT-2 model. Second, we show the text that arises from the fine-tuning process. Lastly, we provide the text after minimal changes have been made by a human editor.

Although in this example the pretrained GPT-2 model yields text that scores relatively high on uniqueness, readability similarity, and naturality similarity (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. Thus, the pretrained GPT-2 model is not well suited for SEO purposes as the quality scores indicate. We explore more on this aspect in web Appendix 1.7 and include additional examples of machine-generated content in web Appendix 1.9.

Application to the Information Technology Service Industry

Experimental Setup

To test the performance of our semiautomated content-generation machine, we collaborate with a midsized international commercial company in the information technology (IT) service industry. Four

Table 1. Abbreviated Example of Generated Content for a Target Landing Page on a Website (Keyword: "IT maintenance support")

Stage	Quality score	Generated content
GPT-2 basic	$s_a = 0.15$ $s_k = 0.11$ $s_d = 0.79$ $s_n = 0.25$ $s_r = 0.72$ $qs_g = 0.002$	IT maintenance support, support staff, and support personnel, including: • the department's technical assistance services (TAS) office, • the department's information technology services management office (ITSMO), • the department's computer-aided design and engineering office, and • the department's human resources office. 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	$s_a = 0.51$ $s_k = 0.57$ $s_d = 0.78$ $s_n = 0.17$ $s_r = 0.77$ $qs_g = 0.03$	**It maintenance support** 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. 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. **System, service, and maintenance maintenance maintenance** System maintenance is maintenance of software applications and hardware that directly affects operations such as core processes, web content management systems, and so on. [. . .]
GPT-2 fine-tuned and revised	$s_a = 0.52$ $s_k = 0.61$ $s_d = 0.81$ $s_n = 0.17$ $s_r = 0.70$ $qs_g = 0.029$	**IT maintenance support** 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. 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. **System, service, and maintenance ■** System maintenance is maintenance of software applications and hardware that directly affects operations such as core processes, web content management systems and so on. [. . .]

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

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 written instructions and clear directions of how to do it), (3) five SEO experts (professionals with at least two months of experience in the SEO industry who received the novices' stimulus), and (4) the semiautomated SEO content-writing machine with revisions made by a company employee who was instructed to keep content changes to a minimum. We provide details of the experimental setup in web Appendix 2.2.

Groups (1)–(3) produced content via an online survey that contained a link to the keyword-specific top 10 search engine ranked content and a word-counter tool that the company uses as part of its content-production workflow. By providing study participants with links to the top 10 search engine results for each keyword, we ensure they have access to the same text used by our semiautomated method. All groups produced content for the same industry-specific keywords (e.g., "IT procurement" or "IT service maintenance"), resulting in 19 pieces of content per experimental group, with the exception of the SEO expert group that produced nine pieces of content for randomly selected keywords due to time and cost considerations. We provide the full set of keywords (and the associated statistics) in web Appendix 2.1.

The company selected the keywords used in our experiment based on its standard procedure (i.e., based on monthly search volume, competition, fit with the firm, and keyword strategy). We ensured the quality of the experiment by using a set of additional experiments and data. We report these in web Appendices 2.2 and 2.3, which include robustness checks and additional measures to ensure experimental quality (e.g., incentives, a realistic simulation of the company's content-production workflow, controls for content length, writing time, and participants' writing skills). We also examine if providing SEO experts with the quality score of their writing and allowing them the opportunity to revise it improves the quality, mirroring an A/B testing setup. Even when the SEO experts are provided with such feedback, they are unable to revise content that improves upon it (see web Appendix 2.4).

Search Engine Rankings Performance

Each piece of content was published on its own page at day 0 on the company website in December 2019. All pages were composed of the same elements and structure, and each URL consists of the keyword and a random alphanumeric suffix. To compare the performance of the semiautomated content to human-generated content, we monitor the top 300 search engine rankings for each keyword (i.e., 30 pages of the search engine results per keyword) for 412 days after the texts were posted.

Figure 4 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 results (black bars). As shown in Table 2, in stark contrast to all human groups ($\chi^2(3) = 1{,}137.98$, $\eta^2 = 0.69$, $p < 0.000$), almost all semiautomated content ranks in the search engine with high stability over the observation period. Moreover, in contrast to the human groups ($\chi^2(3) = 1{,}140.41$, $\eta^2 = 0.69$, $p < 0.000$), the semiautomated approach produces more content that appears on the first page of results (a top 10 ranking) during the observation period. Such performance is critical, as lower visibility in search engines can adversely affect the company's performance (Baye et al. 2016). Moreover, beyond performing well on the specified keywords, an additional examination of keyword performance on related

Figure 4. Number of Pages in Ranking and in the Top 10 Search Results per Day

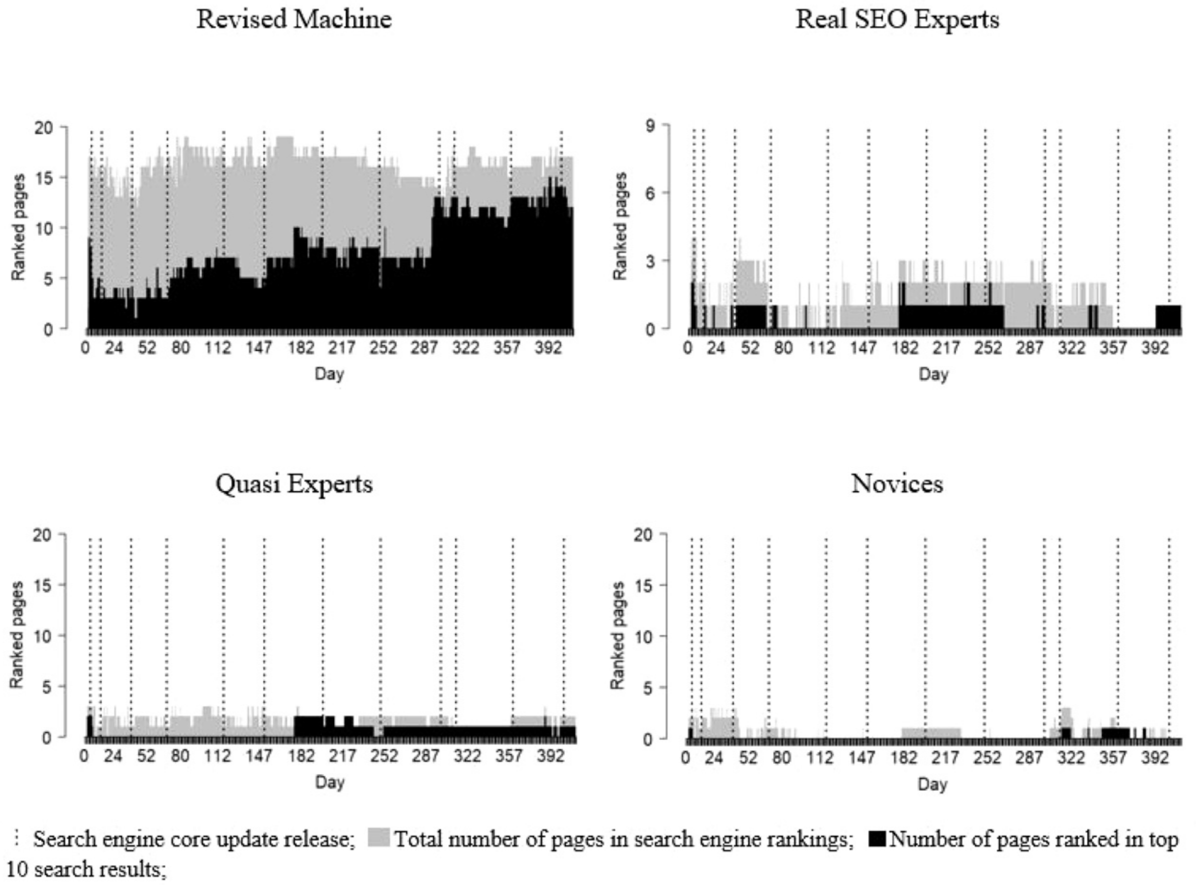


Table 2. Search Engine Rankings Performance Comparison (IT Service Industry)

		Descriptives					Kruskal-Wallis			
Dimension	Group	n_p	Median	(IQR)	Min	Max	χ^2	η^2	*df*	*p*
Pages in ranking/day	Revised machine	19	17.00	(1.00)	12	19	1,137.98	0.69	3	<0.000**
	Real SEO experts	9	2.00	(1.00)	0	4				
	Quasi experts	19	2.00	(1.00)	0	3				
	Novices	19	0.00	(1.00)	0	3				
Pages in top 10/day	Revised machine	19	7.00	(5.00)	0	15	1,140.41	0.69	3	<0.000**
	Real SEO experts	9	0.00	(1.00)	0	2				
	Quasi experts	19	1.00	(1.00)	0	2				
	Novices	19	0.00	(0.00)	0	1				
Mean rankings/day	Revised machine	19	50.37	(25.95)	16.63	132.63	1,237.36	0.75	3	<0.000**
	Real SEO experts	9	243.20	(33.33)	171.22	301				
	Quasi experts	19	271.21	(14.47)	254.47	301				
	Novices	19	301.00	(14.47)	256.47	301				

Notes. Post hoc group comparison tests are in web Appendix 2.3. Compared numbers are daily aggregate numbers. For mean rankings/day, we coded nonranking pages with the value 301 (i.e., one place lower than the maximum ranking). n_p, number of pages per experimental group; n = 412 (days) for each group. IQR, Interquartile Range.

**$p < 0.01$.

subkeywords revealed that the semiautomated content also ranks higher than the human groups (additional information is in web Appendix 2.3).

Consumer Content Perceptions

The capability of our semiautomated procedure to generate content that produces longer-lasting search engine rankings compared with human-written text is important from an SEO perspective. In addition to search engine rankings, the content must also appeal to the human readers. Possible unnatural patterns and related issues with artificial content should be avoided (e.g., Radford et al. 2019), as they may contribute to adverse perceptions among consumers. Our proposed quality score measures are not designed to eliminate such tendencies, because its focus is exclusively on content similarity with the top-ranking search results and does not necessarily capture human perceptions.

To examine the differences in consumer perceptions between the semiautomated and human content, we collect data from English-speaking MTurk participants in the United States (n = 588). We randomly assigned one piece of content to each participant. Following a short introduction and instructions on reading the content, participants rated the content on scales for perceived readability (Pitler and Nenkova 2008), understandability (Kamoen et al. 2013), credibility (Roberts 2010), attitude toward the content (Kamoen et al. 2013), perceived content naturality, consumers' willingness to further inform themselves

Table 3. Consumer Content Perception

	Descriptives (mean, SD)				Kruskal Wallis			
Dimension	Revised machine	Real SEO experts	Quasi experts	Novices	χ^2	η^2	*df*	*p*
Readability	3.81	4.06	3.87	3.87	2.85	0.01	3	0.414
	(1.01)	(0.82)	(0.99)	(1.05)				
Understandability	3.34	3.54	3.51	3.49	3.45	0.01	3	0.327
	(0.95)	(0.89)	(0.96)	(0.99)				
Credibility	3.88	3.99	3.89	3.96	2.11	0.00	3	0.549
	(0.77)	(0.69)	(0.79)	(0.83)				
Attitude toward the content	3.05	3.32	3.21	3.35	7.48	0.01	3	0.058
	(1.04)	(0.86)	(0.93)	(0.96)				
Content naturality	3.23	3.49	3.47	3.43	4.79	0.01	3	0.187
	(1.11)	(1.04)	(1.10)	(1.15)				
Willingness to further inform	48.95	55.12	50.15	56.94	8.39	0.02	3	0.038*
	(30.49)	(30.39)	(29.58)	(29.85)				
Willingness to buy	45.92	52.46	48.36	53.26	5.53	0.01	3	0.137
	(30.87)	(29.15)	(30.30)	(30.18)				

Notes. Dimension scales: for readability, understandability, credibility, attitude toward the content, and content naturality scale range: 1 (bad) to 5 (good); for willingness to further inform and willingness to buy scale range: 0 (bad) to 100 (good); n = 551; for post hoc tests for significant statistical tests see the web appendix.

*$p < 0.05$.

on the service, and willingness to buy the service. Additional details of the MTurk study are provided in web Appendix 2.5. Table 3 shows the perceptions of content by experimental group, with our semiautomated content generally being perceived no differently compared with 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. 2020a) software packages that apply various lexica, analyses, and scales to assess linguistic properties along multiple psychological dimensions. The analyses reveal that differences between the semiautomated and human content are minor. We observe differences in 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 semiautomated content and SEO experts. The full results are reported in web Appendix 2.5.

In addition to comparing performance in search engine rankings, we also investigated consumers' engagement with the website. Consistent with prior research (e.g., Azzopardi et al. 2018, Ghose et al. 2019), a series of χ^2 tests reveal that semiautomated content performs better than human-generated content on the basis of the number of page views ($\chi^2(3) = 257.31$, $p < 0.000$), page views from unique website visits ($\chi^2(3) = 130.52$, $p < 0.000$), and the number of sessions started on the website through the SEO content (76, $\chi^2(3) = 114.21$, $p < 0.000$). 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). Details on this study are reported in web Appendix 2.6.

Reducing Production Costs

While conducting the experiments, we collected responses from all participants on the amount of time needed for content production/revision, the maintenance costs for servers used to host the content-writing machine, as well as the company's time records, which we report in Table 4. The semiautomated approach outperforms all other experimental groups, as an employee just needs to select and revise the output texts, enabling a single employee to significantly increase his or 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 content unit decreases from the company's current cost of 272.81€ to 23.94€ (including labor, server and system maintenance, and initial software development investment cost) using the semiautomated procedure. Since the method runs on an automated basis, a data scientist is not needed on an ongoing basis to fine-tune the model. 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 semiautomated method were available, our proposed method would have resulted in a cost of 10,511€, resulting in a savings of 109,254€ (~91%). Should an organization need to hire a data scientist or programmer to develop the algorithm, this estimate offers an upper bound on the salary that would make it worthwhile. This also demonstrates the opportunity for a cloud-based service to offer such algorithms to support content marketing.

Application to the Education Sector

We conduct a second field study with a large, internationally recognized public business school. In this study, two employees revised 30 pieces of machine-generated content, each targeted at an industry-specific keyword (e.g., "master program in marketing") and replaced the existing content (produced by an SEO expert) that targeted the same keyword. The median amount of content changed by the employees is 83 words (10.51%, median length-revised-machine = 824 words, IQR = 104.5) with a competitive investment in time (median reviser-time-investment = 1.30 hours, IQR = 1.07, min = 0.20, max = 3.13).

After observing the rankings of the SEO expert-generated content for 30 days from December 2019 onward, an employee replaced them with the semiautomated content. Similar to the IT service application, the semiautomated content outperforms human-generated content in search engine rankings. Figure 5 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 96 days after the semiautomated content was posted, the semiautomated content outperforms the previous content based on the mean ranking ($z = -7.06$, $r = -0.64$, $p < 0.000$) and the number of pages that appear on the first page of search results (i.e., the top 10) ($z = 7.98$, $r = 0.72$, $p < 0.000$). Additional details of this field study are provided in web Appendix 3.

Discussion

One of the ways that artificial intelligence has the potential to support marketing is by automating common tasks and consequently reducing the costs associated with completing them (Davenport et al. 2021). Yet, the potential for artificial intelligence to support content marketing has not been empirically assessed in the literature. Through two field studies, we demonstrate that NLG methods can support the

Table 4. Labor Time, Cost and Savings for Content Production

Category	Factor	Revised machine	Company (real)	Real SEO experts	Quasi experts	Novices
Human labor time for content production	Median (hours)	0.55	9.50	4.10	2.58	3.60
	IQR (hours)	0.23	3.69	1.80	2.58	3.55
	Min (hours)	0.28	4.50	1.00	0.90	0.80
	Max (hours)	1.20	21.50	7.00	28.80	12.00
Production output and cost per year	Produced content units p.y.	~1,880	~165	~382	~607	~435
	Production level (%)	~1,040	~100	~132	~268	~164
	Server cost per unit (€)	5	0	0	0	0
	Development cost spread (€)	2,436	0	0	0	0
	Maintenance cost per year (€)	3,480	0	0	0	0
	SEO labor cost per unit (€)	15.79	272.81	117.74	74.09	103.38
	Total cost per unit (€)	23.94	272.81	117.74	74.09	103.38
	Cost for 165 units (€)	~3,949	~45,000	~19,421	~12,221	~17,052
	Cost for 1,880 units (€)	~45,000	~512,756	~221,296	~139,161	~194,306
Possible real financial impact (2015 to 2019)	Produced content units	439	439	439	439	439
	Cost (€)	~10,511	~119,765	~51,688	~32,525	~45,384
	Possible savings (€)	~109,254		~68,075	~87,238	~74,380

Notes. Assumed employees' labor time and salary: 39 hours per week, 1,567 hours per year, 45,000€ per year; Amazon AWS costs per piece of content generated: 5€. Additional programmer for system maintenance: one month of work per year, 3,480€ cost per year. Initial development cost spread: one software developer, seven months of development, cost spread over an assumed system running time of 10 years, 2,436€; Calculation details are in web Appendix 2.7.

production of marketing content. Coupling machine learning and NLG with a content editor, we propose a semiautomated approach for SEO content generation. Not only is the output similar to human-generated content along a number of linguistic dimensions, but it outperforms manual content creation in search engine rankings, production efficiency, and engagement.

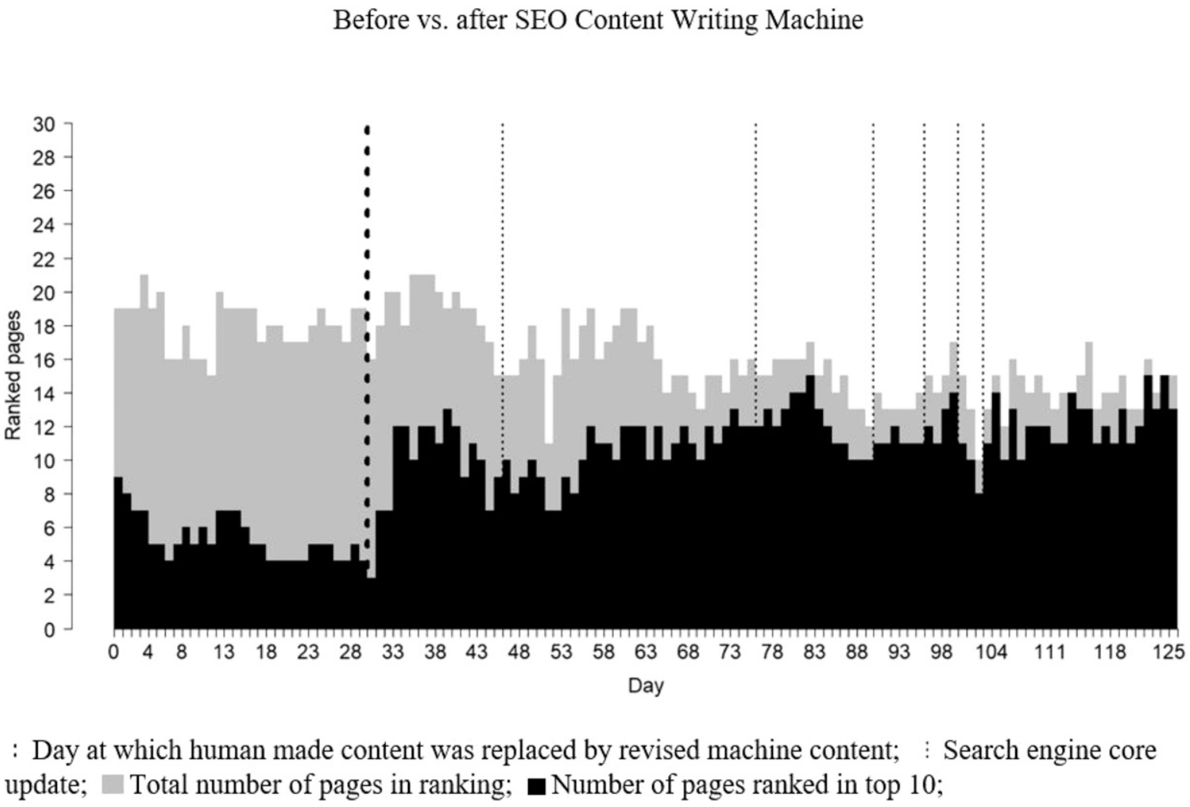
Though a human editor needs to only make minor changes, this individual's role remains essential to add sufficient value to machine-generated text. This perspective is consistent with the position taken by search engines such as Google on automatically generated text.[5] The automated aspect of our approach is designed to mimic the content that performs well in search engines. Based on the search results that are incorporated during the fine-tuning process and the massive corpus on which the language model was developed, it probabilistically outputs text in likely sequences. However, our algorithm does not evaluate the meaning and/or veracity of this output. It is only through a human editor that the content is vetted prior to its use. Moreover, our algorithm currently does not consider factors such as brand personality (e.g., Aaker 1997) or voice (e.g., Carnevale et al. 2017).

Research may extend our methodology by considering multiple textual inputs to inform the substantive content and its tone. A brand's own language, for example, may reveal its personality (e.g., Aaker 1997). There are promising techniques emerging, enabling future model users to possibly accomplish that (e.g., Dathathri et al. 2020). Our method could also be generalized to optimize bundles of multiple keywords simultaneously. This can be achieved, for example, by fine-tuning on the top-ranked content of many related keywords, and modifying the quality score function to differentiate between the main keyword and sub-keywords. Another option would be to modify the quality score components or estimating quality component weights dynamically to account for changes to search algorithms. Future work could also strive to generate content for multiple communication channels.

As consumers become more accustomed to interacting with machine-generated content, it will be important to monitor how consumers react to such interactions. Additional research is needed to understand how consumers react to machine-generated content throughout the customer journey (Puntoni et al. 2021). Although consumers may react favorably to the automation of certain types of content, they may view other types of machine-generated content less favorably. Consumer reactions may also differ based on the industry making use of machine-generated content.

As automation is applied to more marketing tasks, there are broader implications that must be considered. The ability to reduce the costs associated with content marketing suggests that pricing can be reduced or output increased. By aligning the semiautomated process's workflow with a specific objective (i.e., higher search engine rankings), our results demonstrate the potential to increase the return on investment. Given the choice between manually created, semiautomated, and eventually fully automated content, future research should examine the competitive equilibrium in terms of how firms will position themselves with the content they employ. Widespread adoption of NLG could result in

Figure 5. Number of Pages in Ranking and in the Top 10 per Day (Education Sector)



more homogeneous content, which may increase consumer search costs and require that firms find ways to differentiate themselves. This could involve making the brand personality more salient and using richer media. As consumers react to such content, choosing the winners and losers, machine-generated content may require substantial human edits, further underscoring our view that NLG can support (but not replace) content creators.

Semiautomated content production also has workforce implications. As with other forms of marketing automation, the demand for labor to perform some tasks will diminish (Brynjolfsson and Mitchell 2017). Although there will be less demand to produce initial drafts, there may be increased demand for those who can effectively edit automated content. More nuanced and differentiated styles may become an increasingly important component of brand voice. There may also be increased demand for those who can foresee the negative consequences associated with using content that is ill-suited for its purpose (Wilson et al. 2017).

Endnotes

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

[2] We use the open source Generative Pre-Trained Transformer 2 (GPT-2) model as an essential component of our method, which is available at https://github.com/minimaxir/gpt-2-simple. Our approach could be generalized to make use of other language models (e.g., GPT-3) when access is made available.

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

[4] We opt for a value of $T = 10$ because many search engines by default display the top 10 organic search results for a given query on the first result page.

[5] See https://developers.google.com/search/docs/advanced/guidelines/auto-gen-content.

References

Aaker JL (1997) Dimensions of brand personality. *J. Marketing Res.* 34(3):347–356.

Ascarza E, Ross M, Hardie BGS (2021) Why you aren't getting more from your marketing AI. *Harvard Bus. Rev.* (July–August), https://hbr.org/2021/07/why-you-arent-getting-more-from-your-marketing-ai.

Azzopardi L, Thomas P, Craswell N (2018) Collins-Thompson K, Me Q, eds. Measuring the utility of search engine result pages: An information foraging based measure. *SIGIR '18: 41st Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval*, Ann Arbor, MI, July 8–12, 2018 (Association for Computing Machinery, New York), 605–614.

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

Baye MR, De Los Santos B, Wildenbeest MR (2016) Search engine optimization: What drives organic traffic to retail sites? *J. Econom. Management Strategy* 25(1):6–31.

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

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

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

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, et al. (2020) Language models are few-shot learners. *Adv. Neural Inform. Processing Systems* 33:1877–1901.

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

Carnevale M, Luna D, Lerman D (2017) Brand linguistics: A theory-driven framework for the study of language in branding. *Internat. J. Res. Marketing* 34(2):572–591.

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

Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, Yosinski J, Liu R (2020) Plug and play language models: A simple approach to controlled text generation. *ICLR 2020: 8th Internat. Conf. Learn. Representations*, virtual conference. https://iclr.cc/virtual_2020/poster_H1edEyBKDS.html.

Davenport TH, Gua A, Grewal D (2021 How to design an AI marketing strategy. *Harvard Bus. Rev*. (July–August), https://hbr.org/2021/07/how-to-design-an-ai-marketing-strategy.

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

Google (2020) Optimize your content. *Search Engine Optimization (SEO) Starter Guide* (Mountain View, CA). https://support.google.com/webmasters/answer/7451184?hl=en.

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

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

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

Longoni C, Cian L (2020) Artificial intelligence in utilitarian vs. hedonic contexts: The 'word-of-machine' effect. *J. Marketing* 86(1)91–108.

Luh CJ, Yang SA, Huang TLD (2016) Estimating Google's search engine ranking function from a search engine optimization perspective. *Online Inform. Rev.* 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 Sci.* 38(6):937–947.

Marchenko OO, Radyvonenko OS, Ignatova TS, Titarchuk PV, Zhelezniakov DV (2020) Improving text generation through introducing coherence metrics. *Cybernetics Systems Anal.* 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.

Pennebaker JW, Booth RJ, Boyd RL, Francis ME (2015) *Linguistic Inquiry and Word Count: LIWC2015* (Pennebaker Conglomerates, Austin, TX), www.LIWC.net.

Pitler E, Nenkova A (2008) Revisiting readability: A unified framework for predicting text quality. Lapata M, Ng HT, eds. *Proc. 2008 Conf. Empirical Methods Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA), 186–195.

Puntoni S, Reczek RW, Giesler M, Botti S (2021) Consumers and artificial intelligence: An experiential perspective. *J. Marketing* 85(1):131–151.

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) *Improving Language Understanding by Generative Pre-training* (OpenAI), San Francisco.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) *Language Models Are Unsupervised Multitask Learners* (OpenAI), San Francisco.

Roberts C (2010) Correlations among variables in message and messenger credibility scales. *Amer. Behav. Scientist* 54(1):43–56.

Rocklage MD, Rucker DD, Nordgren LF (2018) Persuasion, emotion and language: The intent to persuade transforms language via emotionality. *Psych. Sci.* 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. Ridda M, ed. *Proc. 9th Internat. Conf. Inform. Systems Tech.* (ICIST) (Association for Computing Machinery, New York), 1–8. https://dl.acm.org/doi/10.1145/3361570.3361578.

Sheffield JP (2020) Search engine optimization and business communication instruction: Interviews with experts. *Bus. Professional Comm. Quart.* 83(2):153–183.

Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Sci.* 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. Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. *NIPS'17: Proc. 31st Internat. Conf. Neural Inform. Processing Systems* (Curran Associates Inc., Red Hook, NY), 6000–6010.

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