Original Paper

Harnessing Artificial Intelligence for Health Message Generation: The Folic Acid Message Engine

Ralf Schmälzle, PhD; Shelby Wilcox, BA, MA

Department of Communication, Michigan State University, East Lansing, MI, United States

Corresponding Author:
Ralf Schmälzle, PhD
Department of Communication
Michigan State University
404 Wilson Rd
East Lansing, MI, 48824
United States
Phone: 1 (517) 353 ext 6629
Email: schmaelz@msu.edu

Abstract

Background: Communication campaigns using social media can raise public awareness; however, they are difficult to sustain. A barrier is the need to generate and constantly post novel but on-topic messages, which creates a resource-intensive bottleneck.

Objective: In this study, we aim to harness the latest advances in artificial intelligence (AI) to build a pilot system that can generate many candidate messages, which could be used for a campaign to suggest novel, on-topic candidate messages. The issue of folic acid, a B-vitamin that helps prevent major birth defects, serves as an example; however, the system can work with other issues that could benefit from higher levels of public awareness.

Methods: We used the *Generative Pretrained Transformer-2* architecture, a machine learning model trained on a large natural language corpus, and fine-tuned it using a data set of autodownloaded tweets about *#folicacid*. The fine-tuned model was then used as a *message engine*, that is, to create new messages about this topic. We conducted a web-based study to gauge how human raters evaluate AI-generated tweet messages compared with original, human-crafted messages.

Results: We found that the *Folic Acid Message Engine* can easily create several hundreds of new messages that appear natural to humans. Web-based raters evaluated the clarity and quality of a human-curated sample of AI-generated messages as on par with human-generated ones. Overall, these results showed that it is feasible to use such a message engine to suggest messages for web-based campaigns that focus on promoting awareness.

Conclusions: The *message engine* can serve as a starting point for more sophisticated AI-guided message creation systems for health communication. Beyond the practical potential of such systems for campaigns in the age of social media, they also hold great scientific potential for the quantitative analysis of message characteristics that promote successful communication. We discuss future developments and obvious ethical challenges that need to be addressed as AI technologies for health persuasion enter the stage.

(J Med Internet Res 2022;24(1):e28858) doi: 10.2196/28858

KEYWORDS

human-centered AI; campaigns; health communication; NLP; health promotion

Introduction

Background

Crafting a successful health message involves a mix of art and science. On the one hand, decades of research in linguistics and communication science provides numerous insights into coherent sentence structure, effective value propositions, and other language-specific factors that promote attention, memory, and engagement [1,2]. On the other hand, translating these abstract factors into an appealing, concrete message that could be used in a campaign still requires a leap that must be fueled by human creativity and intuition [3].

Moreover, as larger and longer-term campaigns usually require a multitude of diverse messages, message creation represents a resource-intensive bottleneck. Although computers are often able to increase efficiency related to message development (ie, information gathering, collaborative environments, and graphic designs), the task of message creation was traditionally beyond their scope. Until a few years ago, computers could analyze a sentence and flag errors; however, they were not able to synthesize a meaningful new sentence. However, the latest advances in machine learning (ML) have equipped computers with the ability to generate language for messages that appear natural and readily comprehensible to humans. This work is highly relevant to health communication in general and campaigns in particular as it could be connected to the task of campaign message generation. Specifically, there is the possibility that language generation methods might help in creating and optimizing messages; however, as there has been little contact between the fields of health communication and language generation, more work is needed to examine this possibility. In this paper, we ask, “How feasible is it to automatically generate on-topic messages that could potentially promote awareness about specific health issues?”

In the following section, we first review how the internet and social networking sites have become part and parcel of health communication. Next, we present the health issue of folate or folic acid (FA) as our test case and highlight the need for campaigns to promote FA awareness. We then introduce recent studies on natural language generation (NLG). This leads to a study in which we use a data set of FA-related social media messages to train a *message engine*, which then generates hundreds of new messages about this topic. We evaluate the clarity and quality of these artificial intelligence (AI)–generated messages compared with human-generated content via a web-based study.

The Potential of Social Media Communication Campaigns to Raise Awareness About Specific Health Topics

Social media has become a key component of communication campaigns [4]. This development has enabled new forms of health communication that are more direct and engaging for users. Social media–based messaging has also led to unprecedented opportunities for optimizing and effectively delivering information to the masses via computationally heavy approaches such as A/B-testing, recommender systems, and targeting receiver characteristics or social network positions [5-10]. Social media can diffuse messages widely across the globe and deeply into interpersonal networks [11,12].

The role of social media within the health communication landscape is still evolving; however, almost all health campaigns have embraced social media as cost-effective and highly scalable channels for raising and sustaining public attention [4,13]. Specific health issues that are affected by a chronic lack of awareness can benefit substantially from social media awareness campaigns. This is perhaps most prominently demonstrated by the success of the amyotrophic lateral sclerosis ice water bucket challenge, which brought substantial public awareness to the disease of amyotrophic lateral sclerosis and encouraged donations to research.

Raising awareness and providing basic information is a critical first step toward prevention, considering that all health communication theories posit that if people are unaware of a specific health risk, they will not take preventive action [14]. Of course, many complex health behaviors involve factors beyond awareness and education, such as shifting norms and attitudes or persuading target audiences to engage in specific behaviors [3,15]. However, for some selected health problems, awareness and knowledge deficits can be the primary campaign goals [16,17], and for many others, raising awareness or keeping the issue on the public agenda [18] is at least a secondary goal. Thus, although we are not claiming that raising awareness is a cure-all solution, we consider it a critical first step for any message generation system.

The Case of FA Awareness

Simply raising awareness and providing essential knowledge can go a long way for prenatal health. Many people who are pregnant are intrinsically motivated to adhere to health recommendations if they know them, as can be measured via self-report and behavioral indicators, such as smoking quitting attempts, reduction in drinking, and changes in exercise and nutrition behaviors [19-21]. This includes eating a folate-rich diet (to minimize the risk of neural tube defects [NTDs]) or avoiding rare meat (risk of toxoplasmosis) and certain cheeses (to reduce the risk of listeria infection). However, awareness about FA and knowledge about FA-rich diets among women of childbearing age remain too low [22-24]. This is problematic as most pregnancies are only noticed after NTDs occur, such that once people learn about effective prevention behaviors during, for example, a physician’s visit, it may be too late [25,26]. Therefore, the issue of FA awareness will serve as a proof-of-concept example to demonstrate the potential of AI-generated messages that could potentially be used to raise awareness by providing a steady feed of on-topic but novel messages in long-term health campaigns.

Folate is a vitamin that is required for the body to build cells [27]. Many fruits, vegetables, and other natural foods contain folate, and the synthetic form, *FA*, is used as a dietary supplement or food additive. A folate or FA deficiency during early pregnancy can lead to severe embryonal NTDs [28]. Thus, the World Health Organization and the Centers for Disease Control and Prevention (CDC) recommend that all women of childbearing age consume 400 µg of folate per day [29,30].

Lack of awareness about FA represents a problem that is, at least to some degree, preventable via health communication and education [22,31-33]. As argued above, most people who are pregnant are motivated to achieve FA supply but will only be able to follow the guidelines if they are aware of them in the first place. Moreover, the recommended steps are relatively easy to follow for many people. However, that is not to say that by simply raising awareness, all positive downstream effects would follow. As with most health behaviors, they are embedded in a biopsychosocial context, requiring, for example, availability of food or FA supplementation, cultural factors, and so forth. However, the basic problem constellation of lack of awareness, paired with a relatively high spontaneous motivation and high self-efficacy and response efficacy, suggests that mass media health campaigns are a promising strategy. Indeed, several previous studies support that FA-related campaigns can produce positive effects [22,31,33].

New Challenges for Social Media Communication Campaigns

The key benefit of mass media campaigns on social media is that they can quickly disseminate messages into the homes of millions. Moreover, social media has made it much easier to reach specific audience demographics and keep track of relevant outcomes, such as whether messages are seen, shared, or commented on [3,4].

However, although campaigns are a highly scalable tool, conducting a successful campaign is still far from trivial and requires substantial monetary investment and sustained effort over a longer period [3,15,34]. When it is properly conducted, mass communication is highly cost-effective compared with other approaches [35,36], and most campaigns do not achieve high levels of exposure over a sustained period [35]. For instance, most campaigns only achieve approximately 40% exposure in their target audience [37], which naturally reduces their success as communication effects logically require that messages are seen in the first place [38]. Moreover, in the days of print, radio, and television campaigning, many campaigns comprised only a limited number of messages that were switched at a relatively slow rate (eg, weekly or monthly), if at all. Although the more professional campaigns nowadays feature *feeds* with dozens of messages, if not more [4], maintaining such an effort is very costly and requires dedicated personnel, formative processes, and summative evaluation throughout [39,40]. In summary, campaign creation and maintenance is an effortful business.

However, even campaigns that are executed skillfully have difficulties in reaching their audience. The low 40% exposure rate mentioned above came from a study published in 2004; however, since then, the internet has further exacerbated the competition for attention [41-43]. Specifically, the very nature of today's attention economy on social media requires that health communicators update content frequently. Otherwise, algorithms will downvote the content and make it less likely to be seen by the target audience [43,44]. Similarly, on the side of the audience, switching behavior and searching for novel information are very widespread [45]. In summary, the logistic effort needed to create campaign messages and ensure their constant dissemination, as well as the algorithmic and user-sided information selection decisions, pose challenges for maximizing the potential of health communication campaigns.

Overall, this situation invites new approaches that could help health communicators and practitioners create a large number of awareness messages, which could then be automatically scheduled to ensure a constant and variable feed of appealing and timely messages. The following section introduces how recent developments in NLG, a subfield of machine learning or AI research, offer a potential solution to message development and dissemination limitations.

The Potential of Language Models to Generate Domain-Specific Health Messages

Advances in natural language processing have made it increasingly possible to generate coherent messages [46]. Although enthusiasm and skepticism about using computers for text generation have waxed and waned for decades [47], the advances in the past decade have been particularly impressive as the quality of computer-generated texts is now at a level that makes it often indiscriminable from human-written text [48,49].

A model that attracted substantial public attention is the *Generative Pretrained Transformer-2* (GPT-2) [50]. In brief, GPT-2 is a deep learning–based ML model that performs expertly across several language-related tasks, such as text translation and summarization, question answering, and text generation [51,52]. Approximately 40 GB of data from >8 million webpages were used to train the basic model. GPT-2 comes in 4 sizes ranging from 124M, 355M, 774M, and 1.5B parameters. Humans generally find the output of GPT-2's text generations authentic and interesting. Notably, the model is publicly available and can be adapted to many text-based tasks, such as summarization, question answering, or generation.

Pretrained language models can be fine-tuned to specific domains [53]. Fine-tuning is a form of transfer learning in which an ML model trained on domain-general data is retrained on further domain-specific data to adapt to its particularities. The possibility of using such fine-tuned language models to generate domain-specific text has already been demonstrated across different disciplines [54,55]. However, we are not aware of any effort to examine this in the context of health communication. Thus, the question is whether fine-tuning GPT-2 to the domain of FA messages will enable it to generate new messages that are of sufficient clarity and quality to be useful for a potential social media health campaign.

Present Study

This research examines the capability of language models to generate realistic messages about FA, which could serve as suggestions for a potential health campaign. Furthermore, we ask whether this is realistic in the context of public health communication, a situation often characterized by a lack of funds and computational resources. In brief, we use messages, or tweets, from the popular message-sharing platform Twitter to fine-tune a GPT-2 model. Although the same approach can be used with other social media platforms, Twitter offers a relatively straightforward, mainly text-based message format with a 280-character limit and easy access to existing messages, making it the most promising candidate for piloting such a system. After downloading messages and training the message engine, we use the fine-tuned model as a *message engine* to generate a large number of new FA-specific tweets. We then examine the characteristics of the generated messages to identify the preconditions for success and the current limitations of these generated messages. Finally, we wanted to know how AI-generated messages would be compared against human-generated messages. To this end, we conduct a web-based study in which human judges evaluate the AI-generated and human-generated tweets in terms of clarity and quality.

Methods

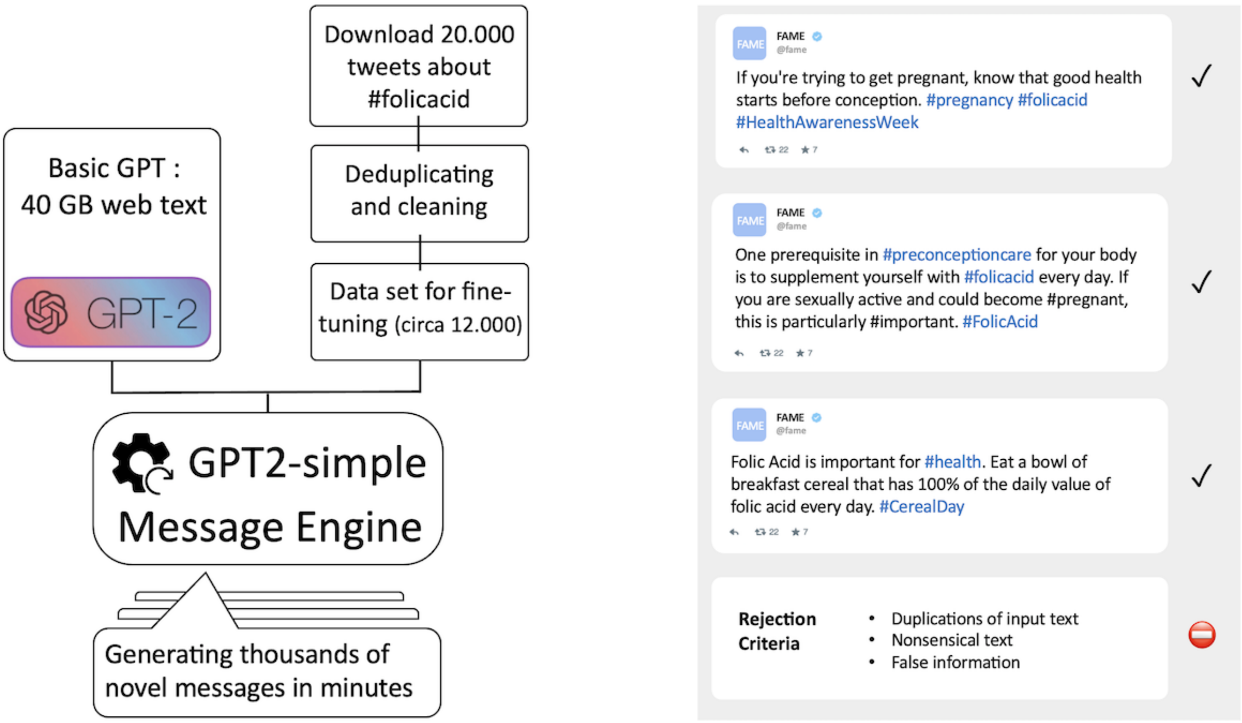
The FA Message Engine: Overall System Description

Overview

In this study, we harnessed the latest advances in AI to build a system that can generate a near-infinite number of health messages to promote FA awareness—the FA Message Engine. What will further be called the *message engine* is essentially an instantiation of the GPT-2 simple system, a Python package dedicated to fine-tuning OpenAI's GPT-2 text generation model [56]. We used the medium-sized GPT-2 model (355M hyperparameter versions trained on 40 GB of web text) and fine-tuned it with a data set of autodownloaded tweets about *#folicacid*. The resulting model was used to generate new messages (Figure 1).

Our specific steps are discussed in the following sections.

Figure 1. The left panel provides a schematic overview of the message engine construction and message generation workflow. The right panel illustrates a few examples of candidate messages. GPT: Generative Pretrained Transformer; GPT-2: Generative Pretrained Transformer-2.


Download 20.000 tweets about #folicacid
Deduplicating and cleaning
Data set for fine-tuning (circa 12.000)
Basic GPT : 40 GB web text
GPT-2
GPT2-simple Message Engine
Generating thousands of novel messages in minutes
FAME
@fame

If you're trying to get pregnant, know that good health starts before conception. #pregnancy #folicacid #HealthAwarenessWeek ✓

One prerequisite in #preconceptioncare for your body is to supplement yourself with #folicacid every day. If you are sexually active and could become #pregnant, this is particularly #important. #FolicAcid ✓

Folic Acid is important for #health. Eat a bowl of breakfast cereal that has 100% of the daily value of folic acid every day. #CerealDay ✓

Rejection Criteria

- Duplications of input text
- Nonsensical text
- False information

Scraping Tweets for Model Retraining

To obtain a data set to fine-tune the model, we used the Twitter Intelligence Tool [57] to scrape a large number of tweets that mentioned *#folicacid* in their text. Specifically, we downloaded 25,304 tweets posted between 2010 and 2020 that mentioned *#folicacid* and extracted the raw text of the tweets. After removing duplicates, non-English tweets, and tweets that mainly promoted nutritional supplements (stopwords: buy, order, or sale), we ended up with a data set of 11,311 unique tweets for fine-tuning.

Fine Tuning the GPT-2

After downloading and cleaning the messages to create a data set for fine-tuning, the next step involved preparing and retraining the GPT-2 model. Specifically, we submitted the data set for fine-tuning to retrain the 355M GPT-2 model with recommended default settings of 2000 training steps and a learning rate of α=.0001. Fine-tuning was accomplished via Jupyter Notebook running Python 3 on a computer equipped with a graphics processing unit and executing the gpt2.finetune()-Method from the gpt-simple package [56]. On a standard graphics processing unit–equipped computer, fine-tuning a model of this size takes approximately 1 hour. We also conducted pilot experiments with other model sizes but chose to put only the medium-sized 355M GPT-2 model for a user test. Larger models require advanced hardware, whereas the medium model can work with most cloud-based computing services available to end users. Larger models are also not recommended for generating short text messages, such as tweets. After training, the fine-tuned model, which constitutes the *message engine*, was saved to the disk.

Generating Candidate Tweets via the Message Engine

We used the *message engine* with a default temperature setting of *t*=0.7 to generate 1000 new tweets. Temperature settings influence the randomness of the textual output, with a lower temperature being less random. As for model size, we conducted pilot tests with different temperature settings but noted that higher settings (*t*=1.0) produced very incoherent output, and low settings (*t*=0.3 and *t*=0.5) led to text that was very close to the training data. Given that our goal was to test the engine's output in humans in terms of clarity and quality, we deemed it worthwhile to conduct user testing for this setting, which is also the recommended default setting as per gpt2-simple's documentary [56].

Evaluating AI-Generated Messages: Web-Based Study

To evaluate the clarity and quality of tweets generated by the *message engine* against human-generated tweets, we performed a web-based study. The procedure was devised based on the emerging guidelines for evaluating NLG studies [58] and is described in the following sections.

Message Selection

From the 1000 AI-generated tweets, we drew a random sample of 60 tweets. Next, a human editor curated these tweet suggestions and compiled them into a set of 30 tweets for the web-based study. The human editor rejected AI-tweet suggestions if they contained duplications from the input data, false information according to CDC guidelines, or problematic advice (see the following section for details). A second human curator confirmed this selection without contradictions.

In parallel, we drew a random sample of 30 tweets from a pool of >10,000 real-life tweets. This strategy was chosen as it is not feasible to evaluate thousands of tweets and as it most likely mimics how practitioners would use such a system [49]. Thus, this procedure yielded 2 sets of 30 tweets each—30 AI-generated messages that came from a pool of 60 randomly drawn samples and 30 human-generated messages from Twitter.

Participants

We recruited 150 young adults from a web-based pool at a large Midwestern university to evaluate these messages in terms of clarity and quality. Study participants received course credit as reimbursement for completing the short survey, which lasted approximately 20 minutes and was approved by the local institutional review board. Of the 150 young adults, after excluding data from participants who did not finish the survey or responded unrealistically fast and clicked through the survey, we ended up with a data set of 129 (86%) respondents (mean age 20 years; range 18-28 years). The sample was predominantly female (96/129, 74.4%). Although this sample was not intended to be representative of the population, our participants clearly belonged to the audience of a potential FA awareness campaign. Moreover, given that the goal was to evaluate message clarity and quality rather than message effects on attitudes or behavior, this sample is sufficient for this purpose.

A power analysis suggested that a sample of approximately 100 raters was sufficient to detect a small-to-moderate effect in terms of the mean difference in evaluations of AI-generated and human-generated tweets ($1-\beta=0.9$; $\alpha=.05$; $dz=0.3$) [59]. Moreover, message evaluation studies suggest that evaluations of individual messages stabilize after averaging data from approximately 25 to 30 raters per message [60], which we surpassed with this sample size.

Procedure

The survey was administered via Qualtrics software (Qualtrics International), and participants were asked to evaluate all messages regarding clarity and quality. Participants were told that the study's goal was to examine human evaluations of Twitter messages about FA or folate, such as whether they considered the messages adequate to raise awareness or educate audiences about this health issue. The test messages were presented randomly, and participants were unaware of whether they came from the pool of AI-generated or human-generated messages. Each message was evaluated on 2 questions, 1 focusing on message clarity ("Please evaluate this message in terms of whether it is clear and easy to understand.") and 1 on message quality ("How much do you agree that the content and quality of this message is appropriate to increase public knowledge about folic acid?"). Answers were collected using a 5-point Likert-style response format (*very clear* and *very unclear* and *strongly agree* and *strongly disagree*). At the end of the survey, participants were debriefed about the study's purpose and provided a link to the CDC's website for the most up-to-date information on FA.

Evaluating AI-Generated Messages: Computational Analyses

In addition to inspecting the AI-generated messages and performing a web-based evaluation study, we conducted several computational analyses. Specifically, we computed n-grams and inspected their distribution between AI- and human-generated messages, including visualizations as word clouds. Next, we performed topic modeling analyses to gain additional insights into the semantic structure. Topic modeling is a prominent method for identifying health topics in social media [61] or subtopics within a given health domain [62-64]. Specifically, we used the topicmodels package [65] within the R statistical software to compute the latent Dirichlet allocation topic models [66]. Finally, we assessed the semantic similarity of individual messages via the sentence-transformers package [67]. To this end, we transformed each message into a sentence embedding and compared different messages via cosine-vector similarity.

Results

Overview

We found that the fine-tuned GPT-2 model can act as a *message engine* by creating grammatically correct, coherent, and novel messages centered on the topics of FA, healthy nutrition, and pregnancy. In the following sections, we will first describe the insights gained during the overall procedure and qualitative characteristics of the generated output, followed by the web-based evaluation study's quantitative results.

Feasibility of the System and Qualitative Description of the AI-Generated Messages

Our overall research question focused on whether it is possible to fine-tune a language model such as GPT-2 to a specific health domain to build a *message engine*. The answer is that it is possible. As can be seen from the sample output in Figure 1, the *FA Message Engine* was able to generate 1000 tweets within a matter of minutes, most of which resembled authentic web-based messages in style and content.

We next asked whether training such a system is realistic in the context of public health, where computational resources and specialized coding skills are scarce. The answer to this question is that it is feasible and surprisingly easy to implement. Although developing the scraping, cleaning, and training procedure took some time, now that the system is set up, it can be replicated with little effort. For instance, if we wanted to replace the topic of *#folicacid* with any other health issue, this can be done in >1 hour. The system is also relatively accessible, even to novice users, as long as they are able to execute Python notebooks. Such skill requires only little training, and it would be possible to build a user interface for the system such that the user only enters the topic or search term (eg, *#folicacid*) and, after fine-tuning and generation, receives a sample of 60 message suggestions.

Most critically, we were interested in the characteristics of the generated messages, to which we turn next.

First, we note that the vast majority of the AI-generated tweets appeared natural and contained many elements of the original input tweets that were scraped from Twitter. For instance, the system uses hashtags that co-occurred with the search term *#folicacid*, such as *#pregnancy, #vitamin, #foodfortification #folicacidawarenessweek,* or *#eathealthy*. Second, as with hashtags, the system also tagged accounts that appeared in the input data, such as *@CDC* or *@NHS* (note that by eliminating these accounts from the input data, such information can be suppressed if not wanted).

Another observation is that most of the generated tweets were rather engaging, enthusiastic, or upbeat. This impression may again arise as the input tweets contained elements such as prompts with exclamation marks ("Eat healthy *now*!" and "*Go Folic! Visit [URL]!*") or encouragement, all of which could be interpreted as *cues-to-action* or attempts to raise *self-efficacy* according to the Health Belief Model [68]. This characteristic is likely as GPT-2 was trained on outgoing links with high so-called *karma scores* [50], thereby selectively emphasizing the language that web-based audiences found interesting and engaging.

Beyond resembling the linguistic style and platform-specific cues that are characteristic of today's Twitter environment (eg, upbeat language and hashtags), we observed that the AI-generated tweets reflected the input data's topic distribution. For instance, input tweets could be categorized into several topical clusters, such as nutritional needs during pregnancy, the link between FA and NTDs, political advocacy for mandatory food fortification, and so forth. Most AI-generated messages could also be categorized into coarse topic clusters. Additional results and visualizations of n-grams, word clouds, and results from topic modeling can be found in Multimedia Appendix 1 [65-67,69-73].

Although the overall system and procedure proved feasible, and the quality of many messages appeared comparable with human-generated messages, we made several observations that point to current limitations.

A simple observation is that the system sometimes parrots the training data; that is, it contains either duplicates of raw tweets or specific formulations that appeared in the data set used for fine-tuning (eg, "If you are trying to get pregnant..." and "Thinking of trying for a baby..."). This issue is well-known and follows logically from the fact that language models are essentially giant statistical association machines, which will learn the information contained in the input data. From an intellectual property perspective, this issue can raise questions about the copyright of the generated output. However, in practice, it is easy to sort out such parrot generations through human supervision, n-gram matching, or paraphrase detection algorithms.

A second limiting observation is that even when not directly parroting the training data, many tweets are still *close to* individual input messages, for example, by mixing formulations such as *trying to get pregnant* and *thinking of having a baby* with various combinations like *start taking #folicacid* or *know that good health starts before conception*. This points to mere reformulations and permutations that are not very creative. Again, this represents a direct consequence of the way natural language models work and is thus not necessarily a severe limitation. In fact, variations of a common message on a health topic may improve a campaign's reach by preventing the content from being downvoted by algorithms that select for novelty. Slight variations may also be beneficial in improving existing messages by making them briefer or more engaging, and it is well-known that repeated exposure to messages improves awareness and retention [38,74-76]. Rather than a limitation, this reformulation strategy of message generation can help to more optimally exhaust the space of possible effective messages. Nevertheless, it is clear that mere rewording represents only a minor achievement in message creation.

A third observation is that some generated messages contained false statements about which foods contain which amounts of FA or what medical defects might occur. Although such tweets are easy to spot in practice, this is an actual limitation. A total of 2 factors may underlie such behavior. First, if the input data contain false or problematic health claims, which are pervasive on social media, then the system will learn them. In this case, the system should not be blamed; however, the curation of the data set for fine-tuning should be optimized. However, more critically, the state of the art of current language models implies that they will simply generate tweets that sound linguistically coherent but may not make sense. We have discussed this issue in the *Discussion* section, where we suggest advancements to the system.

These results suggest that human curation and supervision of AI-generated tweets are necessary for practical use cases. Thus, a campaign manager or team would need to monitor the retraining process and eliminate problematic content, which is also what we opted for to select tweets for the web-based evaluation study.

Quantitative Comparison of AI- and Human-Generated Messages

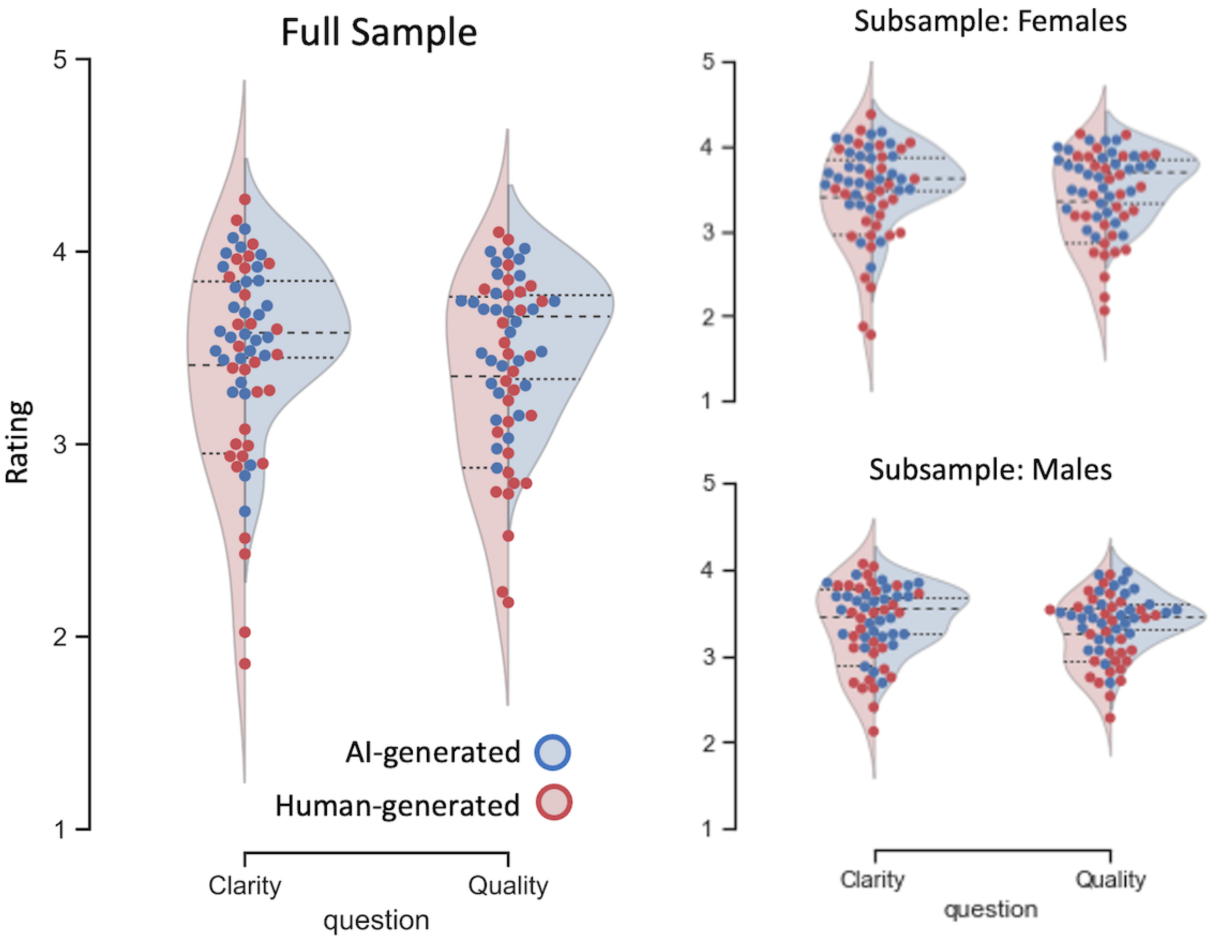
Table 1 shows the results of comparing AI-generated and human-generated tweets in terms of overall clarity or quality. Figure 2 illustrates the results graphically and provides further distributional information as well as analyses by subgroups. As can be seen, overall, the messages were rated as relatively clear and easy to understand, and participants found that their content and quality were appropriate for increasing public knowledge about FA (>3 on a 5-point scale).

Table 1. Means (SDs) from the web-based survey. Scores for message clarity and quality evaluations for 30 AI[a]- and 30 human-generated messages are shown, respectively.

Evaluations	AI-generated messages, mean (SD)	Human-generated messages, mean (SD)	*t* test (*df*)	P value
Clarity	3.58 (0.36)	3.34 (0.6)	1.97 (58)	.05
Quality	3.57 (0.32)	3.3 (0.53)	2.34 (58)	.02

[a] AI: artificial intelligence.

Figure 2. 30 artificial intelligence–generated tweets (blue) and 30 human-generated tweets (red) were evaluated in terms of perceived clarity and quality on a 5-point Likert-style scale. The results revealed very similar evaluations and minor average differences, with a considerable spread within each category. The right panels show analyses separated by gender. AI: artificial intelligence.



Statistical analysis revealed that the AI-generated and human-generated tweets were not rated as different in terms of how clear and easy they were to understand ($t_{clarity}$=1.97; P=.05). A small but statistically significant difference was found in the quality dimension ($t_{quality}$=2.34, P=.02). However, as can be seen from Figure 2, these mean differences were small in light of the variability, and thus, the effect was of a small size. We also zoomed in on the women participants' subgroup as the topic may be more relevant to them or become more relevant in the future with respect to pregnancy. As can be seen from Figure 2, the results were robust, and both subgroups (women and men) exhibited essentially the same pattern of results. However, we noted that, given the sample of college students, the topic of FA might not be very relevant to them, although there are also other health benefits of FA beyond the prevention of birth defects.

Next, we performed analyses at the level of the individual messages. As shown in Figure 2, the differences for both clarity and quality between individual messages were larger than the differences between categories (human- vs AI-generated). Indeed, performing item-wise analyses in which we compared ratings for single messages across raters using dependent-sample tests (because the same raters evaluated all messages) revealed that many messages were rated consistently higher than others. This pattern emerged both within human-generated and AI-generated messages, as well as across categories. Thus, many AI-generated messages were rated much higher than random human-generated messages.

Overall, we took these results as evidence that the *message engine* generated tweets that human raters evaluated as mainly equivalent to real Twitter messages.

Computational Analyses

In addition to the analyses of content (n-grams and topic modeling) and the human evaluation of clarity and quality, we wanted to examine the generated messages using computational methods. Specifically, we compared the 60 messages (30 AI-generated and 30 human-generated) using sentence Bidirectional Encoder Representations from Transformers (BERT), a modification of the pretrained bidirectional encoder representations from transformers model, to derive semantic sentence embeddings, which we then compared using cosine-similarity [67].

We found that across all messages, the average similarity was s=0.35. Within the 30 AI-generated messages, the average similarity was s_{AI}=0.37, and the average similarity between the 30 human-generated messages was $\mathrm{s}_{\mathrm{Human}}$=0.34. The average similarity between AI versus human messages was s_{AI} versus s_{Human}=0.35. Testing for differences between these computational indices of semantic similarities revealed no significant differences in any comparison (AI vs human and within- vs across-classes; all P>.08; for further details, see Multimedia Appendix 1 [65,66]). These results suggest that the sample of AI-generated messages is semantically similar to the sample of human-generated messages.

Discussion

Principal Findings

This study examined whether AI message generation technology can create candidate messages for use in social media health campaigns that focus primarily on raising awareness or increasing knowledge. We found that by retraining a GPT-2 model with thousands of tweets about FA, it is possible to build a *message engine* that can generate novel tweets, which could become part of an actual campaign. Human raters perceived these tweets as broadly similar in terms of clarity and quality to real-world messages. These results suggest that AI-assisted *message engines* could support campaign staff to create more efficient and possibly more effective campaigns for topics that are suitable for awareness-based messaging.

To our knowledge, this study is the first to demonstrate the potential of automated message generation in the context of health communication. Our results are generally positive, suggesting that the *FA Message Engine* can serve as a starting point for more sophisticated AI aides for message generation. Such systems can automatically offer thousands of messages that mimic the style and reflect the substance of existing health messages. Given that message creation is a resource-intensive bottleneck, we see significant application potential for such a system as a catalyst for human creativity [77].

Building a message engine for the topic of FA proved to be surprisingly easy. Our system made use of available tools [56] and could thus be transferred to contexts other than the issue of FA. Although this work did not intend to provide such a general purpose system for end users, it would be only a small step to deploy it as a web application as a turnkey solution.

Although the results of the web-based study demonstrate that the system output achieves good results, and the clarity ratings of AI-generated tweets are even significantly higher, we emphasize that our comparison strategy does not warrant the conclusion that AI-generated messages are superior to human-generated messages. Specifically, we compared a selection of 30 AI-generated messages against a sample of 30 real tweets, which were randomly drawn from a pool of >10,000 tweets. We opted for this procedure as our goal was to test the feasibility of AI-assisted message generation, which is the most realistic use case. In the following sections, we have discussed the significant limitations that currently prevent such a system from operating independently. However, the pool of human-curated AI-generated messages performed on par with or better than the standard tweets, and the analysis of semantic similarities did not reveal any difference. Thus, our approach suggests a simple strategy that might improve the quality of web-based content while saving the time and money of health communication practitioners.

Beyond the practical potential of such a *message engine* in the age of social media, the approach also offers considerable scientific potential. In particular, AI-based message engines might strengthen strategies to analyze message characteristics that underlie successful health communication [78-80]. In its current form, users of the message engine cannot influence the generated text's characteristics other than by what is fed in with the fine-tuning data set.

However, the natural language processing community [55,81,82] strives to gain more control over how the text is generated, and we see this as a promising next step. In particular, a limitation of the current system is that it does not incorporate any theory-based message design principles [1,83], such as barriers, cues to action, and norm or threat appeals. The fact that the current system learned to include some theory-compatible features, such as cues to action, shows promise in this regard; however, a more systematic approach is needed [84-86].

Ideally, this could then set off a virtuous cycle in which one could, via rapid iterations, gather feedback about specific message characteristics that are associated with targeted outcomes (eg, attention, awareness, and message sharing) and thus more clearly identify the message characteristics that facilitate individual outcomes [87]. As these characteristics become more accessible by linking objective message properties to large-scale outcomes, we might expect profound theoretical contributions from this otherwise applied system [88].

Along these lines, the most promising research direction is to fine-tune the fine-tuning process. We simply used the medium-sized GPT-2 model and fine-tuned the model with a set of tweets that were minimally screened. However, as with any manufacturing process, the quality of the input data determines the output. Thus, by fine-tuning the engine with often mediocre tweets, the AI-generated tweets were likely less potent than they would have been with a better training set.

In the future, we envision that one could curate a pool of high-quality tweets to serve as grade-A training material. An option, analogous to the strategy of training the GPT-2 base model with relatively more engaging text content, is to select only those tweets about *#folicacid* that have been retweeted or liked. Another option is to bootstrap messages by having domain experts reword or craft theory-based examples. However, a challenge for this strategy is that fine-tuning requires large amounts of text—a few hundred examples are not enough. Overcoming this challenge is feasible with a large pool of quality input data. In addition, such a message pool could be used to train message engines for domains other than the narrow issue of FA.

We conclude this section by emphasizing again that the primary use case of such systems lies in boosting awareness for selected health problems where awareness is lacking or waning. At this point, a message engine system does not yet solve trickier health communication problems, such as the habitual nature of many negative health behaviors, addressing the socioecological embeddedness of such behaviors, or how to change health-related attitudes [89]. In principle, we see no reason why such systems could not be expanded to contexts beyond social media, especially as reliance on voice assistants such as Amazon's Alexa, Alibaba's AliGenie, or Apple's Siri for information increases. However, the *message engine* presented here is primarily intended for mass communication about public health issues that are affected by low awareness. For such health issues, we envision that this system can improve the cost/benefit ratio and overcome the *message-creation bottleneck* to avoid web-based content from being algorithmically downvoted as it is considered not fresh, dull, or unengaging.

Limitations, Risks, and Avenues for Future Research

This study demonstrates a positive application of NLG technology; however, some risks and limitations are worth mentioning.

A very basic limitation is that our focus was on demonstrating the feasibility of a message engine to generate messages that could potentially become part of a campaign; however, we did not actually conduct such a campaign. Thus, although we are confident that we showed that the generated messages—after going through the human content curation process—are on par with human-generated *baseline* messages, we did not actually show that these messages improved public health and especially not with regard to more distal outcomes such as attitude change and behavior. This should be the topic for future research.

Similarly, we note that our sample comprised college students who were not intended to be representative of the larger population. However, given that our focus was on evaluations of message clarity and quality rather than more idiosyncratically defined responses, this sample seems appropriate. This is also underscored by the fact that evaluations were highly consistent across subgroups of women and men raters. Nevertheless, future work on, for example, message effects on attitudes, beliefs, and other variables beyond basic clarity and quality should also focus on outcomes in specific health audiences, such as people who intend to have a child.

Regarding broader implications and risks, recent events in the political domain have highlighted the danger of algorithmic bots deployed to create or spread misinformation [90-93]. Several malicious actors seem to be using natural language generators to produce fake or divisive messages; thus, several empirical studies have examined the dangers of using NLG technology to create content that is harmful to society [49,94]. The same problems arise concerning the marketing of products that might harm health or use bots to promote certain brands [95-97].

Our study speaks to these issues by showing that it is also possible that benevolent actors can use NLG methods to promote public health. As with all technologies, risk and benefit are correlated, because otherwise the technology would be abandoned [98]. As such, we hope that our study will help explore the potential of AI as a force for promoting positive outcomes. However, this does not mean that we advocate for a laissez-faire strategy. Instead, a discussion of the ethical consequences of these technologies is needed and ongoing [99,100]. However, the field of health communication seems to be a particularly strong example of how human-centered AI could be used for social good.

Another risk and major limitation of the system is its lack of common sense knowledge. People who are not familiar with NLG technology are sometimes ambivalent about the idea of a *message engine*, finding it both magical and critical. Many also expect the systems to operate in a human-like fashion; however, this is not at all the case. On the surface, the generated tweets have a human-like look and feel to them; however, closer inspection reveals that GPT-2–style language models lack the deeper understanding and reasoning capacity that would be necessary for calling them intelligent.

Indeed, some AI-generated tweets are ludicrous, and others contain false information that is presented as fact [101]. For example, a tweet that emerged with little pretraining material was, "Make sure you drink 4 breads of #folicacid per day!" Such examples reflect a lack of common sense knowledge that one cannot drink bread. These examples arise only as GPT-2, although very sophisticated, ultimately boils down to a statistical association machine that links the domains nutrition and FA but does not have knowledge about fluid versus solid substances. The computer programs' inability to draw connections between categories or classify knowledge outside the training set has been a fundamental challenge since the early days of AI research [102-104].

This issue becomes particularly sensitive when generations contain wrong medical advice, such as "Take 4 lbs of FA per day!" Again, this reveals that the language model only picks up on statistical regularities in how words are used; however, it possesses no actual knowledge about pregnancy, nutrition, quantities, the developing fetus, and the causal relations between these concepts. These challenges are relatively easy to overcome with human supervision and insight. However, a trickier issue concerns issues in which the underlying knowledge is still evolving or uncertain. Such situations can provide fertile grounds for health myths or speculations about side effects. Such information will enter the message engine if it is present in the training data set, which again emphasizes the need for human curation.

In addition to the lack of domain knowledge about health and biology beyond that provided in the training set, we have already pointed out that the system also has no theoretical understanding of communication and persuasion. The message engine only mimics and varies word use, albeit very eloquently. For a nonnative speaker who has to learn a foreign language for years, this skill may seem enviable; the ability to swiftly come up with 1000 sentences about FA may also impress and help health campaigners who spend hours coming up with 100 new candidate messages. However, the fact that such language models may talk without real understanding also means that they should only be used under the supervision of medical and communication experts. However, this is also true for more human-centric methods, such as focus groups or user-generated content.

Despite these limitations, the underlying technology can be expected to improve rapidly, and health communication researchers are well-advised to keep an eye on these developments. For instance, a successor model to GPT-2 has already been developed [105]. This model, called GPT-3, significantly improves some of the limitations that characterize GPT-2. Together with systems that are capable of generating persuasive arguments, selecting best-matching arguments for specific groups, and several other advances, we anticipate that the field of AI-assisted health message generation will see significant progress over the next decade [65,106-110].

Conclusions

To conclude, the *message engine* can generate candidate messages for human curators about selected health issues. This is relevant for issues where a lack of awareness is the primary problem, and a rich pool of social media messages is needed. At this stage, human supervision is necessary, and the technology, although very promising for content creation, requires control to select relevant content. Scientifically, this approach may promote a new wave of theoretical insights into the mechanisms of effective health messaging. We foresee that AI-generated messages for health promotion, education, and persuasion will become commonplace. It will be important for health communication researchers and practitioners to develop a strategy to use this technology positively.

Acknowledgments

The authors would like to thank Max Woolf for making the code for retraining the Generative Pretrained Transformer-2 available.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Inspection of word frequency, topic modeling results, and analysis of semantic similarity.
[DOCX File , 1928 KB-Multimedia Appendix 1]

References

1. Cho H. Health Communication Message Design: Theory and Practice. Thousand Oaks: SAGE; 2011.
2. Geeraerts D. Cognitive Linguistics: Basic Readings. Berlin: Walter de Gruyter; 2006.
3. Rice R, Atkin C. Public Communication Campaigns. Thousand Oaks: SAGE; 2012.
4. Shi J, Poorisat T, Salmon CT. The use of social networking sites (SNSs) in health communication campaigns: review and recommendations. Health Commun 2018 Jan 18;33(1):49-56. [doi: 10.1080/10410236.2016.1242035] [Medline: 27858464]
5. Kohavi R, Longbotham R. Online controlled experiments and A/B testing. In: Encyclopedia of machine learning and data mining. Boston: Springer; 2017.
6. Kim HS, Yang S, Kim M, Hemenway B, Ungar L, Cappella JN. An experimental study of recommendation algorithms for tailored health communication. Comput Commun Res 2019;1:103-129 [FREE Full text]
7. Matz SC, Kosinski M, Nave G, Stillwell DJ. Psychological targeting as an effective approach to digital mass persuasion. Proc Natl Acad Sci U S A 2017 Nov 28;114(48):12714-12719 [FREE Full text] [doi: 10.1073/pnas.1710966114] [Medline: 29133409]
8. Yom-Tov E, Shembekar J, Barclay S, Muennig P. Erratum: Author Correction: the effectiveness of public health advertisements to promote health: a randomized-controlled trial on 794,000 participants. NPJ Digit Med 2018 Aug 16;1:38 [FREE Full text] [doi: 10.1038/s41746-018-0047-z] [Medline: 31305589]
9. Gibson LA, Siegel L, Kranzler E, Volinsky A, O'Donnell MB, Williams S, et al. Combining crowd-sourcing and automated content methods to improve estimates of overall media coverage: theme mentions in e-cigarette and other tobacco coverage. J Health Commun 2019;24(12):889-899. [doi: 10.1080/10810730.2019.1682724] [Medline: 31718524]
10. Wang X, Chen L, Shi J, Peng T. What makes cancer information viral on social media? Comput Human Behav 2019 Apr;93:149-156. [doi: 10.1016/j.chb.2018.12.024]
11. Coronel JC, Ott JM, Hubner A, Sweitzer MD, Lerner S. How are competitive framing environments transformed by person-to-person communication? An integrated social transmission, content analysis, and eye movement monitoring approach. Commun Res 2020 Mar 01:009365022090359. [doi: 10.1177/0093650220903596]
12. Katz E, Lazarsfeld P, Roper E. Personal Influence: The Part Played By People in the Flow of Mass Communications. London: Routledge; 2017.

13. Zhao X. Health communication campaigns: a brief introduction and call for dialogue. Int J Nurs Sci 2020 Sep 10;7(Suppl 1):S11-S15. [doi: 10.1016/j.ijnss.2020.04.009] [Medline: 32995373]
14. Schmälzle R, Renner B, Schupp HT. Health risk perception and risk communication. Policy Insights Behav Brain Sci 2017 Aug 24;4(2):163-169. [doi: 10.1177/2372732217720223]
15. Pfau M, Parrott R. Persuasive Communication Campaigns. New York: Allyn and Bacon; 1992.
16. Dutka S, Colley R. DAGMAR, Defining Advertising Goals for Measured Advertising Results. Lincolnwood, IL: NTC Business Books; 1995.
17. Bettinghaus EP. Health promotion and the knowledge-attitude-behavior continuum. Preventive Med 1986 Sep;15(5):475-491. [doi: 10.1016/0091-7435(86)90025-3]
18. Jones K. Agenda setting in health and risk messaging. In: Oxford Research Encyclopedia of Communication. Oxford, United Kingdom: Oxford University Press; 2017.
19. Crozier S, Robinson S, Borland S, Godfrey K, Cooper C, Inskip H, et al. Do women change their health behaviours in pregnancy? Findings from the Southampton Women's Survey. Paediatr Perinat Epidemiol 2009 Sep;23(5):446-453 [FREE Full text] [doi: 10.1111/j.1365-3016.2009.01036.x] [Medline: 19689495]
20. Gaston A, Prapavessis H. Maternal-fetal disease information as a source of exercise motivation during pregnancy. Health Psychol 2009 Nov;28(6):726-733. [doi: 10.1037/a0016702] [Medline: 19916641]
21. Floyd RL, Rimer BK, Giovino GA, Mullen PD, Sullivan SE. A review of smoking in pregnancy: effects on pregnancy outcomes and cessation efforts. Annu Rev Public Health 1993;14:379-411. [doi: 10.1146/annurev.pu.14.050193.002115] [Medline: 8323595]
22. Rofail D, Colligs A, Abetz L, Lindemann M, Maguire L. Factors contributing to the success of folic acid public health campaigns. J Public Health (Oxf) 2012 Mar 03;34(1):90-99 [FREE Full text] [doi: 10.1093/pubmed/fdr048] [Medline: 21727078]
23. Medawar G, Wehbe T, Jaoude E. Awareness and use of folic acid among women of childbearing age. Ann Glob Health 2019 Apr 09;85(1):54 [FREE Full text] [doi: 10.5334/aogh.2396] [Medline: 30977622]
24. Green-Raleigh K, Carter H, Mulinare J, Prue C, Petrini J. Trends in folic acid awareness and behavior in the United States: the Gallup Organization for the March of Dimes Foundation surveys, 1995-2005. Matern Child Health J 2006 Sep;10(5 Suppl):S177-S182 [FREE Full text] [doi: 10.1007/s10995-006-0104-0] [Medline: 16823638]
25. Greene ND, Copp AJ. Neural tube defects. Annu Rev Neurosci 2014;37:221-242 [FREE Full text] [doi: 10.1146/annurev-neuro-062012-170354] [Medline: 25032496]
26. Cameron M, Moran P. Prenatal screening and diagnosis of neural tube defects. Prenat Diagn 2009 Apr;29(4):402-411. [doi: 10.1002/pd.2250] [Medline: 19301349]
27. Crider KS, Bailey LB, Berry RJ. Folic acid food fortification-its history, effect, concerns, and future directions. Nutrients 2011 Mar;3(3):370-384 [FREE Full text] [doi: 10.3390/nu3030370] [Medline: 22254102]
28. Geisel J. Folic acid and neural tube defects in pregnancy: a review. J Perinat Neonatal Nurs 2003;17(4):268-279. [doi: 10.1097/00005237-200310000-00005] [Medline: 14655787]
29. Folic Acid. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/ncbddd/folicacid/index.html [accessed 2022-01-10]
30. Periconceptional folic acid supplementation to prevent neural tube defects. World Health Organization. URL: https://www.who.int/elena/titles/folate_periconceptional/en/ [accessed 2022-01-10]
31. Chivu CM, Tulchinsky TH, Soares-Weiser K, Braunstein R, Brezis M. A systematic review of interventions to increase awareness, knowledge, and folic acid consumption before and during pregnancy. Am J Health Promot 2008;22(4):237-245. [doi: 10.4278/06051566R2.1] [Medline: 18421888]
32. Amitai Y, Fisher N, Haringman M, Meiraz H, Baram N, Leventhal A. Increased awareness, knowledge and utilization of preconceptional folic acid in Israel following a national campaign. Prev Med 2004 Oct;39(4):731-737. [doi: 10.1016/j.ypmed.2004.02.042] [Medline: 15351539]
33. van der Pal-de Bruin KM, de Walle HE, Jeeninga W, de Rover C, Cornel MC, de Jong-van den Berg LT, et al. The Dutch 'Folic Acid Campaign'--have the goals been achieved? Paediatr Perinat Epidemiol 2000 Apr;14(2):111-117. [doi: 10.1046/j.1365-3016.2000.00251.x] [Medline: 10791653]
34. Snyder LB. Health communication campaigns and their impact on behavior. J Nutr Educ Behav 2007;39(2 Suppl):S32-S40. [doi: 10.1016/j.jneb.2006.09.004] [Medline: 17336803]
35. Noar SM. An audience-channel-message-evaluation (ACME) framework for health communication campaigns. Health Promot Pract 2012 Jul;13(4):481-488. [doi: 10.1177/1524839910386901] [Medline: 21441207]
36. Hutchinson P, Wheeler J. The cost-effectiveness of health communication programs: what do we know? J Health Commun 2006;11 Suppl 2:7-45. [doi: 10.1080/10810730600973862] [Medline: 17148098]
37. Snyder LB, Hamilton MA, Mitchell EW, Kiwanuka-Tondo J, Fleming-Milici F, Proctor D. A meta-analysis of the effect of mediated health communication campaigns on behavior change in the United States. J Health Commun 2004;9 Suppl 1:71-96. [doi: 10.1080/10810730490271548] [Medline: 14960405]
38. Hornik RC. Exposure: theory and evidence about all the ways it matters. Soc Marketing Q 2016 Aug 01;8(3):31-37. [doi: 10.1080/15245000214135]

39. Farrelly M, Niederdeppe J, Yarsevich J. Youth tobacco prevention mass media campaigns: past, present, and future directions. Tob Control 2003 Jun;12 Suppl 1:i35-i47 [FREE Full text] [doi: 10.1136/tc.12.suppl_1.i35] [Medline: 12773784]
40. Roditis ML, Jones C, Dineva AP, Alexander TN. Lessons on addiction messages from "The real cost" campaign. Am J Prev Med 2019 Feb;56(2 Suppl 1):S24-S30 [FREE Full text] [doi: 10.1016/j.amepre.2018.07.043] [Medline: 30661522]
41. Simon H. Designing organizations for an information-rich world. In: Computers, Communications, and the Public Interest. Baltimore, Maryland, United States: Johns Hopkins Press; 1971.
42. Atchley P, Lane S. Chapter Four-cognition in the attention economy. In: Psychology of Learning and Motivation. New York: Academic Press; 2014.
43. Citton Y. The Ecology of Attention. London: John Wiley & Sons; 2017.
44. Attention Economy: Understanding the New Currency of Business. Brighton, Massachusetts: Harvard Business Review Press; 2002.
45. Jiang J, He D, Allan J. Searching, browsing, clicking in a search session: changes in user behavior by task over time. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014 Presented at: SIGIR '14: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval; Jul 6-11, 2014; Gold Coast Queensland Australia. [doi: 10.1145/2600428.2609633]
46. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
47. Weizenbaum J. Computer Power and Human Reason: From Judgement to Calculation. New York, United States: W.H.Freeman & Co Ltd; 1976.
48. Seabrook J. Can a machine learn to write for The New Yorker? The New Yorker. URL: https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker [accessed 2022-01-10]
49. Kreps SE, McCain M, Brundage M. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. J Exp Polit Sci 2020 Nov 20:1-14. [doi: 10.1017/XPS.2020.37]
50. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. URL: http://www.persagen.com/files/misc/radford2019language.pdf [accessed 2022-01-10]
51. Alammar J. The illustrated transformer. GitHub. URL: https://jalammar.github.io/illustrated-transformer/ [accessed 2022-01-10]
52. Radford A, Wu J, Amodei D, Amodei D, Clark J, Brundage M. Better language models and their implications. OpenAI Blog. URL: https://openai.com/blog/better-language-models/ [accessed 2022-01-10]
53. Peters M, Ruder S, Smith N. To tune or not to tune? Adapting pretrained representations to diverse tasks. arXiv. Preprint posted online March 14, 2019 [FREE Full text]
54. Frazier S, Al Nahian S, Riedl M, Harrison B. Learning norms from stories: a prior for value aligned agents. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2019 Presented at: AIES '20: AAAI/ACM Conference on AI, Ethics, and Society; Feb 7-9, 2020; New York, USA URL: https://dl.acm.org/doi/proceedings/10.1145/3375627 [doi: 10.1145/3375627.3375825]
55. Ziegler D, Stiennon N, Wu J, Brown T, Radford A, Amodei D, et al. Fine-tuning language models from human preferences. arXiv. Preprint posted online September 18, 2019 [FREE Full text]
56. Woolf M. gpt-2-simple. Github Repository. URL: https://github.com/minimaxir/gpt-2-simple [accessed 2020-10-01]
57. Twint. GitHub. URL: https://github.com/twintproject/twint [accessed 2022-01-10]
58. van Gatt DL, van Wubben ME, Krahmer E. Best practices for the human evaluation of automatically generated text. In: Proceedings of the 12th International Conference on Natural Language Generation. 2019 Presented at: Proceedings of the 12th International Conference on Natural Language Generation; 2019; Tokyo, Japan. [doi: 10.18653/v1/w19-8643]
59. Erdfelder E, Faul F, Buchner A. GPOWER: a general power analysis program. Behav Res Methods Instrum Comput 1996 Mar;28(1):1-11. [doi: 10.3758/bf03203630]
60. Kim M, Cappella JN. An efficient message evaluation protocol: two empirical analyses on positional effects and optimal sample size. J Health Commun 2019;24(10):761-769 [FREE Full text] [doi: 10.1080/10810730.2019.1668090] [Medline: 31543057]
61. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PLoS One 2014;9(8):e103408 [FREE Full text] [doi: 10.1371/journal.pone.0103408] [Medline: 25084530]
62. Pruss D, Fujinuma Y, Daughton AR, Paul MJ, Arnot B, Albers Szafir D, et al. Zika discourse in the Americas: a multilingual topic analysis of Twitter. PLoS One 2019 May 23;14(5):e0216922 [FREE Full text] [doi: 10.1371/journal.pone.0216922] [Medline: 31120935]
63. Zhou S, Zhao Y, Bian J, Haynos AF, Zhang R. Exploring eating disorder topics on Twitter: machine learning approach. JMIR Med Inform 2020 Oct 30;8(10):e18273 [FREE Full text] [doi: 10.2196/18273] [Medline: 33124997]
64. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill 2020 Nov 11;6(4):e21978 [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
65. Grün B, Hornik K. topicmodels: an R package for fitting topic models. J Stat Softw 2011;40(13):1-30 [FREE Full text] [doi: 10.18637/jss.v040.i13]
66. Blei D, Ng A, Jordan M. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1002 [FREE Full text]

67. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. ArXiv.org. 2019. URL: http://arxiv.org/abs/1908.10084 [accessed 2022-01-10]
68. Green E, Murphy E, Gryboski K. The health belief model. In: The Wiley Encyclopedia of Health Psychology. London: Wiley; 2020.
69. ldatuning: tuning of the latent dirichllocation models parameters. CRAN. URL: https://rdrr.io/cran/ldatuning/ [accessed 2022-01-10]
70. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing 2009 Mar;72(7-9):1775-1781. [doi: 10.1016/j.neucom.2008.06.011]
71. Arun R, Suresh V, Veni M, Narasimha MM. On finding the natural number of topics with Latent Dirichlet allocation: some observations. In: Advances in Knowledge Discovery and Data Mining. Berlin Heidelberg: Springer; 2010.
72. Deveaud R, SanJuan E, Bellot P. Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique 2014 Apr 30;17(1):61-84. [doi: 10.3166/dn.17.1.61-84]
73. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. arXiv. Preprint posted online December 14, 2020 [FREE Full text]
74. Baddeley A. Human Memory: Theory and Practice. New York: Psychology Press; 1997.
75. Neubaum G, Krämer NC, Alt K. Psychological effects of repeated exposure to elevating entertainment: an experiment over the period of 6 weeks. Psychol Popular Media 2020 Apr;9(2):194-207. [doi: 10.1037/ppm0000235]
76. Appleton K, Hemingway A, Rajska J, Hartwell H. Repeated exposure and conditioning strategies for increasing vegetable liking and intake: systematic review and meta-analyses of the published literature. Am J Clin Nutr 2018 Oct 01;108(4):842-856. [doi: 10.1093/ajcn/nqy143] [Medline: 30321277]
77. Maiden N. Digital creativity support: designing AI to augment human creativity. In: The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization. Würzburg, Germany: Ergon; 2019.
78. Cappella J. Perceived message effectiveness meets the requirements of a reliable, valid, and efficient measure of persuasiveness. J Commun 2018 Oct;68(5):994-997 [FREE Full text] [doi: 10.1093/joc/jqy044] [Medline: 30479403]
79. O'Keefe DJ. Message generalizations that support evidence-based persuasive message design: specifying the evidentiary requirements. Health Commun 2015;30(2):106-113. [doi: 10.1080/10410236.2014.974123] [Medline: 25470435]
80. Kim HS. How message features and social endorsements affect the longevity of news sharing. Digit J (Abingdon) 2021;9(8):1162-1183 [FREE Full text] [doi: 10.1080/21670811.2020.1811742] [Medline: 34900400]
81. Keskar N, McCann B, Varshney L, Xiong C, Socher R. CTRL: a conditional transformer language model for controllable generation. arXiv. Preprint posted online September 11, 2019 [FREE Full text]
82. Lin Z, Riedl M. Plug-and-blend: A framework for controllable story generation with blended control codes. arXiv. Preprint posted online March 23, 2021 [FREE Full text]
83. Harrington N. Persuasive health message design. In: Oxford Research Encyclopedia of Communication. Oxford: Oxford University Press; 2016.
84. Scott AJ. Persuasive Advertising: Evidence-based Principles. London: Palgrave Macmillan; 2010.
85. Harrington NG. Introduction to the special issue: message design in health communication research. Health Commun 2015;30(2):103-105 [FREE Full text] [doi: 10.1080/10410236.2014.974133] [Medline: 25470434]
86. Tan C, Niculae V, Danescu-Niculescu-Mizil C, Lee L. Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. arXiv. Preprint posted online February 2, 2016 [FREE Full text]
87. Boster FJ, Liu RW, Cheng Y, Kim W, Shaikh SJ. The Suasory force of sticky messages: an application to the application of sunscreen. Commun Stud 2017 Dec 26;69(1):4-22. [doi: 10.1080/10510974.2017.1414067]
88. Fishbein M, Cappella JN. The role of theory in developing effective health communications. J Commun 2006 Aug;56(Suppl 1):1-17. [doi: 10.1111/j.1460-2466.2006.00280.x]
89. Thompson TL, Schulz PJ. Health Communication Theory. London: John Wiley & Sons; 2021.
90. Woolley SC, Howard PN. Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media. Oxford: Oxford University Press; 2018.
91. Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F. Defending against neural fake news. arXiv. Preprint posted online May 29, 2019 [FREE Full text]
92. Subrahmanian V, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, et al. The DARPA Twitter bot challenge. Computer 2016 Jun;49(6):38-46. [doi: 10.1109/MC.2016.183]
93. Coronel J, Bucy E. A cognitive neuroscience perspective on political knowledge, misinformation, and memory for facts. In: The Handbook of Communication Science and Biology. London: Routledge; 2020.
94. Leib M, Köbis N, Rilke R, Hagens M, Irlenbusch B. The corruptive force of AI-generated advice. arXiv. Preprint posted online February 15, 2021 [FREE Full text]
95. Allem J, Ferrara E. Could social bots pose a threat to public health? Am J Public Health 2018 Aug;108(8):1005-1006. [doi: 10.2105/AJPH.2018.304512] [Medline: 29995482]
96. Shi W, Liu D, Yang J, Zhang J, Wen S, Su J. Social bots' sentiment engagement in health emergencies: a topic-based analysis of the COVID-19 pandemic discussions on Twitter. Int J Environ Res Public Health 2020 Nov 23;17(22):8701 [FREE Full text] [doi: 10.3390/ijerph17228701] [Medline: 33238567]

97. Unger JB, Rogers C, Barrington-Trimis J, Majmundar A, Sussman S, Allem J, et al. "I'm using cigarettes to quit JUUL": an analysis of Twitter posts about JUUL cessation. Addict Behav Rep 2020 Dec;12:100286 [FREE Full text] [doi: 10.1016/j.abrep.2020.100286] [Medline: 32637562]
98. Fischhoff B, Slovic P, Lichtenstein S, Read S, Combs B. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sci 1978 Apr;9(2):127-152. [doi: 10.1007/BF00143739]
99. Solaiman I, Brundage M, Clark J, Askell A, Herbert-Voss A, Wu J. Release strategies and the social impacts of language models. arXiv. Preprint posted online August 24, 2019 [FREE Full text]
100. Gibney E. The battle for ethical AI at the world's biggest machine-learning conference. Nature 2020 Jan;577(7792):609. [doi: 10.1038/d41586-020-00160-y] [Medline: 31992885]
101. Frankfurt H. On Bullshit. Princeton: Princeton University Press; 2009.
102. Levesque H. Common sense, the Turing test,the quest for real AI. Boston: MIT Press; 2017.
103. Marcus G, Davis E. Rebooting AI: Building Artificial Intelligence We Can Trust. New York: Knopf Doubleday Publishing Group; 2019.
104. McCarthy J. Programs with common sense. Nature. URL: https://www.cs.rit.edu/~rlaz/files/mccarthy1959.pdf [accessed 2022-01-10]
105. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online May 28, 2020 [FREE Full text]
106. Hoyt RE, Snider D, Thompson C, Mantravadi S. IBM Watson analytics: automating visualization, descriptive, and predictive statistics. JMIR Public Health Surveill 2016 Oct 11;2(2):e157 [FREE Full text] [doi: 10.2196/publichealth.5810] [Medline: 27729304]
107. Ji H, Ke P, Huang S, Wei F, Zhu X, Huang M. Language generation with multi-hop reasoning on commonsense knowledge graph. arXiv. Preprint posted online September 24, 2020 [FREE Full text]
108. Bosselut A, Rashkin H, Sap M, Malaviya S, Celikyilmaz A, Choi Y. Comet: commonsense transformers for knowledge graph construction. arXiv. Preprint posted online June 12, 2019 [FREE Full text]
109. Mazuz K, Yom-Tov E. Analyzing trends of loneliness through large-scale analysis of social media postings: observational study. JMIR Ment Health 2020 Apr 20;7(4):e17188 [FREE Full text] [doi: 10.2196/17188] [Medline: 32310141]
110. Orji R, Moffatt K. Persuasive technology for health and wellness: state-of-the-art and emerging trends. Health Informatics J 2018 Mar;24(1):66-91 [FREE Full text] [doi: 10.1177/1460458216650979] [Medline: 27245673]

Abbreviations

AI: artificial intelligence
CDC: Centers for Disease Control and Prevention
FA: folic acid
GPT-2: Generative Pretrained Transformer-2
ML: machine learning
NLG: natural language generation
NTD: neural tube defect

Edited by R Kukafka; submitted 21.03.21; peer-reviewed by Z Su, R Rice, CY Lin, X Yang; comments to author 26.04.21; revised version received 21.05.21; accepted 21.11.21; published 18.01.22

Please cite as:
Schmälzle R, Wilcox S
Harnessing Artificial Intelligence for Health Message Generation: The Folic Acid Message Engine
J Med Internet Res 2022;24(1):e28858
URL: https://www.jmir.org/2022/1/e28858
doi: 10.2196/28858
PMID: