[1]**PREPRINT:**

Search Term Identification Methods for Computational Health Communication: A Word Embedding and Network Approach for Health Content on YouTube

Abstract

Background: Common methods for extracting content in health communication research typically involve using a set of well-established queries, often names of medical procedures or diseases, that are often technical or rarely used in the public discussion of health topics. Although these methods produce high recall (i.e., retrieve highly relevant content), they tend to overlook health messages that feature colloquial language and layperson vocabularies on social media. Given how such messages could contain misinformation or obscure content that circumvents official medical concepts, correctly identifying (and analyzing) them is crucial to the study of user-generated health content on social media platforms.

Objectives: Health communication scholars would benefit from a retrieval process that goes beyond the use of standard terminologies as search queries. Motivated by this, we put forward a search term identification method to improve the retrieval of user-generated health content on social media. We focused on cancer screening tests as a subject, and YouTube as a platform case study.

Methods: We retrieved YouTube videos using cancer screening procedures (colonoscopy, FOBT, mammogram, pap test) as seed queries. We then trained word embedding models using

[1] Please cite as: Tong, C., Margolin, D., Chunara, R., Niederdeppe, J., Taylor, T., Dunbar, N. & King, A. (in press). Search Term Identification Methods for Computational Health Communication: A Word Embedding and Network Approach for Health Content on YouTube. *JMIR Medical Informatics*. doi:10.2196/37862. http://dx.doi.org/10.2196/37862

text features from these videos to identify the nearest neighbor terms that are semantically similar to cancer screening tests in colloquial language. Retrieving more YouTube videos from the top neighbor terms, we coded a sample of 150 random videos from each term for relevance. We then used text mining to examine the new content retrieved from these videos, and network analysis to inspect the relations between the newly retrieved videos and videos from the seed queries.

Results: The top terms with semantic similarities to cancer screening tests were identified via word embedding models. Text mining analysis showed that the five nearest neighbor terms retrieved content that was novel and contextually diverse, beyond the content retrieved from cancer screening concepts alone. Results from network analysis showed that the newly retrieved videos had at least one total degree of connection (sum of in- and out-degree) with seed videos according to YouTube relatedness measures.

Conclusions: We demonstrated a retrieval technique to improve recall and minimize precision loss that can be extended to various health topics on YouTube, a popular video-sharing social media platform. We discussed how health communication scholars can apply the technique to inspect the performance of retrieval strategy before investing human coding resources and outlined suggestions on how such technique can be extended to other health contexts.

Keywords: health information retrieval; search term identification; social media; health communication; public health; computational textual analysis; NLP; word2vec; word embeddings; network analysis.

Search Term Identification Methods for Computational Health Communication: A Word Embedding and Network Approach for Health Content on YouTube

Introduction

Researchers are increasingly interested in understanding the types and accuracy of health-related messages produced in the public communication environment (PCE) [1–5]. Given the proliferation of online health information sources and social media platforms in which people generate, share, and access information [6], identifying and capturing what message content individuals are likely to see when looking for information about health (i.e., seeking), as well as what information people might encounter while being online (i.e., scanning) [7–9], is crucial in gaining insights into issues including misinformation or inequities in online platforms within the larger PCE.

Identifying appropriate strategies to retrieve this information is nevertheless challenging. To gather data for analysis, researchers often rely on the standard approach of searching for content using keywords, which usually involve a set of technical (e.g., medical) terms that describe a condition or behavior of interest (e.g., “colon cancer” or “diabetes”) [10–12]. However, keyword search strategies that are solely based on technical concepts cannot account for the multifaceted nature of online information. A primary reason is that the messages in the contemporary PCE are often generated by users, and thus often include colloquial terminology rather than medical terminology [7,13–15]. This phenomenon has been well-documented in consumer health vocabularies research, which examines the language gap between official medical texts and user-generated content, such as Q&A sites (Yahoo! Answers) and social media platforms (e.g., Twitter) [16–19].

In addition to messages that do not include technical keywords, another type of content that might be overlooked by the standard retrieval approach is what could be categorized as content that misleads by omission (e.g., messages that describe risky behaviors but fail to name the medical risk it exposes an individual to) [20–22]. For example, messages promoting a fad diet, which might be associated with a specific medical condition but do not mention this risk nor the condition itself, will not be retrieved by keywords naming the condition.

Failure to retrieve these messages could result in the biased identification of content, especially in light of research showing how search results vary according to specific queries [23], and how social media language varies across different geographical locations [24]. In other words, retrieving (and analyzing) only messages produced with the "official" technical language can lead researchers to overlook the information consumed and barriers faced by underprivileged groups [25,26], or users who lack skills and knowledge to correctly use official medical vocabularies to access information [27,28]. For these reasons, public health researchers trying to understand the PCE would benefit from a principled, replicable process for searching for online content relevant to medical terms, but not exclusively restricted to them. Such a process would also inform online users' health information seeking efforts by enabling the retrieval of health-related information from commonly used slang or nontechnical queries.

This paper proposes such a retrieval process for YouTube. Using the platform's application programming interface (API) to retrieve videos and the inferred relatedness between videos determined by YouTube's proprietary algorithm, our process retrieves videos that (i) are frequently relevant to understanding the PCE related to a focal technical term; (ii) are distinct from the videos retrieved directly with the focal term; and (iii) can be easily distinguished from irrelevant videos that could otherwise absorb researchers' attention. Such a search identification approach balances the trade-off between recall and precision [29], identifying content that would not have been found using typical keywords without requiring human coders to sift through large quantities of irrelevant content.

In the following sections, we summarize relevant research on PCE content retrieval, highlighting strengths and weaknesses. We then discuss the rationale for using YouTube before detailing the techniques employed to identify relevant content beyond formal medical concepts. We illustrate the techniques using cancer screening as a case study. We conclude with a discussion of the potential for application of the technique across other topics and platforms.

Challenges of Health-Related Vocabulary Inconsistencies

User-generated health content presents important challenges to researchers attempting to retrieve content from this environment, particularly because (1) researchers may not know the vocabulary users use to discuss health topics, and (2) users can mislead each other by failing to mention relevant information.

Research has shown that patients often do not conceptualize diseases, treatments, or risks in the same terms as health care practitioners [30–32]. Most plainly, the literature on consumer health vocabulary [15–17] shows that the terms used by laypeople are different from those used by health care practitioners. For example, questions about health topics posted on Q&A sites (e.g., Yahoo! Answers and WebMD) by laypeople were found to contain misspelled words, descriptions, and background information, and were more colloquial than texts by health professionals [13,33]. A more recent example is COVID-19, where infodemiologists identified a variety of terms using Google Trends that referred to the virus, including “stigmatizing and generic terms” (e.g., “Chinese coronavirus”, “Wuhan virus”) that had not been identified by other research using more agreed upon and technical language about the virus [34]. These works suggest that user-vocabulary, which is distinct from medical vocabulary, is important for understanding how individuals conceive of their health and the medical vocabulary related to it when looking for or coming upon health information online. More broadly, these different terminologies can reflect different ways of conceptualizing health issues [32,35,36].

It is not surprising, then, that user-vocabulary is important for identifying relevant health related posts to social media, as research indicates retrieval performance significantly change when users' health queries are reformulated using formal, professional terminologies [23]. Thus, if researchers do not know what the user vocabulary is for a given topic, their retrieval strategy will be biased to identify only content posted by users who use medical technical vocabulary. Moreover, this bias is unlikely to be neutral with respect to larger public health concerns. In particular, differences of this nature, like conceptualization of illness and preferred vocabulary, have been shown to be associated with important differences in outcomes [25,26,37]. Such conceptual differences would likely manifest in differences in user-vocabulary.

Problems of Omission in Health Information Retrieval

Another weakness of retrieving user generated health messages with technical terms is that this strategy cannot, by definition, identify information that omits that term. However, this failure to connect risks to outcomes can be precisely what makes user generated content misleading. It is well established that many people lack broad knowledge about risk factors for many leading causes of death in the United States and beyond [38–40], and people routinely receive information that fails to link common risk factors and behaviors to negative health outcomes [41]. Perhaps the best known (and most damaging) example is the failure of tobacco companies to mention that cigarette smoking causes cancer in their promotional materials [42]. This misrepresentation by omitting and distancing from medical terms (e.g., disease) is common for unhealthy products (e.g., alcohol) [43].

In such cases, the PCE misleads by omission because it fails to assign the appropriate words to what is medically accurate in the offline world. This has the potential to mislead the public and makes relevant messages hard to find, because their relevance (to researchers) is defined by what is absent (the mention of the risk). An example is the "Tide Pod challenge" that emerged in 2017 as a popular internet trend. The Tide Pod challenge is dangerous because it fails to connect the terms "Tide detergent" and "eat" with the concept (or concept family) of "poison." A trained medical professional would not discuss "eating" Tide Pods without also mentioning the danger, however users can (and did). Such misleading (and dangerous) user messages cannot be retrieved by strategies that focus on the harm – poisoning.

In the case of well researched and widely understood risks, such as the connection between cigarette smoking and lung cancer, this weakness can be overcome by simply naming the risk factor (i.e., searching for "lung cancer"). But to restrict searches to known and well documented high-risk behaviors is to again return researchers to their cultural bubble [44]. As evidenced by the emergence of the Tide Pod challenge, user-generated content can be extraordinarily inventive, creating new risky behaviors not yet known to the medical community. For example, dangerous fad diets cannot be identified by searching for the risks they pose. Instead, what is needed is a way to identify vocabulary that is "near" to condition of interest, broadening the net so that researchers can identify messages misleading by omission.

For both reasons, researchers should find ways to escape the strictures of official, technical vocabulary when retrieving information to characterize the PCE. Researchers instead need search terms that include culturally relevant colloquial terms that are related to medical terms, as well as terms that identify behaviors or practices "in the neighborhood" of medical terms but which can identify content when those terms are omitted.

YouTube as Public Health Information Source and Site of Inquiry

In this study, we focus on YouTube videos as a meaningful message source of the PCE. We selected YouTube for two reasons. First, YouTube is one of the most widely used online social media and content platforms [45]. Second, YouTube has become increasingly relevant as a source of health information. With its dual function as a reservoir of video content and a social networking platform in which users acquire information through interactions with the content and fellow users, YouTube has served as an informational resource for learning about diverse health topics for users [46,47].

Extant research on medical and health information on YouTube suggests several issues with the quality of YouTube content. A meta-analysis found that YouTube videos tend to contain misinformation prevalently, an implication of which is the potential of the platform to alter beliefs about health interventions [46]. One limitation of these studies (and a weakness shared by many YouTube studies) is the search strategies used to identify relevant content. To address this gap in current research, our project aims to answer two research questions.

The first main research question (RQ1) asks: for a given medical/health term of interest (i.e., a focal term for retrieval), does our proposed search term identification strategy retrieve health messages that (a) are relevant to understanding the public health communication environment related to that seed term and (b) do not explicitly use that term (such that the traditional medical/technical search terms would have failed to retrieve them)? To provide a satisfactory answer to this question, a search strategy must (i) retrieve content relevant to the seed term (called precision) and (ii) find relevant content that is novel, i.e., different from what would be returned by the seed term alone (called recall), without sacrificing too much precision. This leads to our second research question (RQ2): can the derived strategy identify relevant, novel messages with sufficient precision to be practically useful?

Methods

Rationale for cancer screening focal terms

Cancer is one of the biggest public health issues in the United States, thus a topic that requires meticulous attention from multiple stakeholders, including public health practitioners and communicators. A particular challenge to the prevention and management of various cancer types is the persistent disparities in screening, incidence, and mortality rates across different population groups [48]. Given the significance of cancer and important implications of cancer screening disparities, we choose cancer screening as the subject of examination in this paper.

To this end, we first demonstrate our methodological technique using the primary colorectal cancer screening option, the “colonoscopy”, as our focal term. Colorectal cancer is the third most diagnosed and third most deadly cancer in the US, which disproportionately affects Blacks compared to non-Hispanic white Americans [49]. We then replicate the analyses using other cancer screening tests (fecal occult blood test [FOBT], mammogram, pap test) as focal terms, to illustrate how the technique performs in other cancer contexts, including breast and cervical cancer.

Retrieving YouTube videos from the focal term

We collected data from YouTube via the YouTube API v3. Using the “search: list” endpoint (used for search function) allowed us to retrieve two types of data: videos that are most relevant to a search query or set of queries (the “q” parameter with “relevance” sorting), and videos that are related to a specific or set of videos (the “related-to-video-id” parameter) according to YouTube algorithms [50]. We note that collecting data through this API approach bypasses localization and personalization – factors that play important roles in search results that are presented to specific individuals. Since our purpose is to demonstrate a methodology that can be systematically extended to other contexts in future research, we deem this approach to be appropriate in giving us the results as close to a default setting as possible.

On August 22, 2021, using the YouTube Data Tools software [51], we retrieved a set of 250 videos most relevant to the search term “colonoscopy.” These 250 videos comprise our core set. In addition, we retrieved 4,304 videos “related to” this core set, which in total gave us 4,554 videos in the initialization set. We retrieved these videos’ unique identifiers, text data (video titles, descriptions) and metadata (publication date, engagement statistics).

Word Embeddings

Word embedding is an unsupervised method of learning word vectors using a neural networks model [52]. The basic aim of word embeddings is to identify words that appear in “similar contexts” as the focal term. The technique calculates a proximity score, that is, the extent to which two terms are near to one another in a multidimensional space. This score acts as a measure of “semantic similarity.” It is thus a useful way to find texts that discuss a particular concept without explicitly mentioning it. Texts that mention a word’s close neighbors (in the multi-dimensional space) are likely talking about ideas where that word is relevant, too, even if the word itself is not there. We use word embeddings to find YouTube content that is relevant to “colonoscopy,” but which may not mention the word itself.

We apply word embeddings using the word2vec approach on the text data of our initialization set of 4,554 videos. Specifically, we use the text of the 4,554 video titles and descriptions to build a corpus. Then, after preprocessing and standardization steps (including removal of emojis, signs, stop words, performing lowercasing, converting text to ASCII encoding, removing leading/trailing spaces), a word2vec model was trained on the text to identify the terms with the most semantic similarity to the term “colonoscopy” (word2vec R package) [53].

We then used the top six “nearest neighbors” to “colonoscopy” as new search terms to retrieve more videos (250 videos for each neighbor) to inspect the new content.

Human Coding and NLP to Evaluate Recall Improvement

The goal of retrieving new content from the nearest neighbors is the improvement of recall over a direct search – the identification of videos that are relevant to “colonoscopy”, but which would not be found by searching directly for it. To assess this recall improvement, we took a random 10% sample (25 videos for each neighbor) and coded them for relevance. Coding was done by a research team member (the paper’s last author) with expertise in cancer control and cancer communication.

Specifically, a video was coded as relevant if the video content contained: (1) any aspect of screening preparation or procedures (e.g., bowel prep, personal experiences, clinical discussions, etc.) or (2) general information on colorectal cancer or colorectal cancer screening in terms of cancer prevention or early detection. This included content where a patient underwent a colonoscopy, but perhaps for a chronic condition (e.g., ulcerative colitis or Crohn’s disease). Obscure terms identified through this process were also looked up as needed to confirm relevance (e.g., “suprep” – a commercial brand for bowel prep kit).

We evaluated recall in two ways. First, we assessed how many of the relevant “found” videos would have been identified using the search term alone. We did this by counting the number of relevant videos in the newly found set that contain the term “colonoscopy.” Those that did not contain “colonoscopy,” but were nonetheless relevant to it, constituted a recall improvement. Second, we examined whether these newly found videos were substantively different—in terms of content, topics, focus—from the core set. Using the R package quanteda [54], we calculated the average Euclidean distances between the text features embedded in the different video sets. Euclidean distance is a pairwise distance metric that measures to dissimilarities between the text features in different corpuses. We then used hierarchical clustering analysis, with the complete linkage method (*hclust* function in *stats* version 3.6.2), to determine if videos in different sets were substantially overlapping in content.

Network Analysis to Evaluate Precision

Strategies to improve recall are often offset by a substantial loss of precision. In our case, while the nearest neighbors may retrieve many more relevant videos, they could at the same time bring in many irrelevant videos. This introduces a risk of increasing human coding costs or other resource-intensive techniques of classification. Such precision loss needs to be mitigated so that it occurs at a manageable level. To implement this, we used the "related to video id" API endpoint, which reports whether a set of videos are "related to" the others (zero crawl depth), to query the relations between the new videos retrieved from the top neighbor terms and the colonoscopy videos from the core set. Specifically, if video A is related to video B in a set, there is a connection (or link) between them. These relations were used to create a network with videos being nodes and the connections between them being edges.

We then calculated three network measures of relatedness: indegree (videos from the core set that link to a newly found video), outdegree (videos in the core set that each newly found video links to), and total degree (sum of indegree and outdegree). We expect that the newly found irrelevant videos will have few, if any, links to the videos known to be about "colonoscopy," while videos with even loose relevance will have at least some connections to the core set. To examine the extent to which these degree scores were associated with relevance (according to human coding), the corresponding precision and recall statistics at different degree levels were inspected. If our technique worked effectively, there would be some threshold of degree – number of connections between a newly found video and the core set – at which videos with this degree or higher are not only reasonably novel (improving recall over the core set), but also reasonably relevant (maintaining precision at a manageable level).

Results

Word Embeddings

Table 1 provides the list of neighbor terms to the focal term "colonoscopy" and their ranks based on semantic similarity, according to word embedding results.

Table 1. Neighbor terms to "colonoscopy" and similarity scores

Term	Similarity score	Rank
suprep	0.9722890	1
peg	0.9519246	2
sutab	0.9513488	3
plenvu	0.9504289	4
glycol	0.9498276	5
miralax	0.9449067	6
rectal	0.9435940	7
cleanse	0.9422708	8
cologuard	0.9421358	9
colorectal	0.9403084	10

[a] Neighbor terms are terms with the most semantic similarity (with corresponding high similarity scores/low ranks) to "colonoscopy" based on YouTube video data. Score refers to the cosine-similarity metric between word embeddings (i.e., terms) in a multidimensional vector space.

A visual inspection suggests these nearest neighbor terms fit our goals for this method: they contain non-technical terms (e.g., “cleanse,” or brand names like “plenvu”) that are relevant to colorectal health. We selected the top 6 terms (suprep to miralax), retrieved an additional 1,500 videos (250 each) and coded a subset 10% (150 random videos) for the recall analysis.

Human Coding and NLP to Evaluate Recall Improvement

Table 2 displays the retrieval statistics, of which 51 of the 150 videos coded (34%) were deemed relevant. More importantly, 21 of these 51 (~40% of the total and 14% of the coded sample) did not contain the term “colonoscopy,” meaning that identifying them improved recall over what would have been found simply searching for “colonoscopy.” This supported our expectation that the word embedding approach helped address the recall problem inherent in using technical language.

Table 2. Retrieval statistics in the sampled videos for top 6 neighbors of “colonoscopy”

Term	Sample Coded Videos	Relevant	Precision (% Relevant)	Relevant & Mention “colonoscopy”	Relevant & Does not mention “colonoscopy” (Recall Improvement)
suprep	25	18	72%	9	9
peg	25	1	4%	0	1
sutab	25	4	16%	4	0
plenvu	25	23	92%	15	8
glycol	25	0	0%	0	0
miralax	25	5	20%	2	3
total	150	51	34%	30	21

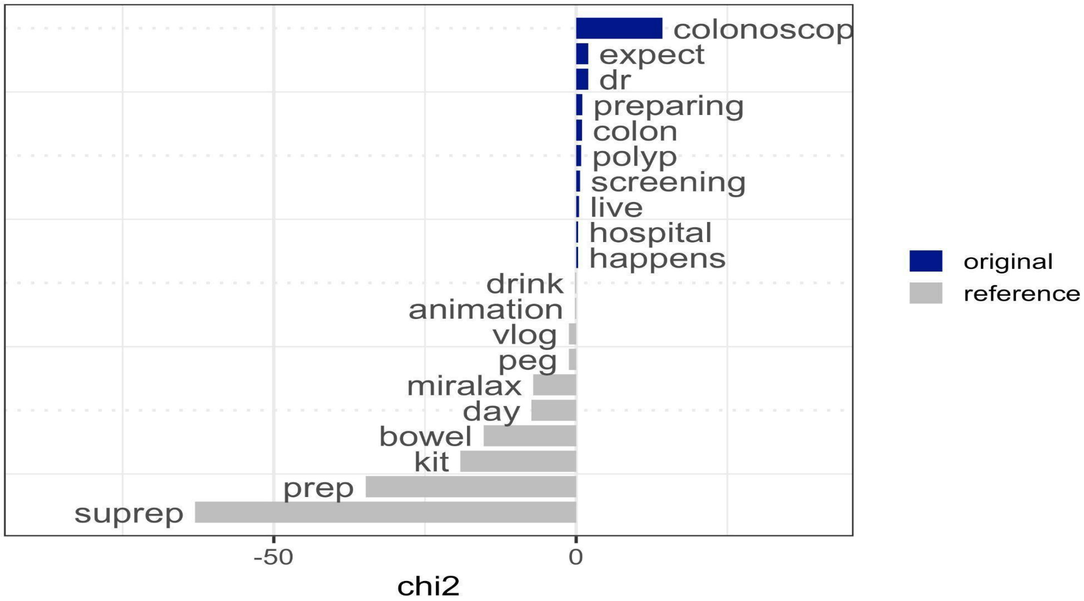
We next assessed whether these newly found videos were substantively different—in terms of content, topics, focus—from what would be retrieved with the typical strategy. To assess this, we compared the Euclidean distances between textual features of the core set (250 videos) and those of the newly found videos (Table 3). Here, higher values meant greater distance. For example, the distance between "miralax" and "peg" was the smallest between our groupings, indicating that videos in these two sets shared the most similar words, compared to other pairs.

Table 3. Euclidean distance between the text features of original colonoscopy video set and video sets generated from top 6 neighbor terms

	1	2	3	4	5	6
1. original	0	255.61	257.97	241.5	248.9	254.68
2. miralax		0	6.32	20.1	21.8	7.14
3. peg			0	22.2	23.1	6.86
4. plenvu				0	20.6	19.08
5. suprep					0	20.57
6. sutab						0

[b] Cell values indicate dissimilarities of the text features belonging to any pair of video sets. Larger values indicate larger distances, and 0 indicates identical text features. "Glycol" was removed due to zero relevant videos retrieved.

Figure 1. Relative frequencies of words in the colonoscopy video set and the combined top 5 neighbor term video set

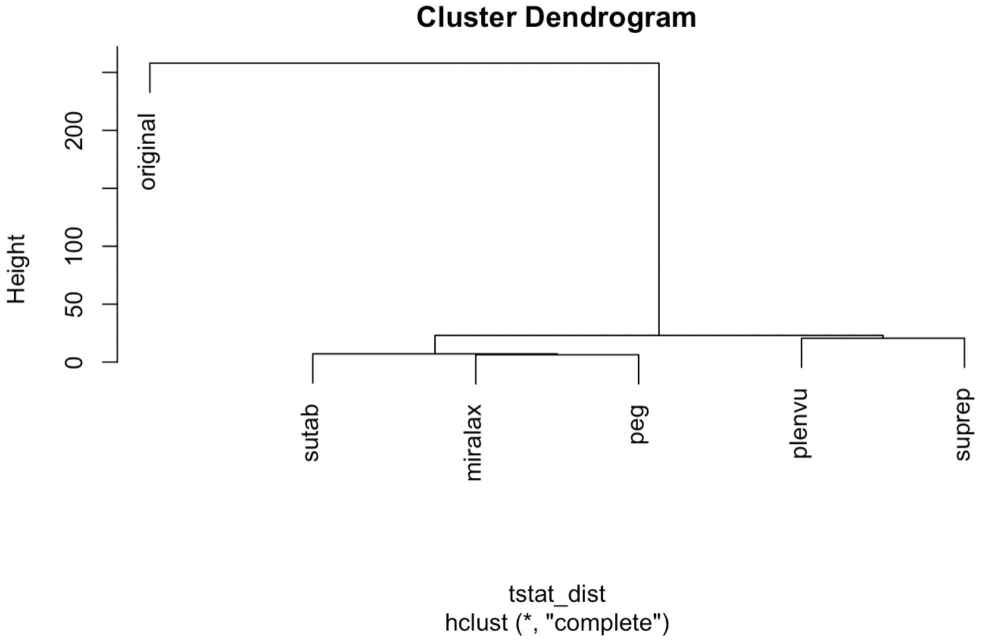


Notes: Words that are "key" to each video set were plotted. Original: set of videos found with search query "colonoscopy". Reference: set of videos found with 5 nearest terms to "colonoscopy" (suprep, peg, sutab, plenvu, miralax).

Relative frequency analysis was used to further illustrate these differences by highlighting the differences in the text features of the core set as opposed to the newly found set. As Figure 1 shows, words such as "colonoscopy", "dr", "preparing", "colon", "polyp" were disproportionately more likely to occur in the core set, whereas words like "suprep", "prep", "kit", "bowel", "miralax" were distinct terms found in the newly found set.

Hierarchical agglomerative clustering performed on the text features of the newly found set and the core set (using complete link method) revealed that the text features in the videos retrieved from neighbor terms (newly found set) were more similar to such from other neighbor terms, than to the core set (Figure 2). In other words, these results show that our approach helped identify videos that are relevant to "colonoscopy" without including the term itself (i.e., improving recall); furthermore, these newly found relevant videos additionally enhanced the topical diversity of our retrieved data (by focusing on preparation brands and procedures).

Figure 2. Visualization of distances between video sets



Notes: Hierarchical cluster analysis indicating dissimilarities and distances between original (set of videos found with search query “colonoscopy”) and sets of videos found with 5 nearest terms to “colonoscopy” (suprep, peg, sutab, plenvu, miralax).

Network Analysis to Evaluate Precision

Table 4 shows the results of the comparison between a found video’s degree of connection to the core set and its associated relevance according to human coding. We first note that new videos that are in other languages than English (n = 28) were found to have no connections with the core set videos. To avoid having this bias our results, we excluded these 28 videos, as well as 8 videos that were already found in the original set and one video where YouTube returned missing meta-data (37 excluded in total). We then performed comparison on the remaining 113 videos (the final total in the “cumulative count of videos”, also the denominator).

The first four columns in Table 4 show the total degree (number of connections) and counts of videos with corresponding total degrees, in comparison with relevance statistics. Specifically, all videos with total degree greater than 7 had been coded as relevant, meaning precision is 100% at or above this threshold. More importantly, though precision was imperfect below this threshold, it remained very high. In fact, if we examined videos of degree 1 or higher, we found that 71% had been coded as relevant. This means that a human coding team choosing to use this liberal threshold (at least 1 connection to any video in the core set) for choosing videos to code would see more than two relevant videos for every irrelevant one, thus expending limited resources examining irrelevant videos.

The cumulative columns on the right-hand side of the table display the trade-offs that would face a coding team. The cumulative count of relevant videos adds up to 37, which is the 51 coded as relevant (see Table 2) less the 8 videos already found in the original dataset (as reported above) and 6 non-English videos that had been coded as relevant. Cumulative precision refers to the relevance of the videos at or above this threshold. Cumulative recall shows the portion of the relevant videos in the set that are preserved at this threshold. As the threshold tighten, precision improves (irrelevant videos are discarded) but recall declines (some relevant videos are discarded, too). For example, if a team chose to examine videos with at least 3 connections to the core set (degree equals or exceeds 3), they would find 32 videos, 28 of which were relevant (88% precision), and miss out on only 9 of the 37 possible (75.7% recall). In other words, this technique provides a basis for researchers to inspect performance of the retrieval strategy before investing human evaluation and coding resources.

Table 4. Relevance of newly found videos by number of links to original set of colonoscopy videos (total degree)

Total Degree	Count of Videos w/ Total Degree	# of Videos coded as "relevant"	% Relevant	Cumulative Count of Non-duplicate Videos	Cumulative Count of Non-duplicate Relevant Videos	Cumulative Precision	Cumulative Recall	Cumulative F1
44	1	1	100%	1	1	100%	2.7%	5.3%
41	1	1	100%	2	2	100%	5.4%	10.3%
26	1	1	100%	3	3	100%	8.1%	15.0%

23	1	1	100%	4	4	100%	10.8%	19.5%
22	1	1	100%	5	5	100%	13.5%	23.8%
21	1	1	100%	6	6	100%	16.2%	27.9%
20	2	2	100%	8	8	100%	21.6%	35.6%
19	1	1	100%	9	9	100%	24.3%	39.1%
18	1	1	100%	10	10	100%	27.0%	42.6%
17	2	2	100%	12	12	100%	32.4%	49.0%
16	1	1	100%	13	13	100%	35.1%	52.0%
15	2	2	100%	15	15	100%	40.5%	57.7%
14	1	1	100%	16	16	100%	43.2%	60.4%
13	1	1	100%	17	17	100%	45.9%	63.0%
12	2	2	100%	19	19	100%	51.4%	67.9%
11	2	2	100%	21	21	100%	56.8%	72.4%
10	1	1	100%	22	22	100%	59.5%	74.6%
9	1	1	100%	23	23	100%	62.2%	76.7%
7	2	1	50%	25	24	96%	64.9%	77.4%
6	1	1	100%	26	25	96%	67.6%	79.4%
5	2	0	0%	28	25	89%	67.6%	76.9%
4	2	2	100%	30	27	90%	73.0%	80.6%
3	2	1	50%	32	28	88%	75.7%	81.2%
2	5	1	20%	37	29	78%	78.4%	78.4%
1	5	1	20%	42	30	71%	81.1%	75.9%
0	71	7	10%	113	37	33%	100%	49.3%

[c] "Total degree": the sum of connections each new video has with the videos in the original colonoscopy video set. "Cumulative precision": The cumulative count of relevant videos divided by the cumulative count of all videos. "Cumulative recall": cumulative count of relevant videos divided by the total number of new and non-duplicate 37 relevant videos. "Cumulative F1": The harmonic mean of cumulative precision and cumulative recall.

Replication: Other cancer screening tests

We extended our analyses to three additional focal terms to illustrate the breadth of the technique's applicability. The first, "FOBT," refers to the fecal occult blood test, another screening method for colorectal cancer. The second and third are "mammogram" and "pap test," screening tests for breast cancer and cervical cancer, respectively. We chose cancer screening as an illustrative case because these are common cancer types that are often discussed on social media [3,55] such that research would benefit from identifying relevant content that does not explicitly mention these technical, formal screening tests.

As shown in the summary statistics in Table 5, the results for these terms were comparable to "colonoscopy." For each focal term, searches using the nearest neighbor terms uncovered through word2vec identified a wide range of new videos that were distinct from the original sets, improving recall (see Supplementary Information for dissimilarity measures of new vs original content). Similar to the results for "colonoscopy," filtering videos based on their degrees of connection to the core set (for the respective focal term) improved precision while maintaining reasonable recall. For both "FOBT" and "pap test", researchers could inspect only videos with degree of 1 or greater and would find few irrelevant videos while maintaining most of the new videos in the set. For "mammogram", the recall statistics of videos with at least one connection is lower (30%); however, even if researchers chose to drop this filter and inspect all videos, they would find that almost 1 in 3 of the new videos found are relevant. Thus, researchers would not be at risk of being overwhelmed with irrelevant content.

Table 5. Summary retrieval statistics for "colonoscopy", "FOBT", "mammogram" and "pap test"

Focal term	Top nearest neighbor terms	Sample Coded Videos	(A) New and Non-duplicate Relevant Videos	(B) Number of videos w/degree >=1	Videos w/degree >=1 and coded as new and relevant (A ∩B)	Precision	Recall
Colonoscopy	suprep peg sutab plenvu glycol miralax	150 (25/term)	37	42	30	75% (30/42)	81% (30/37)
FOBT	iFOBT hemosure immunochemical immunostics	125 (25/term)	50	33	27	82% (27/33)	54% (27/50)

	guaiac						
Mammogram	smartcurve breastcheck biopsy ultrasound breastcancerawareness	250 (50/term)	77	28	23	82% 23/28	30% 23/77
Pap test	Colposcopy Smear ASCUS papsmear STD	250 (50/term)	87	65	59	91% 59/65	68% 59/87

[d] “Videos w/degree >= 1”: videos with at least one connection to the original set of videos resulted from the focal terms.

Discussion

Principal Results

This paper proposes a novel approach to improve the retrieval of user-generated health content. Using medical concepts as focal terms, we employed the similarity-based word embedding approach to detect new search terms related to focal terms but not restricted to technical vocabulary. In line with previous research using similar methods (e.g., word, sentence, or biomedical term embeddings), we identified less widely known terms in user-generated public discourse related to cancer screening tests. Quantitative textual analysis of the newly discovered content returned from the top neighbor terms indicated that these videos were distinct from the original video sets in terms of lexical and topical foci. Network analysis showed that retrieval precision can be improved by detecting videos with at least one total degree, i.e., those with at least one connection to others in the same networks. Researchers could use the technique to inspect the performance of their retrieval strategy before investing additional evaluation resources [56,57]. Beyond suggesting the value of this technique, our analyses provide insight into specific message gaps if user-generated vocabulary is overlooked.

First, our results indicate that commercial speech, particularly tagged by brand names like “suprep” and “miralax,” was particularly prominent and useful for identifying relevant content. In essence, users produced and consumed videos about “prepping”, which could be used for colonoscopies, in reference to branded products. This raises an important follow-up question – do these videos provide accurate information? As reviewed above, the history of corporate actors misleading consumers by omission of risks is substantial [58,59]. While this would be an analysis for a further study, we point out here the importance of retrieving information about medical topics using commercial terms, rather than just medical/technical terms.

Second, we note that our results did not provide examples of de novo slang synonyms (akin to “the sugars”). Rather, when users created terms, they were more likely to be portmanteaus of simple vocabulary, such as “breastcheck” or “papsmear” or even “breastcancerawareness.” This merging of words into one term is unsurprising insofar as it is consistent with the conventions for the creation of hashtags, but this should serve as a caution to researchers to consider these non-standard constructions in their retrieval strategies. In other words, for the terms searched in the present study we found little evidence of colloquial language. But, for any health topic, there is the possibility that such language use is used in less intuitive ways. While we did not find that to be the case for our focal terms, the possibility exists, and this technique could have the potential to identify such in other cases.

More broadly, our analysis reveals that while user generated vocabulary can often be sensibly interpreted after the fact – Plenvu’s website advertises it as a colonoscopy prep technique; “breastcheck” is intuitively related to breast cancer -- the most common terms are not always easy to guess in advance, that is, before analyzing some data. This observation supports the arguments that motivated this research, suggesting that researchers should first learn how users talk about medical topics, then create retrieval strategies to build fuller datasets for analysis of what they are saying. While we do not have explicit evidence here that vocabularies are associated with particular social groups, or marginalized groups in particular, the presence of corporate brand names suggests at the very least that targeted marketing efforts could play such a role for particular medical topics. This is a topic for further research.

Limitations

There are several limitations to this study. First and foremost, our analysis focused only on cancer screening tests as focal terms due to this project's inclusion in a larger project focusing on colorectal cancer screening information in the PCE. Our purpose was to demonstrate a methodological technique in the context of cancer with the understanding that future research will need to assess any unique challenges that might apply to non-cancer screening health topics or medical terminologies of interest (e.g., vaccines or information about diabetes management). While we see no methodological reasons why this technique could not be applied to other keywords and terminologies, future research would be needed to support this expectation.

The second limitation is that the word embedding model was trained on YouTube textual content and our technique relied on YouTube's relatedness data to distinguish relevant from irrelevant videos. This means that the effectiveness of the present approach is limited to YouTube. While there are good reasons to start with YouTube as a prevalent source of health-related information, we encourage future research to consider developing similar approaches for other domains where user-generated texts are found online—including websites, Q&A forum posts, and other social networking sites [21,57]. Importantly, many specific techniques may not be exportable from platform to platform. For example, while YouTube tracks relatedness between videos, messages on Twitter are often related by hashtags. Thus, rather than searching for relevant neighbor words, researchers might focus on identifying relevant neighbor hashtags.

In Q&A forums or other content with threaded replies, researchers might incorporate this hierarchical information to identify the most relevant content (e.g., terms used in top-level posts).

One final limitation is that conducting this process requires some familiarity with available NLP and computational tools. We believe the increasing application of computational methods in social science research, as well as the proliferation of training in R and Python languages for social scientists, increases the likelihood that this technique could be utilized by those with limited NLP proficiency. Nevertheless, health communication is an inherently interdisciplinary field in which we see great potential for collaborations between communication scientists, public health and medical researchers, and data scientists. Still, future work might strive to make this technique more accessible through the creation of specific tools and materials to assist health communicators and public health professionals in applying these approaches in future health promotion and education efforts.

Conclusion

This study demonstrated the potential of using similarity-based word embedding techniques for computational health communication research to improve recall and maintain precision in retrieving content that could be overlooked by standard medical terminologies. The study reveals that there are indeed relevant messages to medical topics in the PCE that do not use medical vocabulary, and that many of these can be identified. While the impact of overlooking these messages on health disparities cannot be determined, these results suggest that further study in this area is warranted.

Acknowledgements

This work was supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under award number R37CA259156. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interests

None declared.

References

1. Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management*. 2018;38(1):1-6. doi:10.1016/j.ijinfomgt.2017.08.002

2. Okuhara T, Ishikawa H, Okada M, Kato M, Kiuchi T. Contents of Japanese pro- and anti-HPV vaccination websites: A text mining analysis. *Patient Education and Counseling*. 2018;101(3):406-413. doi:10.1016/j.pec.2017.09.014

3. Chen L, Wang X, Peng TQ. Nature and diffusion of gynecologic cancer–related misinformation on social media: Analysis of tweets. *Journal of Medical Internet Research*. 2018;20(10):e11515. doi:10.2196/11515

4. Gage-Bouchard EA, LaValley S, Warunek M, Beaupin LK, Mollica M. Is cancer information exchanged on social media scientifically accurate? *J Canc Educ*. 2018;33(6):1328-1332. doi:10.1007/s13187-017-1254-z

5. Hornik R, Binns S, Emery S, et al. The Effects of Tobacco Coverage in the Public Communication Environment on Young People's Decisions to Smoke Combustible Cigarettes. *Journal of Communication*. Published online January 13, 2022. doi:10.1093/joc/jqab052

6. Chou WYS, Oh A, Klein WMP. Addressing Health-Related Misinformation on Social Media. *JAMA*. 2018;320(23):2417-2418. doi:10.1001/jama.2018.16865

7. Zhao Y, Zhang J. Consumer health information seeking in social media: a literature review. *Health Info Libr J*. 2017;34(4):268-283. doi:10.1111/hir.12192

8. Hornik R. Measuring campaign message exposure and public communication environment exposure: Some implications of the distinction in the context of social media. *Communication Methods and Measures*. 2016;10(2-3):167-169. doi:10.1080/19312458.2016.1150976

9. Shim M, Kelly B, Hornik R. Cancer information scanning and seeking Behavior is associated with knowledge, lifestyle choices, and screening. *Journal of Health Communication*. 2006;11(sup001):157-172. doi:10.1080/10810730600637475

10. Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, Barahona M, Ulijaszek SJ. The 'who' and 'what' of #diabetes on Twitter. *DIGITAL HEALTH*. 2017;3:205520761668884. doi:10.1177/2055207616688841

11. Loeb S, Sengupta S, Butaney M, et al. Dissemination of misinformative and biased information about prostate cancer on YouTube. *Eur Urol*. 2019;75(4):564-567. doi:10.1016/j.eururo.2018.10.056

12. Park S, Oh HK, Park G, et al. The source and credibility of colorectal cancer Information on Twitter. *Medicine*. 2016;95(7):e2775. doi:10.1097/MD.0000000000002775

13. Park MS, He Z, Chen Z, Oh S, Bian J. Consumers' use of UMLS Concepts on social media: Diabetes-related textual data analysis in blog and social Q&A sites. *JMIR Medical Informatics*. 2016;4(4):e5748. doi:10.2196/medinform.5748

14. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. *Methods Inf Med*. 2002;41(4):289-298. doi:10.1055/s-0038-1634490

15. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*. 2006;13(1):24-29. doi:10.1197/jamia.M1761

16. Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*. 2011;13(2):e1636. doi:10.2196/jmir.1636

17. Gu G, Zhang X, Zhu X, et al. Development of a consumer health vocabulary by mining health forum texts based on word embedding: Semiautomatic approach. *JMIR Medical Informatics*. 2019;7(2):e12704. doi:10.2196/12704

18. Ibrahim M, Gauch S, Salman O, Alqahtani M. An automated method to enrich consumer health vocabularies using GloVe word embeddings and an auxiliary lexical resource. *PeerJ Comput Sci*. 2021;7:e668. doi:10.7717/peerj-cs.668

19. Lazard AJ, Saffer AJ, Wilcox GB, Chung AD, Mackert MS, Bernhardt JM. E-Cigarette Social Media Messages: A Text Mining Analysis of Marketing and Consumer Conversations on Twitter. *JMIR Public Health Surveill*. 2016;2(2):e171. doi:10.2196/publichealth.6551

20. Ma T (Jennifer), Atkin D. User generated content and credibility evaluation of online health information: A meta analytic study. *Telematics and Informatics*. 2017;34(5):472-486. doi:10.1016/j.tele.2016.09.009

21. Lee K, Hasan SA, Farri O, Choudhary A, Agrawal A. Medical concept normalization for online user-generated texts. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. ; 2017:462-469. doi:10.1109/ICHI.2017.59

22. Chunara R, Wisk LE, Weitzman ER. Denominator Issues for Personally Generated Data in Population Health Monitoring. *Am J Prev Med*. 2017;52(4):549-553. doi:10.1016/j.amepre.2016.10.038

23. Plovnick RM, Zeng QT. Reformulation of consumer health queries with professional terminology: a pilot study. *J Med Internet Res*. 2004;6(3):e27. doi:10.2196/jmir.6.3.e27

24. Relia K, Li Z, Cook SH, Chunara R. Race, Ethnicity and National Origin-based Discrimination in Social Media and Hate Crimes Across 100 U.S. Cities. Published online January 31, 2019. Accessed November 1, 2021. http://arxiv.org/abs/1902.00119v1

25. Fage-Butler AM, Nisbeth Jensen M. Medical terminology in online patient-patient communication: evidence of high health literacy? *Health Expect*. 2016;19(3):643-653. doi:10.1111/hex.12395

26. Kilbridge KL, Fraser G, Krahn M, et al. Lack of comprehension of common prostate cancer terms in an underserved population. *J Clin Oncol*. 2009;27(12):2015-2021. doi:10.1200/JCO.2008.17.3468

27. Deursen AJAM van, Zeeuw A van der, Boer P de, Jansen G, Rompay T van. Digital inequalities in the Internet of Things: differences in attitudes, material access, skills, and usage. *Information, Communication & Society*. 2021;24(2):258-276. doi:10.1080/1369118X.2019.1646777

28. Din HN, McDaniels-Davidson C, Nodora J, Madanat H. Profiles of a Health Information–Seeking Population and the Current Digital Divide: Cross-Sectional Analysis of the 2015-2016 California Health Interview Survey. *J Med Internet Res*. 2019;21(5):e11931. doi:10.2196/11931

29. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans Inf Syst*. 1989;7(3):205-229. doi:10.1145/65943.65945

30. Aidoo M, Harpham T. The explanatory models of mental health amongst low-income women and health care practitioners in Lusaka, Zambia. *Health Policy Plan*. 2001;16(2):206-213. doi:10.1093/heapol/16.2.206

31. Mill JE. Describing an explanatory model of HIV illness among aboriginal women. *Holist Nurs Pract*. 2000;15(1):42-56. doi:10.1097/00004650-200010000-00007

32. Soffer M, Cohen M, Azaiza F. The role of explanatory models of breast cancer in breast cancer prevention behaviors among Arab-Israeli physicians and laywomen. *Prim Health Care Res Dev*. 2020;21:e48. doi:10.1017/S1463423620000237

33. Zhang Y. Contextualizing consumer health information searching: an analysis of questions in a social Q&amp;A community. In: *Proceedings of the 1st ACM International Health Informatics Symposium*. IHI '10. Association for Computing Machinery; 2010:210-219. doi:10.1145/1882992.1883023

34. Rovetta A, Castaldo L. A new infodemiological approach through Google Trends: Longitudinal analysis of COVID-19 scientific and infodemic names in Italy. Published online October 24, 2021. doi:10.31235/osf.io/6pj5x

35. Ogden J, Flanagan Z. Beliefs about the causes and solutions to obesity: A comparison of GPs and lay people. *Patient Education and Counseling*. 2008;71(1):72-78. doi:10.1016/j.pec.2007.11.022

36. Ogden J, Bandara I, Cohen H, et al. General practitioners' and patients' models of obesity: whose problem is it? *Patient Education and Counseling*. 2001;44(3):227-233. doi:10.1016/S0738-3991(00)00192-0

37. Nielsen-Bohlman L, Panzer AM, Kindig DA. *Health Literacy: A Prescription to End Confusion*. National Academies Press (US); 2004. Accessed November 1, 2021. http://www.ncbi.nlm.nih.gov/books/NBK216029/

38. Niederdeppe J, Levy AG. Fatalistic beliefs about cancer prevention and three prevention behaviors. *Cancer Epidemiol Biomarkers Prev*. 2007;16(5):998-1003. doi:10.1158/1055-9965.EPI-06-0608

39. Wang C, Miller SM, Egleston BL, Hay JL, Weinberg DS. Beliefs about the causes of breast and colorectal cancer among women in the general population. *Cancer Causes Control*. 2010;21(1):99-107. doi:10.1007/s10552-009-9439-3

40. Wardle J, Waller J, Brunswick N, Jarvis M. Awareness of risk factors for cancer among British adults. *Public Health*. 2001;115(3):173-174. doi:10.1038/sj.ph.1900752

41. Jensen JD, Moriarty CM, Hurley RJ, Stryker JE. Making sense of cancer news coverage trends: A comparison of three comprehensive content analyses. *Journal of Health Communication*. 2010;15(2):136-151. doi:10.1080/10810730903528025

42. Brandt AM. Inventing conflicts of interest: A history of tobacco industry tactics. *Am J Public Health*. 2012;102(1):63-71. doi:10.2105/AJPH.2011.300292

43. Petticrew M, Maani Hessari N, Knai C, Weiderpass E. How alcohol industry organisations mislead the public about alcohol and cancer. *Drug and Alcohol Review*. 2018;37(3):293-303. doi:10.1111/dar.12596

44. Margolin DB. Computational contributions: A symbiotic approach to integrating big, observational data studies into the Communication field. *Communication Methods and Measures*. 2019;13(4):229-247. doi:10.1080/19312458.2019.1639144

45. Auxier B, Anderson, Monica. Social Media Use in 2021. Pew Research Center: Internet, Science & Tech. Published April 7, 2021. Accessed November 3, 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

46. Madathil KC, Rivera-Rodriguez AJ, Greenstein JS, Gramopadhye AK. Healthcare information on YouTube: A systematic review. *Health Informatics J*. 2015;21(3):173-194. doi:10.1177/1460458213512220

47. Fat MJL, Doja A, Barrowman N, Sell E. YouTube videos as a teaching tool and patient resource for infantile spasms. *J Child Neurol*. 2011;26(7):804-809. doi:10.1177/0883073811402345

48. Liu D, Schuchard H, Burston B, Yamashita T, Albert S. Interventions to Reduce Healthcare Disparities in Cancer Screening Among Minority Adults: a Systematic Review. *J Racial and Ethnic Health Disparities*. 2021;8(1):107-126. doi:10.1007/s40615-020-00763-1

49. US Preventive Services Task Force, Davidson KW, Barry MJ, et al. Screening for colorectal cancer: US Preventive Services Task Force Recommendation Statement. *JAMA*. 2021;325(19):1965. doi:10.1001/jama.2021.6238

50. YouTube Developer. YouTube Developer Documentation. https://developers.google.com/youtube/v3/docs/search/list.

51. Rieder B. *YouTube Data Tools*.; 2015. Accessed November 4, 2021. https://tools.digitalmethods.net/netvizz/youtube/

52. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*. Vol 26. Curran Associates, Inc.; 2013. Accessed November 3, 2021. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

53. Wijffels J. *Distributed Representations of Words Using Word2vec*. bnosac; 2021. Accessed November 3, 2021. https://github.com/bnosac/word2vec

54. Benoit K, Watanabe K, Wang H, et al. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*. 2018;3(30):774. doi:10.21105/joss.00774

55. Kamba M, Manabe M, Wakamiya S, et al. Medical Needs Extraction for Breast Cancer Patients from Question and Answer Services: Natural Language Processing-Based Approach. *JMIR Cancer*. 2021;7(4):e32005. doi:10.2196/32005

56. Kalyan KS, Sangeetha S. BertMCN: Mapping colloquial phrases to standard medical concepts using BERT and highway network. *Artificial Intelligence in Medicine*. 2021;112:102008. doi:10.1016/j.artmed.2021.102008

57. Subramanyam KK, Sangeetha S. Deep contextualized medical concept normalization in social media text. *Procedia Computer Science*. 2020;171:1353-1362. doi:10.1016/j.procs.2020.04.145

58. Tan ASL, Bigman CA. Misinformation About Commercial Tobacco Products on Social Media—Implications and Research Opportunities for Reducing Tobacco-Related Health Disparities. *Am J Public Health*. 2020;110(Suppl 3):S281-S283. doi:10.2105/AJPH.2020.305910

59. O’Connor A. Coca-Cola Funds Scientists Who Shift Blame for Obesity Away From Bad Diets. NY Times. Published August 9, 2015. Accessed March 7, 2022. https://well.blogs.nytimes.com/2015/08/09/coca-cola-funds-scientists-who-shift-blame-for-obesity-away-from-bad-diets/

Supplementary Replication Analysis

Euclidean distance, i.e., the textual distinction of the new relevant videos based on NLP characteristics (the greater, the more distinct).

		1	2	3	4	5	6
FOBT	1. original	0	522.06	522.74	517.61	522.21	522.82
	2. iFOBT		0	7.14	10.30	9.11	9.27
	3. hemosure			0	9.95	7.87	7.28
	4. immunochemical				0	8.89	11.05
	5. immunostics					0	7.94
	6. guaiac						0

		1	2	3	4	5	6
Mammogram	1. original	0	247.9	241.8	247.8	249.1	243.3
	2. smartcurve		0	23.7	16.3	11.4	43.7
	3. breastcheck			0	22.7	24.3	36.2
	4. biopsy				0	13.5	42.6
	5. ultrasound					0	45.2
	6. breastcancerawareness						0

		1	2	3	4	5
Pap test	1. original	0	690.1	688.27	691.88	668.2
	2. colposcopy		0	38.94	38.91	49.3

	3. smear	0	9.38	30.3
	4. ASCUS		0	35.4
	5. papsmear			0

[e] "STD" was removed due to zero relevant videos retrieved.