



Computational Analysis of Superfood Representations in News Media

Natasha Gandhi ^a, Caroline Meyer ^a, Piotr Bogdanski^b, and Lukasz Walasek ^b

^aBehaviour and Wellbeing Science Group, WMG, University of Warwick, Coventry, UK; ^bDepartment of Psychology, University of Warwick, Coventry, UK

ABSTRACT

What do berries, avocado, quinoa, and ginger have in common? These food items are often regarded as superfoods, a marketing term that overstates the importance of single food items for one's health and wellbeing. In the present paper, we set out to investigate how purported superfoods are represented in the discourse of online news. We use computational language models to extract the unique topics and terms used to discuss superfoods. Our results show that news coverage is dominated by many specific claims about the healing properties of superfoods. The structural topic model further demonstrates that articles mentioning superfoods are more likely to include topics about a) nutrients, physical appearance, and health in the same context, b) retail strategies, and c) scientific research about the health benefits of superfoods. These results illustrate complex representations of superfoods in news media.

KEYWORDS

Superfoods; text analysis; computational methods; food healthiness perceptions; online news articles

Even though many people strive to eat more healthily, there is little consensus about the *what, when, and how* of a healthy diet. Whilst plenty of guidance and recommendations from government and experts exists (Julia et al., 2021), discourse in online media is now dominated by diverse opinions and advice concerning what people “should” be eating to improve their health and wellbeing. Given that people's perceptions and behavior are shaped by the media representation of food healthfulness (Nagler, 2014; Oakes & Slotterback, 2001), it is essential to better understand what messages about food healthfulness online media perpetuate.

It is known that food marketing influences consumers' perceptions of food healthfulness (Chandon & Wansink, 2012; Plasek et al., 2021; World Health Organization, 2021). In fact, food and beverage companies invest heavily in healthy food marketing in response to consumer demand (Samoggia et al., 2020). The most recognized and researched types of food marketing strategies are those found on food packaging, such as health claims and symbols, package design, and branding (Plasek et al., 2020; Silchenko et al., 2020). However, large food and drink companies are also positioning themselves as nutritional educators (Garcia & Proffitt, 2021), often marketing specific “healthy” product categories and ingredients (Chandon & Wansink, 2012; Mintel, 2016). As a result, the online media discourse surrounding foods also contains marketing concepts disguised as healthy eating advice (MacGregor et al., 2021; Samoggia et al., 2020).

These online media messages have led to the marketing term “superfood” being popularized in everyday discourse (Delicato et al., 2019; Roth & Zawadzki, 2018), which is the focus of our paper. Broadly speaking, superfoods refer to foods naturally rich in macro-nutrients and various vitamins and minerals (Jagdale et al., 2021). This is despite no clear evidence linking a given food item to any health outcomes (Cloutier et al., 2013; Siipi, 2013; Thurecht et al., 2018). Moreover, the superfood

CONTACT Lukasz Walasek  L.Walasek@warwick.ac.uk  Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

narrative in the media is at odds with the advice of nutritional experts, who advocate a view of healthiness in terms of overall dietary patterns (Freeland-Graves & Nitzke, 2013; Lobstein & Davies, 2009; Lusk & Shankar, 2019). In 2007 the European Food Safety Authority banned the word “superfood” on food packaging unless accompanied by a specific permitted health claim (EU Regulation No. 1924/2006, 2006). Likewise in USA and Australia, as all food claims must be supported by scientific evidence, superfood is not a claim that would be authorized for use on food packaging (Food Standards Australia New Zealand, 2018; US Food And Drug Administration, 2016). Yet despite this regulation, perceptions about the superiority of superfoods persist, which is likely to originate from the abundance of mentions in magazine and news articles, blogs, and social media channels (Liu et al., 2021; MacGregor et al., 2021). Consequently, representations and implicit beliefs about certain foods having a superfood status are still being reinforced, leaving consumers misinformed (Delicato et al., 2019). Brands are then able to take advantage of this loophole by adding putative superfoods to the ingredient list of snack items (Breen et al., 2020), baked goods (Meyerding et al., 2018), and beverages (Brownbill et al., 2020), as a means of increasing healthiness perceptions and sales of their products (Meyerding et al., 2018; Mintel, 2016).

As superfood is a term often used to promote food items that are from the main food groups (e.g., fruit and vegetables) (Jagdale et al., 2021), one may wonder whether there is any harm to this practice. At a minimum, the superfood discourse feeds into an overly simplistic view of foods as “good” or “bad,” causing people to abandon efforts to improve their overall diet (Freeland-Graves & Nitzke, 2013). The use of such language can also contribute to the prevalence of disordered eating (Douma et al., 2021), where individuals follow strict, restricted diets and experience guilt after perceived food transgressions in pursuit of healthiness (Galfano et al., 2022). Furthermore, the classification of various foods as superfoods can produce a health halo effect (Amos et al., 2019), whereby consumers incorrectly interpret the health benefits and health risks of certain foods (Breen et al., 2020). Indeed, one side effect is often for consumers to underestimate the calorie consumption of foods perceived as healthy (Carels et al., 2006, 2007; Larkin & Martin, 2016; Provencher et al., 2009), resulting in weight gain and therefore an increased risk of diet-related illnesses.

From a psychological standpoint, little is known about what underpins people’s representations of superfoods. Most academic research in this area uses case studies of known superfoods to investigate either consumer demand for superfood products (Graeff-Hönniger & Khajehi, 2019; Groeniger et al., 2017; Loyer, 2016; Meyerding et al., 2018), nutritional evidence for popular health and nutrient claims (Jagdale et al., 2021; Proestos, 2018; Šamec et al., 2019; Štepec et al., 2020), or the environmental and social consequences from the rise in superfood consumption (Bedoya-Perales et al., 2018; Magrach et al., 2020; Reisman, 2020). Of the studies that have explored perceptions and beliefs surrounding superfoods, participants were provided with pre-defined survey measures (Franco Lucas et al., 2021; Liu et al., 2021; Rojas-Rivas et al., 2019), meaning that our understanding is potentially constrained by the researchers’ expectations and hypotheses.

As a complementary approach to the existing literature, here we undertake an exploratory computational analysis of a large collection of written language. Specifically, we analyze superfood representations in online news articles. To date, only two papers have looked at the coverage of superfoods in news articles (MacGregor et al., 2021; Weitkamp & Eidsvaag, 2014), but for different purposes. Weitkamp and Eidsvaag (2014) investigated the influence of scientists on the information reported in news articles about superfoods, and MacGregor et al. (2021) used critical discourse analysis to explore how the marketing of superfoods in the media promotes neoliberal ideologies.

A similar methodological approach has been used to investigate the relationship between various foods and perceived health benefits, however, the superfood status associated with these foods in the media discourse is overlooked. Moreover, these studies were conducted using social media sources such as Twitter (Kåle & Agbozo, 2020; Lynn et al., 2020; Pilař et al., 2021; Samoggia et al., 2020; Vidal et al., 2015), or Reddit (Blackburn et al., 2018). For instance, one study investigating perceptions of kale using Twitter data between January 2020 and June 2020 found frequent mentions of health-related concepts (e.g., “anti-inflammatory,” “immune boost”) and a lack of reference to taste or

hedonistic concepts in its research (Kåle & Agbozo, 2020). However, the question remains about whether this finding is replicable when analyzing the discourse in mainstream news media, and if it would generalize to other known superfoods.

In this paper, we leverage the latest methods from natural language processing (NLP) to analyze online news articles about known superfoods in different contexts. Our chosen corpus of online news articles (News on the Web – NOW) captures the representation of superfoods in the US and UK media over a 10-year period (2010–2020). To establish unique features of superfood-specific language, we focus on news coverage that contain mentions of foods typically associated with superfoods. We then compare language use between those articles where the word superfood occurs, with those where it is absent. In summary, we conduct three analyses to extract unique feature of the superfood-specific language. First, we explore commonalities in language between superfood articles and descriptions of superfoods provided by a sample of participants. Next, we use two computational techniques to establish predictive features of articles written about superfoods. The first of these computational techniques is text classification. For this analysis, we formally compare articles written in a superfood context with closely matched articles written in a non-superfood context. Therefore, our classifier is forced to rely on subtle differences in language when making predictions as to whether an article is about superfoods or not. An advantage of this approach is that we can uncover the most important words, used as predictors by the trained classifier, to distinguish the representation of a food item in a superfood-specific context. Next, to gain further insight into the interaction between various words and concepts related to superfoods in online news media, we used structural topic modeling (STM). Topic modeling allows us to formally identify latent themes and topics in our sample of news articles about various foods. We use structural topic modeling specifically as it allows us to compare the likelihood of each topic appearing in superfood or non-superfood articles. Finally, as an additional comparison, we replicated both of the above-mentioned computational analysis techniques using articles mentioning “organic” in place of “superfood.” The purpose of this is to ascertain whether we can successfully capture the unique representation of superfoods, rather than a general concept of healthiness.

In short, the following paper makes three novel contributions. First, it presents a unique computational analysis of news media, offering a detailed insight into how the concept of superfood is portrayed. This is important given the lack of consensus about the definition of this term on the one hand, and its prominence in everyday discourse and marketing on the other hand. Second, this paper employs classification and structural topic modeling to directly compare superfood-specific language with the content pertaining to organic foods. This analysis allows us to explore the unique meaning ascribed to the term “superfood” in the broader space of categories and labels used in food-related discourse. Lastly, our study demonstrates the value of using language models to study latent representations of psychologically relevant constructs and topics (cf. Demszky et al., 2023; Gandhi et al., 2022; Wulff & Mata, 2023).

Method

Corpus selection

The corpus of online news articles used in this paper was taken from the NOW Corpus (<http://corpus.byu.edu/now/>), which is a collection of online newspaper and magazine articles, maintained by Mark Davies at Brigham Young University. This is the only English-speaking corpus that is larger than a billion words (Davies, 2017), and so was most suitable for exploring a niche topic like superfoods. The metadata for this corpus includes “article ID,” “word count,” “date,” “country,” “news outlet,” “URL” and “title.” Specifically, for our analysis, we used a static local copy of the NOW Corpus, accessed in May 2020, which covers the period between January 2010 to February 2020. Only articles published in the United States of America and Great Britain were analyzed.

Identifying food names

We compiled a list of food names by downloading data from the U.S. Department of Agriculture (2019) Food Composition Data, and McCance and Widdowson's Composition of Foods Integrated Dataset (Public Health England, 2021), as both are the official sources of information about commonly consumed foods in the USA and UK respectively. This list was used in an initial pre-processing step to filter articles only containing food names.

Participant survey data

We conducted a short online survey on Prolific Academic, in which we asked participants a series of questions about their perceptions of superfoods. Details of this study are provided in the Appendix, but here we note that we used some of our participants' responses to identify food names that are most associated with superfoods. More specifically, we asked each participant in our study (but only those who indicated to be familiar with the term superfood) to name at least five superfoods. We then selected 25 most frequent responses to select online news articles for our computational analyses.

Transparency and openness

All data and code relating to the participant survey are available at <https://tinyurl.com/39p8uwff>. This study's design and its analysis were not pre-registered. The NOW Corpus can be purchased from the <http://corpus.byu.edu/now/>

Data pre-processing and article selection

All steps to extract and clean the relevant sample of online news articles were performed using R, version 4.1.0 (R Core Team, 2021). The R package "spacyr" (Benoit & Matsuo, 2020), was used for data pre-processing. We started with a sample size of 13,871,016 news articles published in the United States of America and Great Britain during the period specified. First, we removed HTML tags, URL links, and non-alphabetical characters (e.g., special characters, numbers, and all punctuation except for hyphens), then standardized all text and titles of the news articles to lowercase. Next, we selected only news articles that contained the word "food" or "diet" either in the text or title. At this stage, there were 779,919 news articles. We then removed all articles that either had duplicated article IDs or that had identical article titles from the same news outlet. Following this, we filtered articles that mentioned at least one of the food names in our food name list (see Appendix for detail), resulting in 547,568 food-related news articles. Our next data pre-processing steps involved tokenization (splitting article texts into single word units, for example, breaking down the sentence "Kale is a superfood" into ["Kale," "is," "a," "superfood"]), removing stop words (frequent words that provide little information e.g., "and," "in" and "the"), and part of speech tagging (to identify a word's function within a sentence, e.g., "green" as an adjective). Part of speech tagging allowed us to select the most meaningful parts of speech (nouns, adjectives, and verbs) from the articles. We also performed lemmatization, turning words representing the same concepts into their base form (e.g., "energizing" and "energized" both became "energize"). Next, we replaced all possible spellings of the word superfood (e.g., "superfood," "super-food," "super foods") with the former spelling for consistency. This then allowed us to identify the 1,169 articles that mentioned the word "superfood" at least once in the article text or title, and the 216,769 that did not. Consequently, we further subset the data to only include articles that specifically mentioned the 25 superfood names given by participants in the online survey mentioned above. This resulted in a final sample of 57,853 news articles, with 872 articles that mentioned the word superfood at least once (henceforth referred to as superfood articles). In the last step, all articles were tagged with a dummy variable to denote their status as either a superfood or non-superfood article.

Computational methods

Word frequency: a bag-of-words model

As an initial exploration of our corpus, we used a word counting technique known as a bag-of-words model to provide a simplified representation of our superfood articles sample (see Kowsari et al., 2019). The bag-of-words approach was chosen over other known frequency statistics (such as *tf_idf* and weighted logs odd ratio) because these alternative methods would give priority to unique and obscure words even when found in only a small number of articles. Instead, we were interested in finding the most common words across all articles in our chosen context.

For this analysis, we included additional data pre-processing steps. The same steps were replicated when pre-processing participants' survey responses to two questions asking them to describe or define a superfood, for comparative purposes. Further details are provided in the Appendix.

Text classification

Text classification is a supervised machine learning technique that allows us to predict whether an article is written in a superfood context or not. An offshoot of this approach is that we can subsequently identify the terms that underlie the classifier's predictions. Consequently, text classification can enable us to make inferences about concepts that are most likely to be associated with superfoods in a sample of articles discussing the same group of foods.

For text classification, it is optimal to have a balanced distribution of articles in each group. However, our dataset was highly unbalanced, with the superfood articles making up only 1% of the sample. In addition, the much larger 'non-superfood' class likely included articles that only scarcely referenced the food items of interest. Therefore, to down-sample the non-superfood articles and balance the corpus, we used propensity score matching (Ho et al., 2011; Rubin & Rosenbaum, 1985). We obtained the propensity scores for each sample using a generalized linear model with a logit link function. We regressed the article class (i.e., superfood mentioning or not) on the counts of the top 25 superfoods selected by survey participants. The premise was that the logit model could estimate the probability of each article being a superfood or non-superfood article, and the predicted probabilities (propensity scores) would reveal how likely each non-superfood article could serve as a viable counterfactual (or replacement) for that superfood article. All superfood articles were paired with their nearest neighbor; the non-superfood article with the closest propensity score. Note that each match was independent, and thus the same non-superfood article could be matched to several superfood articles (greedy matching). All unmatched non-superfood articles were then discarded from the sample, allowing for maximum homogeneity between the two comparison groups and a reduced sample size. In total there were 872 pairs of articles in our text classification sample, 10% of which were always held-out for model evaluation during cross-validation (as explained below).

Next, we trained a text classification model to discriminate between the superfood and non-superfood articles in our balanced sample. Specifically, we used the logistic regression classifier (also known as the maximum entropy classifier); a linear model often used for text classification tasks due to its interpretability. We chose to encode the data with unigrams, bigrams, and trigrams. Each type of an *n*-gram is an expression of *N* consecutive words combined. Previous studies have shown that increasing the *n*-gram range leads to better performance in a variety of text classification tasks (Bharadwaj & Shao, 2019; Shah et al., 2018). For the purpose of our research question, more complex *n*-grams may also be better suited for capturing unique contextual information that defines discourse surrounding superfoods. For example, the meaning of the word "superfood" in "healthy superfood" and "expensive superfood" is the same if only unigrams are used. However, the two bigrams can have a different meaning, despite both mentioning superfoods. We represented the data using a document-feature matrix, where each column represented an *n*-gram, each row represented a news article, and the observations for each of the documents corresponded with the count of the word's occurrence. Take a single hypothetical sentence "Superfoods are very good." This sentence would be represented by 10 columns in the document-feature matrix: three corresponding to the unigrams ("superfoods,"

“are,” “very,” and “good”), three corresponding to the bigrams (“superfoods are,” “are very,” and “very good”), and two for the trigrams (“superfoods are very,” “are very good”).

A common issue in machine learning models is overfitting. To mitigate against this, we applied an L1 penalty, which shrinks majority of irrelevant coefficients to 0.¹ When training the regularized logistic regression, it was important to choose the optimal strength of the penalty imposed on the L1 norm of the model’s coefficients. We tuned the regularization strength parameter by running a grid search over a range of values and evaluating the average F1 score on a 10-fold cross-validation split for each of them. The F1 score is the harmonic average of the model’s precision and recall, further discussed in the results section. The entire fitting procedure was implemented in the “glmnet” package in R (Simon et al., 2011).

As an additional analysis, we repeated all the aforementioned steps replacing “superfood” with the word “organic.” The reason was to ascertain whether the representation of superfoods is unique or reflects an overarching perception of food healthiness. Specifically, using the sample of 25 superfoods given by participants, we marked news articles that mentioned the word “organic” vs. those that did not.

Topic modeling

Finally, to examine whether the differences identified by the classifier generalized to more broad, latent themes found in the entire subset of the news articles corpus, we used the Structural Topic Model (STM) (Roberts et al., 2014, 2016). Topic modeling refers to a family of unsupervised statistical learning techniques that identify underlying latent semantic structures characterized by the frequent occurrence of a vocabulary subset in a corpus of natural texts. In the past, topic modeling has been applied to analyze corpora from a variety of areas, such as social media discourse (Zamani et al., 2020), financial news (Bybee et al., 2020), and historical texts (Barron et al., 2018).

Topic modeling relies on several assumptions, which enabled us to extract topics from our newspaper corpus. One such assumption is that each document is composed of a mixture of topics, and each topic is formed using a probability distribution of multiple words. In the same manner as the bag of words model, it also assumes that there is no order to the words in a document and that documents are independent. The distinguishing feature of STM is that, while building on the basic idea of probabilistic topic modeling, it allowed us to incorporate document-level covariates (or metadata) into the model’s structure (Roberts et al., 2019). STM was therefore most appropriate for our goal of quantifying the effect of our dummy variable (superfood article or not) on the topical structure of the news articles. As such, STM allowed us to identify topics present across all articles that were more prominent when the term “superfood” was used.

To facilitate model estimation, we further subset our data to only include articles mentioning any of the superfood items specified by the survey participants more than twice, resulting in a sample of 18,219 non-superfood articles and 577 superfood articles. Since the number of the latent topics to be estimated must be specified a priori, we conducted a grid search over a range of values and chose the highest number ($K = 12$) that offered a notable improvement in the exclusivity score calculated over the top 10 words in each of the topics. A word is said to be exclusive to a given topic if it has a high probability of appearing in the topic and a low probability of appearing in the other topics estimated by the model (Roberts et al., 2014). The exclusivity of a model is the aggregated exclusivity of the top N words for each of the topics. The details of the parameter search can be found in [Figure A1](#) of the Appendix.

¹Formally, logistic regression solves the following optimization problem:

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} - \left[\frac{1}{N} \sum_{i=1}^N y_i \cdot (\beta_0 + x_i^T \beta) - \log \left(1 + e^{(\beta_0 + x_i^T \beta)} \right) \right] + \lambda \beta_1$$
, while the L1-regularized variant changes that to $\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} - \left[\frac{1}{N} \sum_{i=1}^N y_i \cdot (\beta_0 + x_i^T \beta) - \log \left(1 + e^{(\beta_0 + x_i^T \beta)} \right) \right] + \lambda \beta_1$, where $\lambda \beta_1$ is the regularization penalty imposed on the model’s coefficients.

Results

Superfood names

In total, there were 115 unique superfood names listed by our survey participants (for the full list, see the OSF repository associated with this project: <https://tinyurl.com/39p8uwff>), with 51 given by at least two participants. As can be seen from Figure 1, these 51 foods identified as superfoods belong to a wide range of food categories from Vegetables and Vegetable Products (e.g., kale and spinach) to Spices and Herbs (e.g., ginger and turmeric), and Sweets (e.g., dark chocolate). Unsurprisingly, most of these foods were from the “Vegetables and Vegetable Products” category and the “Fruit and Fruit Juices” category with “blueberry,” “avocado,” and “kale” being the most frequently mentioned (53, 46 and 42 mentions, respectively). The top 25 food names mentioned by participants, and the food names subsequently used for the computational analysis, are highlighted in bold on the X-axis of Figure 1. These foods are listed in descending order of word frequency and include blueberry, avocado, kale, goji, quinoa, spinach, chia seed, nut, broccoli, acai, ginger, egg, berry, fish, sweet potato, beet, green tea, spirulina, almond, pomegranate, salmon, turmeric, wheatgrass, yogurt, and oats.

Word frequency

Our first analysis concerns the distribution of the most frequent words in superfood articles. These are presented in Figure 2 alongside data from our survey participants, separately for adjectives, nouns, verbs, and bigrams. Even at a glance, the concept of health is most prominent, but we can also see several references to the sensory properties of foods, naturalness, weight control, and scientific research.

The relationship between superfoods and health is demonstrated by the disproportionate prevalence of health-related terms in superfood articles. Explicit references to health-related terms (e.g.,

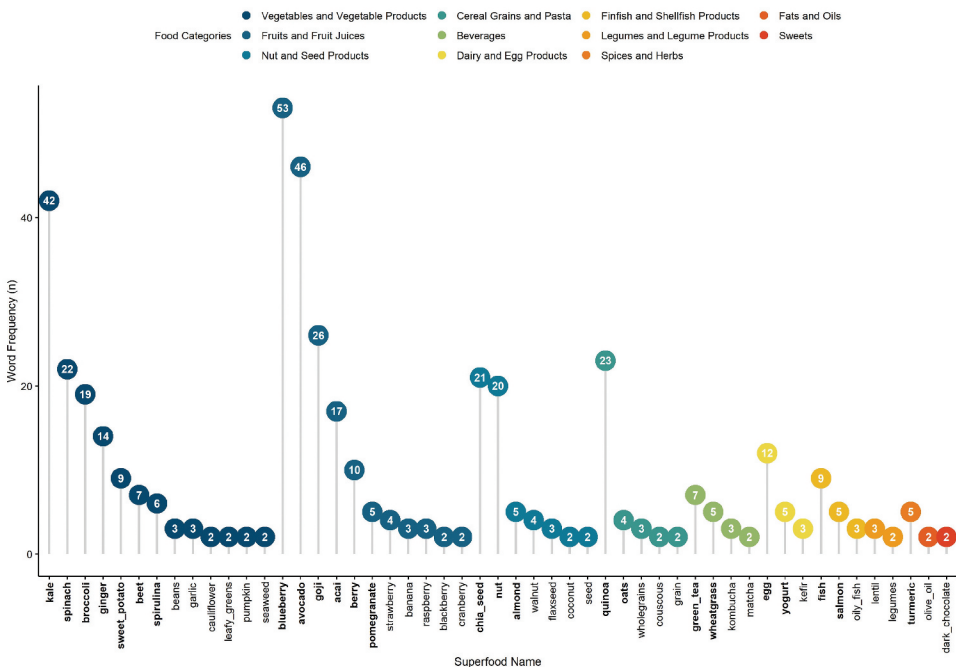


Figure 1. All food items, that participants associate with the “superfood” label, mentioned more than once (using word frequency). For clarity, foods are separated by food category. The top 25 foods, subsequently used in the computational analysis, are highlighted in bold on the X-axis.

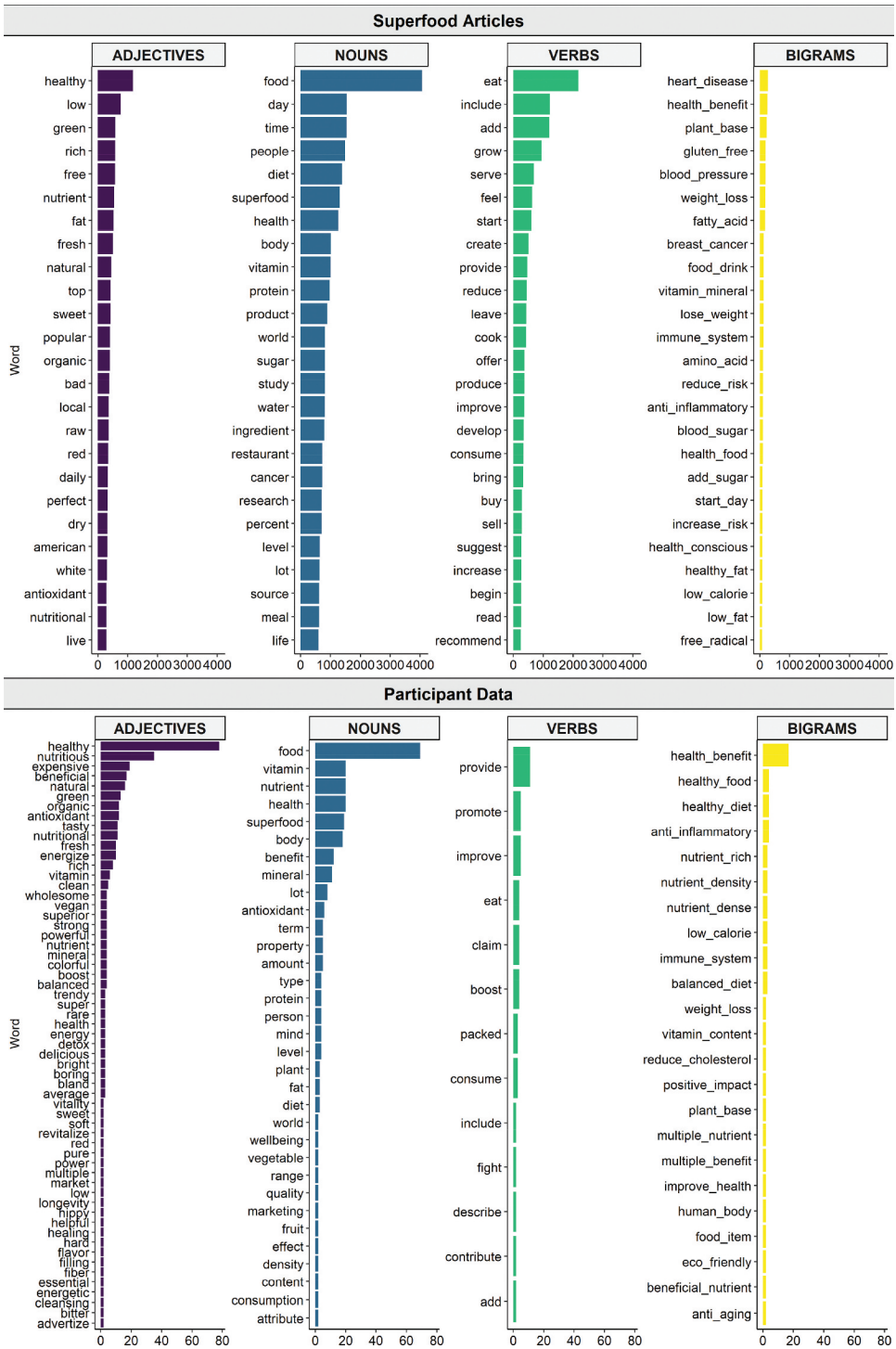


Figure 2. A comparison of the top 25 adjectives, nouns, verbs, and bigrams taken from online news articles written in a superfood context with the adjectives, nouns, verbs, and bigrams mentioned by more than one participant to describe or define a superfood.

“healthy” ($N = 1,179$), “health” ($N = 1,261$), and “health_benefit” ($N = 261$)) are evident in the adjectives, nouns, and bigrams plots generated from the superfood article data. Broader connections to health were also made with mentions of positive and negative nutrients such as “vitamin” ($N = 1001$), “protein” ($N = 976$), “sugar” ($N = 822$), “antioxidant(s)” ($N = 296$), “fatty_acid(s)” ($N = 184$), “amino_acid(s)” ($N = 115$), and “healthy_fat” ($N = 89$). Relatedly, a link with specific diseases is apparent in the media coverage, with an abundance of mentions of “cancer” ($N = 725$), “breast_cancer” ($N = 142$), and “heart_disease” ($N = 278$). Moreover, texts included numerous references to medical concepts and terms such as “blood_pressure” ($N = 195$), “immune_system” ($N = 124$), “anti_inflammatory” ($N = 107$), and “blood_sugar” ($N = 105$). Together with common bigrams such as “reduce_risk” ($N = 113$), “increase_risk” ($N = 96$), and verbs (e.g., “provide” ($N = 465$), “reduce” ($N = 447$), “offer” ($N = 372$), “improve” ($N = 359$) and “increase” ($N = 265$)), these words highlight a discourse where superfoods are likened to medicine.

Reference to the sensory properties of superfoods can be seen from the most frequently used adjectives in superfood articles. Color is most mentioned, with “green” ($N = 592$) the third most frequent adjective, while red ($N = 349$) and white ($N = 314$) are also present. Surprisingly, the only taste descriptor in the top 25 adjectives was “sweet” ($N = 426$). Instead, most adjectives in superfood articles refer to the concept of naturalness (e.g., “fresh” ($N = 508$), “natural” ($N = 453$), “organic” ($N = 406$), and “raw” ($N = 361$)). The theme of weight loss is most apparent when looking at bigrams, with frequent collocations in the superfood articles including “weight_loss” ($N = 194$), “lose_weight” ($N = 127$), and “low_calories” ($N = 85$). Moreover, we can also see words (e.g., “study” ($N = 820$), “research” ($N = 703$), “percent” ($N = 701$)) that allude to the use of scientific evidence as a persuasive technique for the promotion of superfoods in the media.

When comparing these frequent words from the superfood articles to the participant data, we find a similar prevalence of concepts. First, the healthiness aspect of superfoods is most emphasized, with “healthy” ($N = 78$) being the top adjective, many references to health benefits and medically related terms (e.g., “anti_inflammatory” ($N = 4$), “immune_system” ($N = 3$) and “reduce_cholesterol” ($n = 2$), and multiple nutrients named (e.g., “vitamin,” “antioxidant,” “protein”). Similarly to the superfood articles, the most common terms for sensory properties reflected the appearance of superfoods (“green” ($N = 13$), “red” ($N = 2$), “bright” ($N = 2$) and the more general “colorful” ($N = 2$)). Taste was also referred to, but both in a positive and negative way (e.g., “tasty” ($N = 11$), “delicious” ($N = 3$), “bland” ($N = 3$)). Participants also associate superfoods with naturalness, with 16 individuals mentioning the term “natural” explicitly. Plus, other naturalness related words such as “organic” ($N = 12$), “fresh” ($N = 10$), “clean” ($N = 5$), “wholesome” ($N = 4$) and “pure” ($N = 2$), were mentioned. A few participants also referred to aspects of weight control. For example, “weight_loss” ($N = 2$) was explicitly mentioned but also “low_calorie” content ($N = 3$), “detox” ($N = 3$) and “cleansing” ($N = 2$), which are often used in the context of weight loss. Interestingly, no science or research terms were referenced more than once, suggesting that participants may either not remember this persuasion technique or deem it unimportant.

To uncover the specific words that underlie predictions about an article being about superfoods, we used text classification. Note that the use of propensity score matching means each of the superfood articles was compared against a non-superfood article that was highly similar in context (determined by similar counts of the top 25 foods selected by survey participants). For robustness purposes, we replicated this analysis by comparing articles that mentioned the word “organic” (organic articles) vs those that did not (non-organic articles). Table 1 shows the performance of the classifier model on the

Table 1. Text classification validation performance.

	Precision	Recall	Train size	Test size	Accuracy
Non-superfoods	0.663	0.701	785	87	0.672
Superfoods	0.683	0.644	785	87	
Non-organic	0.688	0.803	4156	462	0.720
Organic	0.764	0.636	4156	462	

out-of-sample dataset. We summarize the results in terms of three commonly used classification metrics – accuracy, precision, and recall. Accuracy is simply the number of cases predicted to be in their true respective class. Precision refers to the number of examples correctly predicted as belonging to a given class, as a proportion of all examples belonging to that respective class. Recall (also known as sensitivity) represents the number of observations correctly predicted as a given class divided by the total number of observations truly belonging to that class. As seen in the Table, our classifier has an accuracy rate of 67% for superfood articles and 72% for organic articles, meaning that it can classify these online news articles better than chance. As a result, we can conclude that our classifier model can sufficiently pick up linguistic differences between articles written about foods in a superfood context or not, as well as in an organic context or not.

Text classification

A benefit of our approach is that we can use the classifier model to find the n-grams with the highest probability of being in a news article classified as a superfood article or an organic food article. Figure 3 presents the top 100 n-grams, scaled to be proportional to the log-odds of the corresponding coefficients for each group.

Looking first at the superfood classifier n-grams in Figure 3, the first thing to note is an overlap between the n-grams produced using this approach and the words obtained from the simpler word-counting technique in Figure 2. The relationship between superfoods and health is yet again prevalent. For example, many n-grams reference the nutritional composition of foods such as “source (of) potassium,” “high protein,” “nutrient-rich,” “mineral (and) vitamin,” “vitamin c (and) e,” “monounsaturated fat,” “low sugar,” “anti-oxidant,” and “good/bad cholesterol.” The classifier also picks up on the discourse surrounding superfoods and illnesses, with n-grams predictive of superfood articles including “heart disease percent,” “cancer percent/fight,” “boost immune system,” “carcinogen” and “reduce stress.”

Additionally, the unigram “premature” is highly predictive but it includes two uses of the term in two separate contexts: premature death (for 16 of the superfood articles) and premature aging (for 13 of the superfood articles). Other n-grams that show a relationship with beauty include “protect skin,” “facial,” “hydrating,” “regime,” all of which were not seen in the word frequency (bag-of-words) analysis.

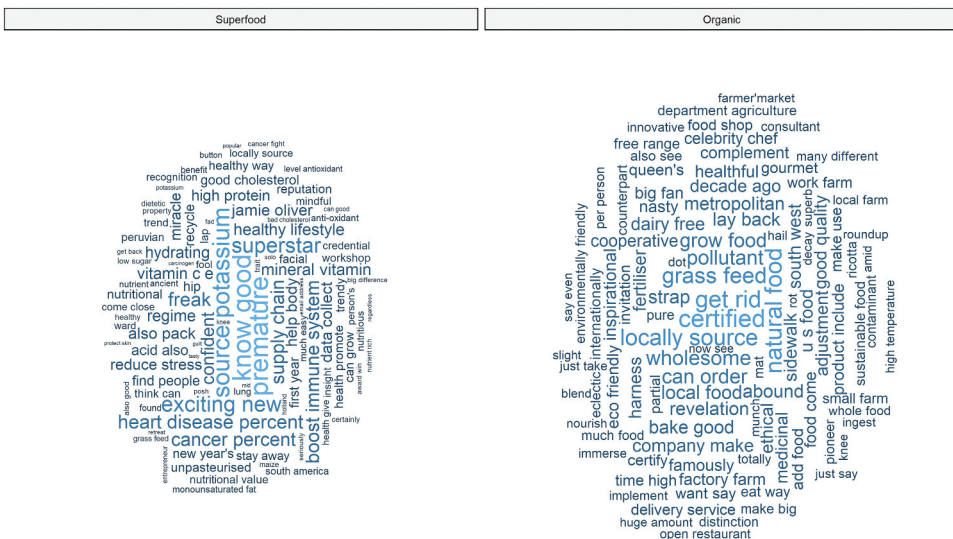


Figure 3. Word cloud of the 100 n-grams most predictive of the terms “superfood” and “organic” occurring in the sample of online news articles.

In terms of sensory attributes, “tasty” is the only n-gram predictive of a superfood article, and the only reference to natural content is “unpasteurized” or “grass feed.” We also detect nuances in language such as the paradoxical nature of foods marketed as superfoods, with predictive n-grams indicating a local origin (“locally sourced”) but also their exotic nature (“ancient,” “Peruvian” and “South America”).

In comparison, there is little crossover between n-grams predictive of superfood articles and organic foods, even though both terms are typically perceived as cues to healthiness. Instead, articles written in the context of organic foods are predominately centered around food production methods, the concept of naturalness, and foods’ environmental impact. Nonetheless, the most predictive n-gram was “certified”, suggesting that articles about organic foods may highlight the necessary government standards and regulations that must be met for a food to be labeled organic. This is followed by “natural food”, which along with “wholesome”, “pure”, “whole food” as well as “decay” and “rot”, suggest a strong link between organic foods and naturalness. Similarly, there is an emphasis on the local aspect of foods with n-grams like “locally source(d)”, “local food/farm”, “small farm”, and “farmer”(s) market” also being predictive of news articles about organic foods. Convenience or ease of access is also implied, as “delivery service”, “open restaurant” and “metropolitan” may suggest. The environmental and ethical impact is another theme that stands out, with direct references to “eco/ environmentally friendly”, “sustainable food”, “ethical”, “good quality” and “cooperative”. Moreover, while the health benefits of foods are alluded to with descriptions like “healthful” and “medicinal”, the focus is on the presence or absence of chemicals (e.g., “fertilizer”, “pollutant”, “contaminant”). From this, we can conclude that our model successfully picks up on representations of healthiness that are unique to superfoods, rather than all health-related terms like organic.

Topic modelling

To confirm and extend the findings from the text classification, we used STM to extract topics. Essentially, STM enables automated discovery of differences in latent themes and topics between articles written about the same 25 foods from different contexts. Figure 4 summarizes the estimated effect of the presence of the “superfood” or “organic” term on the difference in proportion of a latent topic appearing in the articles. A positive value on the X-axis indicates a larger prevalence of a given topic in superfood articles (Panel A) or organic articles (Panel B). Our two reported topic models were fit with the same 12 topics as determined by the grid search (see Methods for detail). Topic labels were assigned by using the top 10 keywords most associated with the given topic.

In articles written about the same 25 foods, what group of words (constituting a topic) is most likely to occur when the term “superfood” is mentioned? As shown in Panel A of Figure 4, the topic relating

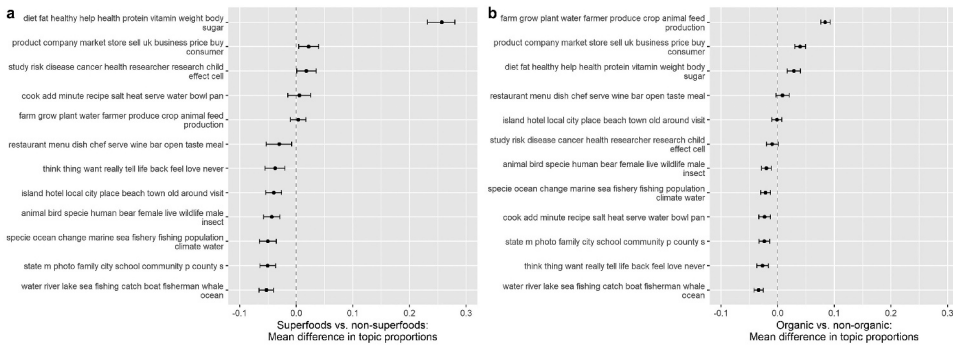


Figure 4. Estimated differences (and 95% confidence interval) in topic probabilities between superfood and non-superfood articles (left panel – A), and organic food and non-organic food articles (right panel – B) Each plot is ordered by the topics most prevalent in the superfood or organic articles.

to diet and weight stands out, with a 25.72% (95% CI [23.20%, 28.15%]) higher likelihood of occurring in superfood articles. The words most indicative of this topic were “diet, fat, healthy, help, health, protein, vitamin, weight, body, and sugar.” Notably, this topic includes both terms related to food nutrients (“vitamin,” “protein,” and arguably “fat”) and appearance (“weight,” “shape”). The fact that these terms appear alongside “health” as one of the most representative words, suggests that the model detected a relationship in discourse between diet, appearance, and health. In comparison, Panel B shows that this topic was the third most prevalent in organic articles relative to non-organic articles. Moreover, it was only 2.90% (95% CI [1.78%, 4.03%]) more likely to occur in an organic context. Given that this topic is considerably more prominent in the discussion around superfoods, one could infer that this association is pivotal in the representation of superfoods in the media.

Perhaps less surprising was the higher likelihood of retailing concepts appearing in superfood articles relative to non-superfood articles about the same foods. However, although this topic was the second most prevalent in superfood articles, the mean difference was much smaller compared with the first (diet, appearance, and weight) topic at 0.02 (95% CI [0.00, 0.04]). The 10 words that formed our interpretation of this topic consisted of “product, company, market, store, sell, uk, business, price, buy, and consumer.” This retailing topic was also the second most likely to appear in organic food articles in proportion to non-organic food articles, but was slightly more likely to occur in organic articles than superfood articles (mean difference of 0.04, 95% CI [0.03, 0.05]).

The third topic more likely (by 1.94%, 95% CI [0.04, 3.84%]) to appear in superfood articles vs non-superfood articles was one associated with scientific research. Keywords constituting this topic such as “study, “disease,” “cancer,” “health,” and “research” had also been present in the findings of our previously mentioned computational analysis techniques. Again, having the word “health” within the top 10 words of this topic suggests a discourse where scientific evidence is given to suggest a relationship with these foods and health or disease. Interestingly, despite more debate and scientific research conducted to assess the relationship between organic foods and health, this topic of research was less likely (by -0.97%, 95% CI [-1.92%, 0.03%]) to occur in organic news articles than non-organic articles. Moreover, this contrast in discourse demonstrates another distinction between representations of superfoods and organic foods in the media.

Our topic modeling approach also reveals that the concept of cooking, “cook, add, minute, recipe, salt, heat, serve, water, bowl, pan,” was slightly more likely to be found in articles written about superfoods, by 0.75% (95% CI [-1.14%, 2.71%]). This was the topic with the largest difference in ranking between superfood and organic contexts, shown from a visual comparison between the superfood (Panel A) and organic (Panel B) plots. We can also see that this cooking topic was ranked higher than the topic on eating out “restaurant, menu, dish, chef, serve, wine, bar, open, taste and meal,” a topic found to be less likely to appear in superfood articles than non-superfood articles (by -3.26%, 95% CI [-5.36%, -1.05%]). Thus, these findings suggest superfoods are more likely to be promoted as ingredients for home cooking, rather than as a treat when dining out. Conversely, the opposite is true for organic foods, where organic is more likely (by 0.84%, 95% CI [-0.37%, 2.03%]) to be discussed in the same context as eating out, but less likely (by -2.28%, 95% CI [-3.38%, -1.23%]) to be mentioned in references to recipes.

Contrary to expectations, a topic relating to naturalness was only 0.34% (95% CI [-1.03%, 1.90%]) more likely to occur in superfood articles relative to non-superfood contexts. The words used to define this topic were “farm, grow, plant, water, farmer, produce, crop, animal, feed, production.” On the other hand, this topic was ranked first for likelihood (at 8.30%, 95% CI [7.41%, 9.17%]) in organic vs non-organic articles, supporting the assumption that this finding is due to representation differences between these two contexts. It is also worth noting that the keywords relating to naturalness all appear to relate specifically to food production methods, with terms relating to the rawness or purity not detected by the topic model.

One may also wonder about the topics least likely to occur about the 25 foods in a superfood context, and how this differs from topics least likely to occur in an organic context. [Figure 4](#) shows us that the majority of the topics (7 out of the 12) were found in articles where the term superfood was not

mentioned, with the same being true for the organic comparison. Of these topics, the one referring to the fishing industry was least likely (mean difference of -0.05 , 95% CI $[-0.07, -0.04]$) to be mentioned in a superfood context, defined by the topic label “water, river, lake, sea, fishing, catch, boat, fisherman, whale, ocean.” This was also true of the topic least likely to occur in articles about organic foods relative to non-organic foods, but with a mean difference of -0.03 (95% CI $[-0.04\%, -0.02\%]$). The second least likely topic in superfoods articles (mean difference of -0.05 , 95% CI $[-0.06, -0.04]$) consisted of words such as “state, photo, family, city, school, community, country,” which appears to reflect a discussion of social factors surrounding the consumption of our sample of foods. The same topic was ranked relatively similarly (third-least likely) in the organic comparison (mean difference of -0.02 , 95% CI $[-0.03, -0.02]$). Interestingly, the third least likely topic (mean difference of -0.05 , 95% CI $[-0.06, -0.04]$) to occur in a superfood context demonstrates a focus on the unsustainability of fishing practices and environmental consequences (consisting of unique words such as “climate, change, fishery, population”). Along similar lines was a topic referencing human-wildlife coexistence (mean difference of -0.04 , 95% CI $[-0.06, -0.03]$) with the words most representative of the topic being “animal, bird, specie(s), human, bear, female, live, wildlife, male, insect.” These two topics were also less likely to occur in an organic context but had a slightly higher mean difference (-0.02 , 95% CI $[-0.03, -0.01]$ and -0.02 , 95% CI $[-0.03, -0.01]$ respectively) in comparison to superfoods vs non-superfood articles. The next topic slightly more likely to be mentioned outside of a superfood context (by -4.02% , 95% CI $[-5.82\%, -2.32\%]$), seems to reflect a discourse relating to food tourism, with words including “island, hotel, local, city, place, beach, town, old, around, visit.” For organic articles, this food tourism topic was equally likely to occur relative to non-organic articles (mean difference of 0.00 , 95% CI $[-0.01, 0.01]$). Last was a more abstract topic of words denoting the writer’s personal perspective (“think, want, really, life, back, feel, love, never”), which had a small mean difference of -0.04 (95% CI $[-0.05\%, -0.03\%]$). However, by comparison, this same topic was the second least likely to appear in an organic context (mean difference of -0.03 , 95% CI $-0.04, -0.02]$).

Discussion

Our computational approach makes several contributions to our understanding of superfood representation in online news media. Through a series of comparisons (with participant data, and with articles about the same foods in either a non-superfood context or organic context) we extracted the words, concepts, and topics most strongly associated with the term superfood. First, we found a unique emphasis on the relationship between individual foods and health benefits in superfood articles. Second, we observe a distinct use of medical terminology (such as “cancer,” “immune system,” “heart disease,” “risk”) in a superfood context, which is notably absent in the representation of organic foods. Third, against our expectations, terms stressing the naturalness and environmental impact of these foods were infrequent in the characterization of superfoods. As a whole, considering superfood has no official definition, our findings offer a deeper understanding into the concept. Given the role media plays in shaping people’s beliefs and attitudes, our results also contribute to our understanding of the origins of misconceptions concerning the health and well-being benefits of superfoods.

Although a link between superfoods and health benefits is expected, our result provides support for previous research findings in a data-driven manner (Franco Lucas et al., 2021; Loyer, 2016; Rojas-Rivas et al., 2019). All three of our bottom-up approaches found “health,” “healthy” and specific nutrients such as “protein,” “sugar,” and “vitamin” to have the largest association with the term “superfood.” Such consistency of findings between our three bottom-up techniques demonstrates the robustness of computational approaches in uncovering superfood representations. Indeed, drawing attention to isolated compounds present in foods is not new to superfood advertising (Scrinis, 2013); in fact, it was a highly successful marketing strategy on food packaging until unfounded nutrient claims were banned in the 1990s (Goldberg & Sliwa, 2011; Silchenko et al., 2020). Nonetheless, a strength of our topic modeling approach is that we can now see the

extent of the association between this nutrient-focused conceptualization of health and superfoods. As such, we observe how the representation of superfood with health considerably exceeds the relationship of organic and health, and even organic and naturalness in media discourse. While we cannot use present results to claim that this superfood representation in the media directly influences consumer perceptions, similar language in participant responses does suggest individuals are at least aware of the same association. Furthermore, we found that mentions of health, various nutrients, weight, and appearance, emerge within the same topic (taken from our topic modeling analysis). This implies a discourse where superfoods are touted for weight loss as part of health messaging (Rodney, 2018; Sikka, 2019), despite scientific evidence establishing that weight is a poor indicator of health (Frederick et al., 2020; Saguy & Almeling, 2008). Considering that nutrient-focused marketing detracts from the recommended “total diet approach” to healthy eating (Freeland-Graves & Nitzke, 2013), plus the implications of a suggested linear relationship between weight and health for disordered eating behaviors (Frederick et al., 2020; Pilař et al., 2021), our findings highlight a need to further extend health and nutrition claim regulation to the online media marketing of food items.

Our results also draw attention to a medicinal representation of foods, centered around disease prevention, which is more likely to occur in a superfood context. Cancer is the most frequently associated disease in our article sample about superfoods, with a slight emphasis on breast cancer. However, in the words of Cancer Research (2020) “there is no good evidence that any one food prevents cancer, including superfoods.” Most of the research conducted on individual foods that are reported in the media are either from animal studies (Jagdale et al., 2021), *in vitro* (outside of a living organism) studies (Šamec et al., 2019), or single studies that should be interpreted with caution (Ladher, 2016). This is also true for the other diseases mentioned in our superfood articles (e.g., heart disease). Thus, the ability of our untrained model to identify a representation of superfoods based on weak evidence (Inoue-Choi et al., 2013), reinforces the role of the media in creating confusion about healthy eating (Hackman & Moe, 1999; Nagler, 2014; Weitkamp & Eidsvaag, 2014). Again, as evident from the language reflected in participant responses, the superfood discourse in the media may therefore help explain the discrepancy between people’s inaccurate beliefs about food healthiness and official dietary guidance.

Contrary to the entwined relationship between health and naturalness found in previous research (Gandhi et al., 2022; Loyer, 2016; Michel et al., 2021; Perkovic et al., 2021; Roman et al., 2017; Siipi, 2013), naturalness was not a concept stressed in the online news article representation of superfoods. This is more surprising because naturalness representation was detected in participant responses, as well as in the comparative analysis of organic food articles. One possible explanation is that the relationship between chosen superfoods and naturalness can be assumed by design, whereas the same food can be sold as organic or not, and thus naturalness associations would need highlighting in organic discourse. Another factor to consider is that superfoods are often sold in supplement form, involving high levels of processing, and so claims about their naturalness may appear contradictory.

The general lack of coverage concerning the social and environmental consequences of superfoods in online news articles may explain why some participants perceived superfoods as “eco-friendly.” However, the existing scientific literature on superfoods reports a detrimental social and environmental impact due to increased demand worldwide (Bedoya-Perales et al., 2018; Loyer, 2016; Magrath et al., 2020). It is perhaps unsurprising that superfoods are spun in a positive light within a media marketing discourse, even if this unbalanced representation further enhances the halo effect of superfoods. However, given the relatively broad range of news articles from a variety of news outlets in this study, one might expect a higher prevalence of this topic in superfood articles than revealed in our topic modeling analysis. As a result, it would be interesting to explore whether differences occur between different news outlets and if this finding is also true of representations in social media. Moreover, as consumers demonstrate a preference for environmentally friendly foods (Franco Lucas et al., 2021), a recommendation for future research is to assess how increased awareness of environmental consequences from global-scale production of superfoods (e.g., water depletion, soil

degradation, reduction in biodiversity, and carbon footprint) might influence perceptions, preferences, and purchase behavior for superfoods.

Our relatively small number of superfood articles, both initially (1,169) and after selecting only articles mentioning 25 known superfoods (872), is unlikely to capture the entirety of online superfood news articles written between January 2010 and February 2020. We chose to prioritize minimizing researcher influence, selecting articles from our corpus using arbitrary means (count of the word “superfood”) rather than adding further news articles from specific news outlets (e.g., The New York Times, or The Guardian). It is also worth noting the existence of related marketing terms that have spawned from the superfood discourse (e.g., “superfruit,” “supergrain,” and “super berry”) (Liu et al., 2021; Loyer, 2016). Unfortunately, too few articles were available in the NOW corpus to extract meaningful themes and representations. Nonetheless, despite some limitations, our approach captures meaningful patterns that are consistent with discourse findings about known superfoods using other corpora (Kāle & Agbozo, 2020).

One limitation of the NOW corpus is that it only offers data for the period of 10 years, which prevents us from drawing any conclusions on how superfood related discourse might have changed over time. We note that there are other data sources (e.g., American Stories database) and pre-trained time-variant language models (e.g., histwords project), that could be used for this purpose. Combined with the insights of the present study, future work could explore how the definitions of superfoods and their relation to organic foods changed over the years (and even across different places).

Overall, we believe that the strength of our approach is that we can uncover and quantify the unique representations of superfoods in the news media. While the term superfood is banned on food packaging, here we demonstrate how this term is prevalent outside of the supermarket environment. More importantly, we demonstrate a number of unique dimensions that make up the representation of superfoods in the media. The next stage for researchers is to ascertain the extent to which these representations influence food perceptions, and ultimately food choices. For now, we recommend advertising regulatory bodies pay close attention to the loopholes being used to produce these misleading and potentially harmful associations.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Engineering and Physical Sciences Research Council (grant EP/N509796/1: project reference no.1939178) and the Research Development Fund from the University of Warwick awarded to Prof Caroline Meyer and Dr. Lukasz Walasek. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ORCID

Natasha Gandhi  <http://orcid.org/0000-0002-4834-1045>
Caroline Meyer  <http://orcid.org/0000-0003-0684-299X>
Lukasz Walasek  <http://orcid.org/0000-0002-7360-0037>

Ethical statement

Ethical approval for this study was obtained from the University of Warwick’s Biomedical and Scientific Research Ethics Sub-Committee (REGO-2018-2268).

References

- Amos, C., Hansen, J. C., & King, S. (2019). All-natural versus organic: Are the labels equivalent in consumers' minds? *Journal of Consumer Marketing*, 36(4), 516–526. <https://doi.org/10.1108/JCM-05-2018-2664>
- Barron, A., Huang, J., Spang, R., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences*, 115(18), 4607–4612. <https://doi.org/10.1073/pnas.1717729115>
- Bedoya-Perales, N. S., Pumi, G., Talamini, E., & Padula, A. D. (2018). The quinoa boom in Peru: Will land competition threaten sustainability in one of the cradles of agriculture? *Land Use Policy*, 79, 475–480. <https://doi.org/10.1016/j.landusepol.2018.08.039>
- Benoit, K., & Matsuo, A. (2020). Spacyr: Wrapper to the 'spacy' 'nlp' library [computer software manual]. <https://CRAN.R-project.org/package=spacyr> (R package version 1.2.1)
- Bharadwaj, P., & Shao, Z. (2019). Fake news detection with semantic features and text mining. *International Journal on Natural Language Computing*, 8(3), 17–22. *IJNLC*, 8(3). <https://doi.org/10.5121/ijnlc.2019.8302>
- Blackburn, K. G., Yilmaz, G., & Boyd, R. L. (2018). Food for thought: Exploring how people think and talk about food online. *Appetite*, 123, 390–401. <https://doi.org/10.1016/j.appet.2018.01.022>
- Breen, M., James, H., Rangan, A., & Gemming, L. (2020). Prevalence of product claims and marketing buzzwords found on health food snack products does not relate to nutrient profile. *Nutrients*, 12(5), 1513. <https://doi.org/10.3390/nu12051513>
- Brownbill, A. L., Braunack-Mayer, A. J., & Miller, C. L. (2020). What makes a beverage healthy? A qualitative study of young adults' conceptualisation of sugar-containing beverage healthfulness. *Appetite*, 150, 104675. <https://doi.org/10.1016/j.appet.2020.104675>
- Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2020). *The structure of economic news* (Working Paper No. 26648). National Bureau of Economic Research. <https://doi.org/10.3386/w26648>
- Cancer Research, U. K. (2020). *Food Myths*. <https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/diet-and-cancer/food-controversies/>
- Carels, R. A., Harper, J., & Konrad, K. (2006). Qualitative perceptions and caloric estimations of healthy and unhealthy foods by behavioral weight loss participants. *Appetite*, 46(2), 199–206. <https://doi.org/10.1016/j.appet.2005.12.002>
- Carels, R. A., Konrad, K., & Harper, J. (2007). Individual differences in food perceptions and calorie estimation: An examination of dieting status, weight, and gender. *Appetite*, 49(2), 450–458. <https://doi.org/10.1016/j.appet.2007.02.009>
- Chandon, P., & Wansink, B. (2012). Does food marketing need to make us fat? a review and solutions. *Nutrition Reviews*, 70(10), 571–593. <https://doi.org/10.1111/j.1753-4887.2012.00518.x>
- Cloutier, K., Mongeau, L., Pageau, M., & Provencher, V. (2013). Food perceptions among adults and registered dietitians: Are they similar? *Food and Nutrition*, 4(10A), 2–8. <https://doi.org/10.4236/fns.2013.410A002>
- Davies, M. (2017). The new 4.3 billion word now corpus, with 4–5 million words of data added every day. In *The 9th international corpus linguistics conference*, University of Birmingham, UK.
- Delicato, C., Salvatore, F. P., & Contò, F. (2019). Consumers' understanding of healthy foods: The evidence on superfoods in Belgium. In *2019 Eighth AIEAA Conference*, June 13–14, Pistoia, Italy No. 300907. Italian Association of Agricultural and Applied Economics (AIEAA). <https://doi.org/10.22004/ag.econ.300907>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., Jones Mitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 1–14. 2023. <https://doi.org/10.1038/s44159-023-00241-5>
- Douma, E. R., Valente, M., & Syurina, E. V. (2021). Developmental pathway of orthorexia nervosa: Factors contributing to progression from healthy eating to excessive preoccupation with healthy eating. experiences of Dutch health professionals. *Appetite*, 158, 105008. <https://doi.org/10.1016/j.appet.2020.105008>
- European Commission. (2006). Regulation (EC) no 1924/2006 of the European parliament and of the council of 20 December 2006 on nutrition and health claims made on foods. *Official Journal of the European Union*, L 404, 9–25. <http://data.europa.eu/eli/reg/2006/1924/oj>
- Food Standards Australia New Zealand. (2018). *Australia New Zealand Food Standards Code – Standard 1.2.7 – Nutrition, Health and Related Claims*. <https://www.legislation.gov.au/Details/F2018C00942>
- Franco Lucas, B., Costa, J. A. V., & Brunner, T. A. (2021). Superfoods: Drivers for consumption. *Journal of Food Products Marketing*, 27(1), 1–9. <https://doi.org/10.1080/10454446.2020.1869133>
- Frederick, D. A., Tomiyama, A. J., Bold, J. G., & Saguy, A. C. (2020). Can she be healthy at her weight? effects of news media frames on antifat attitudes, dieting intentions, and perceived health risks of obesity. *Stigma and Health*, 5(3), 247–257. <https://doi.org/10.1037/sah0000195>
- Freeland-Graves, J. H., & Nitzke, S. (2013). Position of the academy of nutrition and dietetics: Total diet approach to healthy eating. *Journal of the Academy of Nutrition and Dietetics*, 113(2), 307–317. <https://doi.org/10.1016/j.jand.2012.12.013>
- Galfano, V., Syurina, E. V., Valente, M., & Donini, L. M. (2022). When “healthy” is taken too far: Orthorexia nervosa—current state, controversies and future directions. In E. Manzato, M. Cuzzolaro, & L. M. Donini (Eds.), *Hidden and*

- lesser-known disordered eating behaviors in medical and psychiatric conditions (pp. 159–176). Springer International Publishing. https://doi.org/10.1007/978-3-030-81174-7_14
- Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. *Psychological Science*, 33(4), 579–594. <https://doi.org/10.1177/09567976211043426>
- Garcia, C. J., & Proffitt, J. M. (2021). “We are coca-cola and so much more”: Political economic analysis of non-carbonated SSB coke brands. *Food, Culture, and Society*, 25(5), 796–813. <https://doi.org/10.1080/15528014.2021.1922192>
- Goldberg, J. P., & Sliwa, S. A. (2011). Communicating actionable nutrition messages: Challenges and opportunities. *Proceedings of the Nutrition Society*, 70(1), 26–37. <https://doi.org/10.1017/s0029665110004714>
- Graeff-Hönninger, S., & Khajehei, F. (2019). The demand for superfoods: Consumers’ desire, production viability and bio-intelligent transition. In C. Piatti, S. Graeff-Hönninger, & F. Khajehei (Eds.), *Food tech transitions: Reconnecting agri-food, technology and society* (pp. 81–94). Springer International Publishing. https://doi.org/10.1007/978-3-030-21059-5_5
- Groeniger, J. O., van Lenthe, F. J., Beenackers, M. A., & Kamphuis, C. B. (2017). Does social distinction contribute to socioeconomic inequalities in diet: The case of ‘superfoods’ consumption. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1), 1–7. <https://doi.org/10.1186/s12966-017-0495-x>
- Hackman, E. M., & Moe, G. L. (1999). Evaluation of newspaper reports of nutrition-related research. *Journal of the American Dietetic Association*, 99(12), 1564–1566. [https://doi.org/10.1016/s0002-8223\(99\)00384-3](https://doi.org/10.1016/s0002-8223(99)00384-3)
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric pre-processing for parametric causal inference. *Journal of Statistical Software*, 42(8), 1–28. <https://doi.org/10.18637/jss.v042.i08>
- Inoue-Choi, M., Oppeneer, S. J., & Robien, K. (2013). Reality check: There is no such thing as a miracle food. *Nutrition and Cancer*, 65(2), 165–168. <https://doi.org/10.1080/01635581.2013.748921>
- Jagdale, Y. D., Mahale, S. V., Zohra, B., Nayik, G. A., Dar, A. H., Khan, K. A., Abdi, G., & Karabagias, I. K. (2021). Nutritional profile and potential health benefits of super foods: A review. *Sustainability*, 13(16), 9240. <https://doi.org/10.3390/su13169240>
- Julia, C., Fialon, M., Galan, P., Deschasaux-Tanguy, M., Andreeva, V. A., Kesse-Guyot, E., Touvier, M., & Hercberg, S. (2021). Are foods ‘healthy’ or ‘healthier’? front-of-pack labelling and the concept of healthiness applied to foods. *British Journal of Nutrition*, 127(6), 948–952. <https://doi.org/10.1017/s0007114521001458>
- Kåle, M., & Agbozo, E. (2020). Healthy food depiction on social media: The case of kale on twitter. *CEUR Workshop Proceedings*, 2865, 51–62.
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150. <https://doi.org/10.3390/info10040150>
- Ladher, N. (2016). Nutrition science in the media: You are what you read. *BMJ*, 353. <https://doi.org/10.1136/bmj.i1879>
- Larkin, D., & Martin, C. R. (2016). Caloric estimation of healthy and unhealthy foods in normal-weight, overweight and obese participants. *Eating Behaviors*, 23, 91–96. <https://doi.org/10.1016/j.eatbeh.2016.08.004>
- Liu, H., Meng-Lewis, Y., Ibrahim, F., & Zhu, X. (2021). Superfoods, super healthy: Myth or reality? examining consumers’ repurchase and wom intention regarding superfoods: A theory of consumption values perspective. *Journal of Business Research*, 137, 69–88. <https://doi.org/10.1016/j.jbusres.2021.08.018>
- Lobstein, T., & Davies, S. (2009). Defining and labelling ‘healthy’ and ‘unhealthy’ food. *Public Health Nutrition*, 12(3), 331–340. <https://doi.org/10.1017/s1368980008002541>
- Loyer, J. (2016). *The social lives of superfoods* Doctoral dissertation, University of Adelaide. <https://hdl.handle.net/2440/101777>
- Lusk, J. L., & Shankar, B. (2019). Consumer beliefs about healthy foods and diets. *PLoS One*, 14(10), e0223098. <https://doi.org/10.1371/journal.pone.0223098>
- Lynn, T., Rosati, P., Leoni Santos, G., & Endo, P. T. (2020). Sorting the healthy diet signal from the social media expert noise: Preliminary evidence from the healthy diet discourse on twitter. *International Journal of Environmental Research and Public Health*, 17(22), 8557. <https://doi.org/10.3390/ijerph17228557>
- MacGregor, C., Petersen, A., & Parker, C. (2021). Promoting a healthier, younger you: The media marketing of anti-ageing superfoods. *Journal of Consumer Culture*, 21(2), 164–179. <https://doi.org/10.1177/1469540518773825>
- Magrath, A., Sanz, M. J., & Harris, J. (2020). Environmental and social consequences of the increase in the demand for ‘superfoods’ world-wide. *People & Nature*, 2(2), 267–278. <https://doi.org/10.1002/pan3.10085>
- Meyerding, S. G., Kürzdörfer, A., & Gassler, B. (2018). Consumer preferences for superfood ingredients—the case of bread in Germany. *Sustainability*, 10(12), 4667. <https://doi.org/10.3390/su10124667>
- Michel, F., Sanchez-Siles, L. M., & Siegrist, M. (2021). Predicting how consumers perceive the naturalness of snacks: The usefulness of a simple index. *Food Quality and Preference*, 94, 104295. <https://doi.org/10.1016/j.foodqual.2021.104295>
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. “Optimizing semantic coherence in topic models.” In Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 262–272. 2011.
- Mintel. (2016). *Super Growth for “Super” Foods: New Product Development Shoots Up 202% Globally Over the Past Five Years*. <https://www.mintel.com/press-centre/food-and-drink/super-growth-for-super-foods-new-product-development-shoots-up-202-globally-over-the-past-five-years>

- Nagler, R. H. (2014). Adverse outcomes associated with media exposure to contradictory nutrition messages. *Journal of Health Communication, 19*(1), 24–40. <https://doi.org/10.1080/10810730.2013.798384>
- Oakes, M. E., & Slotterback, C. S. (2001). Gender differences in perceptions of the healthiness of foods. *Psychology and Health, 16*(1), 57–65. <https://doi.org/10.1080/08870440108405489>
- Ooms, J. (2020). HunsPELL: High-performance stemmer, tokenizer, and spell checker [computer software manual]. <https://CRAN.R-project.org/package=hunsPELL> (R package version 3.0.1)
- Pandur, M. B., Dobaša, J., & Kronegger, L. (2020). Topic modelling in social sciences—case study of web of science. In *Central european conference on information and intelligent systems*, Varazdin, Croatia (pp. 211–218).
- Perkovic, S., Otterbring, T., Schärli, C., & Pachur, T. (2021). The perception of food products in adolescents, lay adults, and experts: A psychometric approach. *Journal of Experimental Psychology: Applied, 27*(1), 1–12. <https://doi.org/10.31234/osf.io/uy7m>
- Pilař, L., Kvasničková Stanislavská, L., & Kvasnička, R. (2021). Healthy food on the twitter social network: Vegan, homemade, and organic food. *International Journal of Environmental Research and Public Health, 18*(7), 3815. <https://doi.org/10.3390/ijerph18073815>
- Plasek, B., Lakner, Z., & Temesi, Á. (2020). Factors that influence the perceived healthiness of food—review. *Nutrients, 12*(6), 1881. <https://doi.org/10.3390/nu12061881>
- Plasek, B., Lakner, Z., & Temesi, Á. (2021). I believe it is healthy—impact of extrinsic product attributes in demonstrating healthiness of functional food products. *Nutrients, 13*(10), 3518. <https://doi.org/10.3390/nu13103518>
- Proestos, C. (2018). Superfoods: Recent data on their role in the prevention of diseases. *Current Research in Nutrition and Food Science Journal, 6*(3), 576–593. <https://doi.org/10.12944/crnfsj.6.3.02>
- Provencher, V., Polivy, J., & Herman, C. P. (2009). Perceived healthiness of food. If it's healthy, you can eat more! *Appetite, 52*(2), 340–344. <https://doi.org/10.1016/j.appet.2008.11.005>
- Public Health England. (2021). *Mccance and Widdowson's Composition of Foods Integrated Dataset*. <https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid>
- R Core Team. (2021). *R: A language and environment for statistical computing [computer software manual]*. <https://www.R-project.org/>
- Reisman, E. (2020). Superfood as spatial fix: The ascent of the almond. *Agriculture and Human Values, 37*(2), 337–351. <https://doi.org/10.1007/s10460-019-09993-4>
- Roberts, M., Stewart, B., & Airoidi, E. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111*(515), 988–1003. <https://doi.org/10.1080/01621459.2016.1141684>
- Roberts, M., Stewart, B., & Tingley, D. (2019). Stm: An r package for structural topic models. *Journal of Statistical Software, 91*(1), 1–40. <https://doi.org/10.18637/jss.v091.i02>
- Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. <https://doi.org/10.1111/ajps.12103>
- Rodney, A. (2018). Pathogenic or health-promoting? how food is framed in healthy living media for women. *Social Science & Medicine, 213*, 37–44. <https://doi.org/10.1016/j.socscimed.2018.07.034>
- Rojas-Rivas, E., Espinoza-Ortega, A., Thomé-Ortiz, H., & Moctezuma-Pérez, S. (2019). Consumers' perception of amaranth in Mexico: A traditional food with characteristics of functional foods. *British Food Journal, 121*(6), 1190–1202. <https://doi.org/10.1108/BFJ-05-2018-0334>
- Roman, S., Sánchez-Siles, L. M., & Siegrist, M. (2017). The importance of food naturalness for consumers: Results of a systematic review. *Trends in Food Science & Technology, 67*, 44–57. <https://doi.org/10.1016/j.tifs.2017.06.010>
- Roth, A., & Zawadzki, T. (2018). Instagram as a tool for promoting superfood products. *Annals of Marketing Management & Economics, 4*(1), 101–113. <https://doi.org/10.22630/AMME.2018.4.1.8>
- Rubin, D., & Rosenbaum, P. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38. <https://doi.org/10.1080/00031305.1985.10479383>
- Saguy, A. C., & Almeling, R. (2008). Fat in the fire? science, the news media, and the “obesity epidemic”2. *Sociological Forum, 23*(1), 53–83. <https://doi.org/10.1111/j.1600-0838.2004.00399.x-i1>
- Šamec, D., Urlič, B., & Salopek-Sondi, B. (2019). Kale (brassica oleracea var. acephala) as a superfood: Review of the scientific evidence behind the statement. *Critical Reviews in Food Science and Nutrition, 59*(15), 2411–2422. <https://doi.org/10.1080/10408398.2018.1454400>
- Samoggia, A., Riedel, B., & Ruggeri, A. (2020). Social media exploration for understanding food product attributes perception: The case of coffee and health with twitter data. *British Food Journal, 122*(12), 3815–3835. <https://doi.org/10.1108/BFJ-03-2019-0172>
- Scrinis, G. (2013). *Nutritionism*. Columbia University Press.
- Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the effects of news sentiments on the stock market. In *2018 IEEE international conference on big data (big data)* (pp. 4705–4708). <https://doi.org/10.1109/BigData.2018.8621884>
- Siipi, H. (2013). Is natural food healthy? *Journal of Agricultural and Environmental Ethics, 26*(4), 797–812. <https://doi.org/10.1007/s10806-012-9406-y>

- Sikka, T. (2019). The contradictions of a superfood consumerism in a postfeminist, neoliberal world. *Food, Culture, and Society*, 22(3), 354–375. <https://doi.org/10.1080/15528014.2019.1580534>
- Silchenko, K., Askegaard, S., & Cedrola, E. (2020). Three decades of research in health and food marketing: A systematic review. *Journal of Consumer Affairs*, 54(2), 541–580. <https://doi.org/10.1111/joca.12289>
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1–13. <https://doi.org/10.18637/jss.v039.i05>
- Štepec, D., Tavčar, G., & Ponikvar-Svet, M. (2020). Surprisingly high fluorine content in some exotic superfoods. *Journal of Fluorine Chemistry*, 234, 109521. <https://doi.org/10.1016/j.jfluchem.2020.109521>
- Thurecht, R. L., Pelly, F. E., & Cooper, S. L. (2018). Dietitians' perceptions of the healthiness of packaged food. *Appetite*, 120, 302–309. <https://doi.org/10.1016/j.appet.2017.08.036>
- U.S. Department of Agriculture. (2019). *Fooddata Central*. <https://fdc.nal.usda.gov/>
- US Food And Drug Administration. (2016). *Use of the Term 'Healthy' in the Labeling of Human Food Products: Guidance for Industry*. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-industry-use-term-healthy-labeling-human-food-products>
- Vidal, L., Ares, G., Machin, L., & Jaeger, S. R. (2015). Using twitter data for food-related consumer research: A case study on “what people say when tweeting about different eating situations”. *Food Quality and Preference*, 45, 58–69. <https://doi.org/10.1016/j.foodqual.2015.05.006>
- Weitkamp, E., & Eidsvaag, T. (2014). Agenda building in media coverage of food research: Superfoods coverage in UK national newspapers. *Journalism Practice*, 8(6), 871–886. <https://doi.org/10.1080/17512786.2013.865966>
- World Health Organization. (2021). Implementing policies to restrict food marketing: A review of contextual factors. *World Health Organization*. <https://apps.who.int/iris/handle/10665/345128>
- Wulff, D. U., & Mata, R. (2023, October 12). Automated jingle-jangle detection: Using embeddings to tackle taxonomic incommensurability. <https://doi.org/10.31234/osf.io/9h7aw>
- Zamani, M., Schwartz, H. A., Eichstaedt, J., Guntuku, S. C., Virinchipuram Ganesan, A., Clouston, S., & Giorgi, S. (2020). Understanding weekly COVID-19 concerns through dynamic content-specific lda topic modeling. In *Proceedings of the fourth workshop on natural language processing and computational social science* (pp. 193–198). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.nlpccs-1.21>

Appendix

Participant survey

In this Appendix, we provide additional details about the participant online survey used to ascertain a list of known superfoods as well as participant definitions and descriptions for the term itself.

Participants

One hundred participants were recruited via Prolific Academic in return for a fixed payment of £0.63 for 5 minutes of their time. Only English-speaking adults from the United Kingdom or the United States of America were eligible to take part in our survey. Four participants who answered “no” when asked if they were familiar with the term superfood were removed from subsequent analysis, leaving 96 participants (69% female) in our sample. On average, participants were 35.40 years old (SD = 13.48; range = 18–77). Of these participants, 75% had no dietary restrictions.

Design and procedure

Participants completed a short online survey, consisting of an eligibility question followed by three main questions and three demographic questions. All questions were presented in the same order and read: 1) “Are you familiar with the term “superfood (yes or no)?,” 2) “Name at least 5 superfoods,” 3) “List at least 5 adjectives that you associate with superfoods” and 4) “In your own words, how would you define a superfood.” Participants were then asked about their age, gender, and dietary restrictions (options provided were “Vegan,” “Vegetarian,” “Other (please specify if you wish)” and “None of the Above”).

Data analysis

The list of 25 superfoods (as well as non-food words) was determined using word frequency, identical to the computational method outlined in the section “Word Frequency: A Bag-of-Words Model.”

Data pre-processing

The same data pre-processing steps as for the superfood articles (tokenization, stop word removal, part of speech tagging, lemmatization, and conversion to lowercase) were applied to the participant data, for analysis at the aggregate level. In addition, as part of the standardization for food names, all spaces between open compound words (e.g., peanut butter and sweet potato) were replaced with an underscore. Moreover, bigrams from participants’ open text responses (to question 4) were identified as words that appeared next to each other in a sentence more than once. Furthermore, we chose to concatenate all text given by the same participant to account for, and remove, repeated words (e.g., if the same participant gave the word “healthy” as a response to both questions 4 and 5 it was only counted once). Spelling mistakes and inconsistencies, either introduced by participants or through the lemmatization process (e.g., “broccoli” to “broccoli” and “slimme” to “slimming”), were identified using the “hunspell” package in R and corrected manually (Ooms, 2020). Spellings were manually standardized to US spelling to avoid duplication. Only words tagged as adjectives, nouns, verbs (and manually as bigrams) were kept for the analysis of non-food words.

Additional data pre-processing steps of superfood articles

As mentioned in the main text, we included additional data pre-processing steps for the Word Frequency (bag-of-words model) analysis only. First, we created a manual list of stop words, allowing us to remove all food names from the 872 superfood articles. Second, we identified bigrams as words that appeared next to each other in a sentence within the corpora 50 times or more and replaced spaces between these words with underscores. Next, we tokenized and lemmatized the article text, also removing stop words and tagging parts of speech. Similar to the participant data, all spelling mistakes were identified using the “hunspell” package in R and corrected manually (Ooms, 2020). We then removed sparse terms, which are the words that occur in less than 3% of the 872 superfood articles. This allowed us to plot the most common terms of the whole superfood corpora, rather than those that rarely occur. Subsequently, we filtered the most meaningful parts of speech (adjectives, nouns, verbs, and bigrams). Finally, we subset the top 25 words for each of these part of speech tags, which can be seen, alongside the comparative participant data, in [Figure 2](#) of the main text.

Topic Modelling Grid Search

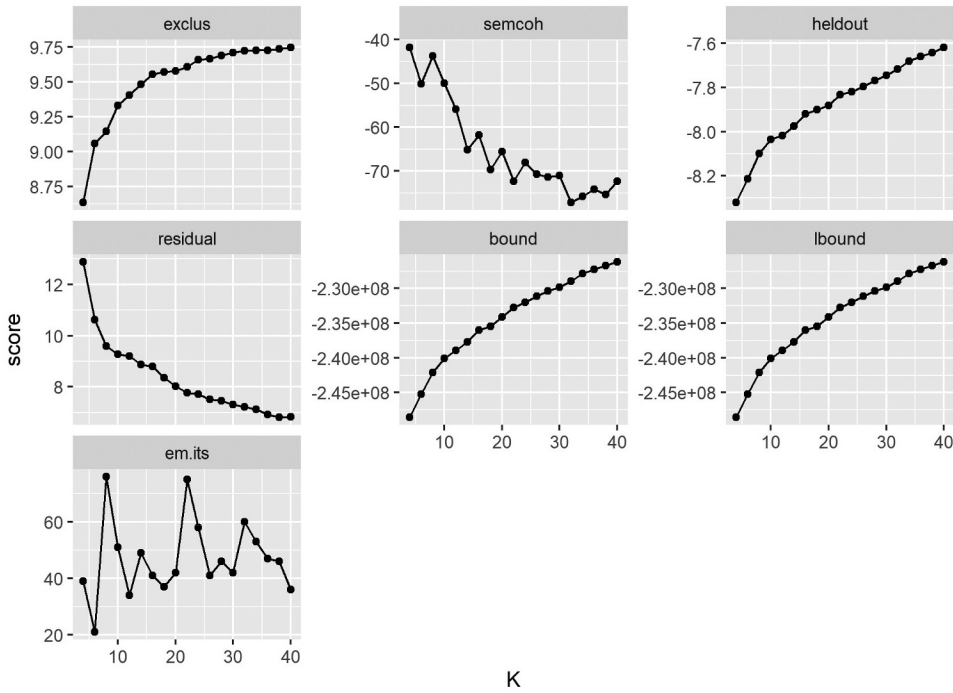


Figure A1. Grid search evaluation results. The model was optimized based on the exclusivity score.

As referenced in the Topic Modelling section of the Computational Methods Section, [Figure A1](#) shows the measures we used to find the optimal number of latent topics. We searched K-values between 4 and 40, shown on the X-axis of [Figure A1](#). From left to right, the plots refer to exclusivity, semantic coherence, held out likelihood, residual, bound, lower bound, and finally “em.its” refers to the total number of EM iterations used in fitting the model (Roberts et al., 2019). Here, our model was set to run at a maximum of 100 iterations. We optimized the model on exclusivity, which compares word distributions between topics to determine the likelihood of the top words of one topic being top words in the other topics (Roberts et al., 2019). Another important measure is semantic coherence, introduced by Mimno et al. (2011), which refers to the probability that the top words of one topic co-occur within our corpora of superfood articles (Pandur et al., 2020). We also consider held-out likelihood estimation, similar to cross-validation, which checks the model’s predictive performance by estimating the probability of words occurring within a document after they have been removed (Pandur et al., 2020). Measuring residuals is useful for determining how much variance remains at a given topic number, and whether more topics would be needed to account for any overdispersion. The bound is a measure of convergence, with the model considered converged when there is a small enough change between iterations (Pandur et al., 2020). The lower bound simply applies a correction to the bound so that the bounds are directly comparable (Roberts et al., 2019).