frontiers in Nutrition

ORIGINAL RESEARCH
published: 09 March 2022
doi: 10.3389/fnut.2022.765794

Dietary Pattern Extraction Using Natural Language Processing Techniques

Insu Choi[1], *Jihye Kim*[2]* *and Woo Chang Kim*[1]*

[1] Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [2] Department of Genetics and Biotechnology, College of Life Sciences, Kyung Hee University, Yongin, South Korea

In this study, we observed the changes in dietary patterns among Korean adults in the previous decade. We evaluated dietary intake using 24-h recall data from the fourth (2007–2009) and seventh (2016–2018) Korea National Health and Nutrition Examination Survey. Machine learning-based methodologies were used to extract these dietary patterns. Particularly, we observed three dietary patterns from each survey similar to the traditional and Western dietary patterns in 2007–2009 and 2016–2018, respectively. Our results reveal a considerable increase in the number of Western dietary patterns compared with the previous decade. Thus, our study contributes to the use of novel methods using natural language processing (NLP) techniques for dietary pattern extraction to obtain more useful dietary information, unlike the traditional methodology.

Keywords: dietary pattern, machine learning, natural language processing (NLP), word embedding, topic modeling

OPEN ACCESS

Edited by:
Kurt A. Rosentrater,
Iowa State University, United States

Reviewed by:
Wisdom Dogbe,
University of Aberdeen,
United Kingdom
Sapna Langyan,
National Bureau of Plant Genetic
Resources (ICAR), India

Correspondence:
Jihye Kim
kjhye@khu.ac.kr
Woo Chang Kim
wkim@kaist.ac.kr

Specialty section:
This article was submitted to
Nutrition and Sustainable Diets,
a section of the journal
Frontiers in Nutrition

Received: *27 August 2021*
Accepted: *04 February 2022*
Published: *09 March 2022*

Citation:
Choi I, Kim J and Kim WC (2022)
Dietary Pattern Extraction Using
Natural Language Processing
Techniques. Front. Nutr. 9:765794.
doi: 10.3389/fnut.2022.765794

INTRODUCTION

Diets are affected by interactions between biological, social, economic, and cultural factors (1). Therefore, people's diet can eventually change. In particular, dietary patterns have rapidly evolved in South Korea due to this country's early westernization compared with most Asian countries (2). The transition of dietary patterns should be monitored because it holds an important risk factor for developing chronic diseases. Traditional methods have been employed to show unique dietary patterns or observe changes in dietary patterns over time in Korean populations (3, 4).

However, epidemiological studies on dietary pattern extraction using traditional methods have many limitations. For example, many subjective decisions are included in food grouping for extracting dietary patterns. Moreover, obtaining information on particular food items has become difficult due to their broad classification.

Therefore, this study applies natural language processing (NLP) techniques to overcome these constraints. NLP techniques have been widely used in many areas. Recently, they were applied in the nutritional domain. Several papers use NLP techniques to analyze their nutrition goals. Tamaddoni-Nezhad et al. (5) demonstrated that logic-based machine learning methods and NLP techniques can be employed to generate food webs from ecological census data. Zhu et al. (6) attempted to obtain a food collocation and investigate its effect using NLP preprocessing methodologies. Korpusik and Glass (7) proposed a novel approach to food journaling that uses speech and language understanding technology to achieve an efficient self-assessment of energy and nutrient consumption. Tao et al. (8) introduced an overview of data sources, computational methods, and applications of text data in the food industry. In addition, Van Erp et al. (9) discussed a food concept and researched on recipes to solve health and sustainability issues addressed in an interdisciplinary manner. They integrated NLP and other AI techniques with historical food research, such as food science and nutrition. Bakhtin et al. (10) presented a text mining study on science and technology in food production from more than 30 million documents. They proposed a methodology that demonstrates the future of food production with each new data that become available and served as an early warning system for a changing technology landscape.

However, studies on modeling-based analysis for food intake data rarely exist. Therefore, we aimed to extract the dietary patterns of Korean adults using NLP techniques and observe the changes in their dietary patterns over the last decade from large-scale national representative data. The remainder of this paper is organized as follows. In Section "Data and Methodology," the methods for data collection and NLP techniques used to analyze them are presented. In Section "Results," the main results of our study are described. The main results of this study are discussed in Section "Discussion." Finally, the concluding remarks are presented in Section "Conclusion."

DATA AND METHODOLOGY

Data

This study used data collected from the Korea National Health and Nutrition Examination Survey (KNHANES) conducted by the Korea Disease Control and Prevention Agency. KNHANES is a clustered, multistage, and stratified sampling design that annually assesses the diet and health of the Korean population. Many studies on the diet and nutrition of the Korean population use this dataset owing to its representation and credibility (4, 11, 12). We used the 2007–2009 and 2016–2018 data from the fourth and seventh KNHANES, respectively. A total of 16,187 and 16,809 study participants were from the fourth and seventh KNHANES datasets, respectively. The demographic characteristics of the study participants are summarized in **Table 1**.

TABLE 1 | Characteristics of survey participants.

Category	2007–2009	2016–2018
All	16,187	16,809
Sex		
Male	6,592 (40.7)	7,144 (42.5)
Female	9,595 (59.3)	9,665 (57.5)
Age		
19–39	4,643 (27.5)	5,296 (32.7)
40–59	6,189 (36.8)	5,814 (35.9)
60+	5,997 (35.7)	5,077 (31.4)
House type		
General	9,750 (60.2)	7,831 (46.6)
Apartment	6,437 (39.8)	8,978 (53.4)
Highest level of education		
Elementary School	4,589 (28.3)	3,288 (19.6)
High School	6,830 (42.2)	6,342 (37.8)
Over Associate Degree/Bachelor Degree	3,666 (20.7)	5,639 (33.5)

The difference between the total number of respondents and the number of respondents per category means no response.

Methodology

Data Preprocessing

We preprocessed the two combined datasets: the 2007–2009 and 2016–2018 datasets. The preprocessing of these datasets is presented in **Figure 1**.

First, data with incorrect column values (e.g., food name in the weight column), which were distorted since they invaded columns on both sides, were deleted. Thus, 6,954 and 17,774 food names were extracted from each dataset. The vocabulary that emerged only once in the subsequent cleaning and normalization process through frequency analysis, was removed, not only because it does not affect the conduct of topic modeling but also because it is seen as an outlier. In addition, the differences in the surveyor's marking were unified and eliminated as much as possible by excluding the English characters, numbers, and special symbols contained in the extracted food names. For example, "Chocolett (marked in Korean pronunciation)" and "Chocolate" were unified as "Chocolate." Thus, all food name data were tested for "Korean spelling" because of differences in spelling (dialects, etc.) or because the typos of each surveyor were not collected systematically. The Python package "Hanspell" based on the NAVER Korean Spellchecker was used for the Korean spelling test. Subsequently, 1,59,552 rows (87.86% of the raw 2007–2009 dataset) and 1,79,846 rows (78.89% of the raw 2016–2018 dataset) of data remained, and 1,178 and 1,653 food name data were used as experimental data. When the two food name data were merged, a total of 1,877 food name data were generated.

For the final food data, fastText, which is a methodology for embedding words developed on Facebook, was used to process food names for analysis. Word embedding involves replacing a word with a word vector containing meanings based on distributed simplicity-based presentation. fastText assumes that the peripheral words of a word with a similar distribution also have similar meanings. A word is represented as a vector with a continuous value in a predetermined dimension on the variance table. Word vectors created in this way contain their meaning and can also be used to analyze the operations and similarities between word vectors. fastText is an advantageous way to cluster similar words. It is another method that complements the disadvantage of hard-to-solve out-of-vocabulary (OOV) problems or unknown word problems. fastText is word-embedded for each *n*-gram of every word in the dataset and has the advantage of being able to calculate similarities to other words for OOV words that are unknown *via* known subwords.

In this experiment, upon grouping the food names using fastText, the grouped foods' formations are similar. Thus, manual processing becomes cheap after our preprocessing procedures. This process is time-efficient and produces more accurate food

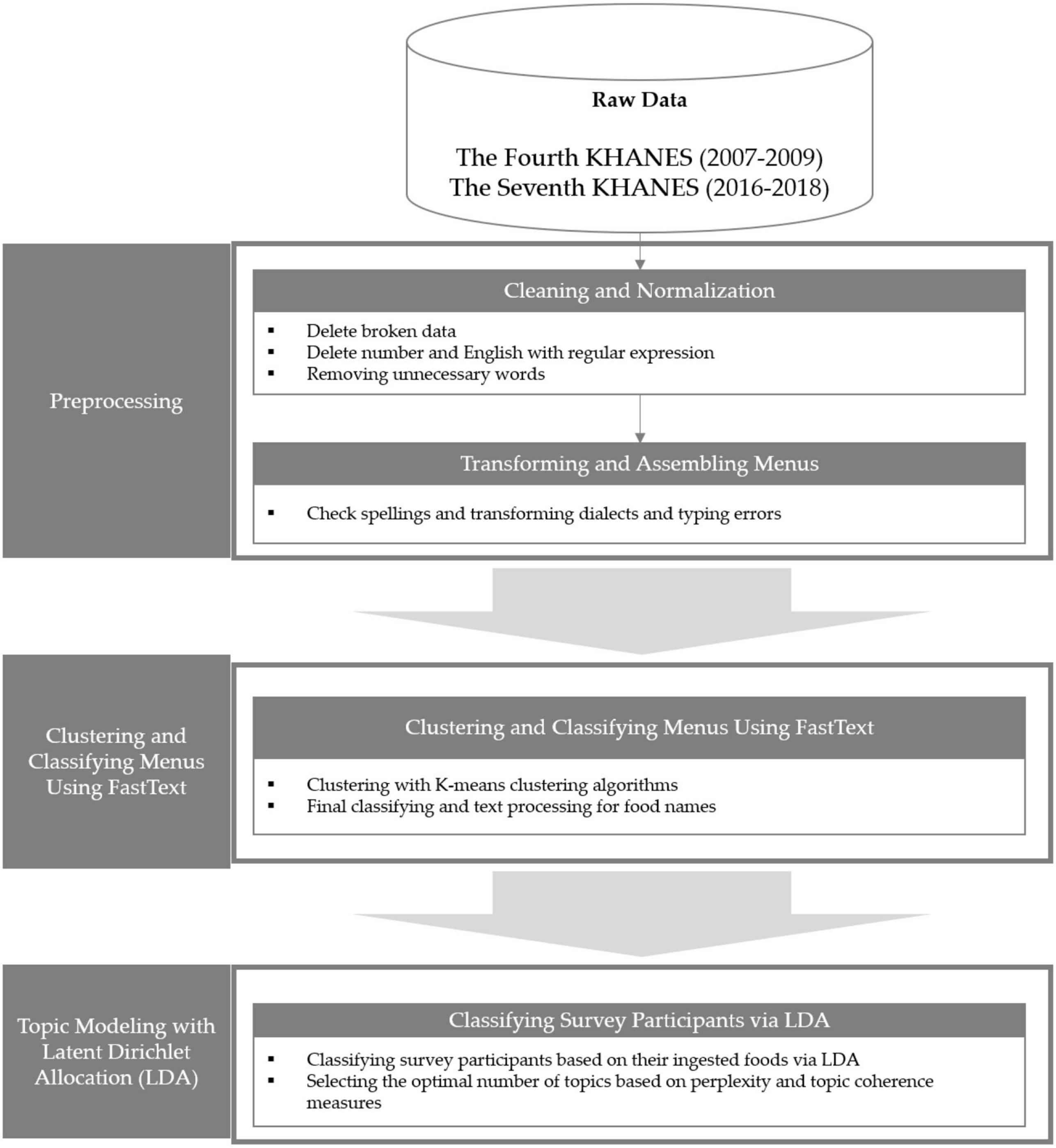
Raw Data
The Fourth KHANES (2007-2009)
The Seventh KHANES (2016-2018)
Preprocessing
Cleaning and Normalization
▪ Delete broken data
▪ Delete number and English with regular expression
▪ Removing unnecessary words
Transforming and Assembling Menus
▪ Check spellings and transforming dialects and typing errors
Clustering and Classifying Menus Using FastText
Clustering and Classifying Menus Using FastText
▪ Clustering with K-means clustering algorithms
▪ Final classifying and text processing for food names
Topic Modeling with Latent Dirichlet Allocation (LDA)
Classifying Survey Participants via LDA
▪ Classifying survey participants based on their ingested foods via LDA
▪ Selecting the optimal number of topics based on perplexity and topic coherence measures

FIGURE 1 | Data preprocessing process.

processing data than by hand. To apply fastText, the Korean Wikipedia, KorQuad, and NAVER movie corpus were used to embed words using fastText as a morpheme-treated corpus. Since the Korean language can be broken down into alphabetical units, consonant-level learning using *MeCab_jamo* was conducted. The proportions of each word-embedding methodology were considered for both syllable and consonant units by treating them at a ratio of 50–50%, respectively. **Figure 2** presents the two-dimensional scatter plot of food name vectors in this study using principal component analysis.

Final modifications were made to some heterogeneous food names in clusters using the designated cluster number of food

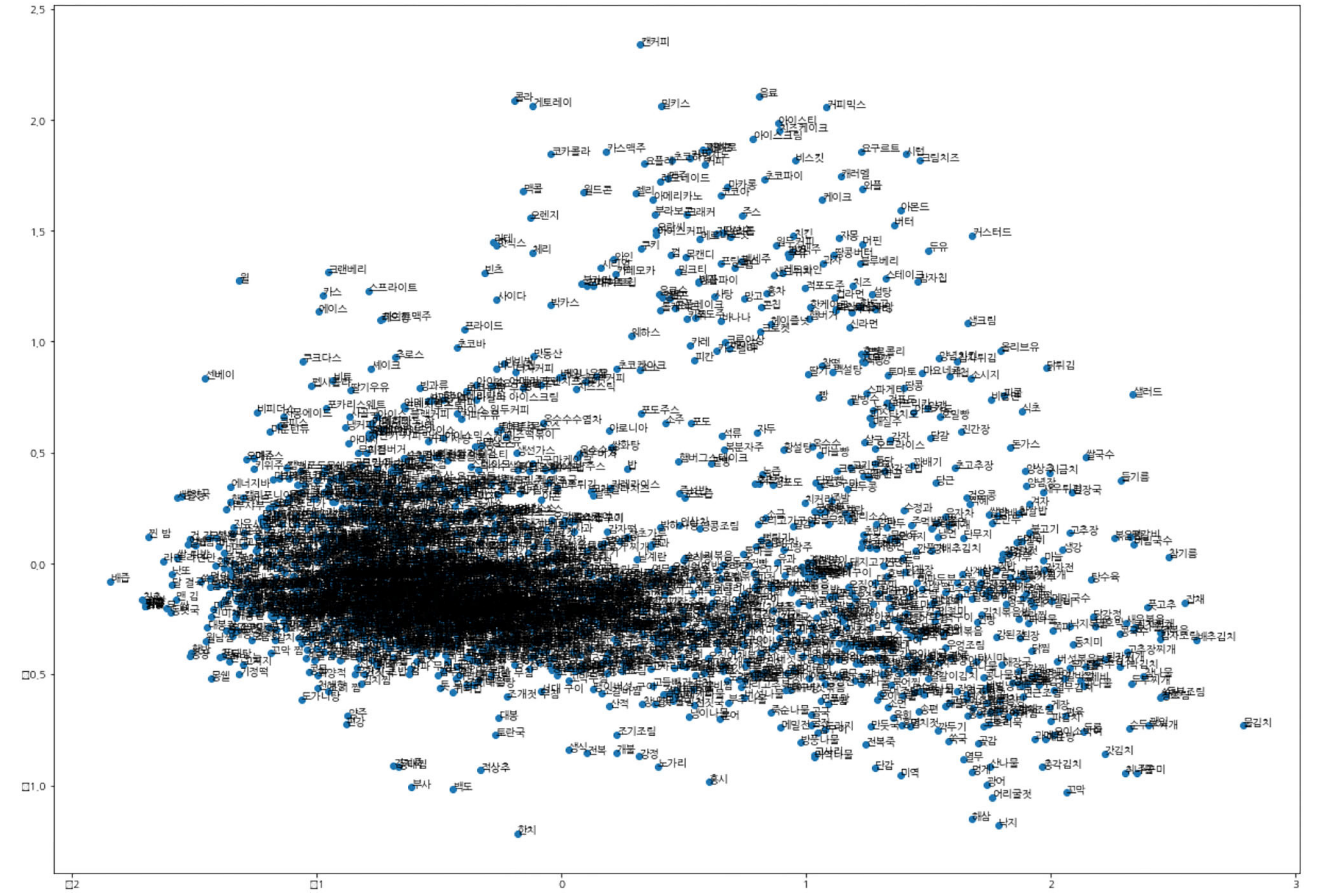

FIGURE 2 | Two-dimensional scatter plot of food name vectors.

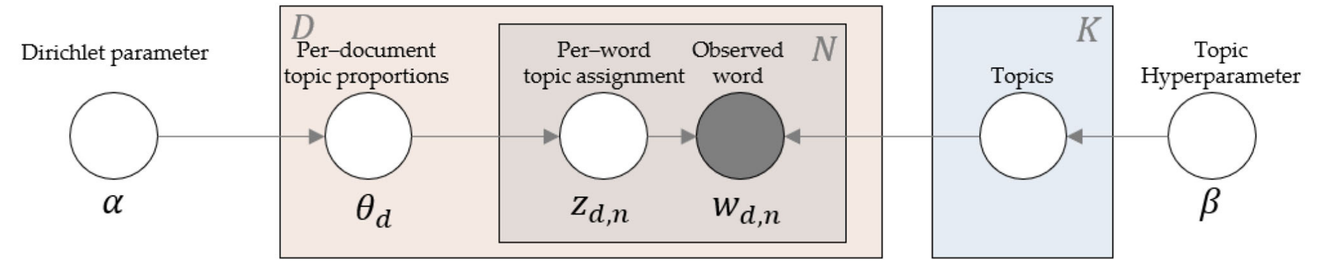

FIGURE 3 | Intuition behind LDA.

names. As a result, 887 food names were derived from the 2007–2009 and 2016–2018 datasets. The final processed food names were treated with a weighted average of food divided by the food intake weight by certain survey participants, and the rounded values were treated with keyword-inclusion frequency.

Latent Dirichlet Allocation (LDA)

Topic modeling methods are powerful, intelligent techniques widely applied in NLP to discover topics and semantic mining from unordered documents. LDA, one of the most popular topic modeling methods, is a generative probabilistic model for collecting discrete data, such as text corpora. It generates a topic per-document model and words per topic model using the Dirichlet distribution. **Figure 3** presents the concept of LDA.

Many studies have applied LDA to topic modeling methods in various fields, such as keyword selection, source code analysis, opinion mining, event detection, music key profiling, image classification, a recommender system, sentiment classification, and even political issues (13).

In this study, we applied LDA based on the Gibbs sampling method owing to its rapid speed compared with the original model. Gibbs sampling is a Markov chain Monte Carlo algorithm for sampling conditional distributions of variables, approximated from an actual distribution when direct sampling is inefficient or difficult. Equation (1) expresses the updated LDA equation using Gibbs sampling for the probability that the k-th topic is assigned to $z_{d,\,i}$, the i-th word of the d-th document.

$$p\left(z_{d,\,i}=k\middle|z_{-i},\,\mathbf{w}\right)=\frac{n_{d,\,k}+\alpha_k}{\sum_{i=1}^{K}n_{d,\,i}+\alpha_i}\,\frac{v_{k,\,w_{d,\,n}}+\beta_{w_{d,\,n}}}{\sum_{j=1}^{V}v_{k,\,j}+b_j}=\mathbf{AB} \tag{1}$$

In Equation (1), z_{-i} denotes leaving the i-th term from the calculation; $\mathbf{w}$, the word vector of documents; $n_{d,\,k}$, the number of times words in the d-th document were assigned to the k-th topic; $w_{d,\,n}$, the n-th word in the d-th document; and $v_{k,\,w_{d,\,n}}$, frequency of the word $w_{d,\,n}$ from the whole corpus in the k-th topic. α_k and β_k are the hyperparameters of per-corpus topic distributions and per-document topic proportions, following symmetric Dirichlet distributions. Equation (1) can be summarized into two parts: **A** and **B**. **A** denotes the relationship between the d-th document and the k-th topic, and **B** indicates the relationship between the n-th words of the d-th document and the k-th topic.

To select the optimal number of topics for the LDA model, we considered perplexity and topic coherence measures: C_V. Perplexity, generally used in language modeling, is the entropy-based measurement of the accuracy of sample prediction by a probability distribution or model. It is algebraically equal to the inverse of the geometric mean per word likelihood. The C_V measure is based on a sliding window, one-set segmentation of the top words, and an indirect confirmation measure that uses normalized pointwise mutual information and cosine similarity.

Moreover, before conducting LDA, we preprocessed the survey data, which were unstructured and unrefined text data, to develop a more elegant and time-saving method than modification by hand. We cleaned and corrected over 4,00,000 food name data in 5 min. Also, we vectorized the food name and clustered them using the K-means clustering algorithm. Finally, we refined the coherent food name data rapidly. These novel methods for preprocessing raw data used in this study are significant to future researchers who will be using food consumption data, such as KNHANES data. Since they should preprocess before analyzing them, our preprocessing method can be employed as an essential process for systematically applying the researchers' methodologies ranging from traditional methods to machine learning studies with high accuracy.

RESULTS

We considered the five indicators of perplexity and coherence score C_V. If two measures have higher values, the data were classified better. As a result, we confirmed three topics as the optimal topic number from **Figures 4**, **5**.

Furthermore, we extracted three dietary patterns from 2007 to 2009 and 2016 to 2018, respectively. We illustrated the

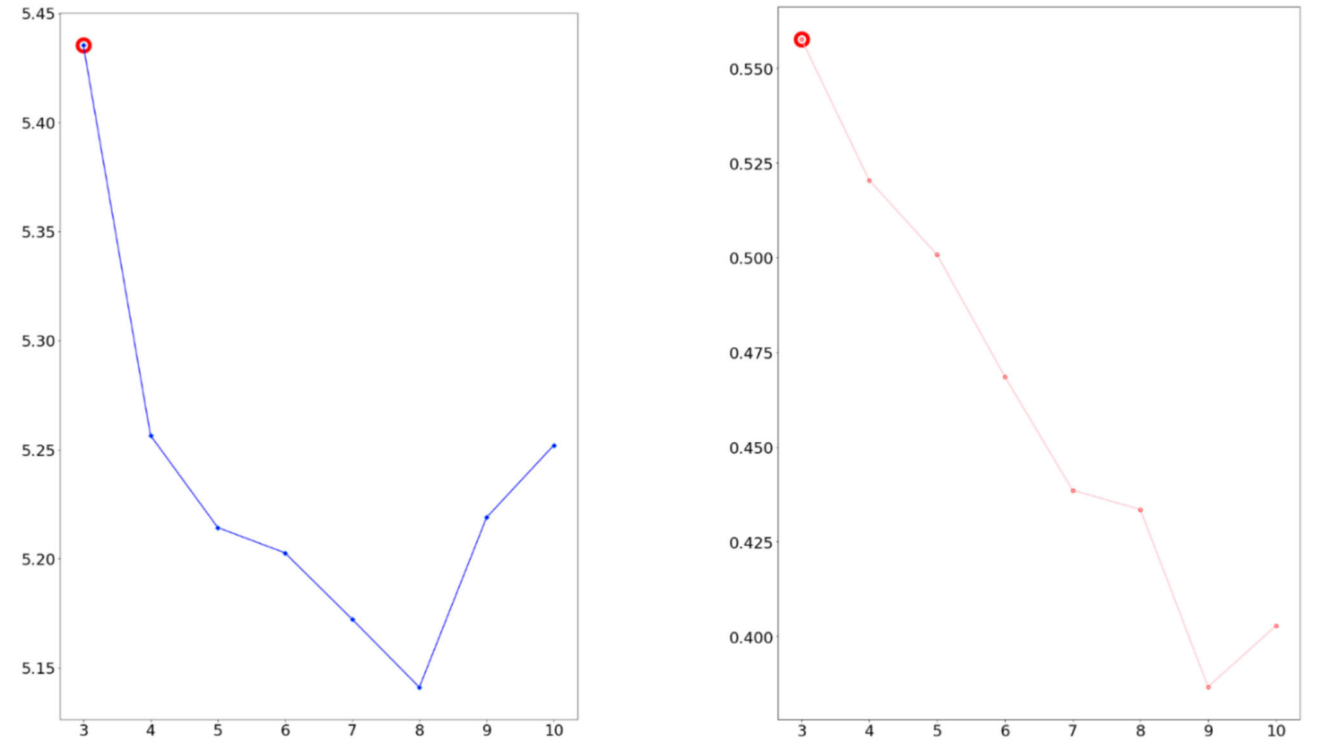

FIGURE 4 | Calculated perplexity (left) and $\mathbf{C_V}$ (right) for selecting optimal number (2007–2009 dataset).

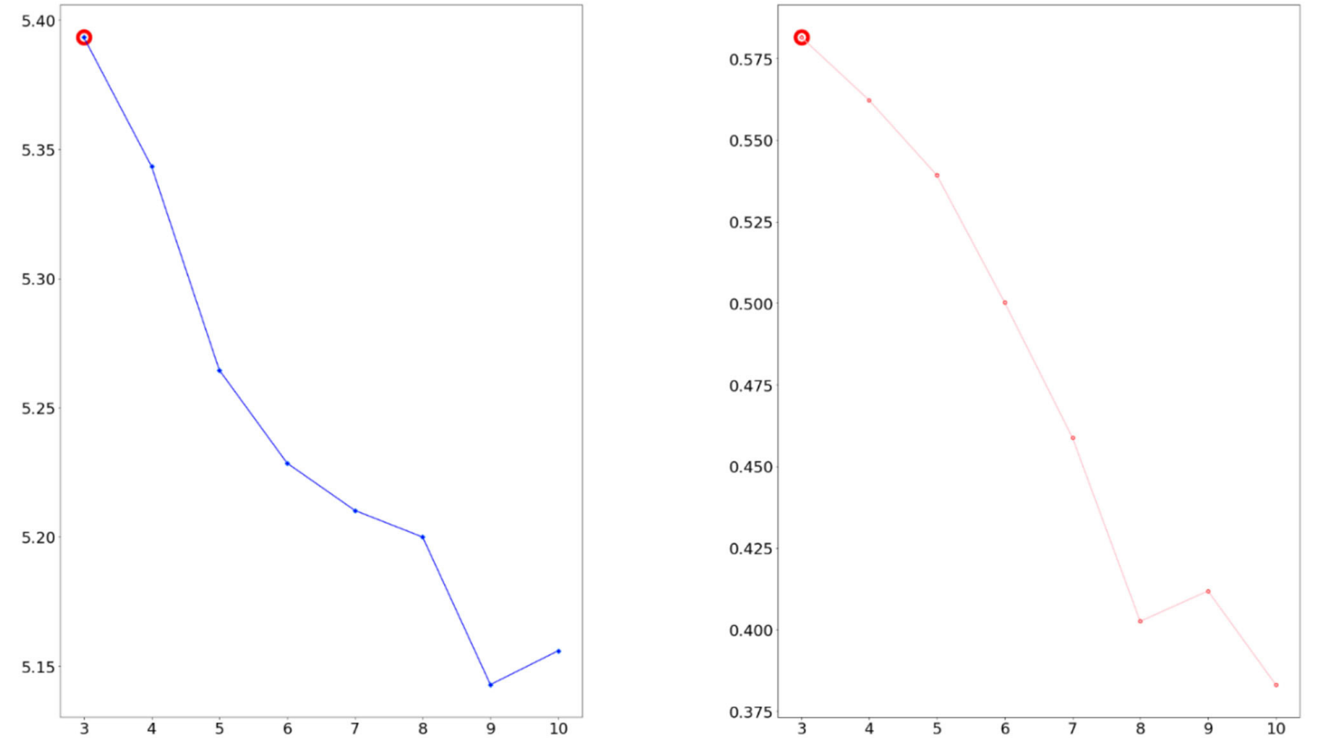

FIGURE 5 | Calculated perplexity (left) and $\mathbf{C_V}$ (right) for selecting optimal number (2016–2018 dataset).

three dietary patterns extracted from the 2007–2009 dataset and showed that the top 10 frequently appearing food names belong to those patterns, as shown in **Table 2**. Topic 1 consists of "Americano," "kimchi," "white rice," "egg," "multigrain rice," and so on. Topic 2 contains "instant coffee," "white rice", "ramen," "kimchi," "pumpkin," and so on. Lastly, Topic 3 comprises "soybean paste soup (doenjang soup)," "mix of red pepper paste and soybean paste (ssamjang)," "cabbage," "kimchi," "white rice," and so on.

These topics are similar to traditional Korean dietary patterns, consisting mainly of rice, soup, kimchi, vegetables, and fish. Each topic accounts for 68.4, 16.2, and 15.3% of the survey participants, respectively. Also, the token proportions of dietary patterns (the number of food names that appeared in each topic) were 51.3, 25.0, and 23.7%, respectively. Similarly, we extracted t three dietary patterns from the 2016 to 2018 dataset and showed that the top 10 frequently appearing food names belong to those shown in **Table 3**. Topic 1 consists of "red chili and soybean paste (ssamjang)," "pork belly," "lettuce," "red pepper," "cold noodle (naengmyeon)," and so on. Topic 2 contains "kimchi," "instant coffee," "milk," "white rice," "multigrain rice," and so on. Lastly, Topic 3 comprises "americano," "fried chicken," "mayonnaise," "fish cake soup," and so on.

Each topic accounts for 17.4, 67.1, and 15.5% of the survey participants, respectively. Moreover, the token proportions of dietary patterns (the number of food names that appeared in each topic) were 15.0, 62.3, and 22.7%, respectively.

TABLE 2 | Topic-based dietary patterns of 2007–2009.

Rank	Topic 1 (Token: 51.3%) (Participant: 68.4%)	Topic 2 (Token: 25.0%) (Participant: 16.2%)	Topic 3 (Token: 23.7%) (Participant: 15.3%)
1	Americano	Instant coffee	Soybean paste soup (Doenjang Soup)
2	Kimchi	White rice	Mix of red pepper paste and soybean paste (Ssamjang)
3	White rice	Ramen	Cabbage
4	Egg	Kimchi	Kimchi
5	Multigrain rice	Pumpkin	White rice
6	Stir-fried anchovy	Cutlassfish	Lettuce
7	Bean sprout	Cucumber	Garlic
8	Sweet potato	Snack	Sesame leaf
9	Marinated meat (Jangjorim)	Cabbage	Red pepper
10	Spinach	Beer	Multigrain rice

The main changes in the dietary patterns of the 2016–2018 dataset compared with that of the 2007–2009 dataset are as follows:

First, an unhealthy dietary pattern occurred in Topic 1 of the 2016–2018 dataset, leading to "americano—fried chicken—ramen—snack—soda—beer," which accounts for 15.5% of

TABLE 3 | Topic-based dietary patterns of 2016–2018.

Rank	Topic 1 (Token: 15.0%) (Participant: 17.4%)	Topic 2 (Token: 62.3%) (Participant: 67.1%)	Topic 3 (Token: 22.7%) (Participant: 15.5%)
1	Mix of red pepper paste and soybean paste (Ssamjang)	Kimchi	Americano
2	Pork belly	Instant coffee	Fried chicken
3	Lettuce	Milk	Mayonnaise
4	Red pepper	White rice	Fish cake soup
5	Cold noodle (Naengmyeon)	Multigrain rice	Ramen
6	Onion	Soybean paste soup (Doenjang Soup)	Snack
7	Soju	Kimchi stew	Chicken breast
8	Grilled mushrooms	Apple	Soda
9	Orange juice	Roasted seaweed	Sausage
10	Duck meat	Stir-fried anchovy	Beer

the survey participants. Second, the new dietary pattern represented as meat and alcohol-oriented dining out pattern occurred in Topic 2, "mix of red pepper paste and soybean paste (ssamjang)—pork belly—lettuce—pepper—cold noodle—onion—soju," accounting for 17.4% of the survey participants.

Figures 6, **7** present the estimated term frequency within the selected topics and the overall term frequency of our results. Suppose the difference between the overall and estimated term frequencies within the selected topic becomes smaller, the uniqueness of the data in the topic becomes larger. In other words, the proportion of biased people toward such a diet increases. Following the previous sentences, as the topic classification for 2016–2018 differs, the diet of the survey participants was more clearly divided between 2016 and 2018. The result indicates that the difference in the dietary patterns rapidly yielded unhealthy dietary patterns, including meat and the aforementioned alcohol-oriented dietary pattern.

Figures 8, **9** present the intertopic distances onto two dimensions from the final topic classification results. The figures originated from topics as circles in the two-dimensional plane, whose centers are determined by computing the Jensen–Shannon divergence (JSD) between topics and then using multidimensional scaling to project the intertopic distances onto two dimensions. Each topic's overall prevalence is encoded using the areas of the circles. The JSD is a method for measuring the similarity between two probability distributions. It is also known as the information radius (IRad) or total divergence to the average. Its square root is a metric referred to as the Jensen–Shannon distance (15, 16). The formula of the JSD is as follows:

Consider the set $M_{+}^{1}(A)$ of probability distributions, where A denotes a set provided with some σ-algebra of measurable subsets. In particular, let A be a finite or countable set with all subsets being measurable. Then, the JSD $M_{+}^{1}(A) \times M_{+}^{1}(A) \rightarrow [0, \infty)$ is a symmetrized and smoothed version of the Kullbeck–Leibler divergence $D(P \parallel Q)$. It is defined by $JSD(P \parallel Q) = \frac{1}{2}\mathrm{D}(\mathrm{P}\parallel M) + \frac{1}{2}\mathrm{D}(\mathrm{Q}\parallel M)$, where $M = \frac{1}{2}(P + Q)$. The geometric JSD yields a closed form for divergence between two normal distributions by taking their geometric mean. The more general definition that allows comparison of more than two probability distributions is as follows:

$$JSD_{\pi_1,\ \pi_2,\ \ldots,\ \pi_n}(P_1,\ P_2,\ \ldots,\ P_n) = \sum_{i=1}^{n} \pi_i D(P_i \parallel M) \tag{2}$$

$$= H\left(\sum_{i=1}^{n} \pi_i P_i\right) - \sum_{i=1}^{n} \pi_i H(P_i)$$

where $M \sum_{i=1}^{n} \pi_i P_i$; $\pi_1, \quad \pi_2, \quad \ldots, \quad \pi_n$ are selected weights from the probability distributions $P_1,\ P_2,\ \ldots,\ P_n$; and $H(P)$ is the Shannon entropy for distribution P. To calculate JSD for illustrating the inter-topic map, we take $P_1 = P$, $P_2 = Q$, $\pi_1 = \pi_2 = \frac{1}{2}$. Thus, for distributions P and Q, the JSD is as follows:

$$JSD = H(M) - \frac{1}{2}(H(P) + H(Q)) \tag{3}$$

DISCUSSION

In a large-scale national survey data, we found three unique Korean dietary patterns from the 2007–2009 and 2016–2018 datasets using AI methodologies. By applying LDA, we observed changes in the Korean dietary patterns over the past decade. In the 2007–2009 dataset, the three major dietary patterns were composed of rice, soup, various vegetables, and fish, which is a traditional dietary pattern. However, in the 2016–2018 dataset, Western dietary patterns based on meat and alcohol (Topic 1), or consisting of fried, processed foods and sugar-sweetened beverages (Topic 3), and a traditional dietary pattern (Topic 2) were observed.

Dietary patterns and changes in food intake among Korean adults were similar to those from previous studies (4, 11, 12, 17). Studies using the KNHANES dataset examined during the same period demonstrated that the major dietary patterns composed of white rice and Kimchi, or grains, vegetables, and fish were extracted using factor analysis (11, 17). Studies using the fifth KNHANES dataset reported the meat and alcohol patterns using a traditional method for dietary pattern extraction (12). Kweon et al. (18) reported a decrease in the intake of grains, vegetables, and carbohydrates and an increase in the intake of beverages (sugar-sweetened beverages), meat, dairy, eggs, and fat. Furthermore, they reported an increase in the intake of snacks, frequency of dining out, and consumption of convenience foods over the past 20 years from the KNHANES datasets between 1998 and 2018, which have similar trends to our results.

Over the last decade, the dietary patterns have changed from traditional to unhealthy patterns among Korean adults. These changes may be influenced by Western diets, and an increase in the frequency of eating out, particularly in young Korean adults. Western dietary patterns are characterized by high intakes

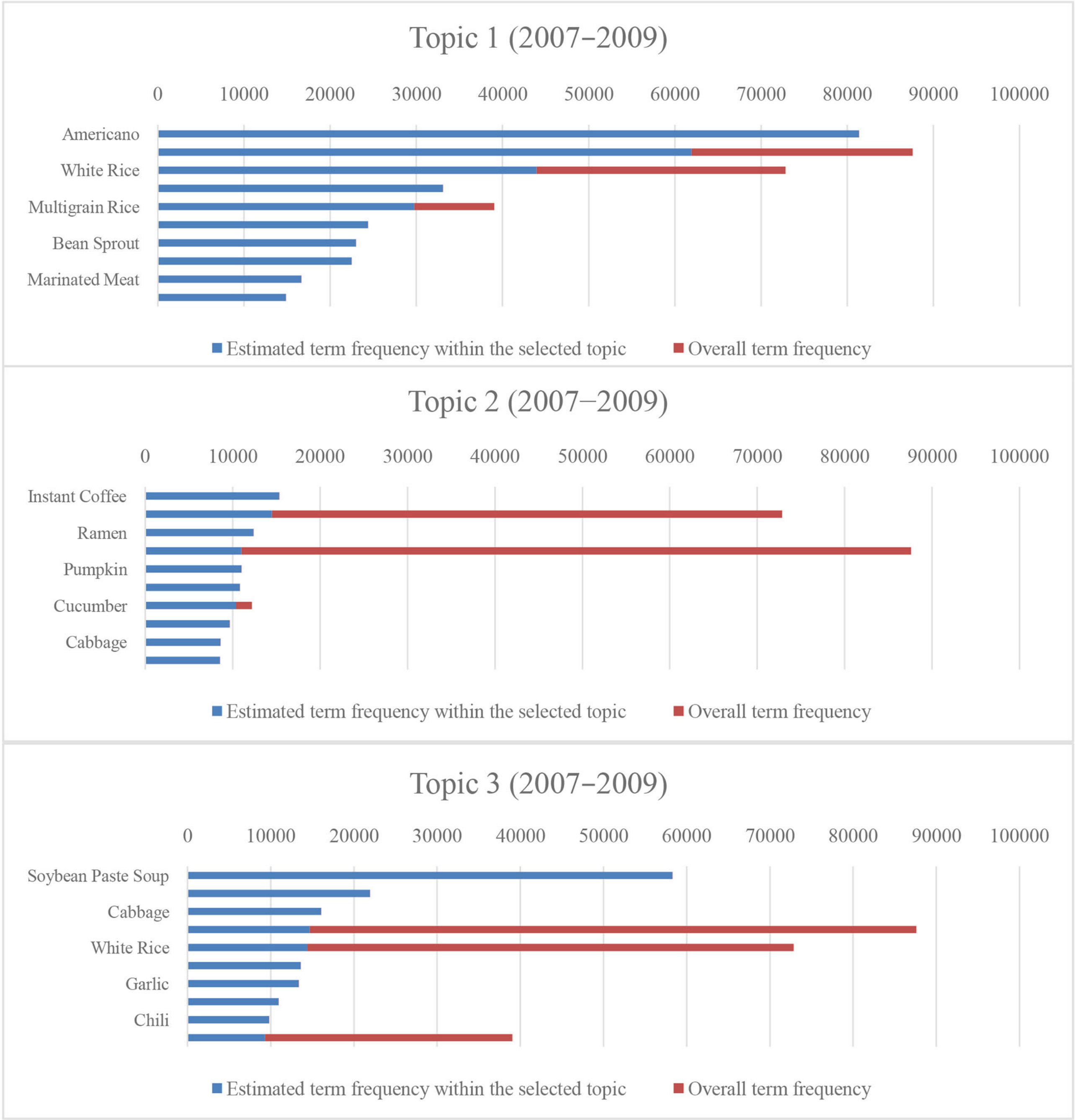

FIGURE 6 | Number of tokens and proportions of entire tokens for the top 10 menus (2007–2009).

of meat, fried foods, refined grains, and added sugars, which may produce unfavorable health outcomes (19–22). Frequent eating out is associated with a higher intake of total energy, fat, saturated fat, cholesterol, and sodium, and a lower intake of dietary fiber (4, 23). Therefore, the shift in dietary patterns should be continuously monitored, and whether the associated changes produce healthy outcomes among Koreans should be checked.

Our findings confirm that the AI and NLP techniques produce similar results to the traditional method previously used for dietary pattern extraction, suggesting that the current NLP-based method can be widely used in nutrition epidemiology.

Our proposed LDA method offers some advantages compared with the traditional method. First, the LDA method provides more specific information on the composition of dietary patterns.

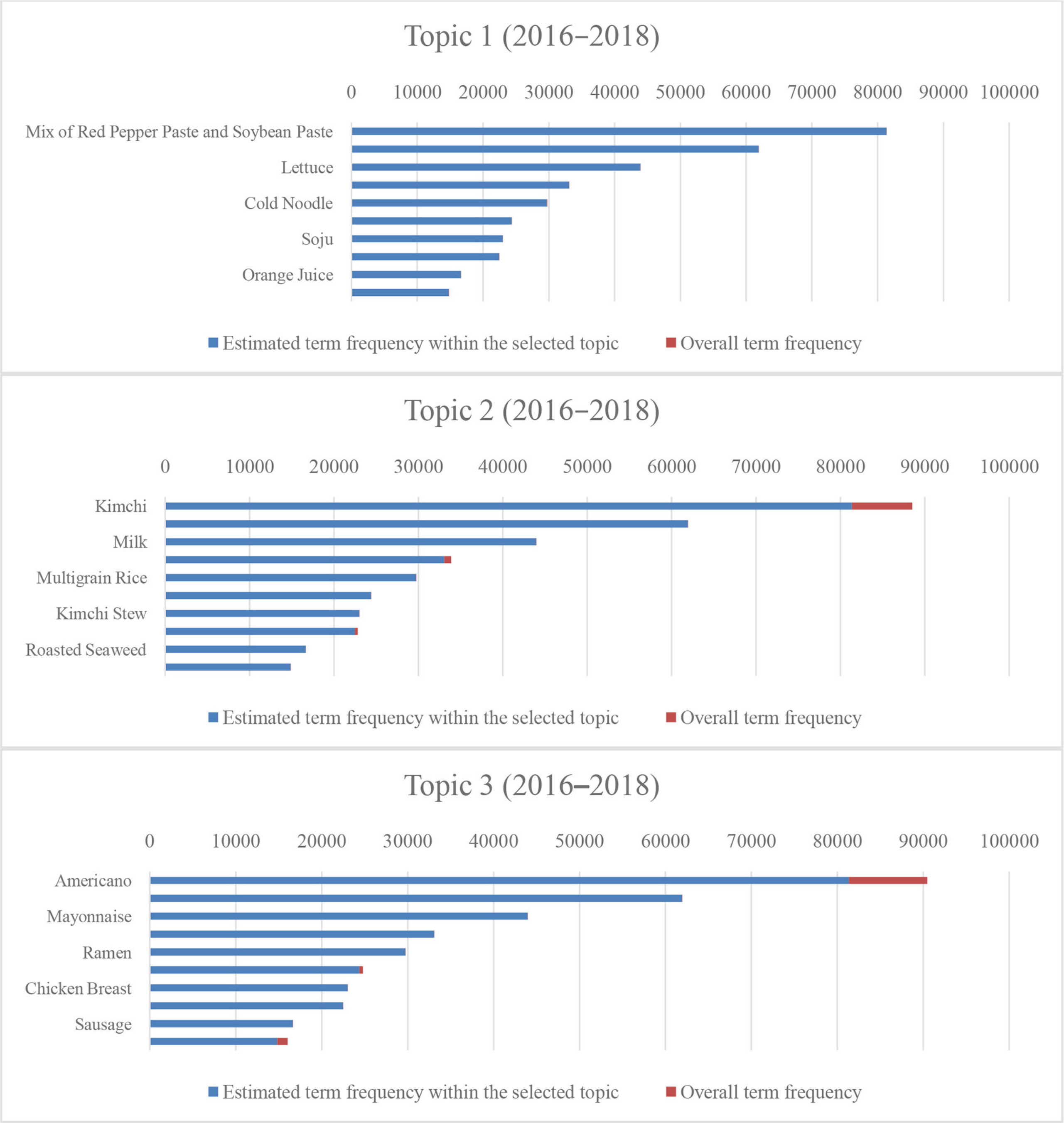

FIGURE 7 | Number of tokens and proportions of entire tokens for the top 10 menus (2016–2018).

For example, in the meat and alcohol pattern (Topic 1), a mix of red pepper and soybean pastes (ssamjang), pork belly, lettuce, red pepper, cold noodle (naengmyeon), onions, and soju is a typical Korean menu for dining out. This result indicates that the recent increase in this dietary pattern compared with 10 years ago may reflect an increase in the frequency of dining out in Korean adults.

However, a traditional method does not provide details about the composition of dietary patterns. Previous studies extracted "meat and alcohol patterns" but did not capture the meat or alcohol types (17). The reason why a traditional method cannot depict details is that the method depends on food item classifications, which vary depending on the investigators (12, 24, 25). For text preprocessing, we developed

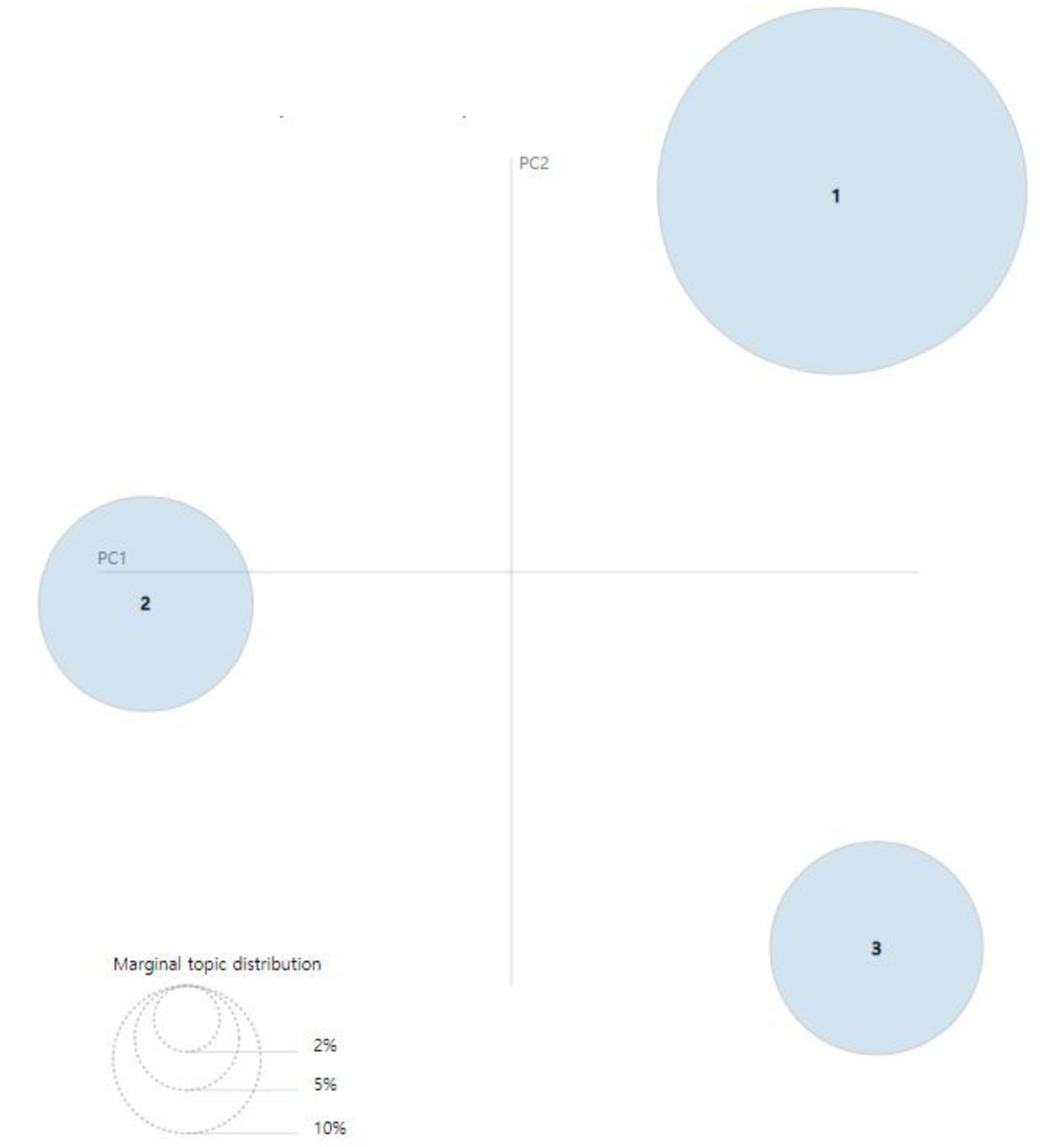

FIGURE 8 | Visualization results of inter-topic distance map *via* multidimensional scaling (MDS) (2007–2009) (14).

the semiautomatic methodology for preprocessing unstructured data, such as dietary recorded text data. Furthermore, we developed the preprocessing method using fastText, the unsupervised or supervised learning algorithm, for obtaining vector representations for words. This method rapidly reduces the preprocessing time for preparing experiments for extracting dietary patterns. The LDA method for classifying topics based on frequently appearing words obtains useful nutritional or dietary information that cannot be obtained by traditional dietary pattern extraction methods. We found that Korean

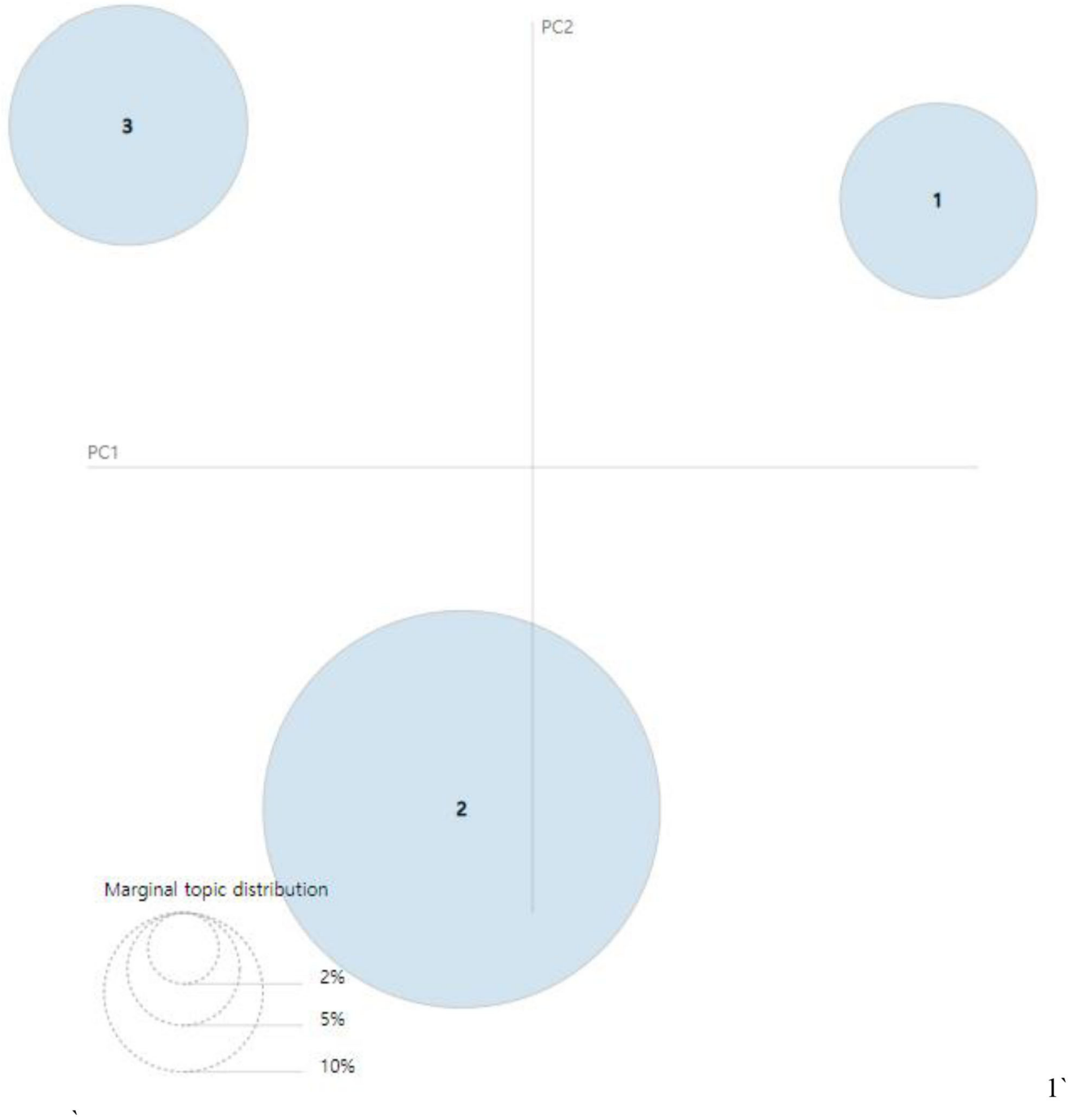

FIGURE 9 | Visualization results of inter-topic distance map *via* multidimensional scaling (MDS) (2016–2018) (14).

adults consumed pork belly and soju frequently using the LDA method. An advantage of the LDA technique is that it needs no assumption to derive dietary patterns (26–28). Without a clue on the dietary pattern, LDA can extract the dietary patterns using the probabilistic model. Furthermore, it is easy to use, less labor-intensive, and more efficient than the traditional method. For instance, suppose more data are added or a new survey is generated, we can easily extend the data, find new patterns *via* topics, and compare them based on frequently appearing data (28–31).

CONCLUSION

In this study, we proposed a novel NLP for extracting dietary patterns. Our results indicate that dietary patterns among Korean adults have evolved from traditional to more Western dietary patterns over the past decade. Given the efficiency and validity of the NLP methods, we recommend that the current NLP-based method be applied in nutrition epidemiology.

Our methodologies and results can contribute to enabling diet research using unstructured data more quickly. In addition, the diary pattern can be identified from a new perspective by performing a nonlinear diary pattern classification based on machine learning rather than a linear dietary pattern classification. Also, the strength of this study is in the application of the novel method using NLP to extract dietary patterns and obtain more useful dietary information, unlike traditional dietary pattern extraction methods. Our results suggest the potential of applying AI-based methods in various fields of nutritional mechanics.

However, the limitations of our research are as follows: First, the experiment results may differ because of the text preprocessing procedure. Second, the last food groups for extracting dietary patterns include subjective opinions of researchers. This means if we set the food groups in different ways, then the conclusion may differ. Therefore, if we improved those limitations, then a more computationally clear and objective dietary pattern may be derived.

Further research on the association between dietary patterns and disease risk can be attempted using machine-learning-based methodologies. In addition, pretrained models for the NLP food name datasets in the nutrition area should be developed.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://knhanes.kdca.go.kr/knhanes/eng/index.do.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WK and JK: conceptualization and project administration. IC: methodology, software, formal analysis, and visualization. IC and WK: data curation and supervision. IC and JK: writing—original draft preparation and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (Ministry of Science and ICT) (No. 2021R1A2C1003211, to JK).

REFERENCES

1. Popkin BM. Nutritional patterns and transitions. *Popul Dev Rev.* (1993) 138–57. doi: 10.2307/2938388
2. Kim S, Moon S, Popkin BM. The nutrition transition in South Korea. *Am J Clin Nutr.* (2000) 71:44–53. doi: 10.1093/ajcn/71.1.44
3. Song S, Kim J, Kim J. Gender differences in the association between dietary pattern and the incidence of hypertension in middle-aged and older adults. *Nutrients*. (2018) 10:252. doi: 10.3390/nu10020252
4. Kim JH, Kim WC, Kim J. A practical solution to improve the nutritional balance of Korean dine-out menus using linear programming. *Public Health Nutr*. (2019) 22:957–66. doi: 10.1017/S1368980018004032
5. Tamaddoni-Nezhad A, Milani GA, Raybould A, Muggleton S, Bohan DA. Construction and validation of food webs using logic-based machine learning and text mining. In: *Advances in Ecological Research Vol. 49*, Academic Press (2013). p. 225–89.
6. Zhu G, Liu H, Zhang S. The extraction method for best match of food nutrition. In: *International Conference on Applications and Techniques in Cyber Security and Intelligence*. Cham: Edizioni della Normale (2017). p. 380–7. doi: 10.1007/978-3-319-67071-3_44
7. Korpusik M, Glass J. Spoken language understanding for a nutrition dialogue system. *IEEE/ACM Trans Audio, Speech, Language Process IEEE-ACM T AUDIO SPE*. (2017) 25:1450–61. doi: 10.1109/TASLP.2017.2694699
8. Tao D, Yang P, Feng H. Utilization of text mining as a big data analysis tool for food science and nutrition. *Compr Rev Food Sci Food Safety*. (2020) 19:875–94. doi: 10.1111/1541-4337.12540
9. Van Erp M, Reynolds C, Maynard D, Starke A, Martín RI, Andres F, et al. Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. *Front Artif*. (2020) 3. doi: 10.3389/frai.2020.621577
10. Bakhtin P, Khabirova E, Kuzminov I, Thurner T. The future of food production-a text-mining approach. *Technol Anal Strateg Manag*. (2020) 32:516–28. doi: 10.1080/09537325.2019.1674802
11. Kim J, Jo I, Joung H. A rice-based traditional dietary pattern is associated with obesity in Korean adults. *J Acad Nutr Diet*. (2012) 112:246–53. doi: 10.1016/j.jada.2011.10.005
12. Kang Y, Kim, J. Gender difference on the association between dietary patterns and metabolic syndrome in Korean population. *Eur J Nutr*. (2016) 55:2321–30. doi: 10.1007/s00394-015-1127-3
13. Choi I, Kim WC. Empirical analysis of politically-themed stocks using text mining techniques and entropy-based network dynamics–focus on the Republic of Korea's case. In: *Proceedings of the 2020 KIIE Fall Conference*. (2020). p. 1518.
14. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. (2014) 63-70. doi: 10.3115/v1/W14-3110
15. Dagan I, Lee L, Pereira F. Similarity-based methods for word sense disambiguation. In: *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. (1997). p. 56–63. arXiv:cmp-lg/9708010. Bibcode:1997cmp.lg.0.0.8010D. Retrieved 2008-03-09.
16. Schütze H, Manning CD. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press (1999). p. 304.
17. Kim J, Jo I. Grains, vegetables, and fish pattern is inversely associated with the risk of metabolic syndrome in South Korean adults. *J Am Diet Assoc*. (2011) 111:1141–9. doi: 10.1016/j.jada.2011.05.001
18. Kweon S, Park JY, Park M, Kim Y, Yeon SY, Yoon L, et al. Trends in food and nutrient intake over 20 years: findings from the 1998–2018 Korea National Health and Nutrition Examination Survey. *Epidemiol Health*. (2021) 43:e2021027. doi: 10.4178/epih.e2021027
19. Van Dam RM, Rimm EB, Willett WC, Stampfer MJ, Hu FB. Dietary patterns and risk for type 2 diabetes mellitus in US men. *Ann Intern Med*. (2002) 136:201–9. doi: 10.7326/0003-4819-136-3-200202050-00008

20. Van Dam RM, Grievink L, Ocké MC, Feskens EJ. Patterns of food consumption and risk factors for cardiovascular disease in the general Dutch population. *Am J Clin Nutr*. (2003) 77:1156–63. doi: 10.1093/ajcn/77.5.1156
21. Newby PK, Muller D, Hallfrisch J, Andres R, Tucker KL. Food patterns measured by factor analysis and anthropometric changes in adults. *Am J Clin Nutr*. (2004) 80:504–13. doi: 10.1093/ajcn/80.2.504
22. Lutsey PL, Steffen LM, Stevens J. Dietary intake and the development of the metabolic syndrome: the Atherosclerosis Risk in Communities study. *Circulation*. (2008) 117:754–61. doi: 10.1161/CIRCULATIONAHA.107.716159
23. Todd JE. Changes in consumption of food away from home and intakes of energy and other nutrients among US working-age adults, 2005–2014. *Public Health Nutr*. (2017) 20:3238–46. doi: 10.1017/S1368980017002403
24. Shin S, Joung H. A dairy and fruit dietary pattern is associated with a reduced likelihood of osteoporosis in Korean postmenopausal women. *Br J Nutr*. (2013) 110:1926–33. doi: 10.1017/S0007114513001219
25. Oh C, No J. The quality of a traditional dietary pattern in relation to metabolic syndrome in elderly South Koreans. *J Obesity Metab Syndr*. (2018) 27:254. doi: 10.7570/jomes.2018.27.4.254
26. Lukins SK, Kraft NA, Etzkorn LH. Source code retrieval for bug localization using latent dirichlet allocation. In: *2008 15th Working Conference on Reverse Engineering*. IEEE (2008). p. 155–64. doi: 10.1109/WCRE.2008.33
27. Tran BX, Nghiem S, Sahin O, Vu TM, Ha GH, Vu GT, et al. Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study. *J Med Internet Res*. (2019) 21:e15511. doi: 10.2196/15511
28. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimed Tools Appl*. (2019) 78:15169–211. doi: 10.1007/s11042-018-6894-4
29. Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation. In: *Proceedings of the Third ACM Conference on Recommender Systems*. (2009). p. 61–8. doi: 10.1145/1639714.1639726
30. Văduva C, Gavăt I, Datcu M. Latent Dirichlet allocation for spatial analysis of satellite images. *IEEE Trans Geosci Remote Sens*. (2012) 51:2770–86. doi: 10.1109/TGRS.2012.2219314
31. Yang S, Yuan C, Hu W, Ding X. A hierarchical model based on latent dirichlet allocation for action recognition. In: *2014 22nd International Conference on Pattern Recognition*. IEEE (2014). p. 2613–8. doi: 10.1109/ICPR.2014.451

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.