



American
Society for
Nutrition

Excellence in
Nutrition Research
and Practice

The American Journal of CLINICAL NUTRITION

journal homepage: www.journals.elsevier.com/the-american-journal-of-clinical-nutrition



Original Research Article

Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods

Guanlan Hu¹, Mavra Ahmed^{1,2}, Mary R. L'Abbé^{1,*}

¹ Department of Nutritional Sciences, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada; ² Joannah & Brian Lawson Centre for Child Nutrition, University of Toronto, ON, Canada

ABSTRACT

Background: Food categorization and nutrient profiling are labor intensive, time consuming, and costly tasks, given the number of products and labels in large food composition databases and the dynamic food supply.

Objectives: This study used a pretrained language model and supervised machine learning to automate food category classification and nutrition quality score prediction based on manually coded and validated data, and compared prediction results with models using bag-of-words and structured nutrition facts as inputs for predictions.

Methods: Food product information from University of Toronto Food Label Information and Price Database 2017 ($n = 17,448$) and University of Toronto Food Label Information and Price Database 2020 ($n = 74,445$) databases were used. Health Canada's Table of Reference Amounts (TRA) (24 categories and 172 subcategories) was used for food categorization and the Food Standards of Australia and New Zealand (FSANZ) nutrient profiling system was used for nutrition quality score evaluation. TRA categories and FSANZ scores were manually coded and validated by trained nutrition researchers. A modified pretrained sentence-Bidirectional Encoder Representations from Transformers model was used to encode unstructured text from food labels into lower-dimensional vector representations, followed by supervised machine learning algorithms (i.e., elastic net, k-Nearest Neighbors, and XGBoost) for multiclass classification and regression tasks.

Results: Pretrained language model representations utilized by the XGBoost multiclass classification algorithm reached overall accuracy scores of 0.98 and 0.96 in predicting food TRA major and subcategories, outperforming bag-of-words methods. For FSANZ score prediction, our proposed method reached a similar prediction accuracy (R^2 : 0.87 and MSE: 14.4) compared with bag-of-words methods (R^2 : 0.72–0.84; MSE: 30.3–17.6), whereas structured nutrition facts machine learning model performed the best (R^2 : 0.98; MSE: 2.5). The pretrained language model had a higher generalizable ability on the external test datasets than bag-of-words methods.

Conclusions: Our automation achieved high accuracy in classifying food categories and predicting nutrition quality scores using text information found on food labels. This approach is effective and generalizable in a dynamic food environment, where large amounts of food label data can be obtained from websites.

Keywords: food categorization, nutrient profiling, food composition database, natural language processing (NLP), pretrained language model, machine learning, food label, nutrition facts table, nutrition quality

Introduction

Unhealthy diets, characterized by high intakes of saturated fat, sodium, and sugar (nutrients of public health concern), are one of the leading modifiable risk factors for the prevention of noncommunicable diseases [1]. Development and evaluation of nutrition interventions toward mitigating noncommunicable diseases require a comprehensive database of

foods. National food composition databases are expensive and challenging to develop, construct, and maintain [2, 3]. The University of Toronto Food Label Information and Price (FLIP) database containing nutrition information for >120,000 branded foods and beverages collected since 2010 was developed to evaluate the changing food supply [4, 5].

Food categorization and nutrient profiling for products in food composition databases are essential for policy and regulation applications

Abbreviations used: BERT, Bidirectional Encoder Representations from Transformers; FLIP, University of Toronto Food Label Information and Price Database; FSANZ, Food Standards of Australia and New Zealand; FVNL, Fruits, Vegetables, Nuts, and Legume; KNN, k-Nearest Neighbor; NLP, Natural Language Processing; t-SNE, t-Distributed Stochastic Neighbor Embedding; TRA, Table of Reference Amounts for Food; XGBoost, eXtreme Gradient Boosting.

* Corresponding author. to MRL e-mail: . Present address: Department of Nutritional Sciences, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada.

E-mail address: mary.labbe@utoronto.ca (M.R. L'Abbé).

<https://doi.org/10.1016/j.ajcnut.2022.11.022>

Received 14 September 2022; Received in revised form 16 November 2022; Accepted 29 November 2022

Available online 23 December 2022

0002-9165/© 2022 American Society for Nutrition. Published by Elsevier Inc. All rights reserved.

(e.g., restricting the marketing of unhealthy foods to children, front-of-package labeling) [6–8]. In Canada, Table of Reference Amounts (TRAs) has been published by Health Canada to standardize the categorization and serving sizes of Canadian foods for nutrition labeling regulations and details 24 major categories and 172 subcategories [9]. There are >300 nutrient profiling models available for nutrition quality assessment (each with specific criteria on categories of food and nutrient thresholds for specific ingredients) [8, 10–13]. Previous versions of FLIP have relied on manual food categorization and nutrient profiling score calculation. Given the number of products and complex nutrient profiling process, manual coding is labor intensive, time consuming, and challenging with a dynamic food supply [14, 15]. Therefore, effective and generalizable automated processes using machine learning are required.

Traditional methods for processing texts on food labels, such as representing ingredient occurrence as binary numbers (0 or 1), have been used to predict nutrient levels not declared on food packages [16, 17]. However, their performance is limited by the input dimension and cannot easily handle other unstructured food label text information such as name, brand, and ingredients. Information from the nutrition facts table was also used in some machine learning models (i.e., k-Nearest Neighbors [KNNs], decision tree) to estimate missing nutrients, such as added sugar and fiber content [18, 19]. Pretrained language model (i.e., Bidirectional Encoder Representations from Transformers [BERT]) is a state-of-the-art language model for natural language processing (NLP) designed to pretrain contextual representations by representing unlabeled text in a deep, bidirectional, and unsupervised way using large text corpora as inputs for pretraining [20, 21]. It can encode a variety of texts on food labels into low-dimensional dense vectors with high performance on multiple tasks similar to large-scale food category classification, product similarity comparisons, and nutrition quality score prediction. Although previous literature has used machine learning algorithms in the classification of recipes and the prediction of nutrient levels not declared on food packages [17–19, 22], no study to our knowledge has applied the pretrained language model to process text information from food labels. There is limited evidence on the application of machine learning for nutrient profiling.

To our knowledge, this study is the first that applied pretrained NLP approach to encode texts on food labels and used supervised machine learning techniques to automate food category classification and nutrition quality score calculation for a large food composition database and compared the performance with traditional bag-of-words ingredients occurrence and structured nutrition facts models. Accuracy was assessed using manually coded data validated by trained nutrition researchers (MSc and PhD level staff and students).

Methods

Food composition database

This study used the University of Toronto Food Label Information and Price (FLIP) database FLIP2017 ($n = 19,720$) and FLIP2020 ($n = 74,445$). Briefly, FLIP is a database of Canadian branded packaged food and beverages developed in 2010 that is updated every 3 to 4 y. It contains food label information (e.g., product name, brand, nutrition facts, ingredients, store, price, and product images) for >100,000 food products from major food retailers that cover approximately 80% of the market share in Canada [23]. The FLIP database, as a national branded food composition database, is essential for in-depth nutritional analyses of the Canadian food environment to understand relationships between the nutritional quality of food products and measurements of policy

impacts and health over time. Previous versions of FLIP before 2020 collected food label information manually or through an iPhone digital collection application (APP). The latest iteration, FLIP 2020, used web-scraping coupled with optical character recognition technology to read food labels and collect real-time food and nutrition information [4].

Data preparation

Packaged foods and beverages were manually categorized by trained nutrition researchers using Health Canada's TRA, which consists of 24 major categories (e.g., A. Bakery products, B. Beverages, D. Dairy products and substitutes, S. Snacks, etc.) and 172 subcategories (e.g., A.4. Brownies, dessert squares and bars, Muffins; D.4. Hard cheese, including grated, such as Parmesan or Romano; D.10. Milk, evaporated or condensed; D.15. Yogurt; S.3. Meat or poultry snack food sticks). The complete TRA category list can be found in Supplementary Materials (Supplementary Methods) section, and a full listing of all categories and subcategories has been published by Health Canada [9]. Manual categorization was completed by trained nutrition researchers, and all categorizations were verified by a different team member. The agreement was >96% overall, varying between approximately 96% and 99% for different categories. When accurate categorization was uncertain, staff at Health Canada who prepared the TRA categories were consulted to ensure accurate categorization.

To determine the nutritional quality and calculate a *nutrition quality score* of products found in FLIP database, the Food Standards Australia New Zealand (FSANZ) method was manually applied based on nutritional composition per 100 g or mL, depending on the unit for which the *Nutrition Facts table* was displayed. The FSANZ score was chosen for this research because it is a complicated nutrient profiling system that calculates a summary score based on the amount of nutrients to limit and components to encourage in food, which includes both positive and negative nutrient attributes and food ingredients and have been used extensively to assess the nutritional quality of foods and beverages [6, 7, 24–26]. FSANZ is a criterion for determining whether foods are eligible to carry health claims on labels [11, 24]. Briefly, foods or beverages were assigned a category, which was used to determine the allocation of scoring cut-offs for products. *Baseline points* were assigned for a product's energy, saturated fat, sodium, and total sugars content. *Modifying points* were awarded by estimating the proportion of the food consisting of fruits, vegetables, nuts, and legumes (FVNLs) ingredients and calculating protein and fiber. In the absence of quantitative ingredient declarations on Canadian food products, a method to estimate the FVNL content of products using the ingredients list was developed and applied, as previously described [14]. All categories were evaluated *as sold*. The process for the FSANZ score calculation is complicated and involves a series of time consuming steps. The FSANZ score of food products in the FLIP2017 and FLIP2020 databases were manually calculated by trained nutrition researchers, following this method (Supplementary Materials).

Figure 1 indicates the data preparation flow for the development and testing of TRA food categorization and FSANZ nutrition quality score prediction algorithms. A total of 74,445 products were extracted from FLIP2020. For TRA categorization algorithms, categories that were not for human consumption (e.g., pet foods, cleaning products) and products that did not contain food name, brand, or ingredient information ($n = 28,425$) were excluded. The final sample size for the FLIP2020 TRA categorization task was 46,020. Food categories that are not applicable for FSANZ calculation (e.g., foods intended solely for children under 4 y of age) and those without nutrient information (n

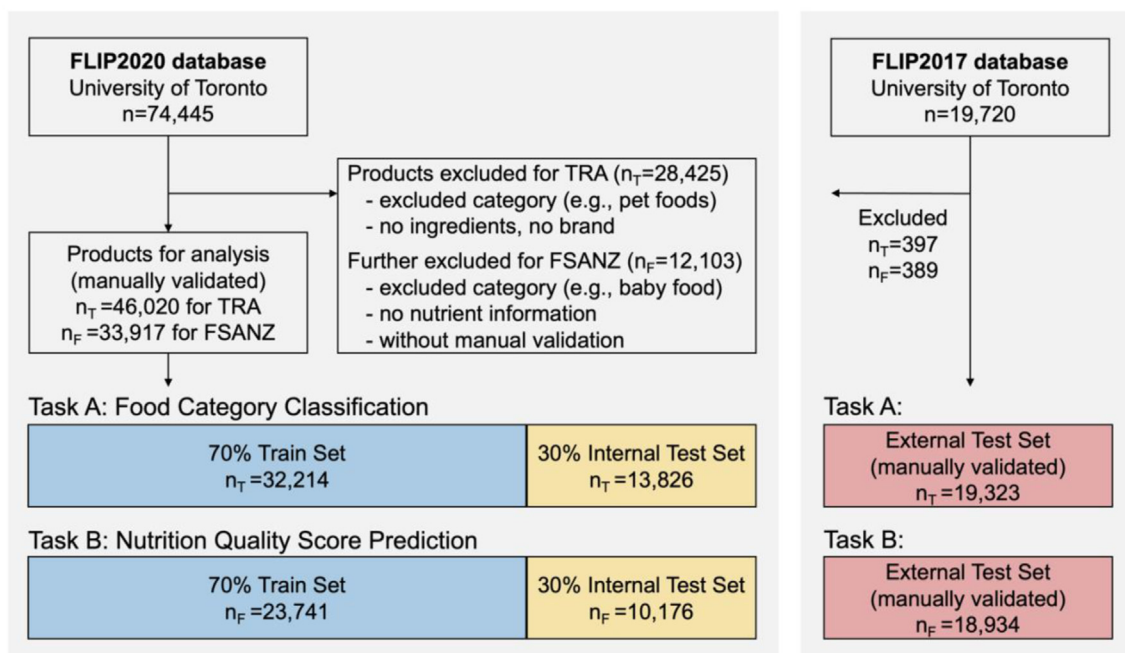


Figure 1. Data preparation flow chart and exclusion criteria used for task A: Table of Reference Amounts food categorization and task B: Food Standards of Australia and New Zealand nutrition quality score predictions and validation.

= 12,103) were further excluded, thereby resulting in a final sample size of 33,917 for FSANZ nutrition quality score prediction algorithms. In addition, the FLIP2017 dataset ($n = 19,323$ after exclusion for TRA and $n = 18,934$ after exclusion for FSANZ), for which TRA categories and FSANZ scores had been determined manually by trained nutrition researchers, was used to further test the algorithms developed based on FLIP2020.

Food product representation and visualization

Pretrained language model.

BERT is a pretrained machine learning model for NLP that learns contextual relations between words from texts. Its key feature is that instead of learning the representation of words from one direction of sentences, BERT takes into account information from both the left and right of texts. This enables researchers to fine-tune BERT for downstream tasks, such as generating embeddings or making predictions. During the pretraining stage of BERT, 2 very large inputs were used: Books Corpus (800M words) [27] and English Wikipedia (2500M words) [20]. This training allows BERT to capture semantic meanings of words more accurately, even for rarely used words and provides additional benefits for nutrition NLP research because texts from nutrition information typically only contain words and phrases, such as brand names and ingredients, rather than complete sentences. In such cases, traditional methods relying on the sentence structure to represent texts will not perform well, and the use of many rarely used words in such a context will make representation even more difficult. By using pretrained models, we can deal with such difficulties easily.

Sentence embeddings using Siamese BERT, a modified pretrained BERT network specialized for encoding text with reduced computing time consumption and maintained accuracy [28], were applied to convert text on food product labels (e.g., product name, brand, and ingredients) into numerical representations. Sentence BERT has the Siamese network architecture, which generates fixed-sized vectors

from the input text and performs efficiently on clustering and semantic similarity search. In our case, sentence BERT encoded texts on food product labels into 384-dimensional dense vectors. These representations were then used for different downstream tasks, such as large-scale food product categorization and similarity comparisons.

Bag-of-words.

To compare the performance of the pretrained language model, the bag-of-words, a simple and commonly used method for text feature extraction, was also applied. Bag-of-words model measured the presence of the words (e.g., each ingredient) in the given text (e.g., ingredients list) without keeping the information about the order or the structure of words in the text [16]. This method is simple to implement and offers flexibility for customizing specific text data, which is ideal for prediction tasks, such as text classification. The list of ingredients of each food product was split by a comma, and the top 100, 500, 1000, or 2000 ingredients appearing most often in all food products were selected. The bag-of-words approach was used to vectorize the occurrence of the top ingredients in the ingredients list (as 1) or not (as 0) into a binary matrix. Each column represented the one-hot-encoding vector of an ingredient.

Structured nutrition facts model.

In addition to the pretrained language model and bag-of-words representations, structured nutrient data as food representations were also used for the nutrition quality score prediction task. The amount of nutrients per 100 units (100 g-basis for solid food products and 100 mL-basis for liquid food products) was calculated based on the nutrition facts table and serving size information provided on the food package.

Visualization.

To visualize the high-dimensional space of food product representation, the t-distributed stochastic neighbor embedding (t-SNE) method

was used to project food products onto a 2D space. The t-SNE algorithm calculates a similarity measure between pairs of instances in the low-dimensional space. It gives each food product a location in a 2D map based on a variation of the Stochastic Neighbor Embedding [29]. By reducing the tendency to crowd points together in the center of the map, t-SNE produces significantly better visualizations and minimizes crowding compared with other techniques.

2.4. Machine Learning Algorithms

Using the inputs from food representations, the elastic net algorithm, KNN algorithm, and extreme gradient boosting (XGBoost) algorithms were applied for multiclass classification (one compared withrest) to predict food TRA major categories and subcategories and were used for regression to estimate nutrition quality score. All the data were split into 70% as a training set and 30% as a testing set, as described in Figure 1.

Elastic net is a novel regularization and variable selection method with good prediction accuracy. We used the stochastic gradient descent classifier with the elastic net penalty that combines penalties of the lasso and ridge methods, namely sparsity and shrinkage, together [30].

The k-nearest neighbors algorithm, also known as KNN, is a nonparametric supervised machine learning algorithm that assumes a similar food product exists in close proximity [31, 32]. The Manhattan distance between the query food product and the other food products was calculated to form decision boundaries. To determine the classification of a specific query food product, we checked 5 neighbors for the KNN classification model. Previous studies have shown that KNN performs very well in nutrition-related tasks, such as predicting added sugar content, dietary fiber content and nutritional phenotypes based on nutrient variables [18].

XGBoost is a scalable end-to-end tree-boosting system that can solve real-world scale problems using a minimal amount of resources [33]. By combining weak learners, the XGBoost reduced the models' residuals and increased the predictive power [31]. It provides parallel tree boosting and performs well on regression, classification, and ranking tasks [33]. In addition, XGBoost does not require input normalization because it is essentially an ensemble algorithm comprised of decision trees. We used the XGBoost one compared withrest, multiclass classification model in this study [33, 34].

Statistical analyses

To evaluate the performance of multiclass classification machine learning algorithms on the given categorization prediction tasks, accuracy and balanced accuracy were calculated as performance indicators, indicating the difference between true values (manually validated TRA categories) and predicted values (by algorithms). Accuracy is the ratio of correctly predicted observations to the total observations, mainly depending on the algorithm's performance in the biggest classes. Balanced accuracy is useful for multiclass classification when classes are imbalanced, and each class has an equal weight in the final calculation [34]. Confusion matrix, precision, recall, macro, and micro F1 scores were also calculated to evaluate the model performance [34].

To assess the nutrition quality regression models' performance, R^2 and mean absolute error were used, indicating the amount of difference between true values (manually validated FSANZ scores) and predicted values (by algorithms). R^2 reflects the proportion of variation in the outcome that is explained by the predictor variables. Mean Squared Error (MSE) measures the average squared difference between the

observed actual outcome values and the values predicted by the model. Higher R^2 and lower mean absolute error represent the better model.

All the analyses were conducted using Python 3.9.

Results

Performance of food label pretrained language model representations for predicting TRA major category and subcategory

Figure 2 illustrates the t-SNE visualization of food products in our FLIP food composition database by TRA category using food label pretrained language model representations of product names and ingredients. The visualization shows that food label pretrained language model representations produced by t-SNE clearly differentiate food products into various TRA categories in FLIP2017 (Figure 2A) and FLIP2020 (Figure 2B), as illustrated by the clusters. In the t-SNE graph, each cluster (i.e., food category) is indicated by a different color. Clusters indicate that different categories of products are closer to each other within the category, whereas more distant to products of different clusters. Each dot corresponds to a vector representing a unique food product. The closer the 2 points are, the more similar the 2 products are. For example, the dark blue clusters represent all bakery products in FLIP2017 and FLIP2020. Cheddar cheese and mozzarella cheese are closer and in the same cluster, but are different from Coca Cola or Sprite, which are in a different cluster.

Overall, food label representations generated by the pretrained language model predicted TRA major categories and subcategories of food products in the FLIP database with high accuracy (Table 1). Of note, accuracy scores of 0.98 for predicting the TRA major categories and 0.96 for predicting the TRA subcategories were reached using the XGBoost multiclass classification algorithm with name, brand, and ingredients as data inputs. XGBoost and KNN classification algorithms performed better on the task of food categorization using food label pretrained language model representations than the elastic net classification algorithm. Based on prediction accuracy and balanced accuracy among TRA categories, using food label pretrained language model representation of product name and ingredients, as well as representation of the name, brand, and ingredients, performed slightly better on food categorization tasks compared with using representations of ingredients alone in elastic net, KNN, and XGBoost algorithms (Table 1).

Figure 3 and Supplementary Table 1 show the prediction performance of food categorization in each TRA major category and subcategory using food label pretrained language model representations. The prediction F1 score of TRA categories ranged from 0.90 to 0.99. Specifically, this method reached an F1 score of 0.99 for dairy, eggs, fats and oils; 0.98 for beverages, desserts, marine and freshwater animals, cereals, fruit, vegetables, soups, and sauces; 0.97 for meat, bakery, etc. (Figure 3). Furthermore, 115 out of 145 TRA subcategories (with a sample size >5) had a prediction F1 score higher than 0.9 (Supplementary Table 1).

Performance of traditional bag-of-words methods for predicting TRA major category and subcategory

The food products in different TRA categories using the top ingredients bag-of-words method are visualized in t-SNE graphs (Figure 4). FLIP database contains about 47k unique ingredients, and the top 100 ingredients account for an average coverage of 89.2% of products, whereas the top 2000 ingredients account for an average

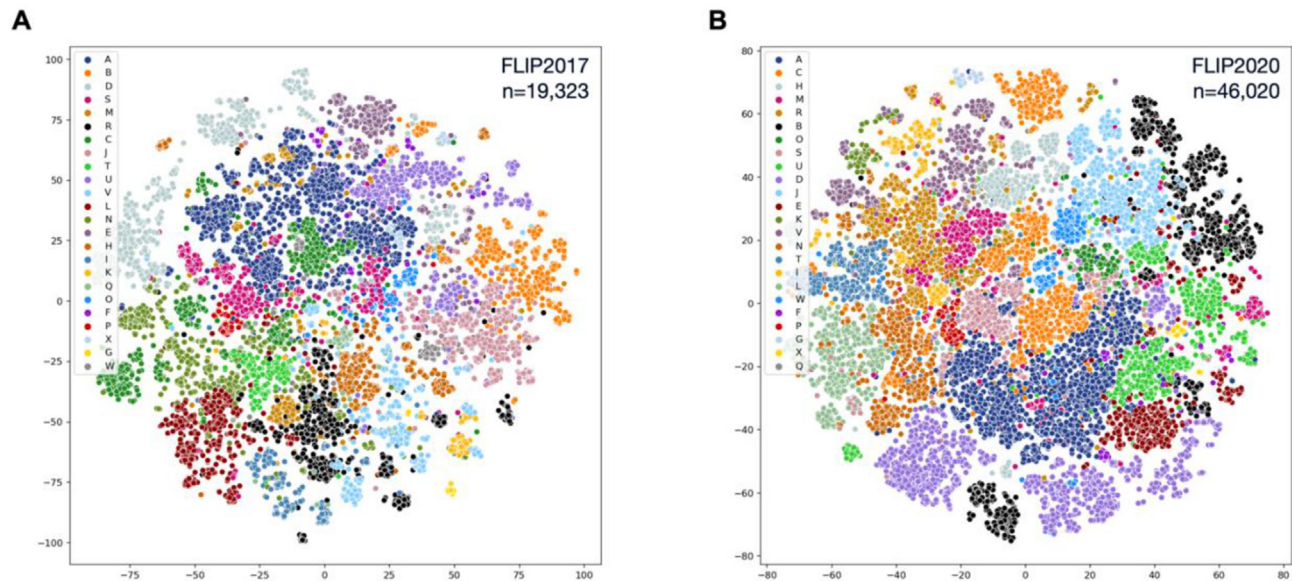


Figure 2. t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of food products in each Table of Reference Amounts (TRA) major category represented by the pretrained language model. (A) FLIP2017 product name and ingredients embedding ($n = 19,323$). (B) FLIP2020 product name and ingredients embeddings ($n = 46,020$). t-SNE (t-distributed stochastic neighbor embedding). TRA, Health Canada’s Table of Reference Amounts. TRA major food categories: A, bakery products; B, beverages; C, cereals and other grain products; D, dairy products and substitutes; E, desserts; F, dessert toppings and fillings; G, eggs and egg substitutes; H, fats and oils; I, marine and freshwater animals; J, fruit and fruit juices; K, legumes; L, meat, poultry, and substitutes; M, miscellaneous category; N, combination dishes; O, nuts and seeds; P, potatoes, sweet potatoes and yams; Q, salads; R, sauces, dips, gravies, and condiments; S, snacks; T, soups; U, sugars and sweets; V, vegetables; W, foods intended solely for children aged <4 y; X, meal replacements and nutritional supplements. (9)

coverage of 99.9% of products in FLIP2020. The top 100 ingredients bag-of-words method did not perform well on several vectorized food products, although this method was reported to well predict nutrients in a previous study [17]. Many food products could not be differentiated, as shown in the red arrow markers (Figure 4A and Figure 4B). However, when increased to the top 2000 ingredients bag-of-words, t-SNE plots showed better food vectorization and TRA separation of almost all products (Figure 4C and Figure 4D).

Similar trends were indicated by the accuracy and balanced accuracy of the TRA major category and subcategory prediction model (Table 2). The highest accuracy of TRA category prediction gradually increased from 0.85 (using top 100 ingredients bag-of-words) to 0.96 (using top 2000 ingredients bag-of-words). XGBoost performed the

best with the highest accuracy compared with the elastic net and KNN algorithms. The TRA subcategory prediction accuracy increased from 0.81 (using top 100 ingredients bag-of-words) to 0.94 (using top 2000 ingredients bag-of-words). Ideally, the more ingredients involved in the algorithm, the higher accuracy can be reached. However, more computing resources and time will be needed.

Regarding prediction performance by TRA category, 38% and 92% of TRA major category predictions reached an F1 score >0.9 using the top 100 ingredients bag-of-words and the top 2000 ingredients bag-of-words, respectively (Figure 5). Only 46 out of 145 TRA subcategory predictions (sample size >5) had an F1 score over 0.9 when using the top 100 ingredients bag-of-words (Supplementary Table 2). Using the top 2000 ingredients bag-of-words, 100 out of 145 TRA subcategory

Table 1
Accuracy and balanced accuracy of TRA major category and subcategory prediction algorithms using different food label pretrained language model representations^{1,2}

Pretrained language model	Algorithm	TRA major category		TRA subcategory	
		Accuracy	Balanced accuracy	Accuracy	Balanced accuracy
Ingredients	Elastic net	0.83	0.76	0.71	0.47
Ingredients	KNN	0.95	0.93	0.90	0.81
Ingredients	XGBoost	0.96	0.94	0.93	0.84
Name & ingredients	Elastic net	0.92	0.89	0.89	0.70
Name & ingredients	KNN	0.97	0.95	0.94	0.88
Name & ingredients	XGBoost	0.97	0.96	0.95	0.88
Name & brand & ingredients	Elastic net	0.95	0.92	0.91	0.76
Name & brand & ingredients	KNN	0.95	0.94	0.91	0.83
Name & brand & ingredients	XGBoost	0.98	0.96	0.96	0.89

¹ KNN; k-nearest neighbors; XGBoost, extreme gradient boosting; TRA, Health Canada’s Table of Reference Amounts (24 major categories and 172 subcategories) (9).

² Accuracy compared with manually categorized and validated categories/subcategories (FLIP2020; $n = 46,020$).

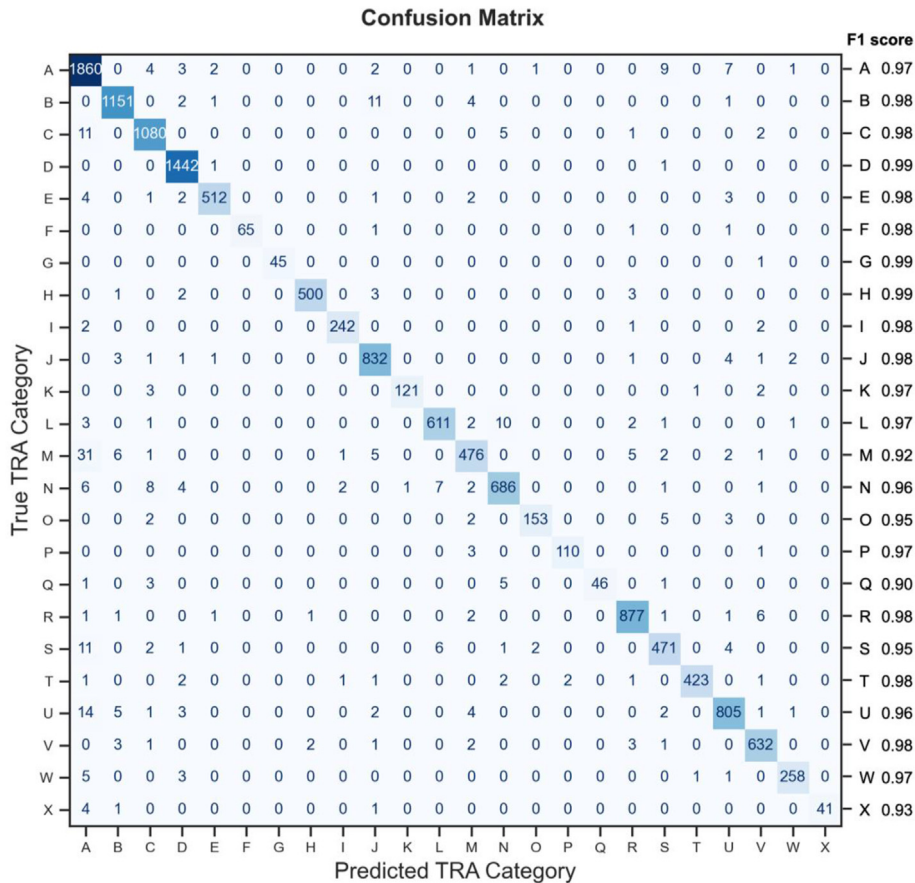


Figure 3. Confusion matrix and prediction F1 score of each TRA major category using food label pretrained language model representations of food product name and ingredients only with XGBoost ($n = 13,806$ in the testing dataset). The number and blue colors indicated the sample size in each category. Health Canada’s Table of Reference Amounts (TRA), including 24 major categories and 172 subcategories (9). TRA major categories: A, bakery products; B, beverages; C, cereals and other grain products; D, dairy products and substitutes; E, desserts; F, dessert toppings and fillings; G, eggs and egg substitutes; H, fats and oils; I, marine and freshwater animals; J, fruit and fruit juices; K, legumes; L, meat, poultry, and substitutes; M, miscellaneous category; N, combination dishes; O, nuts and seeds; P, potatoes, sweet potatoes and yams; Q, salads; R, sauces, dips, gravies, and condiments; S, snacks; T, soups; U, sugars and sweets; V, vegetables; W, foods intended solely for children aged <4 y; X, meal replacements and nutritional supplements.

predictions (sample size >5) had an F1 score >0.9 (Supplementary Table 3).

Comparison of food label pretrained language model representations and traditional methods for predicting TRA category

Food products represented by the food label pretrained language model generally performed better than the bag-of-words method within the same data dimension and computing resources. The food label pretrained language model converted food label ingredients information to a 384-dimensional dense vector space and reached an overall accuracy of 0.96 for TRA major category prediction and 0.93 for TRA subcategory prediction (Table 1). However, using a similar 500-dimensional binary vector from ingredients bag-of-words only reached 0.93 for TRA major category prediction and 0.90 for TRA subcategory prediction (Table 2). The distribution of TRA major category and subcategory prediction F1 scores similarly indicated the same results that food label pretrained language model representations overall led to more accurate predictions than the bag-of-words method. Eighty percent of TRA subcategory predictions using the pretrained language model had an F1 score of >0.9, which outperformed the traditional bag-of-words method (32%–70%) (Figure 5). In addition, the pretrained language model also outperformed the structured nutrition facts

models using nutrients per 100 units as inputs (62% of TRA subcategory predictions had an F1 score of >0.9) (Figure 5).

In addition, when generalizing the algorithms on the FLIP2017 database, the food label pretrained language model performed better (accuracy 0.91 and balanced accuracy 0.85) than the bag-of-words method with top 2000 ingredients (accuracy 0.90 and balanced accuracy 0.84), given the less computing resources needed (768-dimension compared with 2000-dimension) (Supplementary Table 4). Of note, as the food label collection methods and the set food products included were different across the years, the list of the top ingredients also varied. A similar issue exists in many other countries’ food composition databases. Therefore, the food label pretrained language model would be more stable when applied to different sources of databases and when new data is collected.

Comparison of the pretrained language model, ingredients bag-of-words, and structured nutrition facts models for predicting nutrition quality score

The average FSANZ nutrition quality score of the FLIP2017 and FLIP2020 databases that we manually calculated and used for the nutrition quality score prediction task were 7.1 and 6.8, respectively. The performance of FSANZ score prediction using food label pretrained language model representations and ingredients bag-of-words

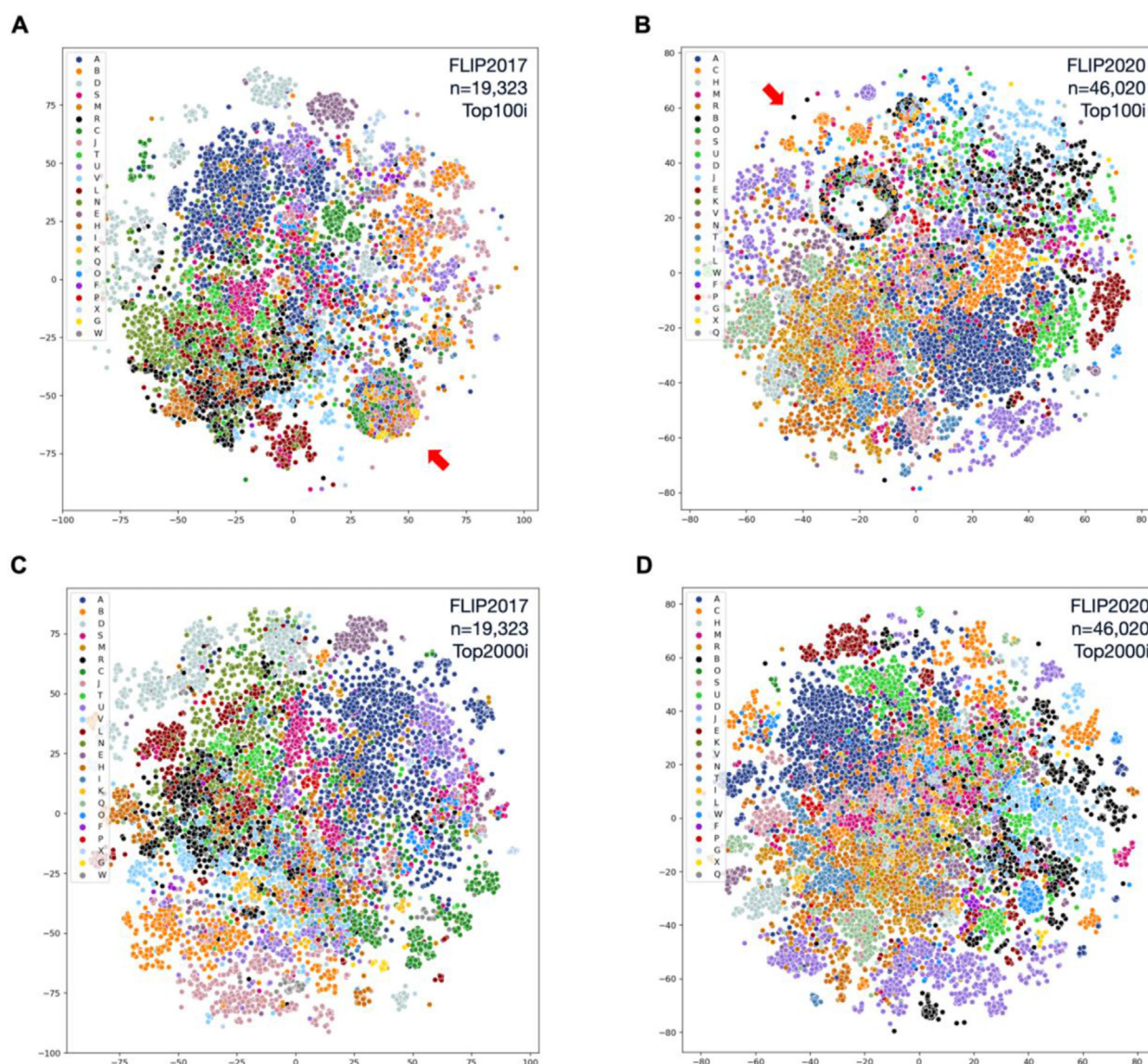


Figure 4. t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of food products in each TRA category represented by bag-of-words. (A) FLIP2017 top 100 ingredients bag-of-words ($n = 19,323$). (B) FLIP2020 top 100 ingredients bag-of-words ($n = 46,020$). (C) FLIP2017 top 2000 ingredients bag-of-words ($n = 19,323$). (D) FLIP2020 top 2000 ingredients bag-of-words ($n = 46,020$). t-SNE, t-distributed stochastic neighbor embedding. TRA, Health Canada's Table of Reference Amounts (9). TRA major food categories: A, bakery products; B, beverages; C, cereals and other grain products; D, dairy products and substitutes; E, desserts; F, dessert toppings and fillings; G, eggs and egg substitutes; H, fats and oils; I, marine and freshwater animals; J, fruit and fruit juices; K, legumes; L, meat, poultry, and substitutes; M, miscellaneous category; N, combination dishes; O, nuts and seeds; P, potatoes, sweet potatoes and yams; Q, salads; R, sauces, dips, gravies, and condiments; S, snacks; T, soups; U, sugars and sweets; V, vegetables; W, foods intended solely for children aged <4 y; X, meal replacements and nutritional supplements.

inputs developed in the previous tasks were calculated based on the XGBoost regression algorithm because it performed better than the elastic net and KNN algorithm. Structured data from the Nutrition Facts table on the food product was also used as an input for comparison because the FSANZ nutrition quality score is calculated by steps based on food product's nutrient levels per 100 units (per 100 g for solid food and per 100 mL for liquid food). As shown in Table 3, food label pretrained language model representations of the food product name and ingredients reached a similar level of model prediction accuracy (R^2 : 0.87; MSE: 14.4) compared with the ingredients bag-of-words method (top 100 ingredients: R^2 : 0.72 and MSE: 30.3; top 2000 ingredients: R^2 : 0.84 and MSE: 17.6). However, models using structured data from the food nutrition facts table

performed much better than the food label pretrained language model and bag-of-words method. The model prediction accuracy reached $R^2 = 0.98$ and MSE = 2.5 using inputs of nutrients per 100 units serving size.

We further applied the FSANZ score prediction algorithm developed from FLIP2020 to our FLIP2017 database for model generalization evaluation. Results showed that using food label pretrained language model representations (R^2 : 0.70; MSE: 31.8) performed similarly to using the bag-of-words method (R^2 : 0.62–0.72; MSE: 40.4–30.0) (Supplementary Table 4). However, the structured nutrition facts model using nutrients per 100 units had the best generalization ability (R^2 : 0.97 and MSE: 2.8). Therefore, the structured nutrition facts model using nutrients per 100 units is the best option for application to

Table 2
Accuracy and balanced accuracy of TRA major category and subcategory prediction algorithms using top ingredients bag-of-words method^{1, 2}

Bag-of-words	Algorithm	TRA major category		TRA subcategory	
		Accuracy	Balanced accuracy	Accuracy	Balanced accuracy
Top 100 ingredients	Elastic net	0.69	0.59	0.64	0.50
Top 100 ingredients	KNN	0.81	0.74	0.75	0.64
Top 100 ingredients	XGBoost	0.85	0.77	0.81	0.71
Top 500 ingredients	Elastic net	0.87	0.82	0.82	0.71
Top 500 ingredients	KNN	0.91	0.87	0.85	0.75
Top 500 ingredients	XGBoost	0.93	0.90	0.90	0.80
Top 1000 ingredients	Elastic net	0.91	0.88	0.87	0.77
Top 1000 ingredients	KNN	0.92	0.89	0.87	0.77
Top 1000 ingredients	XGBoost	0.95	0.93	0.93	0.83
Top 2000 ingredients	Elastic net	0.93	0.91	0.90	0.79
Top 2000 ingredients	KNN	0.93	0.90	0.88	0.79
Top 2000 ingredients	XGBoost	0.96	0.95	0.94	0.86

¹ KNN, k-nearest neighbors. XGBoost, extreme gradient boosting. TRA, Health Canada’s Table of Reference Amounts (24 major categories and 172 subcategories) (9).

² Accuracy compared with manually categorized and validated categories/subcategories (FLIP2020; $n = 46,020$).

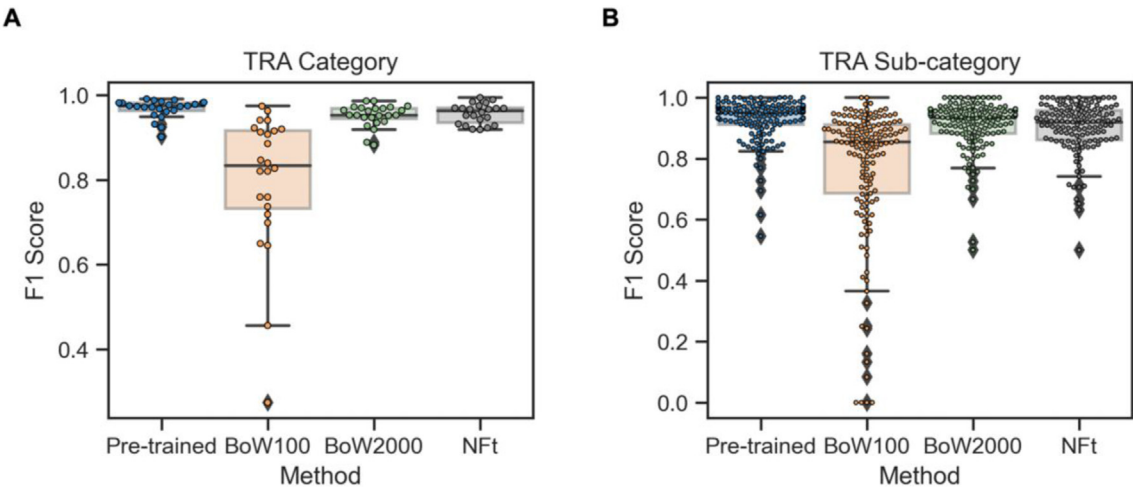


Figure 5. Distribution of food categorization prediction F1 score using different methods. (A) Prediction F1 score of each Table of Reference Amounts (TRA) major category. (B) Prediction F1 score of each TRA subcategory. University of Toronto Food Label Information and Price Database 2020 ($n = 46,020$). Health Canada’s TRA (24 major categories and 172 subcategories) (9). Pretrained: food label pretrained language model representations of product name and ingredients. BoW100: bag-of-words occurrence using top 100 ingredients. BoW2000: bag-of-words occurrence using top 2000 ingredients. Nutrition Facts table: structured nutrition facts model using nutrients per 100 units calculated from Nutrition Facts table.

other food composition databases for FSANZ nutrition quality score prediction.

Discussion

In this study, we used the pretrained language model in NLP to process unstructured text information that appears on food labels and used it with supervised machine learning algorithms to automate food category classification and nutrition quality score calculation. The automation strategy in this paper reached >97% accuracy and enabled fast and large-scale food label processing at a low cost for better evaluation of the impact of food policies on the food environment. To our knowledge, this is the first study that leveraged the state-of-the-art pretrained language model for food label text data and is the first that applied NLP and machine learning methods for nutrient profiling in a large-scale food composition database.

Food composition databases are expensive to develop and maintain, but are essential for a wide variety of applications, such as in epidemiological studies to investigate the associations between nutrient intakes and health outcomes; in clinical settings to plan diets for individuals; for setting reformulation standards; and for monitoring food nutrition composition changes over time [35–38]. Previous iterations of the Canadian FLIP food composition database relied on the manual categorization of foods and complex nutrition quality score calculations by trained nutrition researchers, which is a time and labor-intensive process [4]. This study provides evidence for effective and generalizable automated processes that allow for a more regular assessment and monitoring of the dynamic food environment, particularly when governments are enacting regulations, such as front-of-pack labeling policies which will require monitoring of the nutritional quality of the food supply [39, 40].

Our results show that the pretrained language model using food labels performed fast, high-accuracy, food category classifications, compared with the traditional bag-of-words method relying on most

Table 3

Comparison of the performance of FSANZ nutrition quality score predictions using food label pretrained language model, bag-of-words, and structured data model^{1, 2}

Methods	Input from food label	R^2	MSE
Pretrained language model	Ingredients	0.86	15.2
Pretrained language model	Name & ingredients	0.86	15.1
Pretrained language model	Name & brand & ingredients	0.87	14.4
Bag-of-words	Top 100 ingredients	0.72	30.3
Bag-of-words	Top 500 ingredients	0.82	19.4
Bag-of-words	Top 1000 ingredients	0.82	19.7
Bag-of-words	Top 2000 ingredients	0.84	17.6
Structured data model	Nutrient facts table	0.91	9.8
Structured data model	Nutrients per 100 units	0.98	2.5

¹ FSANZ, Food Standards Australia New Zealand [11]. R^2 , the coefficient of determination. Results from extreme gradient boosting (XGBoost). Nutrients per 100 units, g for solid food and mL for liquid food.

² Accuracy compared with manually calculated FSANZ nutrition quality scores (FLIP2020; $n = 46,020$).

frequent ingredients as inputs. Although similar accuracies were achieved using the bag-of-word model of top 2000 ingredients, this method is less generalizable because the included product types and top ingredient lists vary by different global food composition datasets and year. Results in this study also showed that 80% of TRA subcategory predictions using the pretrained language model had an F1 score of >0.9, which outperformed the traditional bag-of-words method (32%–70%) and structured nutrition facts model (62%). The remaining 15 subcategories for which F1 scores ranged from 0.55 to 0.84 may be due to the insufficient number of products for training (<350 products). As such, the accuracy of the algorithm is highly dependent upon having a large enough sample size in each subcategory to allow for the accurate classification of products.

The pretrained language model and machine learning approaches can be applied to assessing the recently released regulations on front-of-pack labeling and policy for restrictions on the marketing of unhealthy foods to children in Canada [39, 41] and assessing policy impacts in many other countries (e.g., Australia, Chile, Costa Rica, India, United Kingdom) [42]. For example in Chile, there was a significant decrease in the amount of sugars and sodium in several groups of packaged foods and beverages after initial implementation of the *Chilean Law of Food Labeling and Advertising* [40]. Although the machine learning model using structured nutrient facts data as input achieved the highest accuracy of 0.98, the food label pretrained language model representations and/or bag-of-words model can still achieve moderate accuracy in the prediction of nutrition quality score (0.87 and 0.84, respectively), even if the nutrient information is missing. The slight discrepancy in the true and predicted FSANZ score is likely due to identifying specific categories used in FSANZ and the methodology of scoring for FVNL, which are not disclosed quantitatively in Canada and therefore are based on estimates using the descending order of ingredients found in the ingredient list of a product [11, 14]. Improvements refining the data input (e.g., training the technique to learn the percentage contribution of an ingredient) can likely increase the accuracy of estimating FSANZ nutrition quality score on future datasets.

Compared with manual categorization and calculation as well as traditional machine learning approaches, our automation strategies have several advantages. First, a variety of text information (e.g., product name, brand, ingredients, nutrition and health claims,

environment/sustainability claims) is available on food packages but not fully utilized by previous studies [17–19]. A previous machine learning model predicted label nutrients from the ingredients statement in the USDA food composition database but was unable to handle texts outside the current dictionary [17]. The machine learning models for added sugar and fiber contents prediction used available nutrients of packaged food products and reached high accuracy, but were unable to handle products that were missing nutrient information [18, 19]. Our study implemented the pretrained language model to directly encode different texts found on food labels into low-dimensional vectors in a standardized way to bypass the generalization challenges while saving computing resources. Second, the models applied in this study did not need as many computing resources, yet remained highly accurate; therefore, it is cost effective for a more regular assessment and monitoring of the highly dynamic food environment as well as helpful for evaluating food policy and regulations [4, 43–45]. Lastly, this study leveraged previous iterations of FLIP, in which category and nutrition profiling of a large number of food products (>70,000) have been manually determined by trained nutrition researchers, thereby improving the usability of data variables within a model. Given the increased generalizability and the standardized process for handling text data, our automation strategies could be highly adaptable and applicable to other large food and nutrition databases worldwide with different data formats and different food classification and nutrient profiling systems. For example, these features are being implemented for FLIP-Latin America and the Caribbean dataset and could also be implemented for other countries' nutrition databases [46].

The present study is limited by the amount of product-specific information retrieved from grocery retailer websites and the imbalanced sample size. For example, currently, there are no e-grocery food labeling regulations in Canada, as well as globally, to standardize the provision of food product information online, although the Codex Alimentarius Commission and the Canadian government are both currently working on draft guidelines on internet sales/e-commerce [47, 48]. Furthermore, because of the large number of TRA categories, some subcategories have a limited number of products available in the database for model training, which may impact the performance of our algorithm in small food categories. However, despite missing product information and imbalanced sample size, the results of this study indicate a high predictive value using NLP and machine learning techniques. Lastly, although the pretrained NLP model performed well in the task of food categorization and nutrition quality score prediction, it was based on general natural language corpora instead of food-specific text. Further model training to leverage food-specific corpora is needed for food label and recipe-related tasks.

In conclusion, the application of a novel NLP method and machine learning in processing text information on food labels reduces the time needed for manual food categorization and nutrition quality score calculation of a large number of food products. This effective and generalizable automated process is feasible and reliable for food composition databases in other countries. The food category classification and nutrition quality score prediction tasks in this study are just examples, demonstrating that NLP and machine learning are promising methods for the automation of food databases in the future. This study provides evidence of accurate and large-scale, real-time food label processing algorithms that can provide timely analysis of the impacts of food policy on the food supply for governments, health organizations, and researchers and ultimately promote a healthy food environment.

This research was supported by the Canadian Institutes of Health Research (CIHR) grants (PJT-165858 and PJT-152979). GH and ML designed research; GH conducted research and analyzed data; GH and MA wrote the manuscript. ML had primary responsibility for the final content. All authors read and approved the final manuscript. GH received a postdoctoral fellow training award from the CIHR National Healthy Cities SMART Training Program. The funders were not part of the design, implementation, analysis and interpretation of the data.

Data Availability

Data described in the manuscript, code book, and analytic code will be made available upon request, pending application and approval.

Funding

This research was supported by the Canadian Institutes of Health Research (CIHR) grants (PJT-165858 and PJT-152979). The funders were not part of the design, implementation, analysis or interpretation of the data.

Conflicts of Interest

The authors report no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ajcnut.2022.11.022>.

References

- [1] A. Afshin, P.J. Sur, G. Ferrara, J.S. Salama, E.C. Mullany, K.H. Abate, et al., Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 393 (2019) 1958–1972.
- [2] H. Greenfield, D.A. Southgate, Food & Agriculture Org, [Internet]. Food composition data: production, management, and use (2003). <https://www.fao.org/documents/card/en/c/7a6877a2-86b5-588c-a55c-c6a1e63f19c9>.
- [3] J.M. Poti, E. Yoon, B. Hollingsworth, J. Ostrowski, J. Wandell, D.R. Miles, et al., Development of a food composition database to monitor changes in packaged foods and beverages, *J Food Compos Anal* 64 (2017) 18–26.
- [4] M. Ahmed, A. Schermel, J. Lee, M. Weippert, B. Franco-Arellano, M. L'Abbé, Development of the food label information program: a comprehensive Canadian branded food composition database, *Front Nutr* 8 (2022), 825050.
- [5] Government of Canada. The Canadian Nutrient File [Internet]. Available from: <https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/nutrient-data/canadian-nutrient-file-about-us.html>.
- [6] T. Poon, M.É. Labonté, C. Mulligan, M. Ahmed, K.M. Dickinson, M.R. L'Abbé, Comparison of nutrient profiling models for assessing the nutritional quality of foods: a validation study, *Br J Nutr* 120 (2018) 567–582.
- [7] M.É. Labonté, T. Poon, B. Gladanac, M. Ahmed, B. Franco-Arellano, M. Rayner, et al., Nutrient profile models with applications in government-led nutrition policies aimed at health promotion and noncommunicable disease prevention: a systematic review, *Adv Nutr* 9 (2018) 741–788.
- [8] World Health Organization. Nutrient profiling: report of a technical meeting [Internet]. Available from: https://apps.who.int/nutrition/publications/profiling/WHO_IASO_report2010/en/index.html.
- [9] Health Canada. Nutrition labelling—Table of reference amounts for food [Internet]. Available from: <https://www.canada.ca/en/health-canada/services/technical-documents-labelling-requirements/table-reference-amounts-food/nutrition-labelling.html>.
- [10] C.A. Monteiro, G. Cannon, J.C. Moubarac, R.B. Levy, M.L.C. Louzada, P.C. Jaime, The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing, *Public Health Nutr* 21 (2018) 5–17.
- [11] Australian Government. Australia New Zealand Food Standards Code-Standard 1.2.7-Nutrition, Health and Related Claims-F2014C01191[Internet]. Available from: <https://www.foodstandards.gov.au/industry/labelling/Pages/Consumer-guide-to-NPSC.aspx>.
- [12] V. Azais-Braesco, C. Goffi, E. Labouze, Nutrient profiling: comparison and critical analysis of existing systems, *Public Health Nutr* 9 (2006) 613–622.
- [13] Pan American Health Organization. Pan American health organization nutrient profile model [Internet]. Available from: https://iris.paho.org/bitstream/handle/10665.2/18621/9789275118733_eng.pdf.
- [14] L. Vergeer, M. Ahmed, B. Franco-Arellano, C. Mulligan, K. Dickinson, J.T. Bernstein, et al., Methodology for the determination of fruit, vegetable, nut and legume points for food supplies without quantitative ingredient declarations and its application to a large Canadian packaged food and beverage database, *Foods* 9 (2020) 1127.
- [15] R.A. Harrington, V. Adhikari, M. Rayner, P. Scarborough, Nutrient composition databases in the age of big data: foodDB, a comprehensive, real-time database infrastructure, *BMJ Open* 9 (2019), e026652.
- [16] Y. Zhang, R. Jin, Z.H. Zhou, Understanding bag-of-words model: a statistical framework, *Int J Mach Learn Cybern* 1 (2010) 43–52.
- [17] P. Ma, A. Li, N. Yu, Y. Li, R. Bahadur, Q. Wang, et al., Application of machine learning for estimating label nutrients using USDA Global Branded Food Products Database (BFPD), *J Food Compos Anal* 100 (2021), 103857.
- [18] T. Davies, J.C.Y. Louie, R. Ndanuko, S. Barbieri, O. Perez-Concha, J.H.Y. Wu, A machine learning approach to predict the added-sugar content of packaged foods, *J Nutr* 152 (2022) 343–349.
- [19] T. Davies, J.C.Y. Louie, T. Scapin, S. Pettigrew, J.H. Wu, M. Marklund, et al., An innovative machine learning approach to predict the dietary fiber content of packaged foods, *Nutrients* 13 (2021) 3195.
- [20] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805 2018.
- [21] S. Wu, Y. He, Enriching pretrained language model with entity information for relation classification, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, ACM Digital Library, 2019, pp. 2361–2364 [Internet].
- [22] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, et al., Recipe1 M+: a dataset for learning cross-modal embeddings for cooking recipes and food images, *IEEE Trans Pattern Anal Mach Intell* 43 (2019) 187–203.
- [23] USDA Foreign Agricultural Service. Canada: Retail Foods 2021 [Internet]. Available from: https://apps.fas.usda.gov/newgainapi/api/Report/DownloadReportByFileName?fileName=Retail%20Foods_Ottawa_Canada_06-30-2021.pdf.
- [24] W.L. Watson, A. Johnston, C. Hughes, K. Chapman, Determining the 'healthiness' of foods marketed to children on television using the Food Standards Australia New Zealand nutrient profiling criteria, *Nutr Diet* 71 (2014) 178–183.
- [25] A. Kaur, P. Scarborough, S. Hieke, A. Kusar, I. Pravst, M. Raats, et al., The nutritional quality of foods carrying health-related claims in Germany, the Netherlands, Spain, Slovenia and the United Kingdom, *Eur J Clin Nutr* 70 (2016) 1388–1395.
- [26] S.C. Rosentretter, H. Eyles, C. Ni Mhurchu, Traffic lights and health claims: a comparative analysis of the nutrient profile of packaged foods available for sale in New Zealand supermarkets, *Aust N Z J Public Health* 37 (2013) 278–283.
- [27] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, et al., Aligning books and movies: towards story-like visual explanations by watching movies and reading books, Proceedings of the IEEE international conference on computer vision (2015) 19–27, <https://doi.org/10.1109/ICCV.2015.11>. Available from: .
- [28] N. Reimers, I. Gurevych, Sentence-BERT: sentence embeddings using siamese BERT-networks, arXiv preprint arXiv:1908.10084 (2019).
- [29] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J Mach Learn Res* 9 (2008).
- [30] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J Royal Statistical Soc B* 67 (2005) 301–320.
- [31] G. Bonaccorso, Machine learning algorithms, Packt Publishing Ltd, livery place, 35 livery street, Brimingham, B3 2PB, UK, 2017.
- [32] A. Moldagulova, R.B. Sulaiman, Using KNN algorithm for classification of textual documents. 2017 8th international conference on information technology (ICIT), IEEE: Using KNN algorithm for classification of textual documents (2017) 665–671, <https://doi.org/10.1109/ICITECH.2017.8079924>. Available from: .
- [33] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, XGBoost: A Scalable Tree Boosting System, San Francisco, California, USA, 2016, pp. 785–794.

- [34] M. Grandini, E. Bagli, G. Visani, Metrics for multiclass classification: an overview, 2020 arXiv preprint arXiv:200805756.
- [35] S.M. Church, The history of food composition databases, *Nutr Bull* 31 (2006) 15–20.
- [36] S.F. Schakel, I.M. Buzzard, S.E. Gebhardt, Procedures for estimating nutrient values for food composition databases, *J Food Compos Anal* 10 (1997) 102–114.
- [37] A. Delgado, M. Issaoui, M.C. Vieira, I. Saraiva de Carvalho, A. Fardet, Food composition databases: does it matter to human health? *Nutrients* 13 (2021) 2816.
- [38] J.A. Pennington, P.J. Stumbo, S.P. Murphy, S.W. McNutt, A.L. Eldridge, B.J. McCabe-Sellers, et al., Food composition data: the foundation of dietetic practice and research, *J Am Diet Assoc* 107 (2007) 2105–2113.
- [39] Government of Canada. Forward regulatory plan 2022-2024: Regulations amending certain regulations made under the food and drugs act [Internet]. Available from: <https://www.canada.ca/en/health-canada/corporate/about-health-canada/legislation-guidelines/acts-regulations/forward-regulatory-plan/pla n/use-foreign-decisions-pathway.html>.
- [40] M. Reyes, L. Smith Taillie, B. Popkin, R. Kanter, S. Vandevijvere, C. Corvalán, Changes in the amount of nutrient of packaged foods and beverages after the initial implementation of the Chilean Law of Food Labelling and Advertising: a nonexperimental prospective study, *PLoS Med* 17 (2020), e1003220.
- [41] A. Schermel, T.E. Emrich, J. Arcand, C.L. Wong, M.R. L'Abbé, Nutrition marketing on processed food packages in Canada: 2010 Food Label Information Program, *Appl Physiol Nutr Metab* 38 (2013) 666–672.
- [42] Codex Alimentarius Commission, Codex Committee on Food Labelling, CCFL, 2022.
- [43] B. Franco-Arellano, J.T. Bernstein, S. Norsen, A. Schermel, M.R. L'Abbé, Assessing nutrition and other claims on food labels: a repeated cross-sectional analysis of the Canadian food supply, *BMC Nutr* 3 (2017) 74.
- [44] T.E. Emrich, Y. Qi, J.E. Cohen, W.Y. Lou, M.L. L'Abbé, Front-of-pack symbols are not a reliable indicator of products with healthier nutrient profiles, *Appetite* 84 (2015) 148–153.
- [45] M.E. Labonté, T.E. Emrich, P. Scarborough, M. Rayner, M.R. L'Abbé, Traffic light labelling could prevent mortality from noncommunicable diseases in Canada: a scenario modelling study, *PLoS One* 14 (2019), e0226975.
- [46] M.R. L'Abbé, A. Schermel, B. Franco-Arellano, S.J. Vega, J. Arcand, L. Lam, et al., FLIP-LAC user guide, 2018.
- [47] Government of Canada, Consultation on the development of voluntary guidance for providing food information for foods sold to consumers through e-commerce, 2022-07-08.
- [48] Commission Codex Alimentarius, Proposed Draft Guidance on Internet Sales/ E-Commerce - CCFL 45 (2020).