Exploring Knowledge Domain Bias on a Prediction Task for Food and Nutrition Data

Gordana Ispirova[1,2]
[1]*Computer Systems Department*
Jožef Stefan Institute
Ljubljana, Slovenia
[2]*International Postgraduate School Jožef Stefan*
Ljubljana, Slovenia
gordana.ispirova@ijs.si

Tome Eftimov[1]
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

Barbara Koroušić Seljak[1]
[1]*Computer Systems*
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378159

Abstract—**Human understanding and knowledge about food and nutrition is constantly evolving, and has significantly improved recently, one of the main contributor to this is data. The possibilities of gaining knowledge from food and nutrition-related data are yet to be explored. One of the most important information about food is nutrient content, which is very relevant for patients suffering from various diseases, professional athletes, and slowly part of everyday life of many for health or fitness goals. In this paper, we explore the effect of domain bias in a predictive study in the food and nutrition domain. Having a machine learning pipeline for predicting nutrient values with learned vector representations from short text description of recipes, we introduce domain knowledge before the prediction algorithms are applied. On a large corpus of recipe data containing short description and nutrient values we introduce word and paragraph embeddings, learn concept representations for the textual descriptions, introduce domain knowledge for clustering the data, and apply machine learning algorithms for predicting the nutrient content of the recipes. We explore the impact of the domain knowledge by introducing two different criteria of clustering the dataset - using graph embedding of the FoodEx2 codes, and using the traffic light labelling system from the Food Standards Agency; at the end we compare the two different criteria. The results from this study imply that inferring domain knowledge beforehand is crucial for the task of predicting nutrient content.**

Index Terms—**representation learning, unsupervised machine learning, supervised machine learning, domain bias, recipe data, nutrient values**

I. Introduction

Obesity, diabetes, cancer, coronary artery disease, cardiovascular disease, lack of physical activity – are all ”modern day” diseases, i.e. lifestyle diseases. These make up a subset of non-communicable diseases (NCDs), which are a result of the way we – modern humans, live, work, and pretty much go about our everyday life. With the technological revolution our lifestyles have become sedentary, and our diets even more unhealthy [1]. Most of NCDs are linked to poor dietary habits, and the number one step towards making the average human diet healthier is basic understanding of nutrition – what is in our food. Dietary assessment is essential for patients suffering from many diseases (especially diet and nutrition related ones), much needed for professional athletes, and because of the accessibility of nutrient tracking through mobile applications [2], [3] it is becoming part of everyday habits of a vast majority of individuals, for health and fitness goals.

With the spike of morbidity and mortality due to NCDs not only in developed western countries but in middle and low-income countries as well, there is raised public health concern about some subcategories of macronutrients – saturated fats, and added or free sugar, and micronutrients -– sodium, (for individuals suffering from specific diseases like osteoporosis, stomach cancer, and kidney disease), and fiber (for patients suffering from irritable bowel syndrome – IBS). Nutrient content can vary a lot from one food to another, even though they have roughly the same type of ingredients, which complicates the nutrient tracking and calculating, and makes possibility of predicting nutrient content extremely not likely.

In [4], we proposed an approach, called P-NUT, for predicting macronutrient values of a food item considering learned vector representations of text describing the food item. P-NUT is a machine learning (ML) pipeline that consists of three parts: representation learning – learning vector representations from short text descriptions of recipes; unsupervised ML – introducing domain knowledge for obtaining separate clusters of data; and supervised ML – obtaining predictions for the macronutrient values of the recipes.

In this paper, we focus on sensitivity analysis of the unsupervised ML part, i.e. how introducing different types of domain knowledge affect the results from the supervised ML part. The two criteria that we used to explore the food and nutrition domain knowledge bias for predicting macronutrients are – the FoodEx2 classification system [5], and the FSA traffic light system [6]. This two systems have domain knowledge inferred in them, are constructed and developed with food and nutrition domain experts and give us different insights on how to classify and separate foods.

When it comes to macronutrient content, recipes and foods in general can be very unbalanced. In a dataset containing a large variety of foods, one macronutrient can have values from 0 grams to 100 grams per 100 grams of the certain food, for example the content of fat can go from ”fat free” foods to ”fat based” foods (ex. different kinds of nut butters), therefore, a general model for prediction will not be efficient in macronutrient prediction. For this reason, the domain knowledge incorporated in the unsupervised ML part of the methodology is very much needed for obtaining a good prediction model. In P-NUT [4] we introduced clustering as a method to separate foods in order to obtain clusters (groups) of foods with similar characteristics. Subsequently, on these separate clusters we predict the macronutrients with applying supervised ML. Predicting macronutrients is not a task that has been approached in such a manner before, usually nutrient content of food is calculated or estimated from measurements and exact ingredients [7]–[9], which is a multi-step procedure that involves: selection or development of an appropriate recipe, data collection for the nutrient content of the ingredients, correction of the ingredient nutrient levels for weight of edible portions, adjustment of the content of each ingredient for effects of preparation, summation of ingredient composition, final weight (or volume) adjustment, and determination of the yield and final volumes. These calculations can be done only when all the ingredients and measurements are available, when this data is not available, this procedure gets more complicated [7], [8].

The rest of the paper is structured as follows: in Section II first we point out the relevant related work for the topic in matter (Subsection II-A), then the details for the methodology used are given (Subsection II-B), and finally we describe the data used in our experiments (Subsection II-C). In Section III the experimental results, evaluations and further discussion are presented. At the end, in Section IV a summarization of the importance of this methodology and directions for future work are presented.

II. Methods

This section begins with a brief review of the work done involving ML in the direction of predicting nutrient content, continues with an explanation of our recently proposed methodology P-NUT [4], after what a description of the data used in the experiments follows, and how P-NUT has been extended and modified in order to explore the domain bias.

A. Related work

To the best of our knowledge, P-NUT [4] is the first methodology for predicting nutritional content of foods/recipes using only short text description. There has been some work involving ML done in this direction, mainly involving image recognition: employing different deep learning models for accurate food identification and classification from food images [10], dietary assessment through food image analysis [11], calculating calorie intake from food images [12], [13]. All this work is in the direction of predicting total calories, and strongly relies on textual data retrieved from the Web. There are numerous mobile and web applications, for tracking macronutrient intake [2], [3]. Systems like these are used for achieving dietary goals, allergy management or simply, maintaining a healthy balanced diet. The biggest downside is the fact that they require manual imputation of details about the meal/food.

1) P-NUT:

1. Representation learning The first part of the P-NUT methodology is representation learning. Representation learning is learning representations of input data by transforming it or extracting features from it, which then makes it easier to perform a task like classification or prediction [14]. There are two different categories of vector representations: non-distributed or sparse, and distributed or dense, which are currently state-of-the-art. In P-NUT we learn dense vector representations for recipe data, i.e. vector representations for textual data.
   a) Word embeddings – are vector space models (VSM), that in a low-dimensional semantic space (much smaller than the vocabulary size) represent words in a form of real-valued vectors. These distributed representations of words improve the performance of learning algorithms for various natural language processing (NLP) tasks [15]–[20]. Most well-known word embedding algorithm is Word2Vec, introduced by Mikolov et al. in 2013 at Google [21]. It is a neural network based approach with two different architectures: Continuous Bag of Words and Continuous Skip Gram [22]. GloVe – is another word embedding method [23], it is based on co-occurrence statistics from a given corpus.
   b) Paragraph embeddings – Doc2Vec is an unsupervised paragraph embedding method [24] based on Word2Vec. There are two architectures of the Doc2Vec method: Distributed Memory version of Paragraph Vector (PV-DM), and Distributed Bag of Words version of Paragraph Vector (PV-DBOW).
2. Unsupervised machine learning – Foods, being biological materials, can exhibit large variations in their composition, and therefore in their nutrient content (which leads to unbalanced macronutrient content). Any large food dataset, has presumably a broad variety of foods, which implies that the content of a macronutrient can go from one extreme to another. It goes without saying that in order to build better prediction models for macronutrient predictions, the instances of a food dataset should be grouped by some criteria. In order for the grouping to be effective, expert insight from the food and nutrition domain is needed, i.e. domain knowledge. In P-NUT the dataset used in the experiments had FoodEx2 [5] codes available, which already contains a great degree of domain knowledge. These codes were used in the second part of the P-NUT methodology – the unsupervised ML part, in which independently of the representation learning process, we used the method presented in [25], where the FoodEx2 hierarchy is presented as Poincaré graph embeddings [26]. The domain knowledge contained in the FoodEx2 hierarchy is transcended through the graph embeddings, which

later the authors use in order to group the food items from the FoodEx2 system in clusters. The clustering is done using the Partition Around Medoids algorithm [27], and the number of clusters is determined using the silhouette method [28]. This clustering process is performed on the bottom end of the hierarchy, i.e. on the leaves of the graph, and at the end the FoodEx2 codes are clustered in 230 clusters. The dataset used in the experiments of P-NUT was rather small compared to the total number of FoodEx2 codes in the hierarchy, and the fact that when assigned a cluster number some of the clusters contained very few or no elements at all, a post-hoc cluster merging process is conducted. In this merging process the clusters are merged following a bottom up approach – based on their top-level parents, going level deeper until the number of instances per each cluster is as evenly distributed as possible. In P-NUT the dataset contained 3265 food items, which based on their FoodEx2 codes were clustered in 9 clusters.

3. Supervised machine learning
   The third part of P-NUT is the supervised ML part. This part includes training separate predictive models for the macronutrients available (or of choice) – in the case of the dataset in P-NUT: carbohydrates, fat, protein and water. The algorithms of choice are single-target regressions, where the inputs are the learned vector representations of the short text descriptions of the foods, clustered based on their FoodEx2 codes. In a real-time scenario, it is somewhat hard to select the right ML algorithm for the purpose. The overall most accepted approach is to select few algorithms, select ranges for the hyper-parameters for each algorithm, perform hyper-parameter tuning, and evaluate the estimators' performances with cross-validation by the same data in each iteration, benchmark the algorithms and select the best one(s). When working with regression algorithms, the most common baseline is using mean or median (central tendency measures) of the train part of the dataset for all the predictions.

The evaluation of P-NUT in [4] was done on each cluster separately for the predictions of the four macronutrients available – carbohydrate, fat, protein, and water. The results of the predictions were compared to baseline mean and median (mean and median values for each macronutrient in each cluster). The highest accuracy was obtained for carbohydrate predictions – 86%, compared to the baseline – 27% and 36%. The protein predictions yielded the best results across all clusters, 53%–77% of the values fall in the tolerance-level range.

In the evaluation of P-NUT, for comparison reasons, the predictions were also conducted without clustering the dataset based on the FoodEx2 classification system. In this case, again, the embedding algorithms give better results than the baseline mean and median (in this case of the whole dataset), for each target macronutrient. The best results, again, were obtained for the prediction of protein content (62%-64%).

2) StandFood: It is evident from the results in [4] that adding the unsupervised ML part, i.e. implementing the clustering based on the graph embeddings of the FoodEx2 hierarchy, significantly improved the results. Essential for that are, of course, the FoodEx2 codes. A big draw-down of the FoodEx2 classification system is that it is a manual, time – consuming task. StandFood [29] is a method that standardizes foods according to FoodEx2 classification system, and it consists of two parts. The first part is the classification par – the system identifies the type of food being analyzed (raw – r, derivate – d, simple composite – s, or aggregated composite – c) using a ML approach combined with post-processing rules; the second part describes the food using NLP combined with probability theory. The StandFood [29] method, by being a semi – automatic approach has a big advantage over the traditional manual approach of assigning FoodEx2 codes to food items. A tool like StandFood is very much needed for implementing P-NUT when the FoodEx2 codes are not available.

B. Methodology

To explore how including different types of domain knowledge affects the macronutrient prediction task, we propose an extension to the P-NUT methodology, presented in Figure 1.

1) Domain knowledge criteria:

1. FoodEx2 classification system – FoodEx2 is the second version of FoodEx [5], which is a standardized system for food classification and description developed by the European Food Safety Authority (EFSA), it has domain knowledge embedded in it and it contains descriptions of a vast set of individual food items combined in food groups and more broad food categories in a hierarchy that exhibits parent-child relationship. These FoodEx2 codes already contain domain knowledge, and based on them food items are grouped in food groups and broader food categories in the FoodEx2 hierarchy.

2. FSA traffic light system – The Food Standards Agency (FSA) introduced the traffic light system [6], [30] in order to determine how healthy a recipe/food product is. The FSA traffic light system gives independent expert scientific dietary advice in order to help individuals make healthier choices quickly and easily. The main goal of the FSA traffic light system is for the food industry to incorporate this traffic light system and implement it on the food labels, which would quickly give consumers indications about the nutritional content of the food product, and therefore make a more informed decision about their diet.
   There are three colors in this system – green, orange (amber), and red, which follow a hierarchy of 'healthiness'. These colors show at a glance if the food has low, medium or high amounts of fat, saturated fat, sugars and salt.

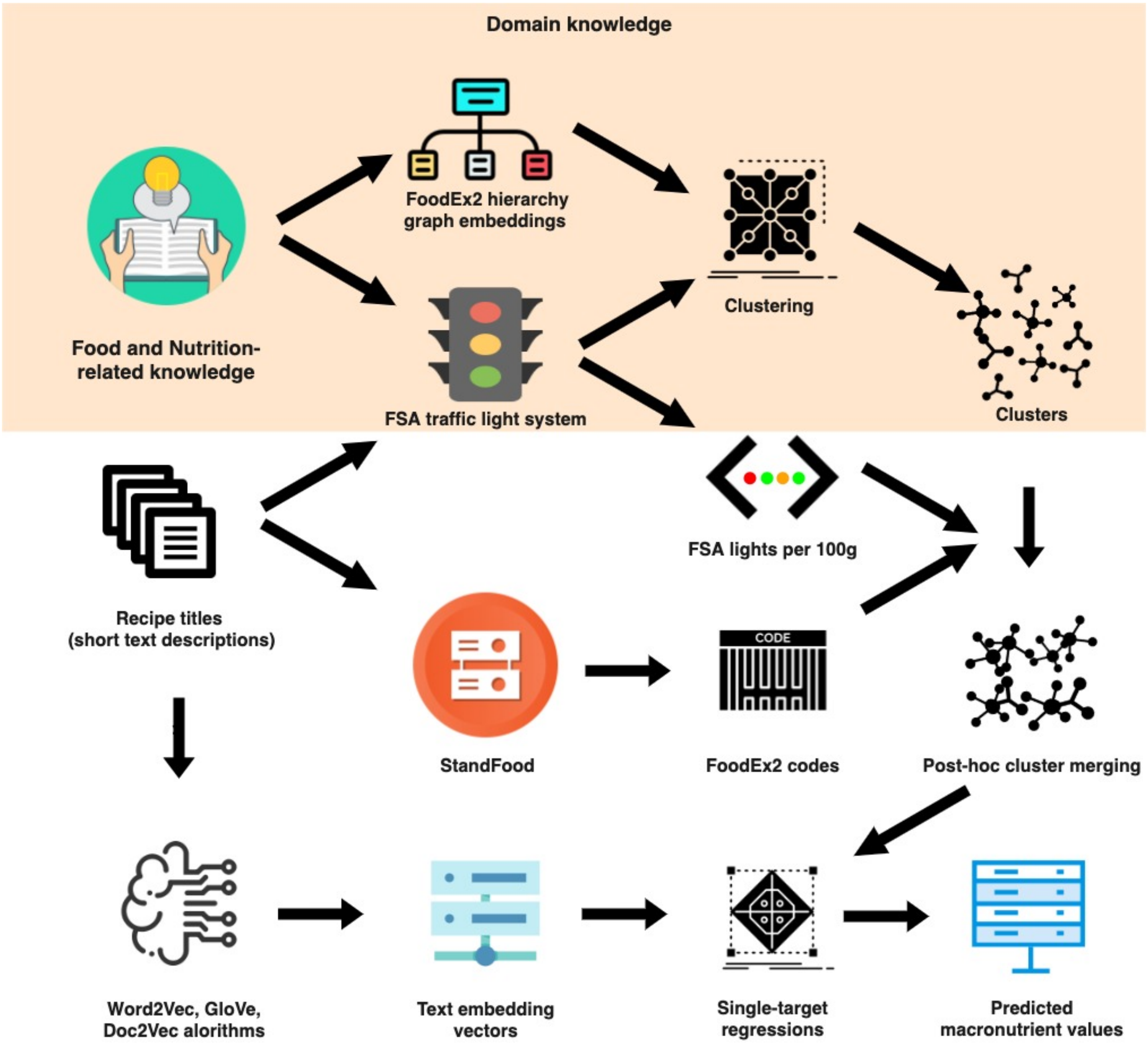
Domain knowledge
FoodEx2 hierarchy graph embeddings
Clustering
Clusters
Food and Nutrition-related knowledge
FSA traffic light system
Recipe titles (short text descriptions)
FSA lights per 100g
StandFood
CODE
FoodEx2 codes
Post-hoc cluster merging
Word2Vec, GloVe, Doc2Vec alorithms
Text embedding vectors
Single-target regressions
Predicted macronutrient values

Fig. 1. Inferring domain knowledge in P-NUT.

2) Tolerance for nutrient values: P-NUT is a ML pipeline with predicting macronutrient values as its main task. Macronutrient values which are expressed in grams, can have tolerances defined by international legalizations and regulations. The European Commission Health and Consumers Directorate General in 2012 published a guidance document [31], with the aim to provide advised recommendations for calculation of the acceptable differences between quantities of nutrients on the label declarations of food products and the ones established in Regulation EU 1169/2011 [32]. These tolerances for the food product labels are important as it is impossible for foods to contain the exact levels of nutrients that are presented on the labels, as a consequence of the natural variations of foods, as well as the variations occurring during production and the storage process. However, the nutrient content of foods should not deviate substantially from labelled values to the extent that such deviations could lead to consumers being misled. From the tolerance levels stated in [31], for our particular case we used the tolerance levels for the nutrition declaration of foods that do not include food supplements, out of which we used the needed information presented in Table I – where the allowed deviations are presented for

TABLE I
TOLERATED DIFFERENCES IN NUTRITION CONTENT IN FOODS BESIDES FOOD SUPPLEMENTS.

Nutrient	Quantity per 100 g	Tolerances (allowed deviations in quantity)
Salt	$<$ 1.25 g per 100 g	±0.375 g
	$\geq$ 1.25 g per 100 g	±20%
Saturates	$<$ 4 g per 100 g	±0.8 g
	$\geq$ 4 g per 100 g	±20%
Fat	$<$ 10 g per 100 g	±1.5 g
	10-40 g per 100 g	±20%
	$>$ 40 g per 100 g	±8 g
Protein, Sugars	$<$ 10 g per 100 g	±2 g
	10-40 g per 100 g	±20%
	$>$ 40 g per 100 g	±8 g

each of the nutrients we are predicting, depending on their quantity in 100 grams of the food in matter. These tolerance levels are included at the very final step in our methodology in the determination on how accurate the predicted nutrient values are.

C. Data

Food- and nutrition-related data comes in many different forms and formats: data about the nutritional composition of food (food composition data); data about actual and recommended food consumption on local, national and even international level (food consumption data); recipe data; and many other kinds of data. When we talk about recipe data we usually mean text descriptions of composite foods, which are readily available on user-based recipe websites. Recipe data can contain: recipe description - which is usually just a short textual description of the recipeIn this dataset for each food item we had available: its name in Slovene and English, its FoodEx2 code, and its nutrient values for: energy, water, fat, carbohydrates, and proteins, recipe institutions - which is a longer text with description of the cooking method and preparation instructions, and list of ingredients - the recipe's ingredients alongside their quantities. This data is unstructured textual data, and can have several variations for conveying the same information, due to the differences in way of expression.

The dataset used for evaluating P-NUT contained nutritional information about food items recently collected in Slovenia for the aims of the EFSA EU Menu project [5], complied of data about simple food products, and recipe dishes – name in Slovene and English, FoodEx2 code, and nutrient values for energy, water, fat, carbohydrates, and proteins.

For this study, we use the dataset Recipe1M, available online [33], which is a large-scale, structured corpus of over one million cooking recipes and 13 million food images.

However, from those one million recipes in our focus are the ones that contain details like macronutrients – a total of 51235 recipes. The following information is available for each of the 51235 recipes:

1) Recipe title – a short textual description of the recipe;
2) Ingredients – list of ingredients;
3) Recipe instruction – long textual description of step by step instructions for preparing the recipe;
4) Nutrient content of ingredients – quantity in grams of fat, protein, saturates, sodium, and sugar per 100 grams of the ingredient for each ingredient;
5) Quantity of each ingredient;
6) Units of measurement per each ingredient – according to the household measurement system (cup, tablespoon, teaspoon, etc.);
7) Weight in grams per each ingredient;
8) Nutrient content – quantity in grams of fat, protein, salt, saturates, and sugars per 100 grams of the recipe;
9) FSA traffic light labels per 100 grams – for each of the four macronutrients (fat, salt, saturates and sugars) one of the three labels from the FSA traffic light system is assigned:
    - red – high content;
    - orange – medium content;
    - green – low content.

In Table II a few example data instances from the dataset are given.

TABLE II
DATASET STRUCTURE AND EXAMPLE DATA INSTANCES

Recipe title	FSA lights per100g			
	fat	salt	saturates	sugars
Salt Free, Low Cholesterol Sugar Cookies Recipe	red	orange	orange	orange
Pacific Wasabi Sauce for Grilled Tuna	red	orange	red	green
Barbeque Brisket Rub	orange	red	green	green
Frozen Banana Smoothie	green	green	green	red

III. Results and discussion

This section explains the evaluation process of the presented methodology – the experimental setup and the results obtained, as well as a discussion about the outcome of the experiments.

A. Pre-processing

The Recipe1M dataset contains detailed information about each recipe. The data of interest here are: the recipe titles, the FSA traffic light labels for fat, salt, saturates, and sugars, and the quantities per 100 grams of the recipe of the five nutrients for prediction.

First, the text descriptions are tokenized, beforehand the punctuation signs and numbers that represent quantities are removed, whereas the percentage values (e.g.: of fat, of sugar, of cocoa…) which contain valuable information concerning the nutrient content, and stop words which add meaning to the description, are kept. Next, the tokenized words are lemmatized [34]. The titles/descriptions of the recipes after this pre-prcoessing are ready to undergo the next steps.

B. Results and evaluation

After the data pre-processing the next step is to apply the algorithms for generating embeddings. We generate the vector representations of the recipe titles in two different ways with three different algorithms:

1) Word embeddings – generating vector representations with the Word2Vec and GloVe algorithms for each word of the description and merging the separate word embeddings into sentence embedding by summing or averaging the separate vectors.
2) Paragraph embeddings - generating paragraph/sentence embeddings for the whole description with the Doc2Vec algorithm.

For the algorithm implementation we use the *Gensim* [**?**] library in Python, and the *Word2Vec* and *Doc2Vec* packages for the algorithms respectively. We run the two algorithms for different values for the vector dimensionality – 50, 100 and 200. Also, for these dimensions we change the 'sliding' window (which is a number that indicates the maximum distance between the current and predicted word within a sentence). For the sliding window we chose 2,3, and 5, because we are dealing with fairly short textual descriptions.

Independently of this process, the data is clustered following the two criteria presented – the FoodEx2 clustering system and the FSA traffic light system. Since we do not have the FoodEx2 codes available, they are obtained beforehand with applying the StandFood [29] method on the dataset. For the FoodEx2 clustering we are using the method presented in [25], where the FoodEx2 codes (based on their graph embedings) are clustered into 230 clusters. By matching the FoodEx2 codes of the clusters and the FoodEx2 codes of our dataset instances obtained by StandFood, the instances in our dataset are clustered, i.e. a cluster number from 1 to 230 is assigned to each recipe. After this initial clustering, because there are some empty clusters, the post-hoc cluster merging is performed, where we merge the clusters following a bottom up approach. For the Recipe1M dataset the parents on the forth level in the FoodEx2 hierarchy are chosen, and with this we obtain 47 clusters. After a close observation, 16 clusters containing very few elements are removed. One criteria that is satisfied as well is – no cluster has less then 200 instances; since it will be contradictory to train models with 200 features and less then 200 instances. The end result is 31 clusters.

For the FSA traffic light system, the clusters are generated by obtaining all the possible permutations of the three colors possible, i.e. calculating permutations with repetitions:

$$_{n}P_{r} = n^{r} \tag{1}$$

Where $n = 3$ are the three colors in the FSA traffic light system – red, orange, and green, and $r = 4$ are the four nutrients for which we have these colors available – fat, salt, saturates, and sugars. Therefore, we have 81 possible combinations, i.e. 81 clusters. After clustering or better said, dividing the recipes from the dataset according to these 81 possible combinations, 66 clusters of recipes are obtained, the rest – 15 combinations are not present in the dataset. Out of the 66 clusters, 10 clusters had very few examples – less then 200, which as stated previously is contradictory to our study. Thus, the final result is 55 clusters according to the FSA traffic light system.

The next step is the actual predictive modeling, i.e. the supervised ML part – applying single-target regressions. This is done according to the following setup:

1) Select regression algorithms – Linear regression, Ridge regression, Lasso regression, and ElasticNet regression (using the Scikit-learn library in Python [35]).
2) Select parameter ranges for each algorithm and perform hyper-parameter tuning – Ranges and values are a priori given for all the parameters for all the regression algorithms. From all the combinations, the best parameters for the model training are selected with GridSearchCV (using the Scikit-learn library in Python [35]). This is done for each cluster separately.
3) Apply k-fold cross-validation to estimate the prediction error – We train models for each cluster using each of the selected regression algorithms. The models are trained with the previously selected best parameters for each cluster and then evaluated with cross-validation. For comparison of the regressors the matched sample approach is chosen, i.e. using the same data in each iteration.
4) Apply tolerance levels and calculate accuracy – The accuracy is calculated according to the tolerance levels in I. If a_i is the actual value of the i^th instance from the test set on a certain iteration of the k-fold cross-validation, and p_i is the predicted values of the same, i^th, instance of the test set, then:

$$d_i = |a_i - p_i| \tag{2}$$

d_i is the absolute difference between the two set values. We define a binary variable that is assigned a positive value if the predicted value is in the tolerance level.

$$allowed = 1, \text{ if } :$$

$$Salt: \begin{cases} a_i < 1.25 \;\&\; d_i < 0.375 \\ a_i \geq 1.25 \;\&\; d_i \leq 0.2 \times a_i \end{cases}$$

$$Saturates: \begin{cases} a_i < 4 \;\&\; d_i < 0.8 \\ a_i \geq 4 \;\&\; d_i \leq 0.2 \times a_i \end{cases}$$

$$Fat: \begin{cases} a_i < 10 \;\&\; d_i < 1.5 \\ 10 \leq a_i < 40 \;\&\; d_i \leq 0.2 \times a_i \\ a_i \geq 40 \;\&\; d_i \leq 8 \end{cases}$$

$$Protein, Sugar: \begin{cases} a_i < 10 \;\&\; d_i < 2 \\ 10 \leq a_i < 40 \;\&\; d_i \leq 0.2 \times a_i \\ a_i \geq 40 \;\&\; d_i \leq 8 \end{cases}$$

At the end we calculate the accuracy as the ratio of predicted values that are in the 'allowed' range, i.e. tolerance level:

$$Accuracy = \frac{\sum_{i=1}^{n} allowed}{n} \tag{3}$$

Where n is the number of instances in the test set. The accuracy percentage is calculated for the baseline mean and baseline median as well – the percentage of baseline values (means and medians from each cluster) that falls in the tolerance level range. In this case – a_i is the actual value of the i^{th} instance from the test set on a certain iteration of the k-fold cross-validation, and instead of p_i in Equation 2 we have:

$$b = \begin{cases} \dfrac{\sum_{i=1}^{m} x_i}{m}, \text{the baseline is the mean} \\ \dfrac{X_{[\frac{(m+1)}{2}]} + X_{[\frac{(m+1)}{2}]}}{2}, \text{the baseline is the median} \end{cases} \tag{4}$$

Where m is the number of instances in the train set, and X is the train set sorted in ascending order. The accuracy percentages are calculated for each fold in each cluster, and at the end for each cluster we calculate an average of the percentages from each fold.

In Figure 2 a graph that represents the highest accuracy percentages obtained with the FoodEx2 clustering method are presented, as well as the baseline mean and baseline median accuracy percentages for the cluster in question. In Figure 3 the same for the FSA traffic light clustering method is presented. In the graphs, for each target nutrient, we give the best result obtained with the embedding vectors and compare them with the baseline mean and median for the particular cluster. In the graphs the embedding algorithm that yields the best results alongside with the parameters and heuristic is given as:

$$E\_h\_d\_w = \begin{cases} h \in [sum, average], \text{is the chosen heuristic} \\ din[50, 100, 200], \text{is the dimension} \\ win[2, 3, 5, 10], \text{is the sliding window} \end{cases} \tag{5}$$

Where, E is the embedding algorithm (Word2Vec, GloVe or Doc2Vec). We can see that which embedding algorithm yields the best results changes, but the in all cases the predictions with the embedding vectors give better results than the baseline methods. From these graphs we can see that both of the clustering methods give much better results than the baseline mean and baseline median. For demonstration purposes in the graphs only the highest obtained accuracies are presented. The range of the percentage of accuracy for the FoodEx2 clustering approach across the 31 clusters is 68% - 99% and for the FSA traffic light clustering approach across the 55 clusters is 71% - 99%. At last in Figure 4 we compare the highest percentages of accuracy obtained by both domain knowledge clustering approaches. As we can see they obtain very similar results, with a little favor for the FSA traffic light approach.

When closer observing the clusters we can state that the benefits to the FSA traffic light clustering approach are that it is based strictly on the quantity of the nutrients, therefore we can have food items that belong to very different food groups, but have similar nutrient labels (contents). Whereas the FoodEx2 clustering approach puts together food items that belong to a similar food group, example – types of fruit based products, types of meat based product, egg dishes, etc. From these results, it is worth arguing that modeling ML techniques on food data with previously considering domain knowledge yields better results than predicting on the whole dataset. If we compare the performances of the three embedding algorithms, it is hard to argue if one outperformed the others, or if one under-performed compared to the other two. This outcome is due to the fact that we are dealing with fairly short textual descriptions. On the other hand, if we compare the results of the two approaches we chose for inferring domain knowledge in the dataset, we can say that both performed very similarly, with a slight favour to the FSA traffic light system, simply because of the fact that it is purely based nutrient values, which is the task in matter for our prediction models.

IV. Conclusion

We live in a modern health crisis. We have a cure for almost everything, and yet the most common causes of the biggest mortality factor – cardiovascular diseases, are nutrition and diet related. Knowing what is in our food, and understanding its nutritional content (macro and micronutrients) is the first step, that is in our power, towards the prevention of diet-related diseases. There is an overwhelming amount of nutrition-related data available, and most of it comes in textual form, structured and unstructured. Data Science can help us utilize this data for our benefit. The application of the P-NUT methodology opens up many possibilities for facilitating and easing the process of calculating nutrient content, which is crucial for many professionals, such as: dietary assessment, dietary recommendations, dietary guidelines, macronutrient tracking, and other such tasks which are key tools for doctors, health professionals, dietitians, nutritional experts, policy makers, professional sport coaches, athletes, fitness professionals, etc. This study applies the P-NUT methodology on the largest publicly available recipe dataset with available nutrient information, and explores the effect that domain bias has over the task of predicting nutrient content. While in [4] we applied the P-NUT methodology and predicted four macronutrient values - carbohydrates, fat, protein and water, in this study for the new dataset we had available five nutrients – salt, sugars, saturates, fat, and protein. We explored the domain bias effect on two ways – clustering the recipes based on the FoodEx2 codes – which we obtained with the application of the StandFood method [29], and the FSA traffic light system codes which are available in the Recipe1M dataset [33]. From the new evaluation we concluded that the FSA traffic light system offers a little better insight when it comes to predicting nutrient values, solely because of the fact that this system classifies foods purely based on the quantity of the nutrients. Both methods obtained accuracy as high as 99%, but the regressors performed better when predicting sugar content on the clusters from the FSA traffic light system. For our future work we intend to extend this methodology with the state-of-the-art Bert Embeddings [36]. Because different food/recipe datasets besides the main nutrients (macronutrients) – protein, fat, carbohydrates, may or may not include also data about other nutrients (e.g.: saturates, fiber, salt), we are also planning on exploring the idea of training models for multiple nutrients on a conglomerate dataset – including many publicly available food/recipe datasets [37], [38]. With this we will be able to obtain a more complete dataset, which is of a great need in the food and nutrition domain.

Acknowledgment

This research was supported by the Slovenian Research Agency (research core grant number P2-0098), and the European Union's Horizon 2020 research and innovation programme (FNS-Cloud, Food Nutrition Security) (grant agreement 863059).

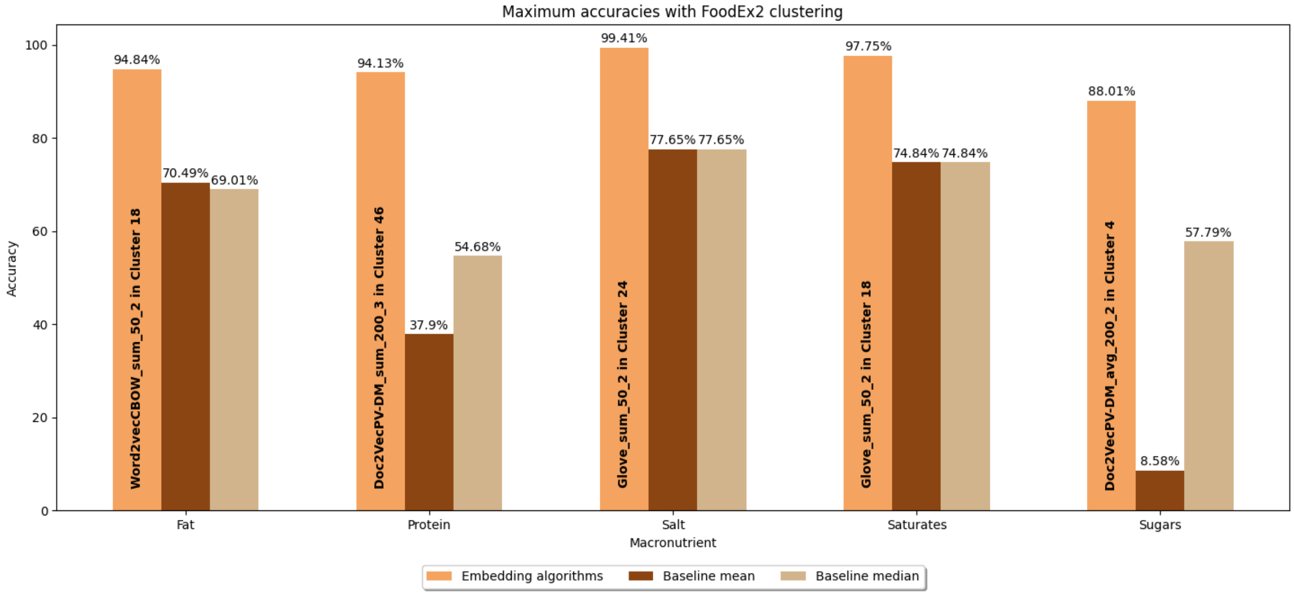

Fig. 2. Highest accuracy percentages obtained with the FoodEx2 clustering method compared to the baseline mean and baseline median.

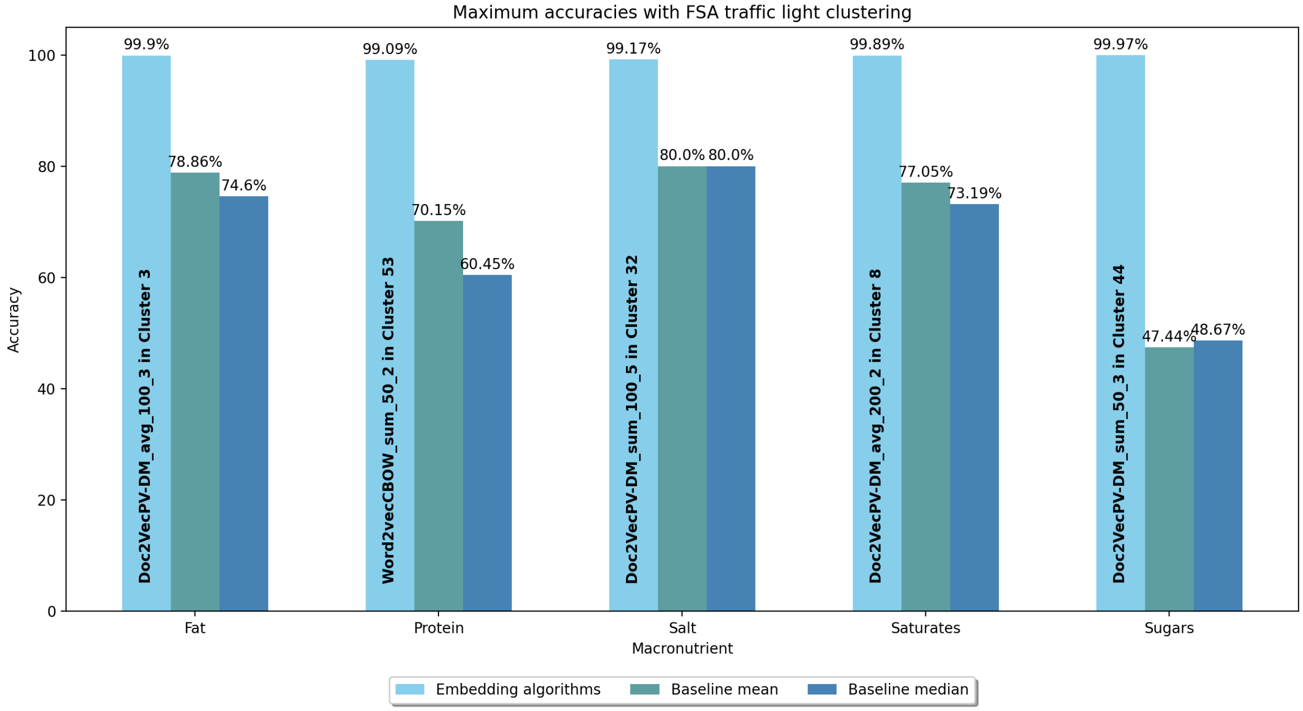

Fig. 3. Highest accuracy percentages obtained with the FSA traffic light clustering method compared to the baseline mean and baseline median.

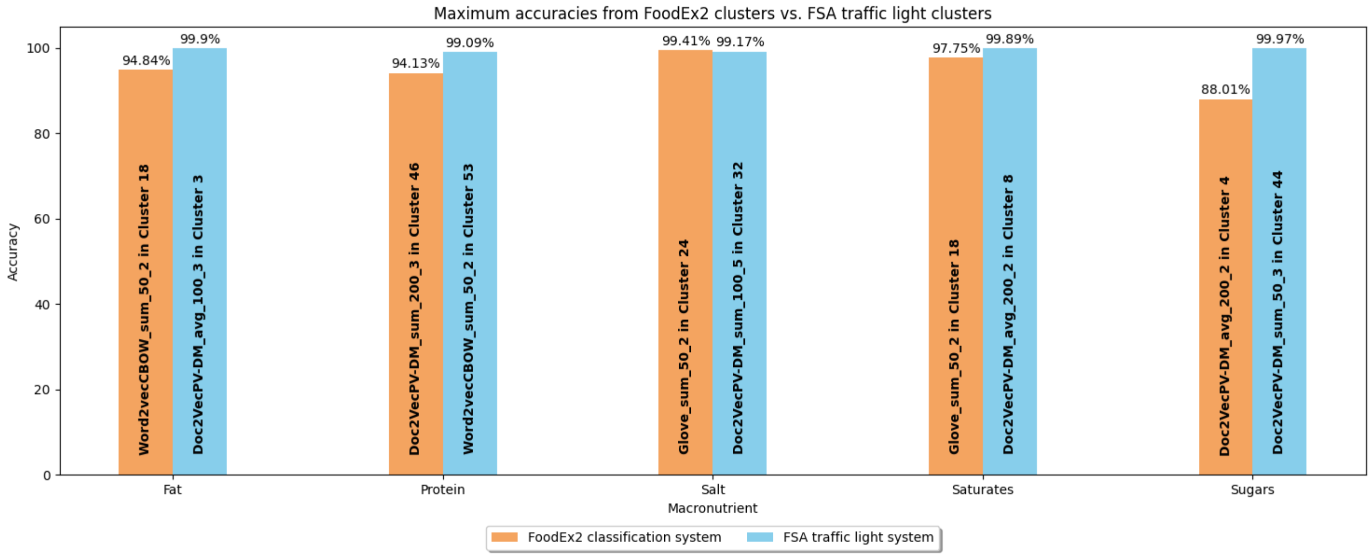

Fig. 4. Comparing the highest accuracies obtained for each nutrient prediction with the FoodEx2 clustering and the FSA traffic light clustering.

References

[1] S. W. Ng, S. Zaghloul, H. Ali, G. Harrison, and B. M. Popkin, "The prevalence and trends of overweight, obesity and nutrition-related non-communicable diseases in the arabian gulf states," *Obesity Reviews*, vol. 12, no. 1, pp. 1–13, 2011.

[2] "MyFitnessPal." [Online]. Available: https://www.myfitnesspal.com/

[3] "Samsung Health (S-Health)." [Online]. Available: https://health.apps.samsung.com/terms

[4] G. Ispirova, T. Eftimov, and B. Koroušić Seljak, "P-nut: Predicting nutrient content from short text descriptions," *Mathematics*, vol. 8, no. 10, p. 1811, 2020.

[5] T. E. F. S. Authority, 2019, (accessed on 11 May 2020). [Online]. Available: https://www.efsa.europa.eu/en/data/food-consumption-data

[6] U. FSA, "Guide to creating a front of pack (fop) nutrition label for prepacked products sold through retail outlets," *Food Standards Agency*, 2013.

[7] W. M. Rand, J. A. Pennington, S. P. Murphy, J. C. Klensin, and others, *Compiling data for food composition data bases*. United Nations University Press Tokyo, Japan:, 1991.

[8] H. Greenfield and D. A. Southgate, *Food composition data: production, management, and use*. Food and Agriculture Org.: Rome, Italy, 2003.

[9] S. F. Schakel, I. M. Buzzard, and S. E. Gebhardt, "Procedures for estimating nutrient values for food composition databases," *Journal of food composition and analysis*, vol. 10, no. 2, pp. 102–114, 1997, publisher: Elsevier.

[10] R. Yunus, O. Arif, H. Afzal, M. F. Amjad, H. Abbas, H. N. Bokhari, S. T. Haider, N. Zafar, and R. Nawaz, "A framework to estimate the nutritional value of food in real time using deep learning techniques," *IEEE Access*, vol. 7, pp. 2643–2652, 2018, publisher: IEEE.

[11] L. Jiang, B. Qiu, X. Liu, C. Huang, and K. Lin, "DeepFood: Food Image Analysis and Dietary Assessment via Deep Model," *IEEE Access*, vol. 8, pp. 47 477–47 489, 2020, publisher: IEEE.

[12] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947–1956, 2014, publisher: IEEE.

[13] T. Ege and K. Yanai, "Image-based food calorie estimation using recipe information," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 5, pp. 1333–1341, 2018, publisher: The Institute of Electronics, Information and Communication Engineers.

[14] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013, publisher: IEEE.

[15] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Twenty-Second International Joint Conference on Artificial Intelligence*, Catalonia, Spain, Jul. 2011.

[16] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, Bellevue, WA, USA, Jul. 2011, pp. 129–136.

[17] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.

[18] P. D. Turney, "Distributional semantics beyond words: Supervised learning of analogy and paraphrase," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 353–366, 2013.

[19] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, pp. 141–188, 2010.

[20] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, USA, Jun. 2013, pp. 746–751.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, Lake Tahoe, Nevada, USA, Dec. 2013, pp. 3111–3119.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, Beijing, China, Jun. 2014, pp. 1188–1196.

[25] T. Eftimov, G. Popovski, E. Valenčič, and B. K. Seljak, "FoodEx2vec: New foods' representation for advanced food data analysis," *Food and Chemical Toxicology*, vol. 138, p. 111169, 2020, publisher: Elsevier.

[26] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in neural information processing systems*, 2017, pp. 6338–6347.

[27] M. Van der Laan, K. Pollard, and J. Bryan, "A new partitioning around

medoids algorithm," *Journal of Statistical Computation and Simulation*, vol. 73, no. 8, pp. 575–584, 2003, publisher: Taylor & Francis.

[28] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987, publisher: Elsevier.

[29] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2," *Nutrients*, vol. 9, no. 6, p. 542, 2017, publisher: Multidisciplinary Digital Publishing Institute.

[30] D. of Health, "Guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlets [internet]," 2016.

[31] E. commission health and consumers directorate general, "Guidance document for competent authorities for the control of compliance with eu legislation on: Regulation (eu) no 1169/2011 of the european parliament and of the council of 25 october 2011 on the provision of food information to consumers, amending regulations (ec) no 1924/2006 and (ec) no 1925/2006 of the european parliament and of the council, and repealing commission directive 87/250/eec, council directive 90/496/eec, commission directive 1999/10/ec, directive 2000/13/ec of the european parliament and of the council, commission directives 2002/67/ec and 2008/5/ec and commission regulation (ec) no 608/2004devlin," Dec. 2012.

[32] E. Commission, "Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers, amending Regulations (EC) No 1924/2006 and (EC) No 1925/2006 of the European Parliament and of the Council, and repealing Commission Directive 87/250/EEC, Council Directive 90/496/EEC, Commission Directive 1999/10/EC, Directive 2000/13/EC of the European Parliament and of the Council, Commission Directives 2002/67/EC and 2008/5/EC and Commission Regulation (EC) No 608/2004," *Off. J. Eur. Union L*, vol. 304, pp. 18–63, 2011.

[33] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba, "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[34] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, Washington, DC, USA, 2004, pp. 625–633.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011, publisher: JMLR. org.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[37] W. Min, L. Liu, Z. Wang, Z. Luo, X. Wei, X. Wei, and S. Jiang, "Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 393–401.

[38] W. Min, B.-K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang, "You are what you eat: Exploring rich recipe information for cross-region food analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 950–964, 2017.