The Impact of Recipe Features, Social Cues and Demographics on Estimating the Healthiness of Online Recipes

Markus Rokicki*
L3S Research Center
Hannover, Germany
rokicki@L3S.de

Christoph Trattner*
University of Bergen
Bergen, Norway
christoph.trattner@uib.no

Eelco Herder
Radboud University
Nijmegen, The Netherlands
eelcoherder@acm.org

Abstract

Apart from taste, people increasingly consider nutritional facts when choosing a recipe or a meal. Researchers and companies alike aim to support health-conscious choices by providing estimates of a meal's calories and levels of sugar, fat, protein and salt. Features that are typically considered for these automatic estimates include a recipe's title, ingredients and cooking directions, as well as photographic material. Little is known based on which features users estimate the healthiness of online recipes themselves. Making use of data derived from a large online food community, and data collected via crowdsourcing, we compare the performance of algorithmic nutritional estimation with the performance of human-provided estimates, and analyze the most influential features used by humans and machines. Our results indicate that simple models already outperform human raters. Basic features such as title, ingredients and cooking directions were more informative than pictures of a recipe or user comments. For human estimates, we observed effects due to age and gender, but not due to dietary preferences or cooking habits. Our quantitative and qualitative results provide guidance for the development and evaluation of methods for nutrition estimation, and give insight in which features are most useful for nudging people into making healthier diet choices.

1 Introduction

How well do we know what we eat? How do we estimate how well-balanced a meal is? And to what extent can technology help us to make better food choices? These are the main questions that we address in this paper.

People typically make around 200 food choices every day (Wansink and Cashman 2006) and increasingly want to know what exactly is on their plates. According to food literature, the most common factors that we take into account include sensory appeal, health-related issues, ethical concerns, convenience, price, and weight control considerations (Steptoe, Pollard, and Wardle 1995).

In addition, governments and national health services actively promote healthy, balanced, leading to stricter rules, better guidelines and higher expectations regarding food labeling[1]. Proper labeling increases awareness on the nutritional value of a meal, in other words the amount of carbohydrates, fat, sugar and salt that people consume.

Whereas ready-made meals, snacks and other products are carefully labeled, people need to guess themselves the nutritional values of home-made meals or dishes offered on a restaurant menu. In addition, many recipes that can be found online are not or poorly labeled in terms of nutritional values. Tech companies currently aim to address this issue, for example by training 'deep learning algorithms' to 'count calories in food photos'[2]. As far as we know, despite significant R&D efforts and ambitious projects, there are currently no reliable, working systems or prototypes available.

Objectives In this paper, we investigate the contribution of common recipe features to the estimation of nutritional properties – and therewith the healthiness – of online recipes and associated meals in a data driven manner.

We build models using different kinds of features that are derived from a recipe's title, ingredient list and cooking directions, from popularity indicators such as the number of ratings and the user comments, as well as from state-of-the-art image analysis methods. Recent research in the area of online recipes has shown that these are important indicators and cues when it comes to online recipes food choices (Elsweiler, Trattner, and Harvey 2017). We also investigate which of these features contribute most to the quality of these estimations.

Further, we directly compare the performance of automated methods (as discussed in the previous paragraph) with human performance (using crowdsourcing, as explained in more detail later in this paper), to find out to what extent automatic estimates would now already improve people's insight in the healthiness of their meal choices. Finally, we investigate which recipe features people take into account for their nutrition estimations.

Research questions To guide and drive our research, we defined the following questions:

*Both authors contributed equally to this work.

[1] see for instance: https://www.nhs.uk/Livewell/Goodfood/Pages/food-labelling.aspx

[2] https://www.popsci.com/google-using-ai-count-calories-food-photos

- *RQ1.* To what extent is it possible to estimate the healthiness of online recipes with simple data driven methods, and what is the contribution of commonly used features and cues, such as title, ingredients and user feedback?
- *RQ2.* Which cues are usually involved when humans try to predict the healthiness of online recipes and to what extent do the estimates compare to the ones provided by our simple data driven models?
- *RQ3.* Finally, can we find difference in performance with respect to nutrition and health estimates among the users, related to their demographics, and eating and cooking habits?

Outline In the following sections, we will review appropriate background literature, introduce the datasets and methodology chosen to address our research questions, and present and discuss the results of our study. Finally, we summarize the findings of our study, discuss the limitations of our study and propose future research directions.

2 Background

Several studies in the field of nutrition science have shown that proper nutrition and health labels help people to make better food choices (Balasubramanian and Cole 2002; Elbel, Gyamfi, and Kersh 2011; Roberto et al. 2010; Neuhouser, Kristal, and Patterson 1999; Caswell and Padberg 1992; Downs, Loewenstein, and Wisdom 2009; Guthrie et al. 1995; Cowburn and Stockley 2005; Sonnenberg et al. 2013). Even though nutrition labels are common – and often obligatory – on packaged food, this information is often missing for other types of food or food listings, including online recipe portals such as Allrecipes.com[3].

Several different methods have been proposed in the literature to estimate nutritional properties and the according health aspects of recipes. Most notable here is the work of Müller et al. (2012), who propose an ingredient matching framework to estimate nutritional facts by employing a learning-to-rank approach and the popular German ingredient database BLS[4]. They show that the proposed solution is close to human *expert* judgment. Other interesting work in this area is (De Choudhury, Sharma, and Kiciman 2016). Similar to Müller et al., they propose a simple ingredient matching method to estimate the nutritional properties of food posts on Instagram, making use of the US Department of Agriculture (USDA) National Nutrient Database[5] for the matching process. Similar approaches were also taken by Fried et al. (2014) and Abbar, Mejova, and Weber (2015).

Kusmierczyk and Nørvåg (2016) proposed a simple solution that only employs title information of online recipes on Allrecipes.com, using Latent Dirichlet Allocation (LDA) with Gibbs sampling. In a number of experiments, they show that their proposed model is close to the nutrition facts as published by Allrecipes.com, employing the ESHA database. We use the same method in in one of the models in our work.

Another strand of research has focused on estimating nutrition facts from images. Pouladzadeh et al. (2012) employ Support Vector Machines (SVMs) to categorize food by using 'shape', 'color', 'size', and 'texture' features. Sudo et al. (2014) predicted nutrition values by using region recognition methods to identify objects and ingredients contained in those regions. Meyers et al. (2015) used convolutional neural networks (CNNs) to estimate the 'size' of the foods, as well as their 'labels'. Chokr and Elbassuoni (2017) performed similar experiments with 'RGB' features and were able to estimate calories with high accuracy. Researchers from QCRI and MIT showed the high potential of employing 'image embeddings' derived from a deep neural network (DNN) to predict the recipe, given an image of a meal (Salvador et al. 2017). We use a similar approach in this work to estimate nutritional properties from images.

Estimating the healthiness of online recipes has only recently been studied. Employing standards as set by the World Health Organization (WHO) as well as the Food Standard Agency (FSA), Trattner and Elsweiler; Trattner and Elsweiler (2017b; 2017a) performed a series of studies to not only show how these standards can be applied to online recipes, but also to reveal in detail how healthy online recipes are. Furthermore, they show how users interact with healthy and unhealthy recipes and what implications this has for online recipe recommender systems. In their latest research (Elsweiler, Trattner, and Harvey 2017) they also show how 'image', 'ingredients' and 'titles' of a recipe influence people in their decision making employing similar features as done in this work.

How people interact with recipes online has also been studied in the context of computational social sciences (Lazer et al. 2009; Strohmaier and Wagner 2014). For instance, West, White, and Horvitz (2013) found correlations between recipes accessed via search engines and incidence of diet-related illness, findings that are in line with (Said and Bellogín 2014). Ahn et al. (2011) mined and analyzed three different large-scale online food community platforms from Europe, the US and China to unveil patterns on how recipes vary between regions and to find out which flavor components make, for instance, Indian food different from the rest of the world. Rokicki et al. (2016) found gender differences in terms of preferences and appreciation of online recipes in German online cooking platform Kochbar.de. Finally, we would like to highlight another work by Rokicki, Herder, and Trattner (2017), showing how different factors such as the social context significantly influences users in their online decision making. In this work, we focus on the estimation of the nutritional value of a meal, a task that is reported to be quite hard for persons without special training (Almiron-Roig et al. 2013).

[3]On Allrecipes.com, only (editorial) recipes on the main site contain nutrition labels, while the larger fraction of recipes published on the users' personal sites do not contain these. Similar is true for other big recipe platforms on the Web, such as Kochbar.de, cookpad.com, food.com, yummly, etc.

[4]https://www.blsdb.de/

[5]https://ndb.nal.usda.gov/ndb/

Differences with previous research Several studies addressed the challenge of nutrition prediction from different angles, either by focusing on a specific feature or feature set, or by making use of an external database. In our study, we use, combine and compare the benefits of different types of information cues in a data-driven manner. Further, even though there have been many studies on how people interact with recipes online, to the best of our knowledge, this study is the first in which algorithmic performance is directly compared with human performance. By doing so, our study does not only provide guidance in how to develop better methods for nutrition estimation, but it also gives insight in which features are most useful for nudging people into making better nutrition choices (Elsweiler, Trattner, and Harvey 2017).

3 Materials

In this work, we make use of a web crawl of the online platform Allrecipes.com. The crawling of the platform was performed between 20th and 24th of July, 2015. We retrieved 242,113 recipes published by 62,100 users between the years 2000 and 2015 through the sitemap that is available in the robots.txt file of the website.

In addition to the core recipe components – such as recipe title, ingredient list, number of servings and instructions – we also collected for each recipe the according image, comments provided by users, rating information and – most important for our research – nutrition facts[6], such as total energy (kCal), protein (g), carbohydrate (g), sugar (g), salt (g), fat (g) and saturated fat (g) content (measured in 100g per recipe). We chose these nutrients, as they are typically used in food research and allow to calculate the healthiness of a recipe via standards as proposed by public health bodies. Focusing, for example, on the macro nutrients 'fat', 'saturated fat', 'sugar' and 'salt' (measured in 100g per recipe) allows us to measure the healthiness of a recipe according to international standards as introduced in 2007 by The Food Standard Agency (FSA) (FSA 2016). The FSA guidelines define a scale for each of these nutrients: green (healthy), amber and red (unhealthy). To derive a single metric we assign an integer value to each color (green=1, amber=2 and red=3) then sum the scores for each macro-nutrient resulting in a final range from 4 for very healthy recipes to 12 for very unhealthy recipes. This metric has been used in related work (as discussed in Section 2) (Trattner and Elsweiler 2017b). Throughout the paper we refer to this metric as 'FSA health score'.

4 Methodology

In this section, we describe in detail the methodology used in our study to answer our research questions. First, we describe the pre-processing steps performed on the dataset. In the following two subsections, we explain in detail how we estimated the nutrition of recipes employing predictive computational methods and human judgment. Finally, we briefly introduce the statistical methods used in this research.

4.1 Data Pre-Processing

Allrecipes.com contains both personal recipes from users and peer-reviewed 'editorial' recipes that are published on the main side. The editorial recipes are the most popular recipes on Allrecipes. In addition, the latter category of recipes also contains proper and peer-reviewed nutrition statistics – which is not the case for personal user recipes. In total there are 60,983 of such editorial recipes with nutrition information.

There are 24 main categories on Allrecipes.com that a recipe can be assigned to, with as the most popular category 'main dishes' contains 13,188 recipes. Previous research (Trattner and Elsweiler 2017b) in the area has shown that the nutritional properties and health aspects of recipes vary significantly across categories. In order to limit variation or bias introduced by certain more specialized categories, we focus in our work only on recipes published in this main dish category, containing valid, peer-reviewed nutritional information. Further, we require recipe images, information on preparation duration, as well as user feedback to be available, which is the case for 9,766 recipes.

4.2 Predictive Computational Modelling

We take a machine learning approach (more specifically, several different regression methods) to understand the benefits of different types of information cues for estimating the nutritional properties and health aspect of an online recipe. More formally, given a recipe and a set of information cues – 'Image', 'Title', 'Ingredients', 'Directions' and 'User Feedback' cue, each modeled with a set of features that are briefly introduced in this section – we aim to predict the healthiness of a recipe (as measured through the FSA health score) as well as the nutritional properties 'Calories', 'Fat', 'Sat. Fat', 'Sugar', 'Carbs', 'Protein' and 'Salt' per 100g of a recipe.

Feature Engineering In total, we derived 547 features for modelling the five information cues 'Image', 'Title', 'Ingredients', 'Directions', 'User Feedback', cues that are typically involved in the process of online recipe selection (Elsweiler, Trattner, and Harvey 2017). Below, we briefly summarize the features to model these cues:

- **Title:** For this cue, we derived 136 different features. Four are simple text metrics, e.g. number of words and characters or text entropy (Pitler and Nenkova 2008). We also measured the sentiment of the title and counted the words appearing in the Oxford English Dictionary (McKean 2010). Furthermore, we employed Mallet's[7] implementation of Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) employing Gibbs sampling on the recipe title, as proposed by (Kusmierczyk and Nørvåg 2016), us-

[6] Allrecipes.com estimates the nutritional facts for an uploaded recipe by matching the contained ingredients with those in the ESHA research database (ESHA 2016). The ESHA system is used by popular companies such as MCDonald's and Kellogs.

[7] http://mallet.cs.umass.edu/

ing 100 topics[8]. Also, we extracted the top-n words from the set of all titles of varying length. Similar to the LDA topics, we tested different ranges here and obtained rather good results with only 10 words. Also for the models, we chose ten words as a maximum. The remaining 'Title' features involve POS-tags[9], such as the number of adjectives and nouns.

- **Image:** For the 'Image' cue, we extracted 169 different features in total. The first feature set involves deep convolutional neural network (CNN) features from a pre-trained VGG16 model (Simonyan and Zisserman 2014), as proposed by related work (Salvador et al. 2017). In total, we induced 4096 image embedding features, which we compressed with Principal Component Analysis (PCA) (Wold, Esbensen, and Geladi 1987) down to 100 features, showing better results than the original 4096 feature model. Furthermore, we derived 5 features that capture image sharpness, brightness, colorfulness, contrast and entropy (San Pedro and Siersdorfer 2009). The rest of the 64 features are image histogram features that capture the different color aspects of the images, as proposed by related work (Pouladzadeh et al. 2012).
- **Ingredients:** For the ingredients cue, we extracted in 116 features in total. Similar to the 'Title' features, four are simple text metrics, such as the number of words and characters, or text entropy. Also the number of ingredients was captured, as well as the number of ingredients that appear in the Oxford English Dictionary. Furthermore, we employed LDA. The rest of the features refer to the top-n ingredients in the set of all recipes.
- **Directions:** We created a model based on the cooking directions provided for each recipe. The model contains 121 features in total, capturing among others the number of servings, preparation time, cooking time and the number of preparation steps. Furthermore, we extracted text features, similar to the ones in the 'Title' and 'Ingredients' model, including number of words, characters, entropy, POS tags and top-n words. The last set of features again involves LDA topics.
- **User Feedback:** Finally, we create a model based on the users' feedback, containing 5 features. We used popularity indicators such as number of ratings and bookmarks as well as appreciation measures, e.g. average rating, sentiment (via comments) provided by users in Allrecipes.com and number of images uploaded for a recipe as a predictor. We used that information cue in our studies as recent research found a strong correlation between number of ratings and bookmarks with health aspects of online recipes (Trattner and Elsweiler 2017b).

[8]We tested several different configurations here, from 10 to 1000 topics, also for the other models. At the end we decided to preset the number of topics to 100, as this gave us close to optimal performance while preserving the number of features and keeping computational cost low.

[9]POS-tags were calculated with the popular Stanford NLP tagger, see http://nlp.stanford.edu/software/tagger.shtml. As title strings are short, we employed the GATE English POS-tagger model, see https://gate.ac.uk/wiki/twitter-postagger.html

Regression Setup The regression experiments were conducted with the help of the R Statistical Computing software. Regression models employed for the experiments were Linear Regression, Random Forest, Ridge Regression and LASSO from the Caret package[10]. For space reasons, we report only Ridge regression results, as this was the best method overall among the regression models that we investigated. The evaluation protocol employed was 10-fold cross-validation. All models were tuned to their optimum.

Regression performance was evaluated by means of median absolute error (MdAE) and symmetric mean absolute percentage error (SMAPE), two commonly used measures.[11] Given a set of predictions $\hat{y}_i$ and corresponding ground truth values y_i, for $i \in [1, n]$ median absolute error is defined as:

$$MdAE = median(|y_i - \hat{y}_i|).$$

SMAPE is a relative error between 0 and 2, given by:

$$SMAPE = \frac{1}{n} \sum_{1}^{n} \frac{|y_i - \hat{y}_i|}{\frac{1}{2} \cdot (|y_i| + |\hat{y}_i|)}.$$

4.3 Human Judgment

To observe how (well) humans estimate healthiness and nutrition of online recipes, we relied on a crowdsourcing-based study setup on Crowdflower[12], a popular microtask crowdsourcing platform. Workers were shown complete online recipes and were asked to estimate calories, as well as the macro-nutrients covered by the FSA front-of-package labeling system: fat, saturated fat, sugar and salt – information that constitutes the basis for assessing the healthiness of a meal (see Section 3). The two other macro-nutrients, carbs and protein, are omitted to avoid overburdening the workers.

Data Selection For our crowdsourcing study, we selected a subset of 60 'main dish' recipes from our dataset to be judged by human crowdworkers. In line with recent research in the area (Elsweiler, Trattner, and Harvey 2017) we chose a setup where each of the recipes could be judged by a sufficient number of workers. In addition, we made sure to have a sufficient number of recipes of various levels of difficulty with respect to estimating the nutrients. To this end, we ranked the recipes according to prediction errors made by our machine models.We selected the 20 highest ranked recipes, in other words the ones of which the nutrients were, on average easiest to predict. Conversely, we also selected the lowest-ranked recipes, those that were hardest to predict. We also selected 20 medium recipes from the middle part of the list. Note that, as a result, it is expected to observe a slightly lower machine performance on this sample compared to the whole data set.

[10]https://cran.r-project.org/web/packages/caret/caret.pdf

[11]We employ median absolute error rather than mean absolute error due to the skewed distributions of nutrients in our dataset (see Figure 1).

[12]https://www.crowdflower.com/

Crowdsourcing Design Our crowdsourcing tasks consisted of two parts: 1) a survey on workers' demographics and familiarity with the food domain and 2) the nutrition estimation tasks themselves. We randomly split the 60 recipes across 6 batches of tasks and let workers complete 11 rows for each of the tasks: estimates for 10 recipes and the survey. This setup was chosen to ensure availability of survey information for all workers who complete a task, without requiring them to provide estimates for all 60 recipes at once, which would have taken a considerable amount of time.

In the survey task, we asked workers to provide the following information:

- Age range (<18, 18-24, 25-34, 35-44, 45-54, $\geq$55) (optional)
- Gender (male, female, other) (optional)
- Recipe website usage (daily, weekly, monthly, rarely)
- Frequency of (being involved in) preparing home-cooked meals (daily, weekly, monthly, rarely)
- Cooking enjoyment (likert scale 1-5)
- Dietary preferences (Vegan, Vegetarian, Pescatarian, Omnivore, Carnivore)

The estimation task was designed as follows. Workers were shown screen captures of recipes [13] that contained the following information cues: recipe *title*, an *image* of the dish, the list of *ingredients*, preparation *directions* (including basic information like servings and preparation time), and *user feedback* in the form of (average) ratings and a selection of reviews. Nutrition information was manually removed from the screen captures of the recipes.

For each of the recipes, workers were asked to provide estimates for calories, fat, saturated fat, sugar and salt for 100g of the meal. Short descriptions of each nutrient were given in the task description. For each nutrient, in addition to providing a real-valued estimate, workers were asked to estimate healthiness levels corresponding to FSA guidelines ('low', 'medium', or 'high'[14]). The discrete inputs had a two-fold purpose: a) they provided the workers with basic reference points for value ranges to expect for the nutrients and b) they allowed us to check for consistency of inputs as a precaution against spammers. In addition, they were asked for their familiarity with the dish in general, on a scale from 1 (unfamiliar) to 5 (familiar).

Costs, Workers, and Judgments We recruited level 2 workers[15] and paid 0.44 USD for completing the task, which required about 10 minutes of work on average. In total, 108 workers provided 1420 nutrition estimates and 142 survey responses[16] for a total cost of 91 USD, including fees. For our analysis, we consider answers from workers who have filled out the survey and have given estimates for at least 10 recipes in total[17]. In addition, estimates were checked manually for consistency and inputs from 8 workers were discarded. Observed unreliable inputs included obviously unrealistic values (e.g. 50 calories and 50 grams of salt for all recipes) or inconsistent inputs (e.g. selecting high calories but giving a low exact estimate).

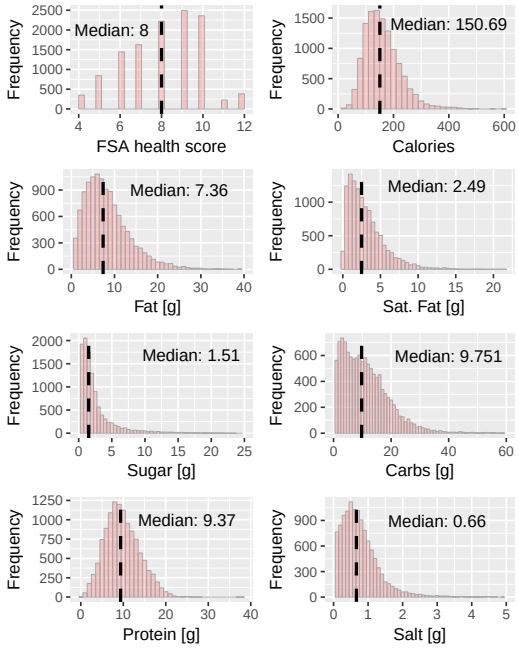

Figure 1: Distribution of nutrition and health scores (FSA) in our dataset.

After filtering based on above criteria, we were left with a dataset consisting of 1242 estimates made by 82 workers from 23 different countries.The reported familiarity with dishes across estimates was 3.04 on average. The majority of workers (63.4%) were male and the most common age range was '25-34' (45.1%), followed by '18-24' and '35-44' (20.7% each) – typical characteristics for crowdsourcing platforms (Ross et al. 2010).

In terms of familiarity with recipe websites and the food domain, 42.7% use recipe websites daily or weekly and 85.4% stated to prepare or help prepare home-cooked meals either daily or weekly. When asked to rate to what extent they enjoy cooking, almost two thirds (63.4%) rated their experience as positive (4 or 5 stars), with a mean rating of 3.73. The vast majority of workers stated to be omnivores or carnivores. Only 8 workers were vegans, vegetarians or pescatarians.

[13]Recipes were displayed as images in order to impede copying the title to search for nutrition information online.

[14]For calories, since there is no FSA recommendation, we defined thresholds of below 100 kcal/100g for 'low' and above 200kcal/100g for 'high', based on the distribution in our dataset.

[15]On Crowdflower, level 2 workers are reliable workers who have maintained a very high accuracy in at least 100 test questions.

[16]Workers were allowed to participate in multiple task batches.

[17]As a side effect of how tasks are distributed on Crowdflower, some workers were shown only a subset of items. Consequently, these workers were filtered out to ensure a sufficient number of estimates per worker and availability of survey information.

Table 1: Regression performance on all main dish recipes (N = 9766) in terms of MdAE (SMAPE) for the ridge regression model. Performances for each of the considered nutrients, as well as the FSA health score, are reported using all cues, individual cues, as well as no cues (corresponding to predicting the average value). The best machine performances for each nutrient and cue are highlighted in bold.

	MdAE (SMAPE)							
Information Cue	FSA health score	Calories	Fat	Sat. Fat	Sugar	Carbs	Protein	Salt
Title	$1.28^{***}_{\dagger\dagger\dagger}$ ($0.185^{***}_{\dagger\dagger\dagger}$)	$36.06^{***}_{\dagger\dagger\dagger}$ ($0.284^{***}_{\dagger\dagger\dagger}$)	$3.10^{***}_{\dagger\dagger\dagger}$ ($0.481^{***}_{\dagger\dagger\dagger}$)	$1.36^{***}_{\dagger\dagger\dagger}$ ($0.594^{***}_{\dagger\dagger\dagger}$)	$1.18^{***}_{\dagger\dagger\dagger}$ ($0.733^{***}_{\dagger\dagger}$)	$4.37^{***}_{\dagger\dagger\dagger}$ ($0.565^{***}_{\dagger\dagger\dagger}$)	$2.42^{***}_{\dagger\dagger\dagger}$ ($0.315^{***}_{\dagger\dagger\dagger}$)	$0.12^{***}_{\dagger\dagger\dagger}$ ($0.544^{**}_{\dagger\dagger\dagger}$)
Image	$1.33_{\dagger\dagger\dagger}$ ($0.189_{\dagger\dagger\dagger}$)	$38.29^{***}_{\dagger\dagger\dagger}$ ($0.295^{***}_{\dagger\dagger\dagger}$)	$3.19^{**}_{\dagger\dagger\dagger}$ ($0.493^{*}_{\dagger\dagger\dagger}$)	$1.47^{***}_{\dagger\dagger\dagger}$ ($0.616^{***}_{\dagger\dagger\dagger}$)	$1.31^{***}_{\dagger\dagger\dagger}$ ($0.764^{***}_{\dagger\dagger\dagger}$)	$4.77^{***}_{\dagger\dagger\dagger}$ ($0.590^{***}_{\dagger\dagger\dagger}$)	$2.63^{***}_{\dagger\dagger\dagger}$ ($0.334^{***}_{\dagger\dagger\dagger}$)	$0.13_{\dagger\dagger\dagger}$ ($0.555_{\dagger\dagger\dagger}$)
Ingredients	$\mathbf{1.12}^{***}_{\dagger\dagger\dagger}$ ($\mathbf{0.167}^{***}_{\dagger\dagger\dagger}$)	$\mathbf{30.48}^{***}_{\dagger\dagger\dagger}$ ($\mathbf{0.248}^{***}_{\dagger\dagger\dagger}$)	$\mathbf{2.69}^{***}_{\dagger\dagger\dagger}$ ($\mathbf{0.440}^{***}_{\dagger}$)	$\mathbf{1.07}^{***}_{\dagger\dagger}$ ($\mathbf{0.511}^{***}_{\dagger\dagger\dagger}$)	$\mathbf{0.91}^{***}_{\dagger\dagger}$ ($\mathbf{0.659}^{***}_{\dagger\dagger\dagger}$)	$3.85^{***}_{\dagger\dagger\dagger}$ ($0.525^{***}_{\dagger\dagger\dagger}$)	$2.07^{***}_{\dagger\dagger\dagger}$ ($0.277^{***}_{\dagger\dagger\dagger}$)	$\mathbf{0.10}^{***}$ ($\mathbf{0.491}^{***}_{\dagger\dagger}$)
Directions	$\mathbf{1.12}^{***}_{\dagger\dagger\dagger}$ ($0.171^{***}_{\dagger\dagger\dagger}$)	$30.87^{***}_{\dagger\dagger\dagger}$ ($0.251^{***}_{\dagger\dagger\dagger}$)	$2.81^{***}_{\dagger\dagger\dagger}$ ($0.451^{***}_{\dagger\dagger\dagger}$)	$1.17^{***}_{\dagger\dagger\dagger}$ ($0.544^{***}_{\dagger\dagger\dagger}$)	$1.06^{***}_{\dagger\dagger\dagger}$ (0.704^{***})	$\mathbf{3.47}^{***}_{\dagger\dagger\dagger}$ ($\mathbf{0.489}^{***}$)	$\mathbf{2.06}^{***}_{\dagger\dagger\dagger}$ ($\mathbf{0.277}^{***}_{\dagger\dagger\dagger}$)	$0.11^{***}_{\dagger\dagger\dagger}$ (0.516^{***})
User Feedback	$1.29_{\dagger\dagger\dagger}$ ($0.192_{\dagger\dagger\dagger}$)	$41.47_{\dagger\dagger\dagger}$ ($0.311_{\dagger\dagger\dagger}$)	$3.40_{\dagger\dagger\dagger}$ ($0.507_{\dagger\dagger\dagger}$)	$1.62_{\dagger\dagger\dagger}$ ($0.639_{\dagger\dagger\dagger}$)	$1.45_{\dagger\dagger\dagger}$ ($0.789_{\dagger\dagger\dagger}$)	$5.60_{\dagger\dagger\dagger}$ ($0.627_{\dagger\dagger\dagger}$)	$2.76_{\dagger\dagger\dagger}$ ($0.347_{\dagger\dagger\dagger}$)	$0.13_{\dagger\dagger\dagger}$ ($0.561_{\dagger\dagger\dagger}$)
All cues	1.05^{***} (0.160^{***})	28.67^{***} (0.234^{***})	2.55^{***} (0.429^{***})	1.01^{***} (0.511^{***})	0.96^{***} (0.721^{***})	3.25^{***} (0.484^{***})	1.86^{***} (0.258^{***})	0.11^{***} (0.508^{***})
No cues	$1.11_{\dagger\dagger\dagger}$ ($0.192_{\dagger\dagger\dagger}$)	$41.44_{\dagger\dagger\dagger}$ ($0.311_{\dagger\dagger\dagger}$)	$3.41_{\dagger\dagger\dagger}$ ($0.508_{\dagger\dagger\dagger}$)	$1.62_{\dagger\dagger\dagger}$ ($0.640_{\dagger\dagger\dagger}$)	$1.46_{\dagger\dagger\dagger}$ ($0.790_{\dagger\dagger\dagger}$)	$5.60_{\dagger\dagger\dagger}$ ($0.628_{\dagger\dagger\dagger}$)	$2.78_{\dagger\dagger\dagger}$ ($0.349_{\dagger\dagger\dagger}$)	$0.13_{\dagger\dagger\dagger}$ ($0.562_{\dagger\dagger\dagger}$)

Cues compared to no cues: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$
Cues compared to all cues: $_{\dagger}p < .05$; $_{\dagger\dagger}p < .01$; $_{\dagger\dagger\dagger}p < .001$

Table 2: Performance of machine (ridge regression) and human (crowd worker) estimates for selected main dish recipes. We report overall MdAE (SMAPE) of all machine and human estimates (N = 1242 estimates). In addition, we compare subsets of human estimates depending on which cues the estimates were based on with corresponding machine estimates based on the same cues. The best machine performances for each nutrient are highlighted in bold and the best human performances in blue color.

	MdAE (SMAPE)					
	FSA health score	Calories	Fat	Sat. Fat	Sugar	Salt
Machine vs Human (Overall)						
Machine	1.16^{**} (0.174^{***})	34.4^{***} (0.274^{***})	3.81^{***} (0.481^{***})	1.1^{***} (0.521^{***})	0.55^{***} (0.637^{***})	0.273^{***} (0.534^{***})
Human	2 (0.282)	96.1 (0.617)	7.63 (0.863)	2.54 (0.856)	3.65 (1.11)	0.644 (0.847)
Machine vs Human (Based on information cue processed)						
Title (Machine)	1.33 (0.183^{*})	48.2^{***} (0.369^{***})	4.15^{***} (0.565^{***})	1.69^{**} (0.617^{***})	1.1^{***} (0.763^{***})	0.366^{***} (0.663^{***})
Title (Human)	2 (0.272)	89.7 (0.586)	8.62 (0.9)	2.57 (0.892)	3.73 (1.1)	0.723 (0.862)
Image (Machine)	1.44^{**} (0.193^{***})	51.8^{**} (0.416^{***})	5^{*} (0.633^{***})	1.99^{*} (0.684^{***})	1.45^{***} (0.848^{***})	0.426^{*} (0.686^{***})
Image (Human)	2 (0.3)	96 (0.588)	7.51 (0.851)	2.56 (0.862)	3.43 (1.1)	0.607 (0.839)
Ingredients (Machine)	1.41^{**} ($\mathbf{0.179}^{**}$)	34.9^{***} ($\mathbf{0.3}^{***}$)	3.54^{***} ($\mathbf{0.528}^{***}$)	$\mathbf{1.04}^{***}$ ($\mathbf{0.56}^{***}$)	$\mathbf{0.868}^{***}$ ($\mathbf{0.72}^{***}$)	$\mathbf{0.282}^{***}$ ($\mathbf{0.539}^{***}$)
Ingredients (Human)	2 (0.27)	98 (0.622)	7.69 (0.856)	2.57 (0.858)	3.85 (1.09)	0.659 (0.843)
Directions (Machine)	$\mathbf{1.35}^{**}$ (0.19^{**})	$\mathbf{30.6}^{***}$ (0.313^{***})	$\mathbf{3.48}^{***}$ (0.539^{***})	1.48^{*} (0.622^{***})	1.21^{***} (0.81^{***})	0.3^{***} (0.589^{***})
Directions (Human)	2 (0.287)	95.3 (0.599)	7.43 (0.856)	2.55 (0.847)	3.94 (1.13)	0.673 (0.867)
User Feedback (Machine)	1.58^{**} (0.195^{***})	57.5 (0.422^{***})	4.87 (0.64^{***})	1.96 (0.669^{**})	1.59^{*} (0.833^{***})	0.434^{*} (0.674^{***})
User Feedback (Human)	2 (0.296)	71.7 (0.515)	7.15 (0.897)	2.44 (0.792)	2.71 (1.05)	0.594 (0.838)

Pairwise comparison: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

4.4 Statistical Analysis

To test for significant differences between groups (e.g. male vs female), Wilcoxon Rank-Sum tests were performed based on (normalized) prediction errors of individual estimates. For space reasons, specific p-values for these statistical comparisons were not included. However, ranges for the p-values are provided at the end of each table in this paper.

5 Results

In the following three subsections, we report the results of our experiments in alignment with our three research questions.

5.1 RQ1: Estimating the healthiness of online recipes with simple data driven methods

Figure 1 shows the distribution of the nutrients of the online recipes investigated in the paper. All nutritional values follow more or less a right-skewed normal distribution, with the majority of recipes having above average values (which is also reflected by the median values). The FSA health scores (see Section 3) are always rounded integers and are skewed to the left (which indicates that the majority of recipes scores relatively low in terms of healthiness).

Table 1 reports regression performance for the ridge regression models for each of the considered nutrients, as well as the FSA health score using information cues derived from different parts of the 9,766 recipes investigated. The regression results show that taking all information cues (and corresponding features) into account is better than employing just one cue at a time. Of the information cues investigated, the 'Ingredients' cue is the best indicator, except for carbs and protein, for which the cooking 'Directions' are slightly more indicative. As the cooking process does not impact carbs and protein levels[18], it is likely that certain preparation methods are associated with foods that are higher or lower in carbs and protein. As also shown in Table 1, 'User feedback' is the worst of all cues to predict nutrition of a recipe, followed by the 'Image' cue, which is surprising as related

[18] see e.g. https://www.livestrong.com/article/552825-what-happens-to-carbohydrates-protein-when-cooked/

Table 3: Distribution of the top-10 information cues and combinations the recipe workers reported to have influenced them in their estimates. Note that workers could report multiple cues at the same time.

Information Cue	N	Percentage
Ingredients	319	25.68%
Image	201	16.18%
Image, Ingredients	161	12.96%
Ingredients, Directions	78	6.28%
Image, Ingredients, Directions	77	6.19%
Title, Image Ingredients	54	4.34%
Title, Image, Ingredients, Directions	54	4.34%
Title	51	4.11%
Directions	43	3.46%
Title, Ingredients	43	3.46%

work has shown good performance using similar image features as we employed (Chokr and Elbassuoni 2017).

The lowest SMAPE value is achieved for the estimation of the FSA health score, indicating that overall healthiness can be predicted quite accurately.

In addition, the results show that calorie content is relatively easy to predict; predicting sugar content is a hard task, most likely due to the natural sugar content of many ingredients. The ongoing confusion and controversy about natural vs added sugar (for more details, see e.g. (Erickson and Slavin 2015)) is likely to be reflected in recipe title and user feedback, leading to overfitting - which is an explanation for the lower performance of all features for sugar.

In summary, the performance of the models and information cues exploited is consistent and in line with what one would expect. While all cues contain useful signals for estimating nutrition facts and healthiness of meals, ingredients and directions are particularly useful. Further, among all macro nutrients, calories are easiest to predict.

5.2 RQ2: Investigating how humans estimate health and nutrition of online recipes

Table 2 compares human and machine estimation performance for calories and the FSA health score as well as related macro-nutrients. It might come as a surprise that the machine (Ridge regression with all features) significantly outperforms human performance. It should be noted, though, that we intentionally recruited *average* human raters, not experts (as was the case in some works discussed in Section 2). For sugar content, the machine model resulted in even 7 times lower average errors compared to the average human, and for calories about 3 times lower as well. Differences in human and machine estimation performance are consistent and significant for each of the individual information cues.

Overall, the smallest gap was found for estimating the FSA health score. Human estimates resulted in a moderate SMAPE of 0.282, suggesting that the average human rater is able to identify unhealthy meals to some degree, which is encouraging.

Table 3 shows the distribution of the top-10 information cues that workers reported to have relied on to form their judgments (selection of multiple options was possible). Most frequently, workers reported to have based their estimates on 'Ingredients', 'Image' or a combination of the two. Interestingly, when investigating Table 2, humans achieved the best results when relying on the cues 'User Feedback' and 'Ingredients' for all nutritional properties, which stands to some extent in contrast to the machine performance, where the best results were again obtained when relying on the 'Directions' or the 'Ingredients' cue.

When investigating just one cue at a time, the 'Ingredient' cue was used in 69.3% of all cases, followed by the 'Image' cue (51.9%) and the 'Directions' cue (29%). Finally, the 'User Feedback' cue was reported as being used only in 5.8% of the cases, although being the most informative one when estimating calories, fat, sat. fat or sugar (see Table 2).

Figure 2 shows human and machine estimation errors and according linear regression lines for different ranges in calories, fat, sat fat, sugar, salt and FSA health score. The plots show that both humans and the machine are similar in overestimating low values and underestimating high values, which comes as no surprise. However, for calories, fat, sat. fat, and sugar the regression lines show that this trend is more pronounced for humans (slopes of the regression lines are steeper). Interesting to note here is that for salt and the FSA health score, the slopes are similar, although we observe again that the machine estimates are less error prone. In particular, human estimates for the FSA health score are overall significantly lower than machine estimates ($M = 7.5$ compared to $M = 8.4$; $W = 50212$, $p < .001$). The observations reported here are in line with previous research in the area showing that humans typically lean towards overestimating the healthiness of meals (Carels, Harper, and Konrad 2006).

We also compared human performance for recipes that were 'Easy', 'Medium' and 'Hard' to estimate by the machine (Ridge regression employing all cues). In Table 4 we show machine and human performance for the 20 most easiest, medium and hardest to predict. In general, the results show that recipes of which the nutritional values were harder to predict by the machine, are also harder to predict by humans. A further observation is that the differences in human estimation performance between 'Hard' and 'Medium' recipes is, for all nutritional values, far bigger than between 'Easy' and 'Medium' recipes.

5.3 RQ3: Demographic differences

As discussed earlier, the spread of human predictions for each single nutritional property of an online recipe was large – which contributed to the relatively poor average performance of human predictions.

Therefore, we also investigated which demographic factors contributed to better or worse performance. Table 5 provides an overview of the results. Among others, we asked the workers how familiar they were with the recipes they rated. Unsurprisingly, those who reported to be familiar with the recipes were slightly better with their estimates than those who reported to be unfamiliar with the dish (see calories, fat and the FSA health scores) .

Table 4: Estimation performance MdAE (SMAPE) of Ridge Regression and human workers (N = 1242 estimates) for recipes of different difficulty levels (with regards to machine performance).

	MdAE (SMAPE)					
	FSA health score	Calories	Fat	Sat. Fat	Sugar	Salt
Machine						
Easy	1.02 (0.153)	$3.74^{***}_{\dagger\dagger\dagger}$ ($0.054^{***}_{\dagger\dagger\dagger}$)	$0.245^{***}_{\dagger\dagger\dagger}$ ($0.068^{***}_{\dagger\dagger\dagger}$)	$0.135^{***}_{\dagger\dagger\dagger}$ ($0.171^{***}_{\dagger\dagger\dagger}$)	$0.124^{***}_{\dagger\dagger\dagger}$ ($0.251^{***}_{\dagger\dagger}$)	$0.0503^{***}_{\dagger\dagger\dagger}$ ($0.098^{**}_{\dagger\dagger}$)
Medium	1.16 (0.174)	$34.4^{\circ\circ\circ}$ ($0.274^{\circ\circ\circ}$)	$3.81^{\circ\circ\circ}$ ($0.481^{\circ\circ}$)	$1.1^{\circ\circ\circ}$ ($0.521^{\circ\circ}$)	$0.55^{\circ\circ\circ}$ ($0.637^{\circ\circ\circ}$)	$0.273^{\circ\circ\circ}$ (0.534)
Hard	1.94 (0.205)	119 (0.53)	11(0.852)	4.53 (0.861)	5.81 (1.12)	0.799 (0.89)
Human						
Easy	$2^{***}_{\dagger}$ (0.261^{***})	$76.4^{***}_{\dagger\dagger}$ (0.613)	$5.65^{***}_{\dagger\dagger\dagger}$ (0.799^{***})	$1.96^{***}_{\dagger\dagger\dagger}$ (0.829)	$2.84^{***}_{\dagger\dagger\dagger}$ (1.1)	$0.472^{***}_{\dagger\dagger\dagger}$ (0.787^{***})
Medium	$2^{\circ\circ\circ}$ ($0.282^{\circ\circ\circ}$)	$96.1^{\circ\circ\circ}$ (0.617°)	$7.63^{\circ\circ\circ}$ (0.863°)	$2.54^{\circ\circ\circ}$ (0.856)	$3.65^{\circ\circ\circ}$ (1.11)	$0.644^{\circ\circ\circ}$ (0.847)
Hard	3 (0.336)	149 (0.689)	13.2 (0.943)	5.23 (0.903)	5.37 (1.03)	1.03 (0.915)

Comparison between Easy and Hard to predict recipes: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.
Comparison between Easy and Medium to predict recipes: $_{\dagger}p < .05$; $_{\dagger\dagger}p < .01$; $_{\dagger\dagger\dagger}p < .001$.
Comparison between Medium and Hard to predict recipes: $^{\circ}p < .05$; $^{\circ\circ}p < .01$; $^{\circ\circ\circ}p < .001$.

Table 5: Estimation performance measured in MdAE (SMAPE) of human workers (N = 1242 estimates) of varying demographics and domain knowledge. Only attributes with significant results are reported.

	MdAE (SMAPE)					
	FSA health score	Calories	Fat	Sat. Fat	Sugar	Salt
Familiarity with the recipe						
Familiar (≥ 3 on likert scale)	2^{*} (0.271^{*})	95.2 (0.592^{*})	7.61 (0.841^{*})	2.5 (0.841)	3.52 (1.11)	0.667 (0.848)
Unfamiliar (≤ 2 on likert scale)	2 (0.306)	101 (0.672)	7.71 (0.91)	2.58 (0.888)	3.85 (1.11)	0.579 (0.847)
Gender						
Female	2^{*} (0.259^{*})	105 (0.682)	8.02 (0.878)	2.5 (0.817)	4.17 (1.07)	0.593^{*} (0.863)
Male	2 (0.293)	92.6 (0.59)	7.49 (0.856)	2.57 (0.873)	3.53 (1.13)	0.666 (0.843)
Age group						
Age < 35	2 (0.297)	97.6 (0.628)	7.86 (0.895)	2.5 (0.874)	3.75 (1.16)	0.586 (0.832)
Age ≥ 35	2^{**} (0.252^{*})	96 (0.595)	6.97^{*} (0.8^{*})	2.58 (0.82^{*})	3.43 (1.02)	0.744 (0.877)
Enjoys cooking						
Yes (≥ 4 on likert scale)	2 (0.286)	95.7 (0.623)	7.46 (0.879)	2.5 (0.878)	3.45 (1.07^{*})	0.655 (0.843)
No (≤ 3 on likert scale)	2 (0.274)	99.7 (0.606)	8.03 (0.831)	2.61 (0.813)	4.04 (1.19)	0.617 (0.856)

Pairwise comparison: $^{*}p < .05$, $^{**}p < .01$

Further, we asked users to provide some demographics and their (culinary) background. Raters who indicated to enjoy cooking were better in estimating sugar content than those who did not indicate so. In terms of gender, male and female workers achieved similar estimates except for salt and the FSA health score, where female workers provided significantly more accurate estimates. As we can only speculate about potential causes, we leave this observation uncommented.

Another factor showing significant results was the age of the workers. We observe that workers aged 35 and older were significantly better able to estimate fat, saturated fat and the FSA health score. At first glance, this result is in contrast to observations made in previous studies showing that elderly people are less capable in estimating the nutritional properties of a meal correctly (Burton and Andrews 1996). However, compared to (Burton and Andrews 1996) the average age in our study was comparatively low. As such, the results would indicate that middle-aged people are better capable of estimating the nutritional properties of a meal correctly than younger ones.

As a final observation, no significant differences were observed for the users' locations, dietary preferences and how often they use recipe websites and prepare home-cooked meals. For dietary preferences, this can be explained by the observed distribution, as most workers stated to be either carnivores or omnivores. However, for the use of recipe websites and number of times of cooking meals at home, this observation is somewhat surprising, as one would expect these people (since more familiar with recipes) to be able to estimate the nutritional properties of online recipes more correctly than those who do not.

6 Discussion and Conclusion

In this paper, we investigated different aspects of nutrition prediction, comparing algorithmic approaches with estimations performed by humans. The answers to our research questions can be summarized as follows:

- *RQ1.* The overall healthiness of a recipe, in terms of FSA health score, can be predicted quite accurately, as well as calorie content. All cues are more or less useful. Interestingly, image cue features and recipe title features, as pro-

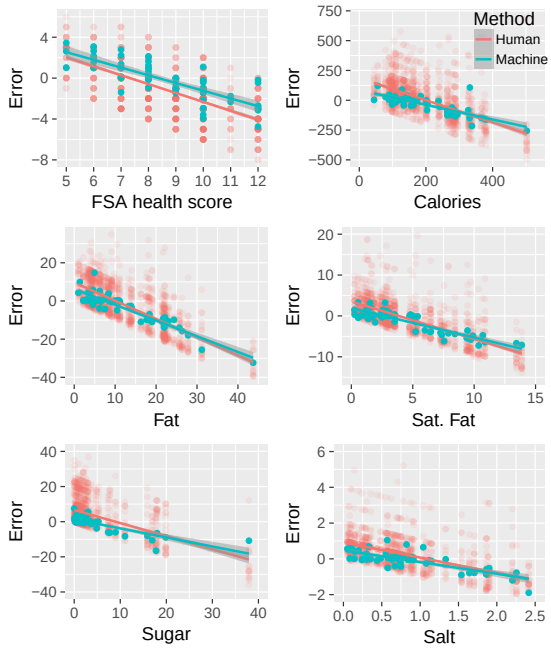

Figure 2: Machine and human estimation errors for different ranges in and FSA health score, calories, fat, sat. fat, sugar and salt.

posed by latest research in the area, were only of limited use.

- *RQ2.* Data-driven methods significantly outperform non-trained humans in estimating nutritional properties. However, the difference is lower for estimating the overall healthiness of a recipe than for sugar, salt or fat content and the overall estimation error is lower as well. Users mainly rely on the ingredient list and the image associated with the recipe, but are most accurate when also relying on user feedback, which hints at additional information not captured by the other cues. Performance for easy, medium and hard recipes is correlated.
- *RQ3.* Differences in demographics and culinary background have an impact on human rating performance. For instance, workers older than 35 were better in estimating the FSA health score, and workers who enjoy cooking better in estimating sugar content. No differences due to dietary preferences or user locations were observed.

In earlier studies (Müller et al. 2012), the performance of machine learning methods for the prediction of nutritional values was shown to equal the performance of human experts. Our results show that, compared to average users, our regression models perform far better. Furthermore, they show similar tendencies, most likely because they take the same features (most of them provided by human authors) into account. This implies that automated estimates are reliable and useful enough to stand in as a substitute for human support, should no nutritional expert be available.

Unfortunately, there is a positive correlation between human and machine performance regarding the difficulty of recipes, probably because they largely rely on the same cues. This would imply that data-driven nutrition estimation is least reliable in situations where humans need it most. For this reason, as future work it would be useful to further investigate which cues are typically used by human experts and to compare data-driven methods with dictionary-based approaches.

In our study, we focused on (all types of) main dishes, as these are the most composite type of dishes and therefore hardest to estimate. Similar approaches could be used for estimating the nutritional properties of breakfast meals, lunches or snacks. It should be noted that our models have been trained with features derived from a recipe website that is mainly populated by native English speakers, most importantly US citizens. It is likely that this focus has biased our results at least to some extent to Western, so-called WEIRD[19] cuisine. We expect that, despite known culinary differences, our observations can also be generalized to non-Western, non-English recipe websites. Preliminary results on two EU-based websites show that this is the case.

To summarize, in cases where nutrition indicators are not available, most users now already would be helped with automatic estimates. Moreover, if for these estimates similar features are used as those considered by humans, these features can be used for automatic explanation and justification of how these estimates were created, for example by comparing a meal or single ingredients with similar, healthier or unhealthier alternatives. Such explanations and justifications would not only contribute to transparency and acceptance of automatic nutrition explanation (Tintarev and Masthoff 2007), but would also serve as nudges for more informed food choices (Elsweiler, Trattner, and Harvey 2017).

Open Science To make the results obtained in this work reproducible, we have made the used data available under https://github.com/rokickim/nutrition-prediction-dataset.

Acknowledgements This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under project GlycoRec (16SV7172).

References

Abbar, S.; Mejova, Y.; and Weber, I. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proc. of CHI '15*.

Ahn, Y.-Y.; Ahnert, S. E.; Bagrow, J. P.; and Barabási, A.-L. 2011. Flavor network and the principles of food pairing. *Scientific reports* 1.

Almiron-Roig, E.; Solis-Trapala, I.; Dodd, J.; and Jebb, S. A. 2013. Estimating food portions. influence of unit number, meal type and energy density. *Appetite* 71:95–103.

Balasubramanian, S. K., and Cole, C. 2002. Consumers' search and use of nutrition information: The challenge and promise of the nutrition labeling and education act. *Journal of marketing* 66(3):112–127.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

[19] https://www.theatlantic.com/daily-dish/archive/2010/10/western-educated-industrialized-rich-and-democratic/181667

Burton, S., and Andrews, J. C. 1996. Age, product nutrition, and label format effects on consumer perceptions and product evaluations. *Journal of consumer affairs* 30(1):68–89.

Carels, R. A.; Harper, J.; and Konrad, K. 2006. Qualitative perceptions and caloric estimations of healthy and unhealthy foods by behavioral weight loss participants. *Appetite* 46(2):199–206.

Caswell, J. A., and Padberg, D. I. 1992. Toward a more comprehensive theory of food labels. *American Journal of Agricultural Economics* 74(2):460–468.

Chokr, M., and Elbassuoni, S. 2017. Calories prediction from food images. In *AAAI*, 4664–4669.

Cowburn, G., and Stockley, L. 2005. Consumer understanding and use of nutrition labelling: a systematic review. *Public health nutrition* 8(1):21–28.

De Choudhury, M.; Sharma, S.; and Kiciman, E. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proc. of CSCW '16*, 1157–1170.

Downs, J. S.; Loewenstein, G.; and Wisdom, J. 2009. Strategies for promoting healthier food choices. *The American Economic Review* 99(2):159–164.

Elbel, B.; Gyamfi, J.; and Kersh, R. 2011. Child and adolescent fast-food choice and the influence of calorie labeling: a natural experiment. *International journal of obesity* 35(4):493–500.

Elsweiler, D.; Trattner, C.; and Harvey, M. 2017. Exploiting food choice biases for healthier recipe recommendation. In *Proc. of SIGIR '17*, 575–584.

Erickson, J., and Slavin, J. 2015. Total, added, and free sugars: are restrictive guidelines science-based or achievable? *Nutrients* 7(4):2866–2878.

ESHA. 2016. Nutrition labeling software. available at http://www.esha.com/. last accessed on 20.6.2016.

Fried, D.; Surdeanu, M.; Kobourov, S.; Hingle, M.; and Bell, D. 2014. Analyzing the language of food on social media. In *Proc. of Big Data '14*, 778–783.

FSA. 2016. Guide to creating a front of pack (fop) nutrition label for pre-packed products sold through retail outlet. available at https://www.food.gov.uk/sites/default/files/multimedia/pdfs/pdf-ni/fop-guidance.pdf. last accessed on 31.6.2017.

Guthrie, J. F.; Fox, J. J.; Cleveland, L. E.; and Welsh, S. 1995. Who uses nutrition labeling, and what effects does label use have on diet quality? *Journal of Nutrition education* 27(4):163–172.

Kusmierczyk, T., and Nørvåg, K. 2016. Online food recipe title semantics: Combining nutrient facts and topics. In *Proc. of CIKM '16*, 2013–2016.

Lazer, D.; Pentland, A. S.; Adamic, L.; Aral, S.; Barabasi, A. L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323(5915):721.

McKean, E. 2010. *The new oxford American dictionary*, volume 3. Oxford University Press New York.

Meyers, A.; Johnston, N.; Rathod, V.; Korattikara, A.; Gorban, A.; Silberman, N.; Guadarrama, S.; Papandreou, G.; Huang, J.; and Murphy, K. P. 2015. Im2calories: towards an automated mobile vision food diary. In *Proc. of ICCV '15*, 1233–1241.

Müller, M.; Harvey, M.; Elsweiler, D.; and Mika, S. 2012. Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *Proc. of PervasiveHealth '12*, 73–80. IEEE.

Neuhouser, M. L.; Kristal, A. R.; and Patterson, R. E. 1999. Use of food nutrition labels is associated with lower fat intake. *Journal of the American dietetic Association* 99(1):45–53.

Pitler, E., and Nenkova, A. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proc. of EMNLP '08*, 186–195.

Pouladzadeh, P.; Villalobos, G.; Almaghrabi, R.; and Shirmohammadi, S. 2012. A novel svm based food recognition method for calorie measurement applications. In *Proc. of ICME '12 Workshops*, 495–498. IEEE.

Roberto, C. A.; Larsen, P. D.; Agnew, H.; Baik, J.; and Brownell, K. D. 2010. Evaluating the impact of menu labeling on food choices and intake. *American journal of public health* 100(2):312–318.

Rokicki, M.; Herder, E.; Kusmierczyk, T.; and Trattner, C. 2016. Plate and prejudice: Gender differences in online cooking. In *Proc. of UMAP '16*, 207–215.

Rokicki, M.; Herder, E.; and Trattner, C. 2017. How editorial, temporal and social biases affect online food popularity and appreciation. In *Proc. of ICWSM '17*, 192–200.

Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proc. of CHI '10 extended abstracts*, 2863–2872.

Said, A., and Bellogín, A. 2014. You are what you eat! tracking health through recipe interactions. In *Proc. of RSWeb '14*.

Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Ofli, F.; Weber, I.; and Torralba, A. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proc. of CVPR '17*.

San Pedro, J., and Siersdorfer, S. 2009. Ranking and classifying attractiveness of photos in folksonomies. In *Proc. of WWW '09*, 771–780.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sonnenberg, L.; Gelsomin, E.; Levy, D. E.; Riis, J.; Barraclough, S.; and Thorndike, A. N. 2013. A traffic light food labeling intervention increases consumer awareness of health and healthy choices at the point-of-purchase. *Preventive medicine* 57(4):253–257.

Steptoe, A.; Pollard, T. M.; and Wardle, J. 1995. Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite* 25(3):267–284.

Strohmaier, M., and Wagner, C. 2014. Computational social science for the world wide web. *IEEE Intelligent Systems* 29(5):84–88.

Sudo, K.; Murasaki, K.; Shimamura, J.; and Taniguchi, Y. 2014. Estimating nutritional value from food images based on semantic segmentation. In *Proc. of Ubicomp '14 Adjunct*, 571–576. ACM.

Tintarev, N., and Masthoff, J. 2007. A survey of explanations in recommender systems. In *Proc. of ICDE '17 Workshops*, 801–810.

Trattner, C., and Elsweiler, D. 2017a. Estimating the healthiness of internet recipes: A cross sectional study. *Frontiers in Public Health*.

Trattner, C., and Elsweiler, D. 2017b. Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems. In *Proc. of WWW '17*, 489–498.

Wansink, B., and Cashman, M. 2006. Mindless eating.

West, R.; White, R. W.; and Horvitz, E. 2013. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proc. of WWW '13*, 1399–1410.

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1-3):37–52.