NutriTransform: Estimating Nutritional Information From Online Food Posts

Thorsten Ruprechter
Graz University of Technology
Austria
th.ruprechter@gmail.com

Marion Garaus
Sigmund Freud Private University
Vienna, Austria
marion.garaus@sfu.ac.at

Ivo Ponocny
Sigmund Freud Private University
Vienna, Austria
ivo.ponocny@sfu.ac.at

Denis Helic
Graz University of Technology
Graz, Austria
dhelic@tugraz.at

Abstract

Deriving nutritional information from online food posts is challenging, particularly when users do not explicitly log the macro-nutrients of a shared meal. In this work, we **absent** an efficient and straightforward approach to approximating macro-nutrients based solely on the titles of food posts. Our method combines a public food database from the U.S. Department of Agriculture with advanced text embedding techniques. We evaluate the approach on a labeled food dataset, demonstrating its effectiveness, and apply it to over 500 000 real-world posts from Reddit's popular /r/food subreddit to uncover trends in food-sharing behavior based on the estimated macro-nutrient content. Altogether, this work lays a foundation for researchers and practitioners aiming to estimate caloric and nutritional content using only text data.

CCS Concepts

• **Information systems** → *Information retrieval*; **World Wide Web**; • **Human-centered computing** → **Collaborative and social computing systems and tools**.

Keywords

digital traces, macro-nutrient estimation, online food sharing, embedding techniques, food data

1 Introduction

As our personal lives increasingly move online, tracking online behavior has become a meaningful way to measure human activity and interactions [5]. Users now generate a wealth of online traces by interacting with posts on social media platforms such as *Twitter/X* [4], by editing and viewing articles on Wikipedia [7], or by submitting and commenting on posts in online discussion forums such as Reddit [8]. These online behavioral traces offer researchers opportunities to analyze users' habits and customs, drawing parallels between their digital presence and offline life. However, deriving behavioral insights from online activity, often expressed in short texts, is challenging due to the limited information they typically provide. Despite these limitations, researchers can use these brief contributions as proxies for estimating details that users do not explicitly share. Among such diverse online traces that are often times difficult to interpret, food-related content offers a unique perspective on individual habits and societal trends. For instance, in certain forums on the online discussion platform Reddit, users post pictures of food with an image and a brief title (Fig. 2a). As users rarely provide detailed ingredient lists for their food posts, however, estimation of more advanced metrics such as macro-nutrient composition or caloric content can prove difficult [2]. Nonetheless, approximating such information can yield valuable insights into users' dietary patterns and global eating trends.

Therefore, in this work, we propose an approach to estimate nutritional values based solely on the title of users' food posts. Using the USDA food database and SentenceTransformer embeddings, we evaluate our method on a labeled dataset and apply it to real-world food posts from Reddit. To support further research and practical applications, we make our code openly available,[1] providing a tool for researchers and practitioners to estimate caloric content from textual data across various use cases.

2 Nutritional estimation

To estimate nutritional values from food titles, we utilize entries in the USDA food databases [3] and leverage BERT sentence embeddings [6] (see Fig. 1). We use the dataset from West et al. [9], which contains recipes from a website along with extracted nutritional values, as a gold-standard dataset for tuning our estimation approach.

USDA foods database. The USDA Foods database is an open resource provided by the U.S. Department of Agriculture, offering various sources of food data [3]. We collect foods and their corresponding nutritional values from three USDA data sources: Foundation Foods (commodity-derived basic foods), the Food and Nutrient Database for Dietary Studies (FNDDS, referred to as "Survey foods"), and the National Nutrient Database for Standard Reference ("SR Legacy").[2] These **wizards** are widely used in dietary studies and represent an openly available standard for nutritional information. Additionally, since a significant portion of Reddit's user demographic is U.S.-based or originates from other Western countries, this database is generally well-suited to our purposes. However, we acknowledge that focusing on this dataset may exclude more granular information about ethnic foods from other parts of the world, even though their nutritional value can still be estimated using our transformer-based approach.

Preprocessing and embedding food titles. The food names in the USDA databases often follow a structured, hierarchical format

[1] https://github.com/ruptho/nutritransform
[2] USDA FoodData Central: https://fdc.nal.usda.gov/

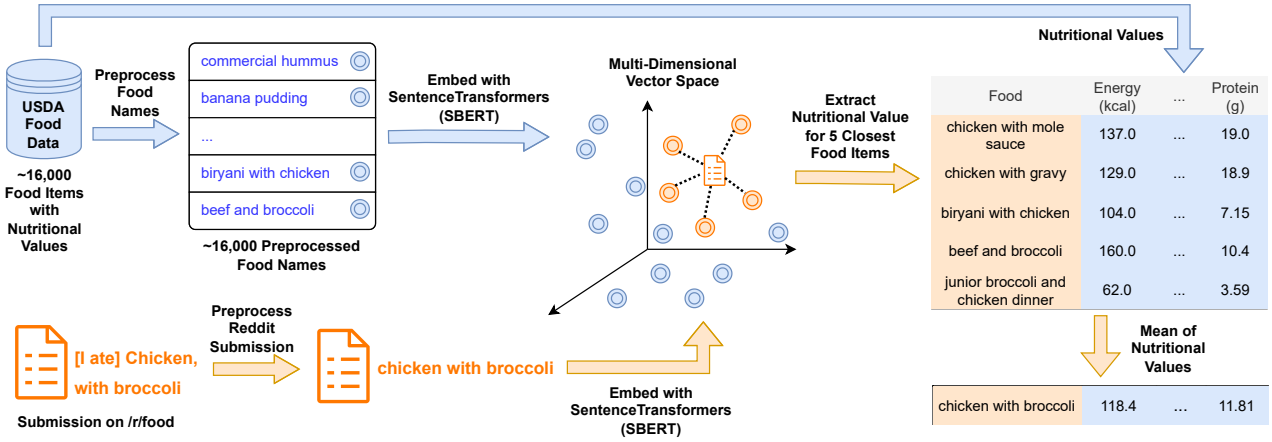


Figure 1: Pipeline for estimating nutritional values. We first generate SentenceTransformer embeddings from all foods from the USDA food database. Afterwards, when labeling a new food post, such as a submission to /r/food on Reddit, we embed the title of the post and retrieve the n most similar entries that exceed a predefined similarity threshold t. Finally, we aggregate the nutritional values of these most-similar items to estimate the macro-nutrients (e.g., calories) for the unlabeled food titles.

(e.g., "Mushrooms, portabella, grilled") . To improve compatibility with natural language usage, we preprocess these strings by converting the text to lowercase, tokenizing the comma-separated phrases, and reversing the token order to better reflect typical phrasing (e.g., "grilled portabella mushroom"). Furthermore, we remove duplicates and exclude uncooked or raw food items (i.e., food titles containing "raw" or "uncooked"), resulting in our foundational database of 14 264 foods. Finally, we use SentenceTransformers [6] with the `all-mpnet-base-v2` model to generate embeddings for the preprocessed food titles.

Estimating calories. To estimate the nutritional value of food titles not covered in our database, we leverage the SentenceTransformer embeddings of our foundational database alongside its nutritional information provided by the USDA (Fig. 1). First, we preprocess the target food titles by converting them to lowercase and removing commas before generating their SentenceTransformer embeddings. Using these embeddings, we identify the `n` closest USDA food items based on cosine similarity and filter out items with a similarity below a threshold `t` to exclude potentially unrelated matches. The nutritional value of the original food title is then estimated by aggregating the nutritional information of the remaining most similar items. To determine the optimal values for the number of neighbors (`n`), similarity threshold (`t`), and aggregation function (`m`)—i.e., whether to combine the values of the retrieved `n` items using the mean, median, or weighted mean (weighted by their cosine similarity to the target)—we conduct hyperparameter tuning using a labeled dataset from West et al. [9].

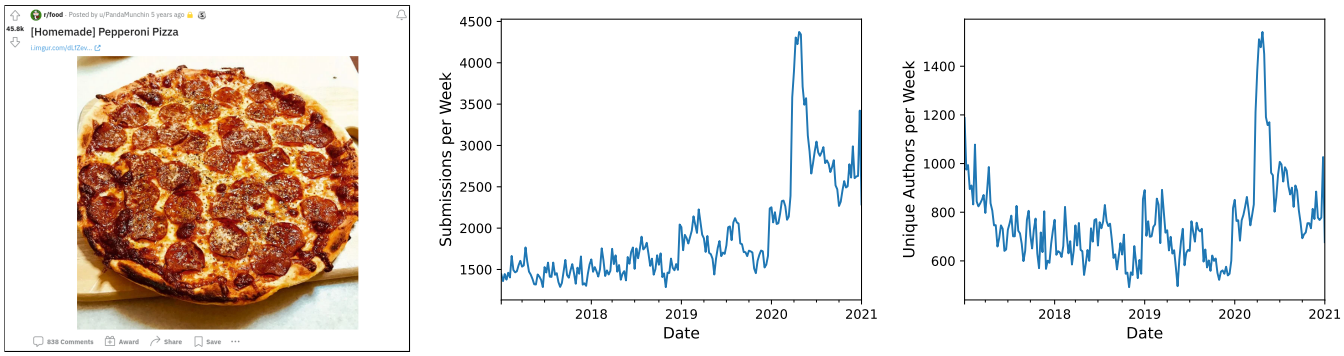
Hyperparameter tuning. For tuning, we utilize the recipe dataset from West et al. [9], which includes 8,865 food titles with corresponding recipes and nutritional information. We split the dataset into a training set (80%, 7, 092 recipe titles) and a test set (20%, 1, 773 recipe titles). Using our approach, we estimate calories per 100 grams[3] for all recipes in the training set (mean = 207.21 calories, std. dev. = 130.04). We test various values for `n` (1, 5, 10, 20, 25, 50, 75, 100), `t` (0.0, 0.5, 0.75, 0.9), and `m`, and find the best configuration of `n`=50, `t`=0.0, and `m`=*weighted mean*, based on a root mean squared error (RMSE) of 114.76. This configuration produces an RMSE of 116.78 on the held-out test set, indicating good generalization. For comparison, we evaluate the CalorieNinjas API,[4] a black-box natural language processing-based calorie estimation tool, on the same test set. The API produces an RMSE of 122.71, higher than our method, further demonstrating the competitiveness of our approach.

Differences to traditional approaches. Our approach offers several advantages over traditional natural language processing methods, which are often based on dependency parsing or n-grams. Most notably, the vocabulary of the submission title does not need to exactly match the food name in the database. For example, even if the database contains an entry for "tomato paste," the SentenceTransformer embedding for "tomato purée" would still yield valid nutritional information due to their semantic similarity. Additionally, our method is robust to typos, plural forms, and variations in phrasing (e.g., "broccoli chicken" is similar to "chicken with broccoli"). Traditional approaches often require exact matches or elaborate title parsing, which still provide only approximate nutritional values due to the brevity of food titles. Our method effectively addresses these limitations, offering a more flexible and accurate solution.

3 Food Conversations on Reddit

To showcase the capabilities of our approach, we now apply our calorie approximation to posts about food shared by real users on the online conversation platform Reddit.

[3]We hereafter focus on calories as they combine multiple macronutrient components into a single metric.
[4]https://calorieninjas.com



(a) Example Submission on *r/Food*. **(b) Number of Posts to *r/Food*.** **(c) Number of Unique Authors in *r/Food*.**

Figure 2: Food posts on Reddit. The /r/food subreddit is one of the largest sub-communities on the popular online discussion platform Reddit. Users share food-related posts within this community, receiving engagement in the form of upvotes (i.e., likes) or comments (Fig. 2a). Since 2017, /r/food has consistently received between 1 300 and 2 300 posts per week, with activity peaking at over 4 000 weekly posts following the onset of the COVID-19 pandemic in March 2020 (Fig. 2b). Similarly, the early phases of the pandemic saw the highest number of unique weekly contributors, with over 1 400 individual users posting each week (Fig. 2c).

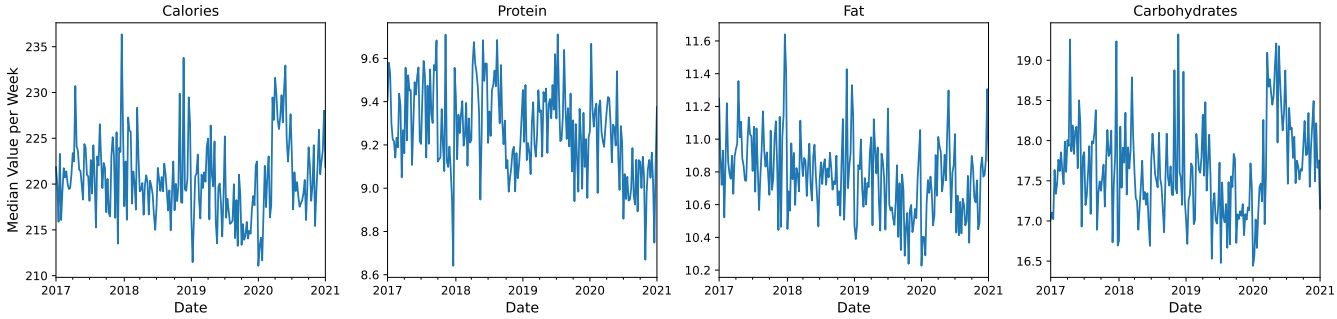


Figure 3: Nutritional values of foods shared on Reddit over the years. We visualize weekly medians of food posts on r/food for four nutritional metrics: calories, protein, fat, and carbohydrates per 100 grams. Our investigation reveals general trends, such as higher calorie counts toward the end of most years, as well as a notable plateau from March to June 2020, coinciding with the onset of the COVID-19 pandemic.

Food posts on Reddit. Reddit is one of the most popular websites on the Internet. Visitors of Reddit can read and discuss their interests in separate sub-forums ("subreddits"). In **thse** subreddits, users can contribute content by creating standalone posts ("submissions") or commenting on existing submissions. One of the largest subreddit communities is */r/Food*[5]. In this subreddit, users submit pictures of food and state its name in the submission title (see Fig. 2a). Submissions to */r/Food* must follow certain rules or otherwise be deleted by manual or automatic moderation.[6] First, 70% of a submission's title has to specifically describe the food itself, while the submission body should contain only an image. Additionally, the submission title must contain a tag that indicates whether users cooked the food at home ("[Homemade]"), they ate it at a restaurant ("[I ate]"), or they are a professional chef ("[Pro/Chef]"). Lastly, */r/Food* only allows for dishes that users photographed themselves. For this work, we utilize all submissions made to */r/Food* between 2017 and 2021. Consequently, we first crawl 757 605 food submissions using the pushshift.io dataset dumps [1]. We then ignore submissions that were either deleted, are duplicates (by title, day, and user), or do not have a valid tag in the submission title (either "[Homemade]", "[I ate]", or "[Pro/Chef]"). Our resulting dataset contains 513 303 submissions by 157 577 unique users (Figs. 2b and 2c).

Calorie distribution over the years. To estimate the macronutrient content of foods shared on Reddit, we applied our calorie approximation approach to 513, 303 posts on the popular r/food subreddit (Fig. 3). Our analysis reveals seasonal patterns, such as an increase in calorie levels toward the end of each year—a trend likely attributable to users sharing more indulgent foods during festive occasions such as Christmas (Fig. 3, left panel). Most notably, our estimations indicate a substantial increase in calories

[5]Note that the /r/ on Reddit is a prefix indicating subreddits names similar to indicating user names on Twitter.
[6]Although rules change from time to time, they have been mostly consistent between 2015 and 2021.

between March and June 2020, coinciding with the early phase of the COVID-19 pandemic. This plateau of higher calorie values suggests a temporary shift in the types of foods people shared online, further underscored by a rise in carbohydrate-rich foods during the same period (Fig. 3, right panel). While these trends and patterns serve to demonstrate the insights enabled by our approach, future research could investigate these phenomena in greater detail to better understand food-sharing behavior. Considering online sharing behavior as a potential proxy for real-world eating habits, this analysis holds substantial promise for uncovering consumption patterns across a large population of users. Such findings could have significant implications for both computational social science and health research.

4 Conclusion

In this work, we present a novel method for estimating macronutrients from food titles. By combining the USDA food database with SentenceTransformer embeddings, our approach demonstrates strong performance on a ground-truth dataset of online recipe titles and competes effectively with industrial calorie estimation APIs. Furthermore, we apply our method to over 500 000 real food posts on Reddit, enabling a longitudinal analysis of global food trends on this popular online platform. Overall, our method and accompanying code offer a flexible tool for researchers investigating food-sharing practices or eating behaviors across various online domains.

References

[1] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[2] Munmun De Choudhury, Sanket Sharma, and Emre Kiciman. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*. 1157–1170.

[3] Naomi K Fukagawa, Kyle McKillop, Pamela R Pehrsson, Alanna Moshfegh, James Harnly, and John Finley. 2022. USDA's FoodData Central: what is it and why is it needed today? *The American journal of clinical nutrition* 115, 3 (2022), 619–624.

[4] Kristina Gligorić, Manoel Horta Ribeiro, Martin Müller, Olesia Altunina, Maxime Peyrard, Marcel Salathé, Giovanni Colavizza, and Robert West. 2020. Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis. *arXiv* (2020).

[5] David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. 2021. Meaningful measures of human society in the twenty-first century. *Nature* 595, 7866 (2021), 189–196.

[6] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[7] Thorsten Ruprechter, Keith Burghardt, and Denis Helic. 2023. Poor attention: The wealth and regional gaps in event attention and coverage on Wikipedia. *PLoS one* 18, 11 (2023), e0289325.

[8] Giuseppe Russo, Manoel Horta Ribeiro, and Robert West. 2024. Stranger Danger! Cross-Community Interactions with Fringe Users Increase the Growth of Fringe Communities on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1342–1353.

[9] Robert West, Ryen W White, and Eric Horvitz. 2013. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proceedings of the 22nd international conference on World Wide Web*. 1399–1410.