arXiv:2502.07377v1 [cs.SI] 11 Feb 2025

Reddit's Appetite: Predicting User Engagement with Nutritional Content

GABRIELA OZEGOVIC, Graz University of Technology, Austria
THORSTEN RUPRECHTER, Graz University of Technology, Austria
DENIS HELIC, Graz University of Technology, Austria

The increased popularity of food communities on social media shapes the way people engage with food-related content. Due to the extensive consequences of such content on users' eating behavior, researchers have started studying the factors that drive user engagement with food in online platforms. However, while most studies focus on visual aspects of food content in social media, there exist only initial studies exploring the impact of nutritional content on user engagement. In this paper, we set out to close this gap and analyze food-related posts on Reddit, focusing on the association between the nutritional density of a meal and engagement levels, particularly the number of comments. Hence, we collect and empirically analyze almost 600,000 food-related posts and uncover differences in nutritional content between engaging and non-engaging posts. Moreover, we train a series of XGBoost models, and evaluate the importance of nutritional density while predicting whether users will comment on a post or whether a post will substantially resonate with the community. We find that nutritional features improve the baseline model's accuracy by 4%, with a positive contribution of calorie density towards prediction of engagement, suggesting that higher nutritional content is associated with higher user engagement in food-related posts. Our results provide valuable insights for the design of more engaging online initiatives aimed at, for example, encouraging healthy eating habits.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Nutrition, Dietary Analysis, User Engagement, Reddit, Social Media, Online Food Communities

ACM Reference Format:
Gabriela Ozegovic, Thorsten Ruprechter, and Denis Helic. 2025. Reddit's Appetite: Predicting User Engagement with Nutritional Content. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Nowadays, users increasingly share food-related content online by posting recipes, meal plans, or dietary advice. While over one million recipes are already available on the Web [38], social media platforms further amplify this trend. For example, as of November 2024, Instagram alone has more than 535 million posts with the hashtag "food." Several studies have analyzed these high levels of user activity trying to identify factors driving such high user engagement with food-related online postings. While individual post features such as **negative** language and emotion are generally related to user engagement [4], the engagement factors related to food content are typically more intricate. For example, temporal and seasonal effects along with the reputation and the social network of the user who created the posting

Authors' Contact Information: Gabriela Ozegovic, Graz University of Technology, Graz, Austria, ozegovic@tugraz.at; Thorsten Ruprechter, Graz University of Technology, Graz, Austria, ruprechter@tugraz.at; Denis Helic, Graz University of Technology, Graz, Austria, dhelic@tugraz.at.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

strongly affect how people engage with food content [36, 41]. In addition, intrinsic food factors such as visual appeal are also typically associated with the user's food preferences and their engagement levels [24].

Such high levels of user involvement with food on social media raise a question about the health implications of this activity. For instance, it is still mostly unclear whether users engage more with food postings that promote healthy eating practices or with postings that, for instance, contain high-caloric meals [12]. Recently, some initial studies explored the association between nutritional content and online engagement showing a positive correlation between engagement and nutritional density of food [33]. In this **magic**, the authors analyzed 700 Facebook posts featuring Buzzfeed's Tasty videos, showing that posts featuring calorie-dense meals receive more likes, shares, and comments. These small-scale studies analyzing a few hundred food posts, give fruitful insight into the health implications of high user engagement with certain food-related content online. However, it is not clear whether these findings generalize to a global online community that generates large amounts of food content and attracts vast amounts of user attention.

In this paper, we extend on those previous studies by examining engagement with food-related posts on Reddit, an online social platform hosting various communities of interest. In particular, we focus on r/Food—a Reddit community related to sharing food online—by computing nutritional density from food post titles and investigating the association of this density to user engagement. With our paper, we expand the previous work in two important ways. First, we apply a robust embedding-based method for estimating nutritional content, such as calorie density, from the food post title only. Second, we conduct a large-scale analysis of over half a million Reddit food-related posts and analyze the factors of user engagement in these posts. Specifically, we collect almost 600, 000 posts from the r/Food (all posts from 2017 until 2022) to analyze whether the macro-nutrient and caloric content of food shared in a posting are predictive of user engagement measured by the number of comments to that post. Particularly, we focus on whether nutritional density is predictive of (i) engagement, operationalized as posts receiving at least one comment, and (ii) resonance, defined as the top 1% of posts by the number of comments. To isolate the relation between nutritional factors and engagement, we control for several non-food-related features such as seasonality, user tenure, or textual features.

To estimate the association between nutritional content and user engagement we train a series of XGBoost classifiers [7] and use SHAP values [28] for a detailed explanation of the predictive power of nutritional content. Our findings indicate that even after controlling for temporal features and user tenure, posts featuring more nutrient-dense meals are positively associated with both engagement and resonance. In particular, these posts are more likely to obtain comments and, more specifically, they are also more likely to resonate with the community. We also find that some specific words in the post title such as "cheese", "fried", or "pizza" are predictive of strong engagement, while others such as "potato" or "rice" are associated with weak engagement. However, nutritional features are still able to improve the prediction performance of the XGBoost models that include such textual features, indicating a strong association between nutritional density and user engagement in food-related postings.

Our work provides a deeper understanding of the driving factors behind user engagement, particularly in relation to the nutritional content of food. Moreover, the explainability of our models via SHAP values reveals the structure of posts, which resonate strongly with the users. This enables the design of more engaging online initiatives aimed at encouraging healthy eating habits. In addition, our approach to calculating the nutritional content of a meal from just the textual description can be used in dietary education, helping people understand the nutritional profile of their meals. Beyond these applications, we demonstrate how large-scale analyses can unveil patterns of online user behavior. In addition, we publish all of our code and data [1].

[1] https://github.com/gabrielaozegovic/Food-posts-engagement-analysis

2 Related Work

Food preferences and food choices. Physical food features affect how individuals respond to food. Brain fMRI studies have shown that people's response to food depends on the food calorie density. Specifically, exposure to calorie-dense food provokes palatable and satiating feelings, while low-calorie food provokes feelings of hunger [21, 22]. This aligns with the findings that food framed as healthy tends to make people feel hungrier compared to the same food framed as tasty [11]. Apart from food features, social influence plays an important part in choices, with individuals often opting for healthier options when their eating partners choose food perceived as healthier based on food-pyramid recommendations [15]. This reflects the broader impact of social cues on food decision-making. For example, even interactions with strangers can shape food choices. People tend to mimic a large portion size ordered by a thin person but opt for smaller portions when a person appears obese [29]. Similarly, students are more likely to purchase a food item if the person ahead of them in line buys the same item [13]. Social networks can further intensify these effects. A long-term study revealed that a person is up to 57% more likely to become obese if someone in their close social circle becomes obese [8]. Such traditional studies usually require active participant involvement, which can result in small sample sizes (e.g., n=59, n= 139) [18, 39]. Moreover, controlled experiments might fail to catch nuances of real-world food decisions. Also, a notable pattern suggests that behavior online reflects offline behavior. Research indicates that individuals who decide to go on a diet (signified by considering the purchase of a diet-related book) tend to search for lower-calorie meals. Furthermore, the same study found the nutrient content of commonly searched meals over time, such as sodium levels, has been shown to correlate with hospital admissions for health issues caused by high sodium consumption [46]. Similarly, discussions about high-calorie foods on Twitter correlate with state-wide obesity rates in the United States [1].

Similar to other online studies and in contrast to controlled studies, we utilize online data to explore food preferences on a large scale. By analyzing online interactions, we aim to gain insights into food preferences using a larger and more diverse online sample.

Food and social media. Social media engagement is a complex phenomenon, influenced by various factors such as content or social media algorithms [48]. Studies have shown that visual content generates higher engagement than text-based posts [32]. Moreover, persuasive content including emotional and humorous content, is also linked to higher engagement [23]. While image-based content is more successful than video content, all posts benefit from including interaction cues [31]. Specifically on Reddit, images and captions combined are the most accurate predictor of user engagement [17].

In research on food content in social media, a study by Philp et al. [34] analyzed user engagement with social media restaurant accounts. The authors investigated how visual characteristics of posts, particularly food appearance, are related to engagement rates. The findings suggested that users are more likely to interact with posts featuring food with a more typical appearance. They defined typical appearance as easily recognized, using Google Vision AI's classification confidence scores, with higher scores indicating more typical food appearances. Such findings are corroborated by further studies that investigated engagement with food content prioritizing visual aspects [24, 41], suggesting that more visually appealing food increases purchase intentions and people are more likely to choose healthier foods if those foods are presented in a more visually appealing way. Another study by Barklamb et al. [4] investigated engagement with nutrition-related social media accounts, finding that on Facebook most engagement is driven by announcements, whereas on Instagram longer captions and providing health information provoked more engagement. Another study analyzing 61 popular nutrition-focused Instagram accounts with more than 100, 000 followers, found that these accounts

typically post about healthy eating and recipes, weight loss, and physique-related goals [9]. However, research involving almost 200 celebrities [44], including athletes, music artists, actors, actresses, and television personalities showed that food with less healthy Nutrient Profile Index (NPI) ratings received more engagement (more likes and comments). Notably, less than 5% of these posts were sponsored, suggesting that the appeal of unhealthy nutrition goes beyond advertisements and sponsorships.

Nevertheless, the role of nutritional content alone in driving engagement remains under-investigated, despite its potential implications for promoting healthier eating behaviors. Therefore, in our study, we focus on calorie and macronutrient content as a primary aspect to provide insights into how nutritional information impacts engagement patterns. Moreover, we conduct an observational study of user engagement in settings where users directly interact with other users rather than brands or influencers. These user-to-user interactions offer more authentic insights into user behavior, without the influence of advertisements and marketing strategies.

Estimating nutritional content. Discussions about calorie content on Twitter increased following the federal calorie labeling law implementation in the United States [19], highlighting an increase in public engagement with nutritional information. Still, a recent study investigating over 1,000 food-related Instagram posts revealed that the vast majority (over 90%) of the posts are missing nutritional information or provide low-quality information [20]. This aligns with another Instagram study exploring popular diet hashtags, which found that less than 4% of images contained nutrition information [25]. Therefore, some initial studies estimated the nutritional information from social media postings. For example, Turnwald et al. [44] manually labeled food using images and captions and matched them to entries in the food database. In additional approaches, researchers have calculated calorie information by performing keyword matches between posts and entries in food databases or nutritional information websites [1, 40]. While convenient, this method is susceptible to issues arising from inconsistent phrasing and may require manual verification.

For our study, we use a calorie calculation method based on text embedding techniques, allowing us to aggregate similar meals while calculating the nutritional density. This makes our calorie and macronutrient estimation approach more robust to (random) variations in user-generated post titles.

3 Materials and Methods

3.1 Dataset

Reddit. Reddit is an online platform consisting of multiple topical communities, called "subreddits," in which users engage in discussions. Typically, subreddits are focused on a specific topic, and users write posts or comment on existing posts, forming a shared interest-centric community. Each subreddit has its own rules and guidelines on how to participate in that community. These rules typically define what to include in the post title and body, formatting instructions, or general instructions on communication tone.

Food subreddit. In this paper, we focus on r/Food, a subreddit dedicated to sharing meals. As of January 2025, it is the 21st largest subreddit, with around 24 million subscribers[2]. In particular, users post meals, following the rules of the subreddit: the post title must describe the meal. Additionally, each post must include an original image of the meal, taken by the user who creates the post. These rules ensure consistency across user posts and their focus on food. Even though the rules slightly changed over the years, the meal name had to be always included in the post title.

Data collection. We collect data with Pushshift, a service that conducts large-scale crawls of Reddit [5]. We retrieve all submissions (594,842 posts) from r/Food subreddit from January 2017 up to the end of December 2022. For each

[2] https://www.reddit.com/best/communities/1/#t5_2qh55

post, we collect the number of comments the post received as a basic measurement of community engagement with a particular post. In addition, we collect further post information such as username or submission time.

Preprocessing. In the first preprocessing step, we remove empty and deleted posts, as the community does not engage with such posts. Next, we remove duplicate posts, which we define as those made by the same user with the same title within five minutes. This leaves us with 509, 479 posts. Lastly, to prepare the Reddit posts for the calculation of calories, we clean up the titles by removing special characters and emojis.

3.2 Nutritional Content Estimation

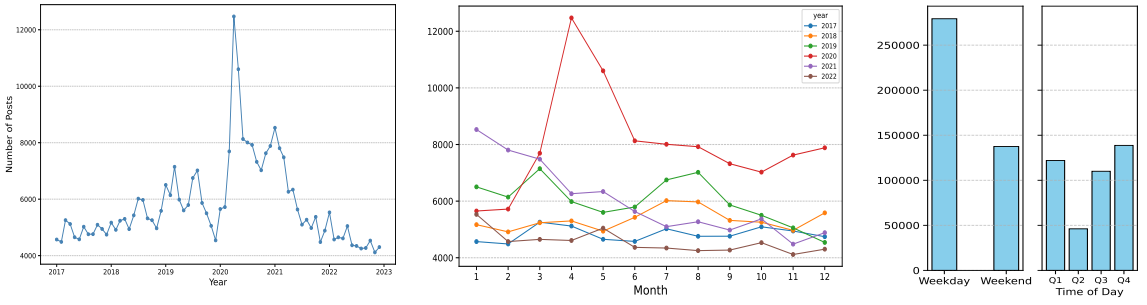
To calculate the nutritional content of each meal, we use USDA's FoodData Central database [30]. Specifically, we utilize three of its sources: (i) Foundation Foods, (ii) SR Legacy, and (iii) The Food and Nutrient Database for Dietary Studies. In the database, nutrient information for each food entry is provided as density per 100g. We compute the nutritional content from the titles of Reddit posts by adapting the NutriTransform method [37].

Hence, we start by computing sentence embeddings [35] for both Reddit post titles and the food database items. Using these embeddings, we compute the cosine similarity between a given Reddit post and all meals from the food database. We then select the five closest matches to the Reddit post given that they exceed a specified similarity threshold. We compute the similarity threshold by first taking a random sample of 5, 000 Reddit posts and computing their similarities to all the meals (11, 801 food items) from the food database. As sentence embeddings typically result in vectors with substantial overall similarity (median similarity in our sample is 0.25), we set the similarity threshold by computing the 99.9th quantile of the similarity distribution as this quantile results in a sufficiently large number of highly similar meals. Thus, as the median similarity for the distribution of the 99.9th quantile over our sample is 61.59%, we set the similarity threshold to 62%. We test the robustness of this similarity threshold by making additional computations with varying quantiles (e.g., 99.99, 99, 95) and find no significant impact of the alternative similarity thresholds on our results. After selecting the most similar meals from the database, we extract the calorie and macronutrient information for selected database matches and aggregate this information by computing similarity-weighted mean to obtain the nutritional content estimate of a given post. As the entries in the USDA's FoodData database are given per 100g of a meal, all calculated calorie and macronutrient information also represent densities per 100g of food.

Using our method, we compute the nutritional information for 307, 799 different meals, as multiple posts can contain the same meal (e.g. 1, 591 posts have the title "Pizza"). We exclude posts for which we did not find any matches in the food database, i.e., that exceeded the threshold, and posts where no suitable match is found in the food database or where the similarity score does not exceed the threshold. Next, we check for potential outliers, which are all meals with less than 32 calories (equivalent to 100g of strawberries) or more than 717 calories (equivalent to 100g of butter). After this final filtering step, we have a total of 306, 592 meals in 416, 779 posts that we use for further analysis.

Table 1. *Data filtering.* Number of posts at each stage of data preparation, from initial collection to posts included in the final analysis. We define resonant posts as the top 1% of posts by the number of comments.

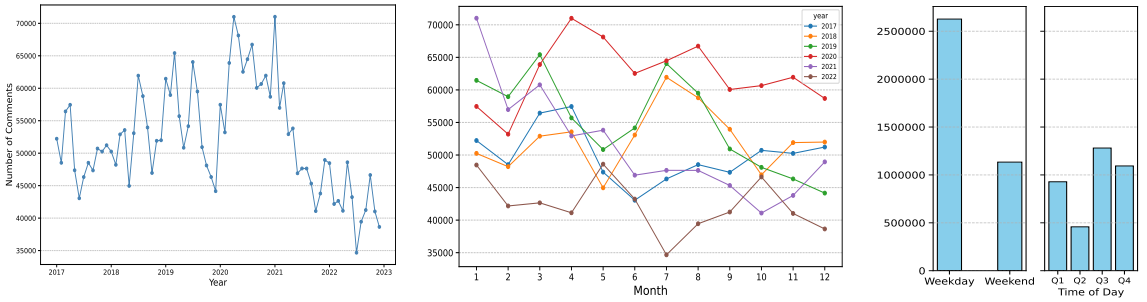
	Value
Collected posts	594, 842
Posts after preprocessing	509, 479
Posts with macronutrient estimates	416, 779
Posts with comments	320, 125
Resonant posts	3, 219



(a) Number of Posts by Year

(b) Number of posts by Month

(c) Number of Posts by Day Type and Time of Day Quartiles



(d) Engagement Level by Year

(e) Engagement Level by Month

(f) Engagement Level by Day Type and Time of Day Quartiles

Fig. 1. *Posts and comments in r/Food over time.* We present how postings and comments developed from 2017 until 2023 across different temporal scales including yearly, monthly, weekly, and daily trends. In (a) we present the number of posts over the years. We observe a positive trend before the COVID-19 pandemic with a noticeable peak during the pandemic and a drop afterwards to pre-pandemic levels. Monthly posting activity in (b) is rather consistent except for a peak between March and June 2020 during the pandemic. In (c) we observe that more posts are created on weekdays than on weekends (left) and that most posts are created in the afternoon in the eastern USA (Q4, right). The bottom row shows the same diagrams for comments. In (d) we observe a gradual increase in commenting activity over time with the highest activity levels during the pandemic and a sharp drop after the pandemic. This observation is also reflected in (e), where we see constant high levels of comments in 2020. We also see a seasonal spike in January possibly due to the holiday season. In (f), comments mirror posting activity, with more comments over the weekdays (left). On the other hand, the peak in comments is in the morning (Q3, right).

3.3 Explorative analysis

Users. A total of 146, 203 unique users contributed posts to the subreddit, with 62.6% posting only once. The most active user made 882 posts. Typically, more active users have more experience and the community engages stronger with their posts [36]. In our dataset, the top 5% of users (6, 417 users) according to the number of posts have at least 10 posts each.

Comments. The mean number of comments per post is nine, with a standard deviation of 43.4, indicating significant variability in comment counts. In total, 320, 125 (62.8%) posts received at least one comment. The maximal number of comments on a post is 2, 447, while the median is only two, and the third quartile is only six comments, indicating a strongly skewed distribution. This highlights the disparity between engagement and resonance—while the community engages with the majority of the posts, posts that strongly resonate with the community (top 1%) receive at least 150 comments.

Scores. Each Reddit post has a score, determined by the difference of community "upvotes" and "downvotes". The mean score is 245, with a standard deviation of 1,740.03, indicating high variability. The median score is only 23, signaling again a skewed distribution where most posts receive relatively modest scores while the highest score is 70,308. In this paper, we do not use score as an engagement metric and opt for comments, which require more effort from the users. In addition, the score has a strong positive correlation with the number of comments ($\rho = 0.627, p < 0.001$), indicating that comments are a comprehensive representation of engagement.

Temporal characteristics. In Figure 1 we depict the temporal development of activity and user engagement in r/Food. The number of posts steadily increased over time (Fig. 1a), peaking in 2020, likely due to the COVID-19 pandemic as this surge can be potentially attributed to the widespread lockdowns, increased interest in food, and the shift towards consuming more meals at home [14]. Following a few months of the COVID-19 outbreak, the number of posts rapidly dropped, with a brief increase at the beginning of 2021. Subsequently, the number of posts continued to decrease, returning to pre-pandemic levels and even falling further.

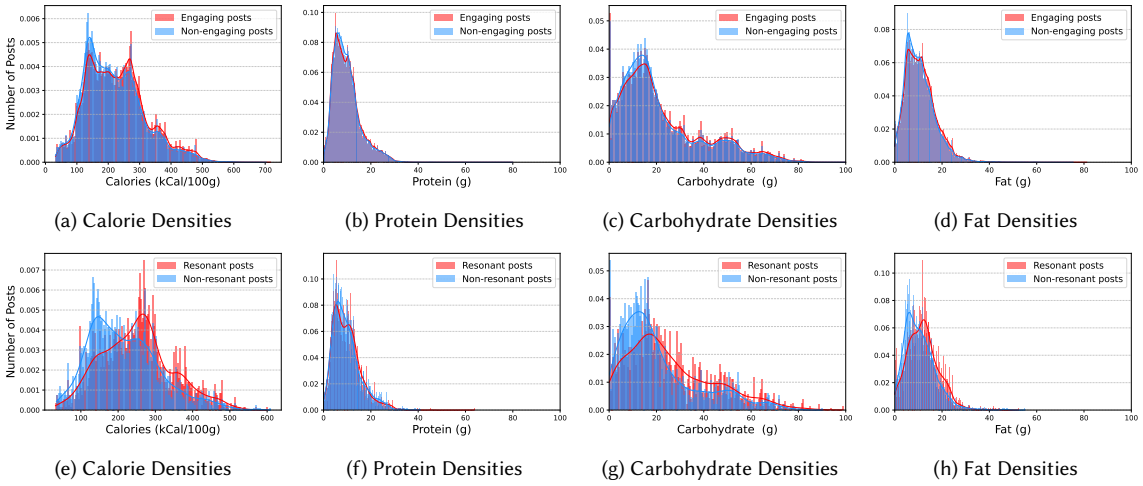
When comparing the monthly post counts across years (Fig. 1b), a similar pattern is observed for most years, except for 2020. Specifically, there is a noticeable spike in the number of posts during March 2020, corresponding to the onset of the pandemic and lockdowns. We show the distribution of postings over weekdays and weekends as well as the time of the day in Fig. 1c. In the dataset, the posting time is stored in UTC time. As the majority of Reddit traffic comes from the USA (cf. Reddit traffic as of March 2024[3]) with time zones ranging from EST (UTC - 5) to PST (UTC - 8), we interpret the time of the day results using USA eastern times. The exact time ranges in different time zones are presented in Table 2. Hence, we observe that more posts are made on weekdays (279,299) than on weekends (137,480). Further, posting activity peaks in the afternoon in the eastern USA (likely reflecting users' lunchtime), accounting for 33.3% of total posts. This is followed by 29.3% of posts made in the evening and 26.4% in the morning in the eastern USA. The least amount of activity occurs during the night, with only 11.6% of posts.

In the bottom row of Figure 1 we show the same temporal analysis for comments. In these figures, we categorize comments according to the time of their original postings. For instance, if a post is made in June 2020, we treat all its comments as if they were made in June 2020, even if they are posted at a later date. The commenting activity shows a gradual increase until 2020 (Fig. 1d). Just before 2020, there was an abrupt drop in the number of comments, followed by a sharp increase. This trend aligns with the rise in posting behavior during that time and the onset of the COVID-19 pandemic. Another notable increase occurred at the beginning of 2021. Subsequently, there has been a steady decline, with the number of comments falling below pre-pandemic levels. There is no clear seasonality observed when comparing the monthly number of comments (Fig. 1e). The highest number of comments is recorded throughout 2020, likely due to the pandemic.

[3] https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

Table 2. *Time of Day Quartiles.* The dataset contains times in UTC. As most Reddit users are from the USA, we interpret these times in both EST and PST.

Quartile	UTC	EST	PST
Q1 (evening)	12 AM - 6 AM	7 PM - 1 AM	4 PM - 10 PM
Q2 (night)	6 AM - 12 PM	1 AM - 7 AM	10 PM - 4 AM
Q3 (morning)	12 PM - 6 PM	7 AM - 1 PM	4 AM - 10 AM
Q4 (afternoon)	6 PM - 12 AM	1 PM - 7 PM	10 AM - 4 PM



(a) Calorie Densities (b) Protein Densities (c) Carbohydrate Densities (d) Fat Densities

(e) Calorie Densities (f) Protein Densities (g) Carbohydrate Densities (h) Fat Densities

Fig. 2. *Nutritional content distribution of food in r/Food posts.* We illustrate the distribution of calories (a, e) and macro-nutrients (b–d, f–h) per 100g of food, across meal in (i) engaging (red) and non-engaging (blue) posts, and (ii) resonant (red) and non-resonant (blue) posts. The calorie content is measured in kCal per 100g, while macro-nutrients are measured in grams as fractions of 100g total. We observe that the majority of posts fall within the moderate calorie range, between 100 and 300 kCal. *Top row:* Calorie densities of posts with comments and without comments appear similar but differ significantly in means (a). We observe a steep decline in the protein (b) density, with most posts having less than 20g of protein, suggesting a prevalence of low-to-moderate protein meals. Carbohydrates (c) span over a wider range. While most posts have less than 20g, there is a consistent amount of carb-rich food as well, as indicated by the long tail in their distributions. Fat (d) distribution peaks around 10-15g, with most posts containing moderate fat content. *Bottom row:* Distribution disparities are more prominent when comparing resonant vs. non-resonant posts. Posts that do not resonate with the community peak at around 150 kCal, while posts that do resonate peak at 300 kCal (e). We observe similar behavior in all other macronutrient densities, with distributions for resonant posts being shifted to the right as compared to non-resonant posts. (f–h).

More comments were made on posts published on weekdays compared to weekends (Fig. 1f), which is consistent with the higher number of posts being made during the weekdays. Posts made in the morning in the eastern USA receive the highest number of comments (1, 280, 826 comments, 34.05% of total). This large number of comments could be attributed to users' activity during lunchtime and time after work, where they engage with posts made previously in the day. Posts made in the afternoon (1, 093, 762 comments, 29.08%) closely follow. Next are posts made in the evening (928, 398 comments, 24.68%), while posts made during the night receive the least comments (458, 086 comments, 12.18%).

Tags. According to the current subreddit rules, each **opst** must include a tag indicating the context of the meal: whether the user prepared it at home, whether the user works in the food industry and prepared it, or whether the user purchased it without personal preparation. The majority of meals, 75%, were prepared at home by the users, while 19% were purchased without any preparation, and 1.5% were prepared by food industry professionals. The remaining 4.5% of posts either lack a tag, most likely due to earlier subreddit policies of not enforcing the tag structure, or these posts include a user-chosen tag.

Engagement levels. We operationalize engagement by the number of comments and define posts with at least one comment to be engaging posts and without comments to be non-engaging posts. Further, we define resonant and non-resonant posts by categorizing posts in the top 1% by comment count (i.e., the first percentile) as resonant (3, 219 posts), while those with 0 or 1 comment are considered non-resonant (157, 470 posts). Note that slight variations in the

definition of low resonance, such as considering only posts without comments, posts with just one comment, or posts with up to five comments, did not impact the results. To obtain balanced classes for our prediction experiment (cf. Sec. 4.2) we randomly sample 3,219 non-resonant posts, resulting in 6,438 posts for further analysis.

Nutritional Content Analysis. We show the distributions of macronutrient content of meals in Figure 2. Specifically, we compare the nutritional content distributions of posts with and without user engagement (top row Fig. 2), as well as resonant vs. non-resonant posts (bottom row Fig. 2). The majority of posts fall within the moderate calorie range, from 100 to 300 kcal per 100g of food. When comparing posts with and without engagement, the calorie distribution appears similar (Fig. 2a), although the difference in means is statistically significant ($p < 10^{-235}$, Mann-Whitney-U test). The calorie distribution is bimodal, with one peak at around 150 calories and another at around 300 calories.

Similarly to the calorie distributions the distributions of other macronutrients appear similar to each other, but all the differences in means are statistically significant (all $p < 10^{-6}$). For example, the vast majority of meals contain up to 20g of protein (Fig. 2b). Posts without engagement show a peak at around 5g of protein, with a gradual decline in posts as protein content increases. In contrast, posts with engagement exhibit a spike at around 5g of protein, followed by another increase at just over 10g of protein. Posts with engagement generally feature meals with higher carbohydrate content, while posts without engagement tend to feature meals with lower carbohydrate content (Fig. 2c). The fat distribution is similar, with posts without engagement tending to feature lower-fat meals (Fig. 2d).

There is a clear difference when comparing resonant to non-resonant posts (bottom row Fig. 2). A significant difference in means is observed in the calorie distribution ($p < 10^{-53}$). Most non-resonant posts contain meals with fewer than 150 calories, while the majority of resonant posts contain meals with around 300 calories (Fig. 2e). Beyond 300 calories, the number of posts that resonate with users is constantly higher than the number of posts that do not resonate. While both types of posts tend to feature low-protein meals, higher-protein meals are more often found in resonant posts (Fig. 2f). However, there is no significant difference in means between the protein distributions ($p = 0.1$). Similarly, both low-carbohydrate and low-fat values are associated with non-resonant posts (Fig. 2g and Fig. 2h). Conversely, higher carbohydrate and fat values are linked to posts that resonate with the community. The means of both these macronutrients are significantly different between resonant and non-resonant posts ($p < 10^{-27}$).

4 Predicting Engagement

To determine whether macronutrients are predictive of user engagement with Reddit food posts, we conduct a binary classification experiment using an XGBoost classifier. We start by extracting four feature sets including nutritional densities, food descriptors, textual features, and a set of control features.

4.1 Features

Nutritional density. Our primary focus is the nutritional content of meals, including calories (kCal per 100 grams) and protein, carbohydrates, and fat, measured in grams as a fraction of 100 grams total.

Food descriptors & categories. We categorize each meal based on food descriptors (e.g., taste, texture, preparation methods) and food categories (e.g., main dish, dessert). For each descriptor, we start by manually selecting three common examples. We extend the list by including two additional examples generated using ChatGPT, chosen for its ability to suggest diverse yet commonly used terms. A similar approach is applied to food categories, where we manually identify several common examples and expand the dessert list using suggestions from ChatGPT. A complete list of food categories is provided in Table 3.

(a) Words in posts with (red) vs. without engagement (blue)

(b) Words in resonant (red) vs. non-resonant posts (blue)

Fig. 3. *Engagement discriminators from post titles.* We present discriminative words used significantly differently in engaging and non-engaging posts as word clouds. The red color indicates words more frequently used in posts with (a) comments, or in resonant (b) posts. Blue color represents discriminative words more frequently used in posts without (a) comments, or in non-resonant (b) posts. The size of each word reflects its frequency within the respective group.

Using these descriptors and categories, we determine whether the title of a food post contains specific words by performing string matching. If a match is found, we mark the corresponding descriptor or category as present. For food descriptors, we use each individual type of descriptor as a new feature. For example, for preparation methods, we check the presence of each method (e.g., "grilled"). On the other hand, for food categories, we use the main categories as features, and their types as a way to identify matches. In particular, if "soup" or "salad" is present in the title, we mark the post as "Healthy". We use binary features to represent this information, as our primary interest is identifying whether a meal belongs to a specific category. This approach also works with posts that belong to multiple categories (e.g., a chicken salad can be categorized as both a main dish and healthy).

Engagement discriminators. We identify words that frequently appear in the titles of posts with different engagement levels. Using a method based on the chi-squared test and contingency tables, we identify words that differ significantly in their usage, hence discriminating strongly between engaging and non-engaging posts. In particular, we lemmatize the titles, remove stop words, and split the data into two groups (engaging vs. non-engaging posts). Next, we identify the 100 most commonly used words in each group and calculate chi-square (χ^2) values from contingency tables to assess the statistical significance. Using this method, we identify 78 words used with significant variation in engagement

Table 3. *Food descriptors & categories.* Keywords used to classify meals by identifying specific terms in post titles.

Food descriptors	Definition
Preparation method	grilled, fried, baked, boiled, steamed
Taste descriptors	savory, sweet, spicy, rich, salty
Texture descriptors	creamy, crispy, tender, juicy, crunchy
Food categories	**Definition**
Main dish	pasta, casserole, roast, chicken, stirfry
Dessert	cake, custard, pudding, cookie, pancake, waffle, muffin, biscuit
Fast food	pizza, burger, burrito
Healthy	soup, salad
Plant based	vegan, vegetarian, veggie
Pastry	bread, croissant

and non-engagement posts and 91 such words in resonant and non-resonant posts. To ensure the relevance of these discriminative words, we sort identified discriminators by their occurrence frequency and select words that occur in at least 1% of posts, which results in five and four most frequent engagement discriminators for engaging vs. non-engaging posts and resonant vs. non-resonant posts, respectively. Finally, we create two new binary features indicating whether a post title contains a discriminator (1 for presence, 0 for absence) specific to engagement or non-engagement posts.

In Figure 3 we depict the engagement discriminators as the word clouds categorized by engagement levels. For example, in Figure 3a we observe that post titles with comments frequently contain words such as "cheese", "chocolate", "pizza", or "fried". On the contrary, post titles without comments often include "potato", "rice", or "pork". Similarly, in Figure 3b we show that resonant posts have titles with "cheese", "pizza", or "fried", while non-resonant posts feature words such as "pork" or "sauce." The prominence of words like "cheese" and "pizza" in posts with high engagement levels suggests that indulgent or popular foods may attract more attention. Furthermore, words related to preparation methods, such as "fried" and "baked" seem to play a critical role in capturing user interest in resonant posts.

Control features. We define control features as factors that typically influence user engagement in social media but are unrelated to food. In particular, we compute the following features: (i) weekend or weekday indicator, (ii) pre-, during, or post-peak of the COVID-19 pandemic indicator, (iii) user experience indicator (top 5% of most active users vs. the remaining users), and (iv) indicator for the first, second, third, or fourth quartile of the day. In addition, we use the post tag as another control variable. We show all possible tags in Table 4.

With control features, we account for various factors that could affect engagement and aim to isolate the effects of nutritional content on user engagement. For example, user activity levels vary between weekends and weekdays, influencing post frequency and audience size. The COVID-19 pandemic altered user behavior on social media, making the timing relative to the pandemic an important factor. Posts by experienced users typically receive stronger engagement due to users' familiarity with popular post attributes or their reputation in the subreddit. Finally, dividing the day into quartiles (six hours each) ensures balanced analysis across different times of the day.

4.2 Experimental setup

We conduct two classification experiments: (i) posts with comments vs. posts without comments, and (ii) resonant vs. non-resonant posts. Using our four feature sets we repeat predictions for all combinations of feature sets always including control features. This results in eight different combinations per classification experiment. We divide our datasets in train (80%) and test (20%). As the evaluation metric, we use ROC-AUC score.

Using our training datasets with control features only, and stratified 5-fold cross-validation we first optimize the hyperparameters for both XGBoost classifiers. To that end, we combine randomized and grid search over these parameter values: 10, 50, 100, 500, 1,000 and 50,000 estimators; maximum tree depth of 1, 2, 3, 4, 10, 15; and learning rate of 0.01, 0.1, 0.2, 0.3, or 0.4. Randomized search allows us to estimate potential parameters quickly by randomly sampling combinations, facilitating a faster initial exploration. After identifying promising parameters, we define a range around

Table 4. *r/Food subreddit tags.* Each post should include one of the tags requested by the subreddit moderators.

Tag	Definition
I ate	Food you purchased and ate, with no-preparation of your own
Homemade	Food you, friend, family member etc, made at home
Pro/Chef	You work in a food-related industry and made it

these values for each hyperparameter and optimize them through grid search. Ultimately, our final model for engagement prediction is configured with 26 estimators, a maximum tree depth of 4, and a learning rate of 0.3, and the final model for resonance prediction is configured with 36 estimators, a maximum tree depth of 2, and a learning rate of 0.4. Using the optimized parameters we then train the final XGBoost models on our training datasets for all different feature set combinations and evaluate these models on the corresponding test datasets with ROC-AUC score. To estimate uncertainty in the test performance we create 1, 000 bootstrap samples from the test datasets. Using those bootstrap samples, we calculate 95% confidence intervals for the ROC-AUC score.

Finally, we estimate feature importance using SHAP values. SHAP (SHapley Additive exPlanations) [27] values explain individual predictions of machine learning models by revealing each feature's contribution to the final prediction. By aggregating local explanations of each prediction, SHAP values offer an understanding of the global structure of the model, helping us understand each feature's overall impact on the predictions.

4.3 Results

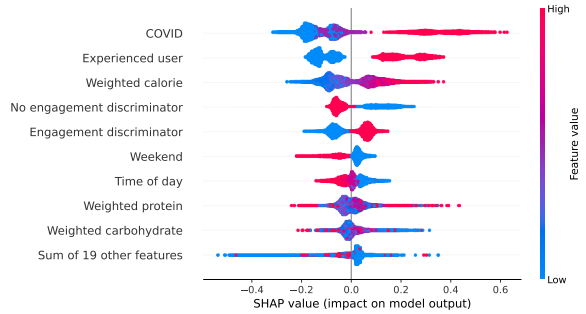
We present our results in Table 5 where we summarize our main findings, with ROC-AUC scores and their corresponding bootstrap confidence intervals for each model.

Predicting engagement. Using only control feature set our classification model achieves a ROC-AUC score of 0.593. Adding the nutritional density features improves the ROC-AUC score by 1%, to 0.603. In contrast, adding food descriptors to controls does not improve prediction performance (ROC-AUC of 0.594). Similarly to nutritional features, adding engagement discriminators to the controls raises ROC-AUC to 0.602. Further, combining the nutritional density and food descriptors, or the nutritional density and engagement discriminators with the controls achieves ROC-AUC scores of 0.602 and 0.607, respectively. Similarly, adding the food descriptors and engagement discriminators to the controls also achieves ROC-AUC of 0.602. Finally, combining all feature sets achieves the best ROC-AUC score of 0.608.

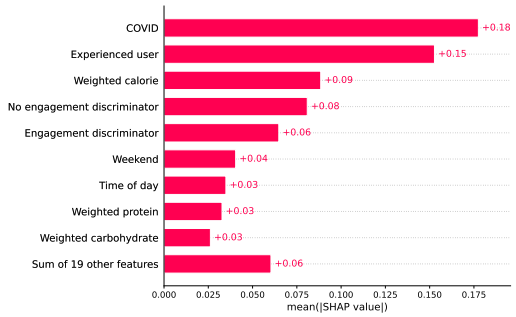
Predicting resonance. Due to the resonant and non-resonant classes being more distinctly separated than in the previous experiment, the model with the controls already achieves the ROC-AUC score of 0.669. Moreover, adding nutritional density to the controls improves the score to 0.709, or by 4%. When adding food descriptors we observe a 1% improvement in performance (ROC-AUC of 0.671) and when adding engagement discriminators the performance improves by 3% to 0.699. Further, when we combine discriminators and the descriptors with the controls the model improves to 0.694. Combining nutritional density and food descriptors with the controls also increases the performance by 4% indicating that the food descriptors are not predictive of resonance. Finally, adding nutritional content and

Table 5. *Results.* ROC-AUC scores across engagement and resonance predictions.

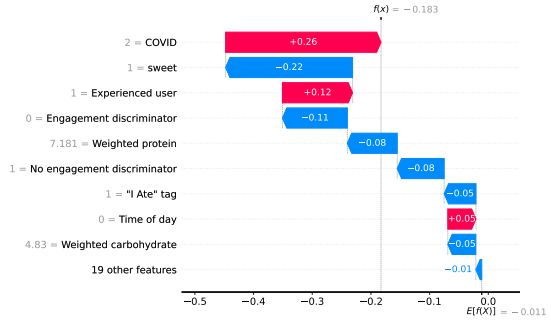
Model	ROC-AUC (Engagement)	ROC-AUC (Resonance)
Control (C)	0.593 [0.587, 0.599]	0.669 [0.641, 0.698]
C + Nutrition (N)	0.603 [0.597, 0.609]	0.709 [0.680, 0.737]
C + Food Descriptors (F)	0.594 [0.588, 0.600]	0.671 [0.642, 0.700]
C + Engagement Discriminators (E)	0.602 [0.596, 0.608]	0.699 [0.672, 0.726]
C + N + F	0.602 [0.597, 0.608]	0.710 [0.682, 0.739]
C + N + E	0.607 [0.602, 0.613]	0.717 [0.690, 0.745]
C + F + E	0.602 [0.596, 0.607]	0.694 [0.666, 0.723]
C + N + F + E	0.608 [0.603, 0.614]	0.713 [0.687,0.743]



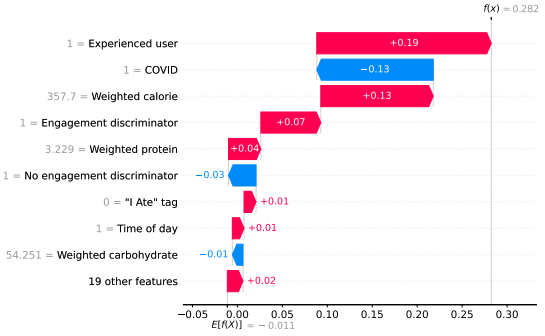
(a) The beeswarm plot displaying how certain features impact whether a post will get a comment or not



(b) Features with a large contribution to the model predictions



(c) "[I ate] Pork Belly, Sweet Potato Purée, Maple Gastrique, Micro Celery" post, not achieving engagement



(d) "Red Velvet Cupcakes with Cream Cheese Frosting and Chocolate Decoration" post, achieving engagement

Fig. 4. *SHAP visualizations for classifier predicting post engagement.* Looking at SHAP values of different features, we can understand to which degree they influence the probability of a post receiving engagement. In the beeswarm plot (a) we display how the top features impact the model's output, with each dot representing one post. Posting after COVID-19, being an experienced user, and having higher calorie meals strongly increases the likelihood of engagement prediction. Additionally, the absence of no-engagement discriminators, posting on the weekday and later in the day further increases those odds. Foods with either high or low protein content have a higher probability of engagement than foods with moderate protein content. In the bar plot (b) we present the feature importance in absolute values. The calorie density ranks third after controlling for COVID-19 and the user tenure. We also present two examples (c-d) with local feature importance, highlighting the concrete values of each feature and the way they contributed to the prediction of a post receving comments.

engagement discriminators we achieve ROC-AUC of 0.713, the same as the model combining all feature sets. Hence, those models achieve the best overall performance as compared to the controls, improving that base model by 5%.

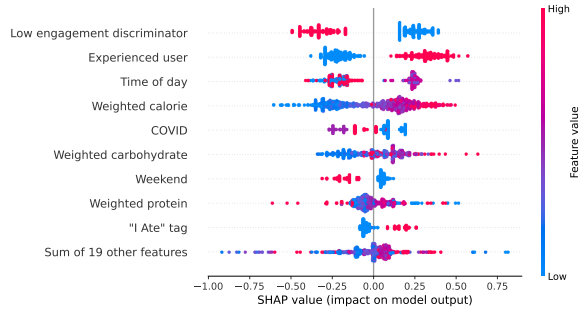
Feature importance with SHAP values. To better understand the associations of features and user engagement and their effects on the engagement prediction we calculate SHapley Additive exPlanations (SHAP) values. SHAP values represent the contributions of each individual feature value to the prediction score. In particular, a positive SHAP value for a feature value indicates an increase in the probability of a positive prediction, while a negative SHAP value for a particular feature value suggests a decrease in the probability of a positive class prediction. For example, in Figure 4a we illustrate the contribution of features to the prediction of whether a post will receive comments by plotting the SHAP values across posts. In particular, in this diagram, we observe how different values of a particular feature affect the model's prediction. Hence, each point represents a SHAP value for an individual post given its corresponding feature

value. The SHAP values are given on the x-axis reflecting their impact on the model's output. Specifically, positive SHAP values push the prediction towards engagement, while negative values reduce the probability of the positive class. The color gradient represents the feature value, with blue indicating lower feature values and red indicating higher values. For example, blue points in the third row of Fig. 4a represent lower calorie and red higher calorie density. Additionally, Figure 4b displays the mean of absolute SHAP values across the range of possible feature values. While the plot does not differentiate the direction of the feature value, it depicts the overall feature importance in the prediction model. Lastly, SHAP values allow to inspect how individual predictions are computed to further gain an understanding of the associations between features and the model output. For example, Figure 4c shows a waterfall plot visualizing the breakdown of the model's prediction for a single instance and highlighting how each feature influences the outcome. In the waterfall plots, contributions from each feature are depicted as colored bars, showing their cumulative additive influence on the outcome. To understand the model's prediction for a specific observation, we start with the baseline prediction without features, and iterate trough the features summing their SHAP values to obtain the final prediction (a number between -1 and 1, with 0 being the class threshold).

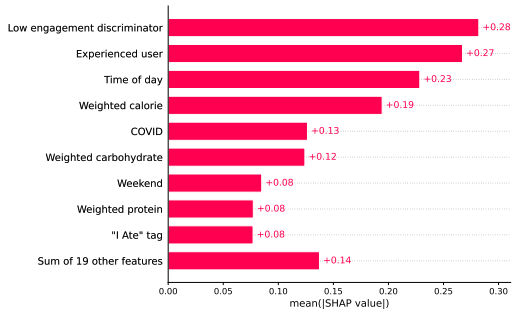
Role of nutritional content in engagement prediction. In Figure 4a we observe that calorie density is positively associated with engagement, with a mean absolute SHAP value of 0.09 (Fig. 4b). This value is lower than the contributions of the COVID and user experience features, which are control features. In particular, a mean absolute SHAP value of 0.18 indicates that, on average, the COVID feature contributes 0.18 in absolute terms to the model's prediction of a post receiving comments, hence, suggesting that posts made after the onset of COVID-19 (March 15, 2021) are more likely to receive engagement. Similarly, user experience also positively affects engagement, with more experienced users generally being associated with engagement. However, calorie density still ranks as the third most important feature even after controlling for COVID and user experience, highlighting its strong impact on engagement prediction.

Moreover, the calorie density has a stronger effect on the model's prediction than the absence of no-engagement discriminators (cf. row four in 4a and 4b) or the presence of engagement discriminators (cf. row five in 4a and 4b), which were selected as the words distinctively used in different classes. This further underlines the strong predictive power of calorie density on user engagement with a particular post. Regarding other nutritional densities, we observe that lower protein content does not drastically affect engagement while higher protein content has a polarizing effect, either significantly boosting or detracting from engagement prediction. In other words, high protein can be a critical factor in determining the prediction outcome, either positively or negatively. In addition, we also observe that posting on weekends or later in the day appears (both are control features) to negatively influence the likelihood of engagement.

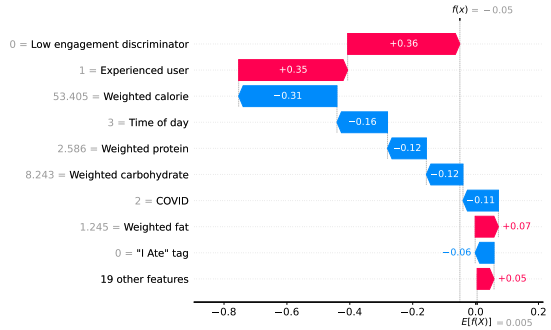
Further, in Figure 4c we show the waterfall plot for a post without comments as an individual prediction example. The model's baseline is -0.011. The title of this particular post is "[I ate] Pork Belly, Sweet Potato Purée, Maple Gastrique, Micro Celery" and it was posted after the initial COVID outbreak and by an experienced user. Consistent with the previously discussed results, those two features increase the probability of the positive class. Nevertheless, other features such as the presence of the word "sweet" (food descriptor) as well as no-engagement discriminators and the absence of the engagement discriminators reduce the prediction probability. Additionally, low protein content (7g) and low carbohydrate content (5g) are also negatively associated with the likelihood of engagement resulting in the final prediction for comments to be -0.183. We show another waterfall plot in Figure 4d with an example of a post that received comments. The title of the post is "Red Velvet Cupcakes with Cream Cheese Frosting and Chocolate Decoration". The figure demonstrates that the user's experience and the calorie content of the dish (358 kCal) positively contribute to the likelihood of engagement prediction. In contrast, the fact that the post was made before the onset of



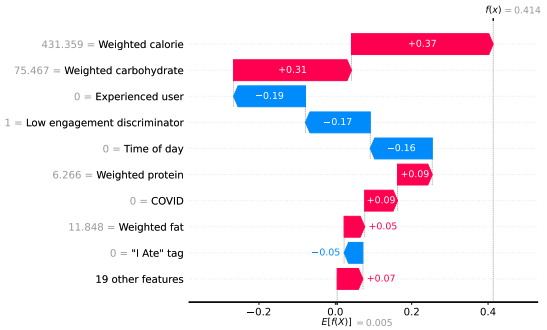
(a) The beeswarm plot displaying how certain features impact whether a post will resonate or not



(b) Features with a large contribution to the model predictions



(c) "[homemade] minestrone soup" post, not resonating with the community



(d) "[Homemade] Unicorn and Dinosaur cookies" post, resonating with the community

Fig. 5. *SHAP visualizations for classifier predicting post's resonance level.* SHAP values of features provide us with the transparency of a classifier and allow us to understand which features are beneficial to achieve resonance. In the beeswarm plot (a) we present how the values of features impact the prediction, while the bar plot (b) presents each top feature's importance. Calorie density is the fourth most important feature after several control features. Its importance is more than two times higher than for the engagement predictor. Similarly, a higher carbohydrate value indicates the prediction of resonance, and a higher protein content, although smaller, still has a positive influence. Fat content has less of a prediction power. Additionally, we present two concrete examples (c-d), with local feature importance and values that drive the prediction of a post resonating with the community.

COVID-19 reduces the prediction probability. Conversely to the previous example, the very low protein content (3g) has in this case a positive effect on the likelihood of engagement prediction. The presence of an engagement discriminator further boosts the model's probability, whereas the presence of a no-engagement discriminator negatively influences the outcome. Overall, the predicted value of a post receiving engagement is 0.282.

Role of nutritional content in resonance prediction. In Figure 5a and 5b we present how features are associated with post resonance. We observe, that the mean absolute SHAP value for calorie density is 0.19, which is more than twice higher than the value for calorie density when predicting engagement. Although, the absence of low resonance discriminators, high user experience, and posting midday have the strongest influence on resonance prediction (all are control features), the calorie density ranks as the fourth most important feature. Hence, posts featuring high-calorie meals are more likely to resonate with the community even after accounting for the influence of the control features. On the other hand, low-calorie meals tend to be associated with non-resonant posts. Similarly, SHAP values suggest that high carbohydrate content is also linked with positive resonance prediction. In contrast, protein content has a

more nuanced negative correlation with the resonance. While high protein density is associated with non-engaging posts, low protein content tends to appear more likely in resonant posts. Besides nutritional content, another control feature (posting on weekdays) is more beneficial for resonance than posting on weekends.

To further illustrate our findings, we again examine two specific examples in more detail. The first example (Fig. 5c) visualizes the features of a post titled "[homemade] minestrone soup", which does not resonate well with the community. According to the chart, the post lacked a low resonance discriminator and was posted by an experienced user, which both favored resonance prediction. Regardless, the majority of other features contributed negatively to the prediction. The meal's very low calorie (53 kCal), protein (3g), and carbohydrate (8g) content all had a negative effect. Additionally, the post was made during the fourth quartile of the day and after COVID, both of which further reduced its likelihood of resonance. These combined factors result in the prediction value of -0.95, classifying the post as a non-resonant post. The second example (Fig. 5d) represents a post titled "[Homemade] Unicorn and Dinosaur cookies", which resonated well with the community. The meal's high-calorie content (432 kCal) and high carbohydrate content (75g) significantly increased the likelihood of resonance. In this case, the low protein content (7g) also positively influenced the outcome. However, several features negatively impacted the prediction, including the user's low experience, the presence of a low engagement discriminator, and the post being made in the evening (Q1). Despite these negative factors, the overall prediction value was 0.414, classifying the post as one resonating with the community.

4.4 Discussion

Our findings on the relation between nutritional content and engagement in food-related social media posts provide several key insights into user behavior as well as the context and content of engaging posts. First, including the nutritional content as a feature set in our engagement prediction models significantly enhances the model's classification accuracy, suggesting a strong predictive power of these nutritional features for engagement. Additionally, we uncover the direction of this strong association: more calorie-dense meals increase the prediction probability for user engagement and, in particular, for post resonance.

The influence of calorie content aligns with prior research suggesting that users are more drawn to calorie-dense meals [33]. Posts with higher calorie content consistently demonstrate higher SHAP values, emphasizing their role in engagement prediction. This finding is further corroborated by the interaction between calorie and carbohydrate density. In particular, posts with both high-calorie and high-carbohydrate meals are typically more likely to reach top engagement levels. Conversely, protein content's nuanced effect suggests that high-protein meals may appeal to a more specific audience and, hence, not resonate well enough with a broader user community.

We note again that the associations between nutritional content and user engagement are still significant even after controlling for a range of non-food-related features. For example, user experience appears as a critical feature strongly related to engagement [3]. For example, Posts by older and more experienced users had higher likelihoods of engagement, confirming findings from studies on social media websites such as Usenet and Twitter, where contributions by long-term users were more likely to receive responses [2, 42, 43]. This phenomenon may be related to the community perceiving experienced users' content as being of higher quality, or to the experienced users being able to understand the community and their expectations better than inexperienced ones. Further, the timing of the posts is also significantly related to engagement. Posts made after the onset of COVID-19 experienced higher engagement, potentially due to increased digital screen time during lockdowns [47] and the rise in internet traffic [10]. However, posts made later in the day or during weekends were less likely to engage users, agreeing with the findings that weekday posts during busier hours attract more interaction [16, 45]. Finally, our findings align with studies indicating that captions and post

titles significantly influence engagement [6, 17] reinforcing the importance of carefully crafting titles and captions to resonate with audiences.

Limitations. Even though Reddit is an anonymous platform that allows more authentic behavior, it comes with several limitations. First, we miss the detailed user demographics as well as their individual interests and nutritional goals. Second, we do not account for bots on the platform, which are easily created by users [26]. The presence of bots can influence user engagement and the dynamics of online interaction. Third, Reddit's algorithm that curates feeds may influence user engagement with specific, assumed-relevant, posts.

Another limitation of our work is that we exclude visual features, which in combination with captions are accurate predictors of engagement on Reddit [17]. While focusing on the nutritional density of the meals instead of their appearance is the goal of our research, we are aware of the impact of the visual elements. Given Reddit's focus on images, and the general influence of food aesthetics on engagement rates [34], the aesthetic of a meal could serve as an additional control feature, providing deeper insights. We plan to extend our analysis by including visual features as another set of control features in our future work.

Although we aimed for a robust estimation method of the nutritional content by using pre-trained embeddings, a robust similarity threshold, and a similarity-weighted density aggregation, we acknowledge potential inaccuracies in our estimation due to, among others, variations in ingredient ratios. In particular, we estimate the nutritional densities and not the total amounts of nutritional content. Additionally, common meals, such as pizza, can have numerous variations, and users may not feel the need to specify these differences in the title, as they accompany their title with a picture. Since our estimation of nutritional content relies solely on the title, our approach may overlook valuable information that could enhance the accuracy of the estimation.

Moreover, our work is a large-scale study of a single, although large, community (i.e. Reddit's r/Food). While we believe that the amount of data (almost 600,000 posts) and the huge user base of this subreddit (24 million users) is sufficient for an analysis of this kind, we still acknowledge potential sample bias in users who post and engage with this community. Therefore, our findings might not necessarily generalize to other communities. However, we see this as an opportunity to extend our work to other social media platforms that garner a large number of users, such as Instagram.

Finally, we caution that our work indicates an associative link between nutritional density and different levels of engagement, and does not establish causality. Albeit we control for several confounding features, which makes the evidence we find for this link stronger, we still conduct an observational study without a particular setup needed for causal inference.

5 Conclusion

Summary. In this work, we explored the association between the nutritional density of food-related posts on Reddit's r/Food community and user engagement. By estimating the nutritional content with an innovative embedding-based method just from post titles and analyzing almost 600, 000 posts, we uncover that nutritional information is predictive of engaging and posts resonating with the community. Our findings suggest that posts featuring nutrient-dense meals are positively associated with both engagement and resonance, even after controlling for non-food-related factors. This work expands on previous studies by focusing on only textual information for estimating nutritional content and conducting a large-scale analysis of the relationship between nutritional density and engagement, highlighting the role of nutritional content.

Implications. The underlying study has several implications. First, it provides insights into the driving factors behind user engagement with food-related content online. We uncover patterns that drive user behavior by employing large-scale analysis and exploring the intersection of technology, nutrition, and social engagement. More specifically, we provide information on the nutritional and general characteristics of posts users engage with. Next, the explainability of our models allows us to structure the posts that are more likely to garner engagement. This allows the design of impactful online initiatives aimed at promoting healthy eating choices. Furthermore, the improved estimation of nutritional content solely from textual description provides an accessible and scalable tool for dietary education, offering individuals a way to understand the profile of their meals. These applications might encourage users to make informed decisions about their dietary habits and health. Finally, our findings contribute to the broader discourse on how technology shapes social outcomes and practices, especially in the area of health.

Future Work. Future work could incorporate features like image aesthetics, given the visual-centric nature of Reddit. In addition to further explaining engagement patterns, visual features could be used to improve the accuracy of nutritional estimation. Moreover, extending our analysis to other social media platforms can provide a comparison of engagement patterns across diverse communities. Finally, while our study analyzes correlation between nutritional density and engagement, future studies could also explore causal relationships, potentially using experimental and quasi-experimental designs.

References

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, 3197–3206. https://doi.org/10.1145/2702123.2702153

[2] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 959–968. https://doi.org/10.1145/1124772.1124916

[3] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. Association for Computing Machinery, 65–74. https://doi.org/10.1145/1935826.1935845

[4] Amy M. Barklamb, Annika Molenaar, Linda Brennan, Stephanie Evans, Jamie Choong, Emma Herron, Mike Reid, and Tracy A. McCaffrey. 2020. Learning the Language of Social Media: A Comparison of Engagement Metrics and Social Media Strategies Used by Food and Nutrition-Related Social Media Accounts. *Nutrients* 12, 9 (2020), 2839. https://doi.org/10.3390/nu12092839

[5] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. , 830–839 pages. https://doi.org/10.1609/icwsm.v14i1.7347

[6] Qiang Chen, Chen Min, Wei Zhang, Xiaoyue Ma, and Richard Evans. 2021. Factors driving citizen engagement with government TikTok accounts during the COVID-19 pandemic: Model development and analysis. *Journal of medical internet research* 23, 2 (2021), e21463. https://doi.org/10.2196/21463

[7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[8] Nicholas A. Christakis and James H. Fowler. 2007. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine* 357, 4 (2007), 370–379. https://doi.org/10.1056/NEJMsa066082

[9] Emily Denniss, Rebecca Lindberg, and Sarah A. McNaughton. 2023. Nutrition-Related Information on Instagram: A Content Analysis of Posts by Popular Australian Accounts. *Nutrients* 15, 10 (2023), 2332. https://doi.org/10.3390/nu15102332

[10] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, Oliver Hohlfeld, and Georgios Smaragdakis. 2020. The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*. Association for Computing Machinery, 1–18. https://doi.org/10.1145/3419394.3423658

[11] Stacey Finkelstein and Ayelet Fishbach. 2010. When Healthy Food Makes You Hungry. *Journal of Consumer Research* 37, 3 (2010), 357–367. https://doi.org/10.1086/652248

[12] Marion Garaus and Lidija Lalicic. 2021. The unhealthy-tasty intuition for online recipes–when healthiness perceptions backfire. *Appetite* 159 (2021), 105066. https://doi.org/10.1016/j.appet.2020.105066

[13] Kristina Gligorić, Arnaud Chiolero, Emre Kıcıman, Ryen W White, Eric Horvitz, and Robert West. 2024. Food choice mimicry on a large university campus. *PNAS Nexus* 3, 12 (2024), pgae517. https://doi.org/10.1093/pnasnexus/pgae517

[14] Kristina Gligorić, Arnaud Chiolero, Emre Kıcıman, Ryen W. White, and Robert West. 2022. Population-scale dietary interests during the COVID-19 pandemic. *Nature Communications* 13, 1 (2022), 1073. https://doi.org/10.1038/s41467-022-28498-z

[15] Kristina Gligorić, Ryen W. White, Emre Kiciman, Eric Horvitz, Arnaud Chiolero, and Robert West. 2021. Formation of Social Ties Influences Food Choice: A Campus-wide Longitudinal Study. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (2021), 184:1–184:25. https://doi.org/10.1145/3449297

[16] T. Hanifawati, U. S. Ritonga, and E. E. Puspitasari. 2019. Managing Brands' Popularity on Facebook: Post Time, Content, and Brand Communication Strategies. *Journal of Indonesian Economy and Business* 34, 2 (2019), 185. https://doi.org/10.22146/jieb.45755

[17] Jack Hessel, Lillian Lee, and David Mimno. 2017. Cats and Captions vs. Creators and the Clock: Comparing Multimodal Content to Context in Predicting Relative Popularity. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 927–936. https://doi.org/10.1145/3038912.3052684

[18] Katrijn Houben, Anne Roefs, and Anita Jansen. 2010. Guilty pleasures. Implicit preferences for high calorie food in restrained eating. *Appetite* 55, 1 (2010), 18–24. https://doi.org/10.1016/j.appet.2010.03.003

[19] Yulin Hswen, Alyssa J. Moran, Siona Prasad, Anna Li, Denise Simon, Lauren Cleveland, Jared B. Hawkins, John S. Brownstein, and Jason Block. 2021. The Federal Menu Labeling Law and Twitter Discussions about Calories in the United States: An Interrupted Time-Series Analysis. *International Journal of Environmental Research and Public Health* 18, 20 (2021), 10794. https://doi.org/10.3390/ijerph182010794

[20] Paweł Kabata, Dorota Winniczuk-Kabata, Piotr Maciej Kabata, Janusz Jaśkiewicz, and Karol Połom. 2022. Can Social Media Profiles Be a Reliable Source of Information on Nutrition and Dietetics? *Healthcare (Basel, Switzerland)* 10, 2 (2022), 397. https://doi.org/10.3390/healthcare10020397

[21] William D.S. Killgore and Deborah A. Yurgelun-Todd. 2006. Affect modulates appetite-related brain activity to images of food. *International Journal of Eating Disorders* 39, 5 (2006), 357–363. https://doi.org/10.1002/eat.20240

[22] William D. S Killgore, Ashley D Young, Lisa A Femia, Piotr Bogorodzki, Jadwiga Rogowska, and Deborah A Yurgelun-Todd. 2003. Cortical and limbic activation during viewing of high- versus low-calorie foods. *NeuroImage* 19, 4 (2003), 1381–1394. https://doi.org/10.1016/S1053-8119(03)00191-5

[23] Dokyun Lee, Kartik Hosanagar, and Harikesh S Nair. 2018. Advertising content and consumer engagement on social media: Evidence from Facebook. *Management science* 64, 11 (2018), 5105–5131. https://doi.org/10.1287/mnsc.2017.2902

[24] Jinha Lee and Heejin Lim. 2023. Visual aesthetics and multisensory engagement in online food delivery services. *International Journal of Retail & Distribution Management* 51, 8 (2023), 975–990. https://doi.org/10.1108/IJRDM-09-2021-0451

[25] Natalie B. Lister, Hannah Melville, and Hiba Jebeile. 2024. What adolescents see on Instagram: Content analysis of #intermittentfasting, #keto, and #lowcarb. *Nutrition & Dietetics* 81, 3 (2024), 316–324. https://doi.org/10.1111/1747-0080.12853

[26] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could You Define That in Bot Terms"? Requesting, Creating and Using Bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 3488–3500. https://doi.org/10.1145/3025453.3025830

[27] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[28] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777. https://doi.org/10.48550/arXiv.1705.07874

[29] Brent McFerran, Darren W. Dahl, Gavan J. Fitzsimons, and Andrea C. Morales. 2010. I'll Have What She's Having: Effects of Social Influence and Body Type on the Food Choices of Others. *Journal of Consumer Research* 36, 6 (2010), 915–929. https://doi.org/10.1086/644611

[30] Kyle McKillop, James Harnly, Pamela Pehrsson, Naomi Fukagawa, and John Finley. 2021. FoodData Central, USDA's Updated Approach to Food Composition Data Systems. *Current Developments in Nutrition* 5 (2021), 596. https://doi.org/10.1093/cdn/nzab044_027

[31] Gillian Moran, Laurent Muzellec, and Devon Johnson. 2020. Message content features and social media engagement: evidence from the media industry. *Journal of Product & Brand Management* 29, 5 (2020), 533–545. https://doi.org/10.1108/jpbm-09-2018-2014

[32] Ludovit Nastisin, Richard Fedorko, Beata Gavurova, and Radovan Bacik. 2024. Examination of Content Types and Social Media Engagement Indicators on Facebook: Case Analysis of a 5-Star Hotels of Visegrad Group Countries. *Marketing and Management of Innovations* 15, 1 (2024), 112–119. https://doi.org/10.21272/mmi.2024.1-09

[33] Ethan Pancer, Matthew Philp, Maxwell Poole, and Theodore J. Noseworthy. 2022. Content Hungry: How the Nutrition of Food Media Influences Social Media Engagement. *Journal of Consumer Psychology* 32, 2 (2022), 336–349. https://doi.org/10.1002/jcpy.1246

[34] Matthew Philp, Jenna Jacobson, and Ethan Pancer. 2022. Predicting social media engagement with computer vision: An examination of food marketing on Instagram. *Journal of Business Research* 149 (2022), 736–747. https://doi.org/10.1016/j.jbusres.2022.05.078

[35] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[36] Markus Rokicki, Eelco Herder, and Christoph Trattner. 2017. How Editorial, Temporal and Social Biases Affect Online Food Popularity and Appreciation. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (2017), 192–200. https://doi.org/10.1609/icwsm.v11i1.

14905

[37] Thorsten Ruprechter. 2025. NutriTransform: Estimating Nutritional Information from Online Food Posts. In *Under review*.

[38] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. *Learning Cross-Modal Embeddings for Cooking Recipes and Food Images*. IEEE Computer Society. 3068–3076 pages. https://doi.org/10.1109/CVPR.2017.327

[39] Monica Serrano-Gonzalez, Megan M. Herting, Seung-Lark Lim, Nicolette J. Sullivan, Robert Kim, Juan Espinoza, Christina M. Koppin, Joyce R. Javier, Mimi S. Kim, and Shan Luo. 2021. Developmental Changes in Food Perception and Preference. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.654200

[40] Sanket S. Sharma and Munmun De Choudhury. 2015. Measuring and Characterizing Nutritional Information of Food and Ingestion Content in Instagram. In *Proceedings of the 24th International Conference on World Wide Web*. Association for Computing Machinery, 115–116. https://doi.org/10.1145/2740908.2742754

[41] Alain D. Starke, Martijn C. Willemsen, and Christoph Trattner. 2021. Nudging Healthy Choices in Food Search Through Visual Attractiveness. *Frontiers in Artificial Intelligence* 4 (2021). https://doi.org/10.3389/frai.2021.621743

[42] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*. IEEE Computer Society, 177–184. https://doi.org/10.1109/SocialCom.2010.33

[43] Safa Enes Turkoglu and Alev Mutlu. 2023. Improving Retweet Prediction via Tweet Features. In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*. IEEE Computer Society, 1–5. https://doi.org/10.1109/HORA58378.2023.10156692

[44] Bradley P. Turnwald, Kathryn G. Anderson, Hazel Rose Markus, and Alia J. Crum. 2022. Nutritional Analysis of Foods and Beverages Posted in Social Media Accounts of Highly Followed Celebrities. *JAMA Network Open* 5, 1 (2022), e2143087. https://doi.org/10.1001/jamanetworkopen.2021.43087

[45] Risqo Wahid and Muhammad Wadud. 2020. Social Media Marketing on Instagram: When is The Most Effective Posting Timing? *EPRA International Journal of Multidisciplinary Research (IJMR)* 6, 7 (2020), 312–321. https://doi.org/10.36713/epra4834

[46] 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 (WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1399–1410. https://doi.org/10.1145/2488388.2488510

[47] Chee Wai Wong, Andrew Tsai, Jost B. Jonas, Kyoko Ohno-Matsui, James Chen, Marcus Ang, and Daniel Shu Wei Ting. 2021. Digital Screen Time During the COVID-19 Pandemic: Risk for a Further Myopia Boom? *American Journal of Ophthalmology* 223 (2021), 333–337. https://doi.org/10.1016/j.ajo.2020.07.034

[48] Yiming Yan. 2024. The evolution and impact of Multi-Armed Bandit algorithms in social media. *Applied and Computational Engineering* 68 (2024), 150–158. https://doi.org/10.54254/2755-2721/68/20241418