Measuring and Characterizing Nutritional Information of Food and Ingestion Content in Instagram

Sanket S. Sharma
College of Computing
Georgia Institute of Technology
Atlanta, GA
sanket@gatech.edu

Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA
munmund@gatech.edu

ABSTRACT

Social media sites like Instagram have emerged as popular platforms for sharing ingestion and dining experiences. However research on characterizing the nutritional information embedded in such content is limited. In this paper, we develop a computational method to extract nutritional information, specifically calorific content from Instagram food posts. Next, we explore how the community reacts specifically to healthy versus non-healthy food postings. Based on a crowdsourced approach, our method was found to detect calorific content in posts with 89% accuracy. We further show the use of Instagram as a platform where sharing of moderately healthy food content is common, and such content also receives the most support from the community.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Instagram, nutrition, social media, food

1. INTRODUCTION

Social media platforms have emerged as popular tools allowing individuals to share the mundane goings on in their lives. An important element of our daily lives is food; it constitutes a central role in not only defining our identity, culture, and lifestyle, but also impacts our health and wellbeing. In this paper, we pursue the research question on how social media content may be leveraged to extract nutritional information of food people ingest, and the dynamics of how the broader community responds to sharing of food and dining experiences on social media.

Language of food and dietary patterns have been studied in the context of social media [2], however prior literature on extracting nutritional information from social media content is limited. Some works have focused on recipe websites and utilized food names in them as a proxy to derive consumption and dietary patterns of individuals [4, 3]. However the coverage of recipe websites can be limited, and the nutritional information may not always be normalized across websites. Close to our work is the work in [1], in which the website http://caloriecount.about.com/ was used to derive nutritional content in Twitter posts. However the method did not distinguish between actual reports of consumption versus cursory references to food items.

In this paper we address the limitations of prior literature by leveraging a novel dataset from a relatively underexplored social media, Instagram. Instagram has emerged as a highly popular photo sharing social platform, where food and ingestion related content are extensively shared. In 2013, the Business Insider reported that food photos are a phenomenon on Instagram[1]. A unique aspect of Instagram use is its ubiquitous nature. The platform can only be used via a smartphone, which has allowed individuals to share photos and videos of the food they are consuming anytime, anywhere. Consequently, the platform provides a reliable way to capture people's dining experiences, and thereby the nutritional information of the food they are consuming.

This paper makes the following contributions. First we leverage the textual content of tags associated with Instagram posts to extract nutritional information without actually analyzing the visual content in photos/videos. Second, via analysis of the patterns of "likes" and comments on posts we examine how the greater community responds to food and ingestion content on the platform.

2. DATA AND METHOD

Data Collection. We describe our data and method of extracting nutritional information from Instagram posts. First, we compiled a set of food-related words which could be used to query Instagram and collect a dataset of food posts. We obtained a list of 564 words (henceforth referred to as canonical food names) from an online food vocabulary word list[2] [2]. Two researchers manually went through the list to remove words which were not indicative of any food item (e.g., "utensils", "roast", "hot"). Additionally we added 24 other words which were likely to be indicative of food posts, based on prior literature [1] and our qualitative inspection of food content sharing on Instagram. Some of them were: "breakfast", "lunch", "yummy", "meal", "foodporn". Generic tags like "food" and "cooking" were considered but not finally included—qualitative inspections of searches of these tags on Instagram yielded noisy posts. The filtered canonical food names were then used as seed tags to get the posts.

Using the list of canonical names, next we collected 1,815,752 posts from Instagram between Nov 27-Dec 9, 2014. We used Instagram's official API[3] to get both image and video posts. For the purposes of this paper, we considered only English language public posts. We did not download the image or video themselves except for validation discussed later.

Extracting Nutritional Information. Following data collection,

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742754
.

[1] http://www.businessinsider.com/instagram-food-photos-are-a-phenomenon-2013-1

[2] http://www.enchantedlearning.com/wordlist/food.shtml

[3] http://instagram.com/developer/

Post tags	Canonical name	Calorie
miami, organic, garden, usa, food, fresh, creole, okra	okra	28.33
bykaila, postre, luneslight, cuantocomer, dessert, flan	flan	177.25
yam, instafood, hamburger, thebird	yam, hamburger	219.93
muesli, granola, easterngranola, localislovely	granola, muesli	330.87
cheesecake, breakfast, cheesecakefactory, redvelvet	cheesecake	402

Table 1: Example posts and calorific content.

we devised a two step approach to measure nutritional information of each of the Instagram posts, based on their tag list. For the purpose, we referred to the official USDA National Nutrient Database for Standard Reference database[4]. This resource provides precise nutritional values of over 30 nutrients for 8618 food items, spanning calorific content, protein, fat, cholesterol etc. Note that the nutritional information is reported based on per 100 grams of serving. Also, food items in the database are described in varying granularities, by a set of words, henceforth referred to as "food descriptors". Our approach proceeded as follows:
(1) We first processed the list of tags in each post by removing stopwords. Then we matched each tag in each Instagram post to the list of canonical food names described above. Wherever possible, mapping to canonical names ensures that we are able to describe each post through names of generic food items, corresponding to which nutritional information may be derived from USDA.
(2) For each of the tags in a post matching a canonical name, thereafter we devised a second matching procedure, in which we compared the tag to the food descriptors in USDA.
(3) From the above compiled set of matching food descriptors, we extracted USDA reported nutrients information listed corresponding to each of them. For posts with more than one match with USDA food descriptors, we computed an aggregate calorie information using either of the following two methods. First, if the standard deviation of calorific content over all matching descriptors was less than the mean, we used the mean as the aggregate calorie. Otherwise, we considered the maximum calorific content in the descriptors as the aggregate for the post. While this is a slightly conservative approach, it ensures that we do not underestimate a post's calorific content when precise information about the actual food item in the post cannot be obtained.

Using this approach we were able to extract calorific information in 93.5% posts in our dataset. Table 1 provides examples of posts with USDA derived calorific information.

Validation of Nutritional Information. Next we adopted a crowdsourced procedure to validate if the nutritional information obtained thus accurately described the food items in the Instagram posts. Specifically, we selected a random sample of 1000 posts with calorific information and engaged two human raters to independently gauge whether the calorific information aligns with human perception using "yes"/"no" labels. The raters were familiar with Instagram and were additionally given access to the associated image/video of the posts in order to help them make an informed judgment. Raters agreed on their judgment for most cases—Cohen's κ was .84. In 89% of the cases, our approach was found to be effective in extracting nutritional information in posts.

3. EMPIRICAL FINDINGS

Figure 1 shows the distribution of posts over calorific content. We observe a non-linear trend in the distribution—the maximum number of posts lie within the calorie range 150-300. Sharing content on high calorie food e.g., beyond 400 is less common among the Instagram sample of users we study. At the same time, while there are relatively more posts in the low calorie i.e., 1-100 range, still they are fewer than ones in the 150-300 range.

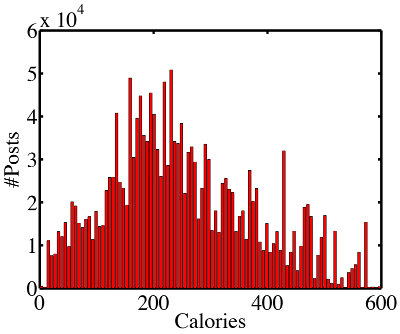

Figure 1: Distribution of number of posts over calorific values.

Next we investigate how the broader Instagram community responds to posts of food with different calorific content—we use the number of "likes" and comments on the posts for the purpose. We generated probability frequency distributions for both these attributes over calorific content of their corresponding posts. Then we compared each of these distributions with the distribution of posts over calories. Normalized mutual information was observed to be .88 and .84 respectively for the "likes" and comments distributions. This shows that "likes" and comments on the posts are higher for posts that are in the moderate calorie range, with less community endorsement and response on posts of very low or very high calorie food. Specifically we found that mean "likes" for posts in the 150-300 calorie range is 22.6 (5.2 for comments); the same is 12.7 for those over 400 calories (1.2 for comments), and it is 16 for posts between 1-100 calories (3.3 for comments). We observed these differences to be statistically significant based on Welch t-tests ($p < .001$).

4. CONCLUSION

Our results showed that reliable and high quality nutritional information (accuracy 89%) may be gleaned from tags associated with food-posts on Instagram. Further, the Instagram community tends to share food photos/videos of moderate calorific content, with high calorie food posts being less common. We further observed that the distribution of number of likes and comments on food posts show high correlated with the distribution of posts, showing that the community tends to like or comment more frequently on posts of moderate calorific content, compared to those of extremely low or extremely high calorie food items.

We note that we did not use any image content analysis in our nutritional information extraction method. Finally, as we relied on English language posts, cross-cultural generalizations cannot be derived. Addressing these limitations are promising directions for future research.

5. REFERENCES

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through twitter. In *Proc. CHI*, 2015.

[2] Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, and Dane Bell. Analyzing the language of food on social media. In *Proc. IEEE Big Data*.

[3] Claudia Wagner, Philipp Singer, and Markus Strohmaier. Spatial and temporal patterns of online food preferences. In *Proc. WWW Companion*.

[4] Robert West, Ryen W White, and Eric Horvitz. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proc. WWW*.

[4] http://ndb.nal.usda.gov/