Virgile Landeiro Dos Reis and **Aron Culotta**
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
vlandeir@hawk.iit.edu, aculotta@iit.edu

Abstract

1 Introduction

2 Related Work

3 Data

3.1 Physically active users

We use the dataset from (Landeiro and Culotta 2015) as our main dataset. This dataset is composed of two groups of users: a group of physically active users, and a group of users who do not exercise. We summarize in a few points how this dataset has been constructed:

1. To detect physically active users, Landeiro and Culotta leveraged existing physical activity tracking applications such as Nike Plus, Runtastic, or RunKeeper. These applications help active users track their progress and give them a summary of their workouts. Additionally, it can automatically tweet a summary of each workout associated with a dedicated hashtag if the user has activated this function. When a user is found to have used one of these dedicated hashtags, it is placed in the treatment group.

2. To build the control group, Landeiro and Culotta used a technique where each user in the treatment group is matched with another user. To find this match for a user i, they first collected the list of accounts F_i that have a mutual-follow relationship with i on Twitter. Then, they filtered out from F_i users that were not of the same gender than i (using the US Census data) and users that were not in the same city or same state as i (using heuristics on the Twitter location field of each user). Finally, they built a cosine similarity score between i and the remaining users in F_i on social media features (number of followers, number of followees, number of posts) and kept the user in F_i with the highest score as the best match to be included in the control group.

3. For each of the 2,322 users in this dataset (1161 in each group), they collected the most recent tweets, up to 3,200.

3.2 Food-related datasets

Because the Twitter dataset built in the previous section does not focus on food tweets, we merge three datasets with food information into one in order to construct a large dataset of food vocabulary.

1. FooDB[1] is an ensemble of resources on food constituents, chemistry and biology. In particular, we use the dataset listing 889 food sources associated with a category (e.g. kiwi is associated with the fruits category and turkey is in the poultry category).

2. In (Abbar, Mejova, and Weber 2015), the authors combined a large data collection of 50M tweets on manually selected keywords, bootstraping techniques, and crowdsourcing to build a dataset of food vocabulary of 461 words with the category they belong to as well as the average amount of calorie per serving for each food.

3. Finally, we use WordNet (Miller 1995), the popular lexical database for English, and we build a dynamic programming algorithm that looks up the ancestors of a word up to the tree root and returns True if at least one ancestor is one of:

   - food: any substance that can be metabolized by an animal to give energy and build tissue.
   - food: any solid substance (as opposed to liquid) that is used as a source of nourishment.
   - edible nut: a hard-shelled seed consisting of an edible kernel or meat enclosed in a woody or leathery shell.

3.3 Exemplar health Twitter accounts

Using a subset of the data built in (Culotta and Cutler 2016), we obtain a list of 407 accounts that tweet about health, called exemplar accounts. These exemplar accounts are collected using Twitter Lists (i.e. manual aggregation of accounts). First, a query (e.g. "health") is sent to Twitter's search engine, which returns Twitter Lists and tweets. If an account appear in at least two of the top 50 Lists, then this account becomes an exemplar account for the given query.

[1] http://foodb.ca/

bite	bootleg	brain	buffalo	cat
enter	centre	coloring	cup	cut
date	dip	dish	dog	feed
game	green	jack	joint	kisses
leg	must	pop	punch	rock
rose	scratch	shoulder	side	snowball
stick	stock	sucker	table	white

Table 1: Example of words detected as food-related that we manually remove from our analysis.

Control	last, wish, right, girl, dr, king, celery, plus, real, thanksgiving, hate, corner, lost, gin, wrong, episode, event, ah, smoked, mozzarella
Treatment	others, state, looks, burrito, starbucks, used, cookies, co, cafe, pic, run, man, drinking, lots, NUMBERk, pub, day, broccoli, going, grill

Table 2: 20 most predictive features for each group.

Control		**Treatment**	
Rank	Feature	Rank	Feature
7	celery	4	burrito
14	gin	7	cookies
20	mozzarella	18	broccoli
21	pickles	23	caramel
34	deer	25	breakfast
40	molasses	30	carrot
43	bananas	31	burritos
50	wrap	32	coconut
57	curds	33	onions
62	frank	34	citrus

4 Experiments and results

To evaluate the differences in the relation to food and health between users who exercise and users who do not exercise, we experimented at two levels. First, we looked at the distribution of food-related words in both groups and studied the most discriminative features between both groups after keeping a subset of users that we detected tweeted about food. Then, we focused on a network-level analysis where we compared the relation that the users in our treatment and control groups have with health exemplar accounts.

4.1 Text-level analysis

To evaluate the vocabulary differences between our two groups of users, we started by filtering out tweets that are not about food. We then trained two models: one to underline the most predictive features of each group (linear SVM), and one to explain similar observations and group them together (Latent Dirichlet Allocation).

Filter food-related tweets In order to remove irrelevant tweets for our analysis, we started by combining the FooDB dataset and the food sources in the dataset developed by Abbar, Mejova, and Weber. In later steps, we noticed that some food words were more often associated with non food-related topics. For example, the word "apple" was more frequently used in reference to the technology company `Apple` and its products than it was to reference the actual fruit. For this reason, we manually created a set of excluded words that we detect as being sources of false positive in the detection of food-related tweets. Initially, this set is composed of the words `apple` (the tech company), `raspberry` (the $30 computer raspberry pi), `blackberry` (the smartphone company), `kevin` (Kevin Bacon), as well as the words displayed in Table 4.1. We will see in some of the next steps that we had to come back to this step and add words to the list of excluded words in order to remove more false positive tweets.

We end up with a list of 1,217 food-related words $food\_list$, and a list of words inclined to yield false positive tweets $excluded\_list$. Then, for each tweet t of each user, we keep t if none of the words in t are in $excluded\_list$ and if at least one of the words in t is in $food\_list$.

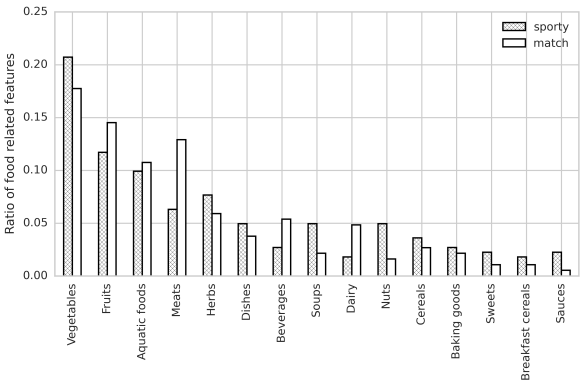
Filter food-conscious users Once we obtained the food-related tweets for each user in both groups, we compute a score for each tweet t as the proportion of terms that are food-related (i.e. number of food terms in t over the total number of terms).

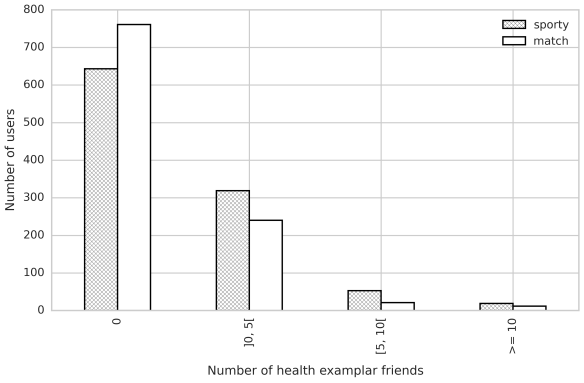
In order to define if a term is a food term, we not only use $food\_list$ and $excluded\_list$ as in the previous step, but we also make use of our algorithm based on WordNet. This allows us to detect a wider range of food terms to compute the score for users. We decided not to use the WordNet-based algorithm in the previous step to avoid detecting too many false positive food-related tweets. Indeed, with our algorithm, all the words that have at least one ancestor being "food"" or "edible nut" in the WordNet graph will be classified as food-related terms. Therefore, a large amount of words that have several meanings and for which the main meaning is not related to food will be classified as food-related (e.g. "must" is a kind of freshly pressed grape juice, "dog" is a way to describe a "hotdog").

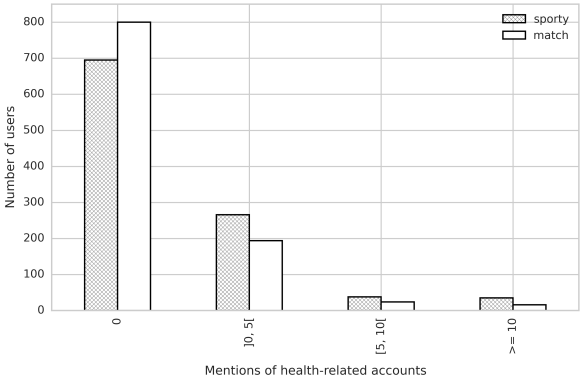
Finally, we keep a physically active user u in our treatment group if u and its match in the control group have at least 15 food-related tweets with a score of .15 or more. This yields 530 users that are talking about food in each group.

Most predictive features for each group Using the food-related tweets of each of the remaining 1060 users, we analyze the differences between the most predictive features of the physically active group and the most predictive features of the matched group. To do so, we create a features vector for each user by removing retweets, tokenizing the remaining tweets using unigrams, removing punctuation signs as well as stopwords. Then, we fit a linear SVM model on these features where the label of each user is the group to which it belongs. In table 4.1, we report the ten most important features for each class

Differences in food categories between groups







4.2 Network-level analysis

Follow relationship

Mention relationship

5 Discussion

References

Abbar, S.; Mejova, Y.; and Weber, I. 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*, 3197–3206. ACM.

Culotta, A., and Cutler, J. 2016. Mining brand perceptions from twitter social networks. *Marketing Science* 35(3):343–362.

Landeiro, V., and Culotta, A. 2015. Using matched samples to estimate the effects of exercise on mental health via twitter. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.