Challenges for using social media for early detection of T2DM

Dane Bell, Daniel Fried, Luwen Huangfu, Mihai Surdeanu, Stephen Kobourov
University of Arizona
Tucson, AZ 85721, USA
{dane, dfried, huangfuluwen, msurdeanu, kobourov}@email.arizona.edu

Abstract

Twitter and other social media data are utilized for a wide variety of applications such as marketing and stock market prediction. Each application and appropriate domain of social media text presents its own challenges and benefits. We discuss methods for detecting obesity, a risk factor for Type II Diabetes Mellitus (T2DM), from the language of food on Twitter on community data, the peculiarities of this data, and the development of individual-level data for this task.

1. Introduction

This project is in aid of implementing a system that can detect individuals who are likely to be at high risk for preventable Type II Diabetes Mellitus (T2DM), a life-shortening disease that generates fatal complications that is common in the developed world. The system is part of an effort to nudge (Thaler and Sunstein, 2008) individuals at risk for T2DM to make changes to their diet and exercise level to prevent or delay the disease's onset. The central hypothesis of this work is that (features of) individuals' tweets about food correlate with their real-world food consumption, which is in turn correlated with their likelihood of developing T2DM. We began by learning to detect obesity, a factor often implicated in the rising rate of T2DM diagnosis in the United States. Through work on community-level data, we found that this hypothesis was supported, but our machine learning model for detecting obesity rates at a state level proved not to transfer well to individuals. For this reason, we sought to engage with tweeters and other individuals on social media sites to help collect individual data through the use of a novel, 20-questions-style quiz generated semiautomatically from a classifier trained on community-level Twitter data. We discuss our approach to these challenges as well as future directions.

2. Community-level detection

In order to begin detecting obesity, we began with models over community-level data, namely cities and US states. Using the Twitter API, we gathered ca. 3.5 million tweets containing relevant hashtags such as #dinner and #breakfast, of which 16% (562,547 tweets) could be assigned a location within a US state (Fried et al., 2014).

As is typical in Twitter data, our tweets required significant preprocessing, most importantly in removing Uniform Resource Locators (URLs) and @mentions of user handles. We experimented with different feature sets, including limiting our features to hashtags, food words, or both. We also used Latent Dirichlet Analysis (Blei et al., 2003) to mitigate sparsity, with 200 topics added to our feature set. In all cases, we used Support Vector Machines (SVM) with a linear kernel (Vapnik and Vapnik, 1998).

A model was trained to predict whether a state was above or below the national median for overweight rate according to a Kaiser Commission on Medicaid and the Uninsured (KCMU) analysis[1]. In addition to predicting community-level obesity at an accuracy of 80%, this dataset was able to predict whether a state had greater or lower than median diabetes rate (69% accuracy). Similar models were able to predict the less obviously related variables of location and political party affiliation.

Different sets of features were optimal for each of these factors, sometimes favoring all words and sometimes only food words, for example, although the addition of LDA topics was beneficial in all cases. Table 1 shows the top 20 features of the SVM model for classification, displaying intuitively appropriate correspondence between diets (*fried* vs. *vegan*) and rate of obesity and diabetes.

3. Transfer to individuals

Although the previous experiments showed that tweets about food contained information about our variables of interest, the performance of the community-trained models on manually annotated individual Twitter accounts was at chance. This made it clear that a corpus of individually annotated Twitter accounts was necessary for accurate prediction, and we devised a 20-questions-style quiz site based on our community-level data to serve two purposes: evaluation on individuals, and data collection for new models.

SVMs do not produce models that are easily converted into natural-language questions, but tree-based classifiers such as random forest classifiers do. Through further experimentation, we discovered that a small number of relatively shallow decision trees with discrete features could perform comparably to our prior SVM model when predicting state-level overweight rates (Bell et al., 2016). The high performance of these models (78% accuracy, compared to 80% of our previous work and baseline accuracy of 51%) came in spite of their simplicity and interpretability: the best performing model used a 7-tree decision forest with maximum depth 3 and three-bin discrete features.

These trees were converted semiautomatically into natural languages questions, so that a feature based on the word *brunch* became "How often do you eat brunch?" with three multiple-choice Likert scale (Likert, 1932) answers such as *Practically never*. Figure 1 illustrates one tree of the decision forest. The questions that were asked depended cru-

[1] http://kff.org/other/state-indicator/adult-overweightobesity-rate/

Class	*Highest-weighted features*
overweight: +	i, day, my, great, one, *American Diet* (chicken, baked, beans, fried), #snack, *First-Person Casual* (my, i, lol), cafe, *Delicious* (foodporn, yummy, yum), *After Work* (time, home, after, work), house, chicken, fried, *B*reakfast (day, start, off, right), #drinks, bacon, call, eggs, broccoli
overweight: -	*You, We* (you, we, your, us), #rvadine, #vegan, make, photo, dinner, #meal, #pizza, *Giveaway* (win, competition, enter), new, *Restaurant Advertising* (open, today, come, join), #date, happy, #dinner, 10, jerk, check, #food, #bento, #beer
diabetes: +	*Mexican* (mexican, tacos, burrito), *American Diet* (chicken, baked, beans, fried), #food, *After Work* (time, home, after, work), #pdx, my, lol, #fresh, *Delicious* (foodporn, yummy, yum), #fun, morning, special, good, cafe, #nola, fried, bacon, #cooking, all, beans
diabetes: -	#dessert, *Turkish* (turkish, kebab, istanbul), #foodporn, #paleo, #meal, *Paleo Diet* (paleo, chicken, healthy), i, *Giveaway* (win, competition, enter), *I, You* (i, my, you, your), your, new, today, #restaurant, *Japanese* (ramen, japanese, noodles), some, jerk, #tapas, more, *Healthy DIY* (salad, chicken, recipe), *You, We* (you, we, your, us)

Table 1: Top 20 highest-weighted features in descending order of importance for each dataset from Fried et al. (2014), for both the positive and negative classes. For example, "overweight: +" indicates the most representative features for being overweight, whereas "overweight: -" shows the most indicative features for **not** being overweight. The features include LDA topics, with manually assigned names (*italicized*) for clarity, and a few of their most common words within parentheses.

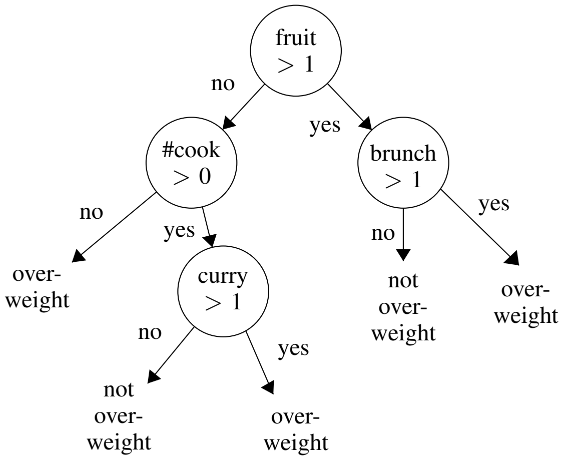

Figure 1: A decision tree from the random forest classifier trained using state-level Twitter data.

cially on which answers the quiz takers provided, as in the tree's binary choices based on three quantized values, 0, 1, and 2. Notice that this conversion relies on our underlying assumption that mentions of a food word are correlated with the consumption of that food. The 20-questions-style quiz based on these questions allowed us to evaluate how well the community-trained classifier applied to individuals, resulting in high accuracy (79%) which was nevertheless lower than the baseline (82%) in our highly biased sample. We interpret this to indicate that, as in our initial experiments, individuals are not highly representative of their regions (or vice versa), meaning that individual-level training information is necessary for good individual classification. Fortunately, by engaging with quiz takers, the site also afforded an opportunity to collect individual-level training data. Quiz takers optionally provided their (public) social media accounts and their height and weight, from which we can calculate body mass index (BMI), as well as other demographic information: location, age, and gender. This will allow data collection from these accounts with permission to train more direct, individual-based classifiers.

4. Discussion

The greatest challenges for obesity detection are familiar from other NLP work. There is the persistent problem of non-human accounts (e.g., businesses, organizations, and bots) which add noise to the training data. The signal fighting against that noise is also imperfect, notably in its sparsity, since the average user has on the order of hundreds of tweets, of which only a very small percentage regard food. However, tweets that do not mention food may still be useful for obesity detection, much in the way that food tweets can significantly predict political affiliation (Fried et al., 2014) through indirect cultural connections.

Future work will include taking more information into account in the models. With individual-level data, we can capitalize on users' locations, photo, user handle, bio, and age, all of which are informative, though optional, parts of a Twitter profile. With these as well as features generated from the tweets themselves, classifiers can be constructed for intermediate factors such as gender which will in turn add valuable features for obesity classification. This in turn will improve our ability to develop a valuable public health tool for detecting and preventing T2DM efficiently through social media.

5. References

Bell, D., Fried, D., Huangfu, L., Surdeanu, M., and Kobourov, S. (2016). Towards using social media to identify individuals at risk for preventable chronic illness. In *LREC 2016*.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., and Bell, D. (2014). Analyzing the language of food on social media. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 778–783. IEEE.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, 140:1–55.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1. Wiley New York.