Towards Using Social Media to Identify Individuals at Risk for Preventable Chronic Illness

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

We describe a strategy for the acquisition of training data necessary to build a social-media-driven early detection system for individuals at risk for (preventable) type 2 diabetes mellitus (T2DM). The strategy uses a game-like quiz with data and questions acquired semi-automatically from Twitter. The questions are designed to inspire participant engagement and collect relevant data to train a public-health model applied to individuals. Prior systems designed to use social media such as Twitter to predict obesity (a risk factor for T2DM) operate on entire communities such as states, counties, or cities, based on statistics gathered by government agencies. Because there is considerable variation among individuals within these groups, training data on the individual level would be more effective, but this data is difficult to acquire. The approach proposed here aims to address this issue. Our strategy has two steps. First, we trained a random forest classifier on data gathered from (public) Twitter statuses and state-level statistics with state-of-the-art accuracy. We then converted this classifier into a 20-questions-style quiz and made it available online. In doing so, we achieved high engagement with individuals that took the quiz, while also building a training set of voluntarily supplied individual-level data for future classification.

Keywords: machine learning, obesity detection, social media

1. Introduction

Data collection in the public health domain is difficult due to privacy concerns and low engagement. For example, people seldom engage with surveys that require them to report their height and weight. However, such data is crucial for training automated public health tools, such as algorithms that detect risk for (preventable) type 2 diabetes mellitus (T2DM, henceforth diabetes). We propose a semi-automated data collection algorithm for obesity detection that mitigates these issues with a game-like quiz that is automatically bootstrapped from a machine-learning model trained over social media data. The resulting quiz is non-intrusive, focusing on “fun” questions about food and food language while avoiding personal questions, which leads to high engagement. We believe this idea contributes to addressing one of the most challenging unsolved public health problems: the high rate of chronic illness resulting from modifiable risk factors such as poor diet quality and physical inactivity. It is estimated that more than 86 million Americans over the age of 20 exhibit signs of pre-diabetes, and 70% of these pre-diabetic individuals will eventually develop T2DM, a chronic and debilitating disease associated with heart disease, stroke, blindness, kidney failure, and amputations (National Center for Health Statistics, 2014; American Diabetes Association and others, 2008). In the United States, the estimated cost of T2DM rose to $245 billion in 2012 (Association, 2013). Yet, 90% of these individuals at high risk are not aware of it (Li et al., 2013).

Our long-term goal is to develop tools that automatically classify overweight individuals (hence at risk for T2DM\(^1\)) using solely public social media information. The advantage of such an effort is that the resulting tool provides non-intrusive and cost-effective means to detect and warn at-risk individuals early, before they visit a doctor’s office, and possibly influence their decision to visit a doctor.

Previous work has demonstrated that intervention by social media has modest but significant success in decreasing obesity (Ashrafian et al., 2014). Furthermore, there is good evidence that detecting communities at risk using computational models trained on social media data is possible (Fried et al., 2014; Culotta, 2014). However, in all cases, classification is made on aggregated data from cities, counties, or states, so these models are not immediately applicable to the task of classifying individuals. Our work takes the first steps towards transferring a classification model that identifies communities that are more overweight than average to classifying overweight (and thus at-risk for T2DM) individuals. The contributions of our work are:

1. We introduce a random-forest (RF) model that classifies US states as more or less overweight than average using only 7 decision trees with a maximum depth of 3. Despite the model’s simplicity, it outperforms Fried et al. (2014)’s best model by 2% accuracy.

2. Using this model, we introduce a novel semi-automated process that converts the decision nodes in the RF model into natural language questions. We then use these questions to implement a quiz that mimics a 20-questions-like game. The quiz aims to detect if the person taking it is overweight or not based on indirect questions related to food or use of food-related words. To our knowledge, we are the first to use a semi-automatically generated quiz for data acquisition.

\(^1\)In the CDC diabetes questionnaire available at http://www.cdc.gov/diabetes/prevention/pdf/prediabetestest.pdf, overweight BMI contributes more than half of the points associated with diabetes risk.
We demonstrate that this quiz serves as a non-intrusive and engaging data collection process for individuals\(^2\). The survey was posted online and evaluated with 945 participants, of whom 926 voluntarily provided supplemental data, such as information necessary to compute the Body Mass Index (BMI), demographics, and Twitter handle, demonstrating excellent engagement. The random-forest model backing the survey agreed with self-reported BMI in 78.7% of cases. More importantly, the differences prompted a spirited Reddit discussion, again supporting our hypothesis that this quiz leads to higher participant engagement\(^3\).

This initial experiment suggests that it is possible to use easy-to-access community data to acquire training data on individuals, which is much more expensive to obtain, yet is fundamental to building individualized public health tools. The anonymized data collected from the quiz is publicly available.

2. Prior work

Previous work has used social media to detect events, including monitoring disasters (Sakaki et al., 2010), clustering newsworthy tweets in real-time (McCreadie et al., 2013; Petrović et al., 2010), and forecasting popularity of news (Bandari et al., 2012).

Social media has also been used for exploring people’s opinions towards objects, individuals, organizations and activities. For example, Tumasjan et al. (2010) and O’Connor et al. (2010) have applied sentiment analysis on tweets and predicted election results. Hu and Liu (2004) analyzed restaurant ratings based on online reviews, which contain both subjective and objective sentences. Golder and Macy (2011) and Dodds et al. (2011) are interested in the temporal changes of mood on social media. Myslín et al. (2013) focus on understanding the perception of emerging tobacco products by analyzing tweets.

Social media, especially Twitter, has been recently utilized as a popular source of data for public health monitoring, such as tracking diseases (Ginsberg et al., 2009; Yom-Tov et al., 2014; Nascimento et al., 2014; Greene et al., 2011; Chew and Eysenbach, 2010), mining drug-related adverse events (Bian et al., 2012), predicting postpartum psychological changes in new mothers (De Choudhury et al., 2013), and detecting life satisfaction (Schwartz et al., 2013) and obesity (Chunara et al., 2013; Cohen-Cole and Fletcher, 2008; Fernandez-Luque et al., 2011).

We focus our attention on the language of food on social media to identify overweight communities and individuals. In the last couple of years, several variants of this problem have been considered (Fried et al., 2014; Abbar et al., 2015; Culotta, 2014; Ardehaly and Culotta, 2015). Fried et al. (2014) collected a large corpus of over three million food-related tweets and use it to predict several population characteristics, namely diabetes rate, overweight rate and political tendency. Generally, they use state-level populations, e.g., one of their classification tasks is to label whether a state is more overweight than the national median. Overweight rate is the percentage of adults whose Body Mass Index (BMI) is larger than a normal range defined by NIH. The classification task is to label whether a state is more overweight than the national median. Individuals’ tweets are localized at state level as a single instance to train several classifier models, and the performance of models is evaluated using leave-one-out cross-validation. Importantly, Fried et al. (2014) train and test their models on communities rather than individuals, which limits the applicability of their approach to individualized public health.

Abbar et al. (2015) also used aggregated information for predicting obesity and diabetes statistics. They considered energy intake based on caloric values in food mentioned on social media, demographic variables, and social networks. This paper begins to address individual predictions, based on the simplifying assumption that all individuals can be labeled based on the known label of their home county, e.g., all individuals in an overweight county are overweight, which is less than ideal. In contrast, our work collects actual individual information through the survey derived from community information.

Even though performing classification at state or county granularity tends to be robust and accurate (Fried et al., 2014), characteristics that are specific to individuals are more meaningful and practical. A wave of computational work on the automatic identification of latent attributes of individuals has recently emerged. Ardehaly and Culotta (2015) utilize label regularization, a lightly supervised learning method, to infer latent attributes of individuals, such as age and ethnicity. Other efforts have focused on inferring the gender of people on Twitter (Bamman et al., 2014; Burger et al., 2011) or their location on the basis of the text in their tweets (Cheng et al., 2010; Eisenstein et al., 2010). These are exciting approaches, but it is unlikely they will perform as well as a fully supervised model, which is the ultimate goal of our work.

3. Method

Fried et al. (2014) showed that states and large cities generate a considerable number of food-related tweets, which can be used to infer important information about the respective community, such as overweight status or diabetes risk. In an initial experiment, we tested this classifier on the identification of overweight individuals. This classifier did not perform better than chance, likely due to the fact that individuals have a much sparser social media presence than entire communities (most tweeters post hundreds of tweets, not millions, and rarely directly about food). This convinced us that a realistic public health tool that identifies individuals at risk must be trained on individual data directly, in order to learn to take advantage of the specific signal available.

We describe next the process through which we acquire such data.

---

\(^2\)Previous work has demonstrated the high engagement of such quizzes. For example, the most popular post in the New York Times for 2013 was a quiz predicting respondents’ locations by features of their dialect such as distinctive vocabulary: [http://www.nytco.com/the-new-york-times-most-visited-content-of-2013](http://www.nytco.com/the-new-york-times-most-visited-content-of-2013)

\(^3\)http://www.reddit.com/r/RandomSize/comments/3bhi3/academic_can_our_automatically_generated/
I love waffles

Unlike Fried et al. (2014), we do not use support vector and T2DM rates (Fried et al., 2014). has already been shown to correlate with higher overweight or dinner trained at least one relevant hashtag, such as #breakfast collected which were localizable to a specific state and containing as Fried et al. (2014): we used the 887,310 tweets they way that can be converted into fun, engaging natural language questions. To this end, we started with the same set-
ing of by asking 20 or fewer yes-or-no questions. Further tuning confirmed that a small number of shallow trees are most effective in accurately partitioning the state-level data. To further increase the interpretability of the model, word and hashtag counts were automatically discretized into three bins (e.g., infrequent, somewhat frequent, and very frequent) based on the quantiles of the training data. Figure 2 illustrates one of the decision trees in the trained random forest, with 0 standing for the infrequent, 1 for the somewhat frequent, and 2 for the very frequent bin of the corresponding word or hashtag. The figure highlights that the tree is immediately interpretable. For example, the left-most branch indicates that a state is classified as overweight if its tweets mention the word “fruit” infrequently or somewhat frequently (fruit ̸∈ 1), and the hashtag “#cook” appears infrequently (̸#cook ̸∈ 0). A state with infrequent mention of the word fruit would take the left branch, then test for the frequency of #cook. If this is not an infrequent token, then the classifier tests for curry; very frequent use of curry would lead to an “overweight” classification (relative to the median state).

3.2. Quiz
We next manually converted all decision statements in the random forest classifier into natural language questions. The main assumption behind this process is that language use parallels actual behavior, e.g., a person who talks about fruit on social media will also eat fruit in real life. This allowed us to produce more intuitive questions, such as How often do you eat fruit? for the top node in Figure 2, instead

![Figure 1: Architecture of the semi-automatic approach for quiz generation from social media data.](https://code.google.com/p/fast-random-forest/)

![Figure 2: A decision tree from the random forest classifier trained using state-level Twitter data.](https://code.google.com/p/fast-random-forest/)

---

4https://code.google.com/p/fast-random-forest/

5Despite its simplicity, the proposed RF model performs better than the SVM model of Fried et al. (2014) on the same data.
How often do you eat fruit?
→ Practically never, Sometimes, Often
What proportion of your meals are home cooked?
→ None or very little, About half, Most or all
How often do you eat curry?
→ Practically never, Sometimes, Often

4.1. Evaluation of random forest classifier

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority baseline</td>
<td>50.89</td>
</tr>
<tr>
<td>SVM (Fried et al., 2014)</td>
<td>80.39</td>
</tr>
<tr>
<td>RF (food + hashtags)</td>
<td>82.35</td>
</tr>
<tr>
<td>Discretized RF (food + hashtags)</td>
<td>78.43</td>
</tr>
</tbody>
</table>

Table 1: Example questions derived from the decision nodes in Figure 2.

We further classified the Reddit comments received into state-level data relative to majority baseline and Fried et al. (2014)’s best classifier. We include two versions of our classifier: the first keeps numeric features (e.g., word counts) as is, whereas the second discretizes numeric features to three bins.

Table 2: Random forest (RF) classifier performance on state-level data relative to majority baseline and Fried et al. (2014)’s best classifier. We include two versions of our classifier: the first keeps numeric features (e.g., word counts) as is, whereas the second discretizes numeric features to three bins.

Table 2: Random forest (RF) classifier performance on state-level data relative to majority baseline and Fried et al. (2014)’s best classifier. We include two versions of our classifier: the first keeps numeric features (e.g., word counts) as is, whereas the second discretizes numeric features to three bins.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority baseline</td>
<td>50.89</td>
</tr>
<tr>
<td>SVM (Fried et al., 2014)</td>
<td>80.39</td>
</tr>
<tr>
<td>RF (food + hashtags)</td>
<td>82.35</td>
</tr>
<tr>
<td>Discretized RF (food + hashtags)</td>
<td>78.43</td>
</tr>
</tbody>
</table>

Table 2: Random forest (RF) classifier performance on state-level data relative to majority baseline and Fried et al. (2014)’s best classifier. We include two versions of our classifier: the first keeps numeric features (e.g., word counts) as is, whereas the second discretizes numeric features to three bins.

4.2. Quiz response

Many of the 945 participants were highly engaged with the quiz: 97.9% volunteered demographic information at the end of the quiz. Many of the participants also left feedback, some on the Reddit page linking to the quiz, as shown in Figure 4, some on the quiz page itself, as shown in Table 3. The feedback comprised mostly comments on the accuracy (or inaccuracy) of the quiz, comments expressing interest in particular questions, and speculation about how the quiz was constructed.

It seems that quiz accuracy was not a prerequisite for commenting on the quiz. On the contrary, participants were more likely to comment when their results were inaccurate. It is unknown whether the up- and down-voting was motivated by the accuracy of the quiz, but researchers making interactive prediction sites may discover that inaccuracy is in fact more engaging in some regards. The perceived stigma of obesity was also evident in the reactions to the quiz, with some negative reactions to a prediction of overweight regardless of its accuracy.

For a better understanding of the feedback received, we performed a post-hoc analysis. Our analysis indicated that while there were 3 comments made about accuracy out of 744 people with correct predictions (0.40% commented), and 13 noting incorrect answers out of 201 with incorrect predictions (6.5% commented). Thus, the participants were 16 times as likely to comment on the quiz’s accuracy if its prediction was incorrect than they were if it was correct. We further classified the Reddit comments received into identical experimental settings as (Fried et al., 2014), i.e., leave-one-out-cross-validation on the 50 states plus the District of Columbia. The table shows that our best model performs 2% better than the best model of (Fried et al., 2014). Our second classifier, which used discretized numeric features and was the source of the quiz, performed 2% worse, but it still had acceptable accuracy, nearing 80%. As discussed earlier, this discretization step was necessary to create intelligible Likert-scaled questions (Likert, 1932).
Figure 3: Demographic data from the present study and from the National Health and Nutrition Examination Survey (NHANES) (National Center for Health Statistics, 2013). Of 945 respondents, 833 provided their age, 864 provided information for their BMI calculation, 846 provided their gender, and 625 provided their location.

six classes: affective comments (7%), hypothesizing comments (17%), cultural comments (17%), result-based comments (53%), constructive criticism (7%), and comments seeking a greater understanding of the quiz (7%)\(^6\), examples of which can be seen in Figure 4. The comments made on the quiz site also frequently included additional diet and demographic information.

4.3. Quiz evaluation

We evaluated the quiz on 945 volunteers recruited at the University of Arizona and on social media, namely Facebook, Twitter, and Reddit’s SampleSize subreddit\(^7\). The results are summarized in Table 4. We evaluated the accuracy of the random forest classifier by comparing each individual’s actual BMI, based on the self-reported height and weight, to the classifier’s prediction. The cutoff boundary BMI for both training and testing was 28.7 – the average US adult BMI according to National Center for Health Statistics (2013). This figure is above the NIH’s definition of overweight (BMI \(\geq 25\)) because the average US resident is overweight by that standard. These results are promising: the quiz had a 78.7% accuracy for the classification of individuals into the two classes: higher or lower BMI than the average US resident.

\(^6\)Numbers sum to more than 100% because of comments with multiple classes of content.

\(^7\)http://www.reddit.com/r/SampleSize/
5. Discussion

It is important to note that the limitations of the sample are considerable: as shown in Table 4 and Figure 3, our initial sample is taller, lighter, and younger than the average US adult, leading to a biased test sample. The strong bias of the sample means that the trivial baseline of predicting no participants to be over BMI 28.7 would have accuracy 82.3%. Moreover, while the overall accuracy of the random forest backing the quiz was 78.7%, the accuracy on participants who reported a BMI over 28.7 was only 16.0%. Our conjecture is that, in general, participants who are overweight are more reluctant to mention food- and health-related topics on social media, which led to lower-quality training data for this group, and the distribution of BMI for participants classified as under 28.7 was not significantly different from that of those classified as over 28.7.

<table>
<thead>
<tr>
<th>Quiz accuracy</th>
<th>78.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy, participants BMI ≥ 28.7</td>
<td>16.0%</td>
</tr>
<tr>
<td>Accuracy, participants BMI &lt; 28.7</td>
<td>92.2%</td>
</tr>
</tbody>
</table>

| Proportion, participants BMI ≥ 28.7 | 17.7% |
| Proportion, participants BMI < 28.7 | 82.3% |

| Mean participant weight | 74.4 kg (164 lbs) |
| Mean US adult weight | 80.3 kg (177 lbs) |
| Mean participant height | 173 cm (5 ft 8 in) |
| Mean US adult height | 167 cm (5 ft 6 in) |
| Mean participant BMI | 24.9 |
| Mean US adult BMI | 28.6 |
| Mean participant age | 26.1 years |
| Mean US adult age | 47.1 years |

Table 4: Results of the quiz evaluation, together with statistics of the adult participants (18 or older) who took the test, compared against average values in the US (National Center for Health Statistics, 2014).

Far from being a problem for this data collection technique, however, the failure of transfer from state-level training to individual-level testing underscores the need for the data collection itself. No existing system has been able to automatically predict individuals’ weight after training on state-level data.

6. Conclusions and future work

We described a strategy for the acquisition of training data necessary to build a social-media-driven early detection system for individuals at risk for T2DM, using a game-like quiz with data and questions acquired from Twitter. Our approach has proven to inspire considerable participant engagement and, in so doing, provide relevant data to train a public-health model applied to individuals.

First, we built a random forest classifier that improves on the state of the art for the classification of overweight communities (in particular US states). We then use this as the basis of a 20-questions-style quiz to classify individuals. Early results are promising: 78.7% accuracy, but the sample does not represent the general population well, and the quiz performs poorly on classifying overweight individuals. The most immediate goal is to obtain a large respondent sample that is more representative of US adults, and to extend the information gathered to longitudinal data. Based on the high engagement observed in this initial experiment, we hope that a large dataset can be constructed at minimal cost. This dataset will be used to develop a public-health tool capable of non-intrusively identifying at-risk individuals by monitoring public social-media streams.

Our long term goal is to use this data to train a supervised classifier for the identification of individuals at risk for type 2 diabetes. The dataset collected through the quiz described here is sufficient for this goal: it includes necessary information for the calculation of BMI (weight, height), demographic information, and social media handles. We plan to explore (public) multi-modal social media information: natural language, posted pictures, etc. From this, we will...
extract and use preventable risk factors, such as poor diet or lack and perceived lack of physical activity. The data will be made available to interested researchers. There is great potential for further improvements of the model by adding calorie count estimates for food pictures associated with individual tweets, by incorporating individual-level demographic features such as gender and age, and by using words and hashtags about physical activities.

7. Resources

- The software used to generate and test the random forest classifier is open-source at http://github.com/clulab/twitter4food/
- The quiz is available from the project’s main page at http://sites.google.com/site/twitter4food/
- Anonymized quiz results are available at http://git.io/vZY5U. They detail the responses of each participant to the quiz questions, the system’s prediction, its accuracy, and the participants’ height, weight, location, age, gender, and (anonymized) comments.

8. References


