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, in so doing, 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, which garnered state-of-the-art accuracy. We then converted this classifier into an engaging 20-questions-style quiz, which was made 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). 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). Yet, many of these individuals at high risk are not aware of it.

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 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. The quiz is publicly available[2]. To our knowledge, we are the first to use a semiautomatically generated quiz for data acquisition.

3. We demonstrate that this quiz serves as a non-intrusive

[1] `http://www.cdc.gov/diabetes/prevention/pdf/prediabetestest.pdf`

[2] `http://sites.google.com/site/twitter4food`

and engaging data collection process for individuals[3]. 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[4].

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. We publicly release the anonymized data collected by the quiz[5].

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), 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. There are several similarly motivated works (Fried et al., 2014; Abbar et al., 2014; 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 normal range defined by NIH. The classification task is to label whether a state is more overweight (and, similarly, more diabetic and Democratic) than the national median. Users' tweets are localized at state level as a single instance to train several classifier models, and performance of models are evaluated on 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. (2014) 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 (friendship and mention). 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 online users (Bamman et al., 2014; Burger et al., 2011) or their regional origin (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 users write 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.

[3]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 their dialectal features: `http://www.nytco.com/the-new-york-timess-most-visited-content-of-2013`

[4]`http://www.reddit.com/r/SampleSize/comments/3hbiz3/academic_can_our_automatically_generated/`

[5]`http://git.io/vZY5U`

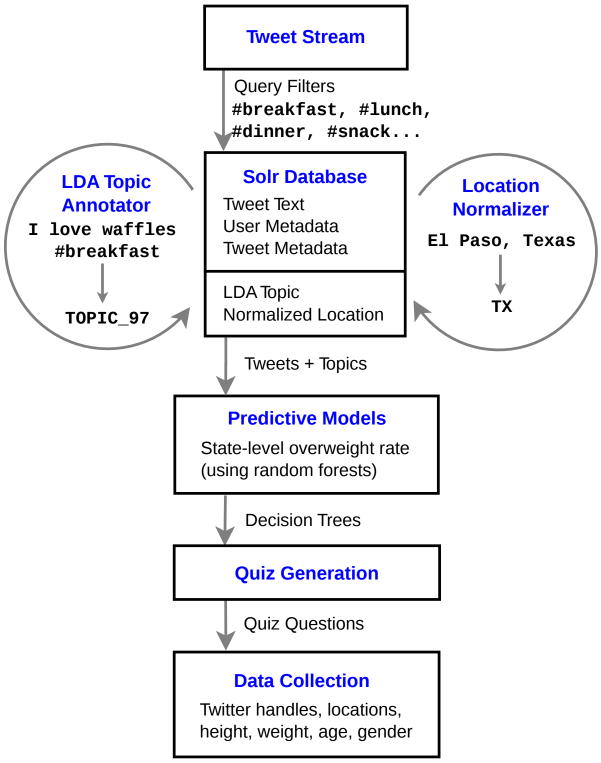


Figure 1: Architecture of the semi-automatic approach for quiz generation from social media data.

3.1. An interpretable model for community classification

Our main data-collection idea is to use a playful 20-questions-like survey, automatically generated from a community-based model, which can be widely deployed to acquire training data on individuals.

Our approach is summarized in Figure 1. As shown in the figure, the first step is our approach is to develop an *interpretable* predictive model that identifies communities more overweight than average, and which can be converted into fun, engaging natural language questions. To this end, we started with the same settings as Fried et al. (2014): we used the 887,310 tweets they collected which were localizable to a specific state and contained at least one relevant hashtag, such as `#breakfast` or `#dinner`. Each state was assigned a binary label (more or less overweight than the median) by comparing the percentage of overweight adults against the median state. For each state, we extracted features based on unigram (i.e. single) words and hashtags from *all* the above tweets localized to the corresponding state. To mitigate sparsity, we also included topics generated using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and all tweets collected by Fried et al. For example, one one of generated topics contains words that approximate the standard American diet (e.g., *baked*, *beans*, *chicken*, *fried*), which was shown in our previous work to correlate with higher rates of overweight population and T2DM (Fried et al., 2014).

Unlike Fried et al. (2014), we do not use support vector machines, but rather a random forest (RF) classifier[6]. The motivation for this decision was interpretability: as shown below, decision trees can be trivially converted into a series of *if* ... *then* ... *else* ... decision statements, which formed the building blocks of the quiz. To minimize the number of questions, we trained a random forest with 7 trees with maximum depth of 3, and we ignored tokens that appear fewer than 3 times in the training data. These parameter values were selected to make the quiz a reasonable length, approximately 20 questions. This is to mirror the popular game "20 questions", in which one player must guess what object the other is thinking of by asking 20 or fewer yes-or-no questions, and as the result of tuning, which revealed that a small number of shallow trees was most effective in partitioning the state-level data accurately. To further increase the interpretability of the model, word and hashtag counts were automatically discretized into three bins (e.g., practically never, sometimes, and often) based on the quantiles of the training data. The code used to generate and test this classifier is available at `http://github.com/clulab/twitter4food/tree/twitterforest`.

Figure 2 illustrates one of the decision trees in the trained random forest, with 0 standing for infrequent, 1 for somewhat frequent, and 2 for very frequent mentions 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 do not mention the word "fruit" very frequently (*not* > 1), and the hashtag "#cook" appears infrequently (*not* > 0).[7] A state with infrequent mention of the word *fruit* would take the left branch, then testing for the frequency of *#cook*. If this is not an infrequent token, then the classifier tests for *curry*; very frequently use of *curry* would lead to an "overweight" classification (relative to the median state).

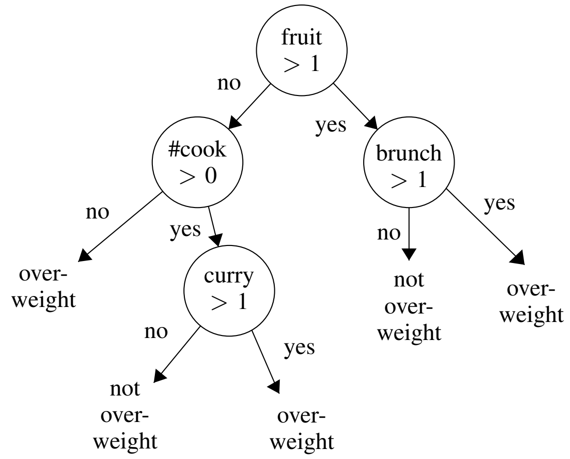


Figure 2: A decision tree from the random forest classifier trained using state-level Twitter data.

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*

[6] `https://code.google.com/p/fast-random-forest/`

[7] Despite 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*	

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

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 of *How often do you mention "fruit" in your tweets?* Table 1 shows the questions and corresponding answers we crafted for the three left-most decision nodes in Figure 2. Conversion to natural language questions was as consistent as possible. For example, whenever the relevant feature's word was a food name *x*, the question would be formulated "How often do you eat *x*?" with an accompanying picture of the food named. When the relevant word was not a food, such as *hot* or *supper*, or a topic, such as the cluster containing *diner*, *bacon*, *omelette*, etc., the question was formulated in terms of proportion of meals rather than frequency.

In all, we generated 33 questions that cover all decision nodes in the random forest classifier. However, when taking the quiz, each individual participant will answer between 12 and 24 questions, depending on their answers and the corresponding traversal of the decision trees.

This quiz serves to gather training data, which will be used in future work to train a supervised model for the identification of individuals at risk. To our knowledge, this approach is a novel strategy for quiz generation, and it serves as an important stepping-stone toward our goal of building individualized public health tools driven by social media. With respect to data retention, we collect (with participants' permission) the following additional data to be used for future research: height, weight, sex, location, age, and social media handles for Twitter, Instagram, and Facebook. We only downloaded public posts using these handles. This data (specifically height and weight) is also immediately used to compute the participant's BMI, to verify whether the classifier was correct.

4. Empirical results

4.1. Evaluation of random forest classifier

Table 2 lists the results of our RF classifier on the task of classifying overweight/not-overweight states. We used the identical experimental settings as (Fried et al., 2014), i.e., leave-one-out-cross-validation on the 50 states plus 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).

Model	*Accuracy*
Majority baseline	50.89
SVM (Fried et al., 2014)	80.39
RF (food + hashtags)	82.35
Discretized RF (food + hashtags)	78.43

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.

Comment type	*%*	*Example*
affective	11	This is awesome. Good luck and keep up the great work!
hypothesizing	20	I probably stumped your system because I love all types of food, but I know how to portion correctly.
cultural	5	was the question about "supper" meant to isolate people from the Midwest?
result-based	13	Surprised that this was correct, I eat like a fat person. Good job!
demographic	25	Single, Caucasian, eat fast food 3-5 times a week
constructive criticism	26	Further breaking down the options would be better.

Table 3: Comments submitted with the demographic questionnaire on the quiz site.

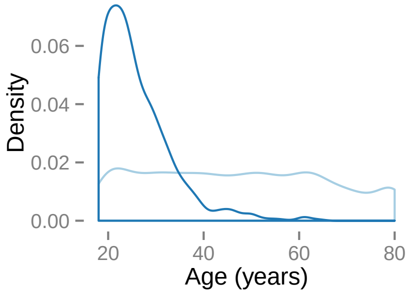
.

4.2. Quiz response

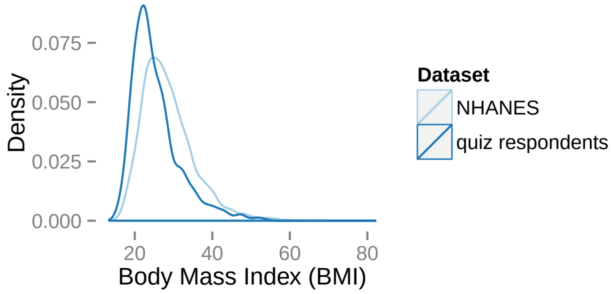
Participants were highly engaged with the quiz: at the end, 97.9% of respondents volunteered demographic information. Many users left feedback, some on the Reddit page linking to the quiz, as shown in Figure 4, some on the quiz page itself, as shown in Table 3. The feedback comprised mostly comments 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 three times 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.

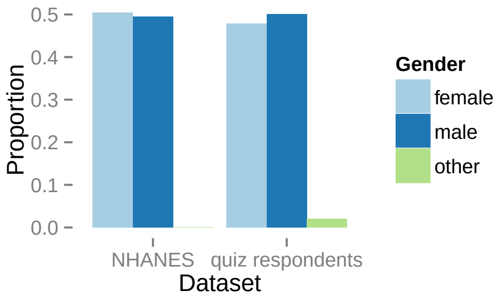
For a better understanding of the feedback received, we performed a post-hoc analysis. Our analysis indicated that while there were 3 comments made about accuracy out of 744 people with correct predictions (0.40% commented), and 13 noting incorrect answers out of 201 with incorrect predictions (6.5% commented). This means that partici-



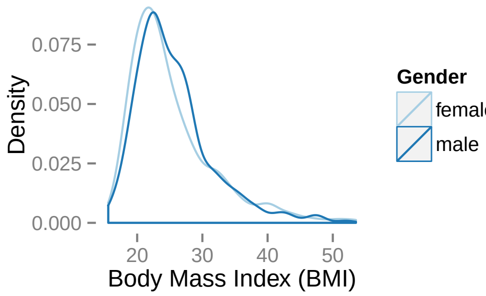
(a) Age distribution in the NHANES survey and in the present study's online quiz.



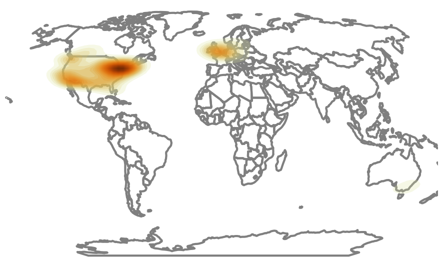
(b) BMI distribution in the NHANES survey and in the present study's online quiz.



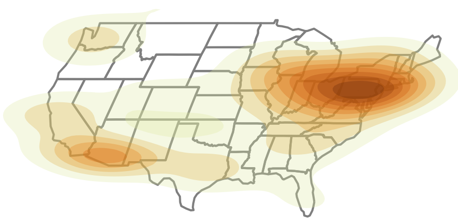
(c) Gender distribution in the NHANES survey and in the present study's online quiz.



(d) BMI distribution by gender in the present study. Too few respondents marked a gender of *other* to allow a density plot.



(e) Geographical distribution of survey respondents worldwide. Of 625 respondents providing location information, 421 (67.4%) provided a location within the US.



(f) Geographical distribution of survey respondents in the US.

Figure 3: Demographic data from the present study and from NHANES. Of 945 respondents, 833 provided their age, 864 provided information for their BMI calculation, 846 provided their gender, and 625 provided their location.

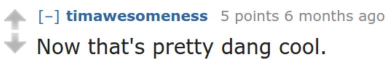
pants were 16 times as likely to comment on the quiz's accuracy if its prediction was incorrect than they were if it was correct. We further classified the Reddit comments received into six classes: affective comments (7%), hypothesizing comments (17%), cultural comments (17%), result-based comments (53%), constructive criticism (7%), and comments seeking a greater understanding of the quiz (7%)[8], examples of which can be seen in Figure 4. The comments made on the quiz site also frequently included additional diet and demographic information.

[8]Numbers sum to more than 100% because of comments with multiple classes of content.

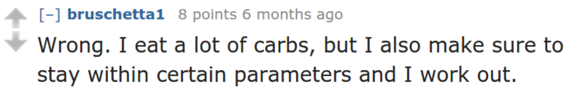
4.3. Quiz evaluation

We evaluated the quiz on 945 volunteers recruited at the University of Arizona and on social media, namely Facebook, Twitter, and Reddit. 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 (with a cutoff boundary BMI of 28.7 – the average US adult BMI according to National Center for Health Statistics (2014)) to the classifier's prediction. 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.

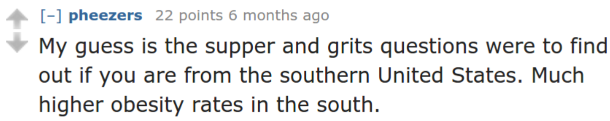
Anonymized results are available at `http://git.io/vZY5U`. They detail the responses of each participant to the



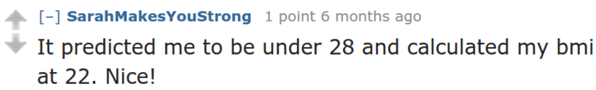
(a) affective



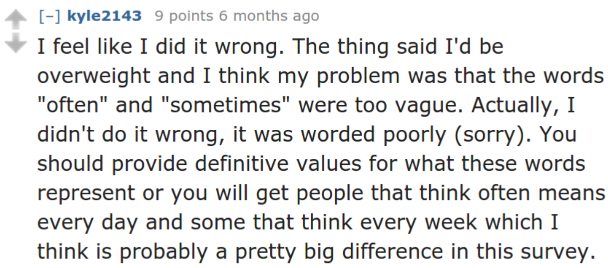
(b) hypothesizing



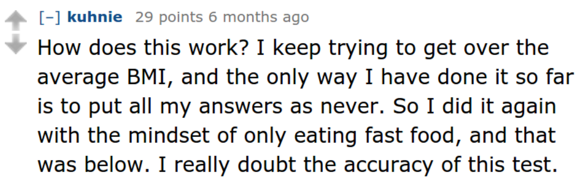
(c) cultural



(d) result-based



(e) constructive criticism



(f) understanding

Figure 4: Some comments of various classes from Reddit users in response to the quiz.

quiz questions, the system's prediction, its accuracy, and the participants' height, weight, location, age, gender, and (anonymized) comments.

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

Quiz accuracy	78.7%
Participants with BMI $\geq$ 28.7	17.7%
Participants with 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).

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.

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 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. References

Abbar, S., Mejova, Y., and Weber, I. (2014). You tweet what you eat: Studying food consumption through twitter. In *arXiv preprint arXiv:1412.4361*.

American Diabetes Association and others. (2008). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 31(Supplement 1):S55–S60.

Ardehaly, E. and Culotta, A. (2015). Inferring latent attributes of twitter users with label regularization. In *North American Chapter of the Association of Computational Linguistics*.

Ashrafian, H., Toma, T., Harling, L., Kerr, K., Athanasiou, T., and Darzi, A. (2014). Social networking strategies that aim to reduce obesity have achieved significant although modest results. *Health Affairs*, 33(9):1641–1647.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Bandari, R., Asur, S., and Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. *CoRR*, abs/1202.0332.

Bian, J., Topaloglu, U., and Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of CIKM Workshop on SHB*, pages 25–32.

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

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of EMNLP*, pages 1301–1309.

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of CIKM*, pages 759–768.

Chew, C. and Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS one*, 5(11).

Chunara, R., Bouton, L., Ayers, J. W., and Brown-stein, J. S. (2013). Assessing the online social environment for surveillance of obesity prevalence. *PloS one*, 8(4).

Cohen-Cole, E. and Fletcher, J. M. (2008). Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics*, 27(5):1382–1387.

Culotta, A. (2014). Estimating county health statistics with twitter. In *CHI*.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*, pages 1277–1287.

Fernandez-Luque, L., Karlsen, R., and Bonander, J. (2011). Review of extracting information from the Social Web for health personalization. *Journal of Medical Internet Research*, 13(1):1382–1387.

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.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Golder, S. A. and Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Greene, J. A., Choudhry, N. K., Kilabuk, E., and Shrank, W. H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of General Internal Medicine*, 26(3):287–292.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

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

McCreadie, R., Macdonald, C., Ounis, I., Osborne, M., and Petrovic, S. (2013). Scalable distributed event detection for twitter. In *International Conference on Big Data, 2013*, pages 543–549. IEEE.

Myslín, M., Zhu, S.-H., Chapman, W., and Conway, M. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *J Med Internet Res*, 15(8):e174, Aug.

Nascimento, T. D., DosSantos, M. F., Danciu, T., DeBoer, M., Holsbeeck, H., Lucas, S. R., Aiello, C., Khatib, L., Bender, M. A., Zubieta, J., and DaSilva, A. F. (2014). Real-time sharing and expression of migraine headache suffering on Twitter: a cross-sectional infodemiology study. *Journal of Medical Internet Research*, 16(4).

National Center for Health Statistics. (2014). 2014 National Diabetes Statistics Report. `http://www.cdc.gov/diabetes/data/statistics/2014StatisticsReport.html`.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P., Ungar, L., and Lucas, R. E. (2013). Characterizing geographic variation in well-being using tweets. In *ICWSM*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Yom-Tov, E., Borsa, D., Cox, I. J., and McKendry, R. A. (2014). Detecting disease outbreaks in mass gatherings using Internet data. *Journal of Medical Internet Research*, 16(6).