Detecting Diabetes Risk from Social Media Activity

Dane Bell[1], **Egoitz Laparra**[2], **Aditya Kousik**[3], **Terron Ishihara**[3], **Mihai Surdeanu**[3], and **Stephen Kobourov**[3]
[1]Department of Linguistics, University of Arizona
[2]School of Information, University of Arizona
[3]Department of Computer Science, University of Arizona
{dane, laparra, adityak, tishihara, msurdeanu, kobourov}@email.arizona.edu

Abstract

This work explores the detection of individuals' risk of type 2 diabetes mellitus (T2DM) directly from their social media (Twitter) activity. Our approach extends a deep learning architecture with several contributions: following previous observations that language use differs by gender, it captures and uses gender information through domain adaptation; it captures recency of posts under the hypothesis that more recent posts are more representative of an individual's current risk status; and, lastly, it demonstrates that in this scenario where activity factors are sparsely represented in the data, a bag-of-word neural network model using custom dictionaries of food and activity words performs better than other neural sequence models. Our best model, which incorporates all these contributions, achieves a risk-detection F_1 of 41.9, considerably higher than the baseline rate (36.9).

1 Introduction

The prevalence of diabetes is increasing in the US, mounting to 30.3 million cases in 2015, of whom 7.2 million were undiagnosed (Centers for Disease Control and Prevention, 2017). Diabetes caused over 79 thousand US deaths in 2015, in addition to $245 billion in economic costs in 2012 (American Diabetes Association, 2013). Along with genetic factors, lifestyle factors such as diet and physical activity are one of the important drivers of risk for Type 2 Diabetes Mellitus (T2DM), the most common type of diabetes. At the same time, the widespread use of social media has produced a digital record of these factors, offering potential insight into how these factors interact to contribute to health risk over time. These publicly available data present an opportunity to detect diabetes risk and similar health risks at scale.

This work shows that the detection of *individuals'* diabetes risk solely from their public Twitter activity is possible, demonstrating that at-risk individuals use language differently from less at-risk individuals. Importantly, this detection is a first, crucial component in a larger battery of social media-based, public-health intervention tools that will work toward disease prevention on a large scale. Specifically, our contributions are:

(1) We introduce a process that creates a novel dataset, which pairs individuals' T2DM risk with their social media activity. We measured individuals' T2DM risk using a well-established, validated questionnaire (Bang et al., 2009), and aligned the result with the corresponding Twitter accounts. To our knowledge, this is the first dataset that directly links T2DM risk with social media activity.

(2) We introduce the first machine learning (ML) approach for classifying individuals' T2DM risk based solely on their Twitter activity. Our deep learning approach has several novel contributions: (a) following previous observations that language use differs by gender, it captures and uses gender information through domain adaptation[1] (Daumé, 2007); (b) it captures recency of posts under the hypothesis that more recent posts are more representative of an individual's current risk status; and, lastly (c) it demonstrates that in this scenario where words representing real-life risk factors are sparsely represented in the data, a bag-of-word (BOW) model that uses custom dictionaries of food and physical activity words is a better solution than recurrent neural networks (RNN). Our best model, which incorporates all these contributions, achieves a risk-detection F_1 of 41.9, considerably higher than the baseline rate (36.9). In comparison, a realistic ceiling model based on the true age, gender, and Body Mass Index (BMI, $\frac{kg}{m^2}$) of each respondent, achieves only 62.7 on this task.

[1]In our experiments, domain adaptation for age did not improve performance.

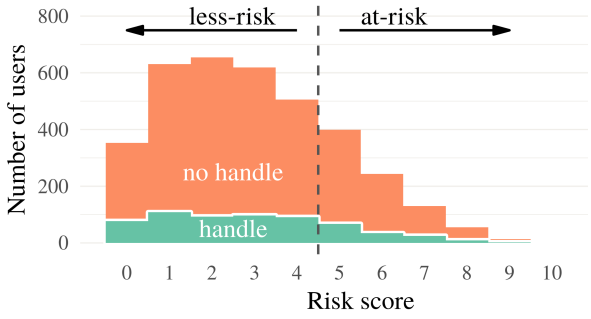

Figure 1: Histogram of respondents' risk scores, with labels as assigned based on Bang et al. (2009).

(3) We provide a feature analysis based on Layerwise Relevance Propagation (Bach et al., 2015; Binder et al., 2016; Arras et al., 2016, 2017), revealing that relevance aligns, albeit inconsistently, to expected food and activity values on average.

2 Data

We collected the dataset used in this work on a voluntary basis through a Qualtrics survey.[2] Participants self-selected by following an URL in an invitation tweet, and after consenting to participate, provided their Twitter handles, demographic information, and answers to an established questionnaire that estimates T2DM risk (Bang et al., 2009). The questionnaire provides an easy-to-understand measure of diabetes risk from data such as age and physical activity level, ranging from 0 to 10, with a score of 5 or higher representing elevated risk. Each participant received a risk assessment, including a summary of the sources of their risk, an explanation of how to get diagnosed (i.e., through a blood test), and a link to further information.

Of the 3,612 respondents who completed surveys, 736 (20.4%) supplied a Twitter handle. After removing respondents who provided no handle, an obviously false handle,[3] or a handle with no public tweets, 604 (16.7%) respondents with handles remained. The relatively modest dataset size is a natural consequence of the complexity of the data and the sensitivity of its collection. The distribution of risk scores among respondents is summarized in Fig. 1.

The complex relationship between height, weight, and risk score is illustrated in Fig. 2. Although BMI is a major risk factor for diabetes, the existence of other factors means that there is considerable risk variation within BMI categories, and the discretization of BMI into categories necessarily obscures variation within categories. Many respondents would change BMI categories if an inch were added to or subtracted from their height, for example.

[2]The collection and analysis was approved by an institutional research board (IRB).

[3]These were inspected manually. Some examples of excluded handles are `@jack` (the example handle given), `@realdonaldtrump`, and `@no`.

	less-risk	*at-risk*
accounts	467	137
tweets (mean)	893 K (1,912)	282 K (2,059)
tokens (mean)	15.2 M (32.5 K)	5.1 M (37.0 K)
# women (%)	312 (67%)	73 (53%)
mean age	36.4	51.1
mean BMI	25.6	34.8

Table 1: A summary of the size and qualities of the *less-risk* and *at-risk* accounts in the dataset collected for this work. *BMI*: Body Mass Index, $\frac{kg}{m^2}$.

We used the Twitter API to collect the tweet and profile text for each handle. The tweets and profile descriptions were tokenized and part-of-speech tagged using ARK Tweet NLP (Owoputi et al., 2013). Each account was labeled *at-risk* if the owner's questionnaire risk score was 5 or greater, or *less-risk* otherwise. A summary of account statistics is shown in Table 1.

3 Approach

We predict individual-level T2DM risk from individual-level data (i.e., individual Twitter accounts), as opposed to transferring from community level statistics (e.g., county diabetes rate as dependent variable; all tweets in that region as input). Intuitively, using a community-level model should be a viable strategy: much more data is available for training; previous work has shown that exploring this data leads to good community-level estimations (Fried et al., 2014). However, our initial experiments showed that individual variation *within* communities was considerable, overshadowing the variation *across* communities and limiting the effectiveness of such methods. In our preliminary experiments the community-level model did not perform better than chance for estimating individual risk.

As a result of this initial analysis, in this work we focus on predicting T2DM risk from individual Twitter accounts. To this end, we propose a neural network (NN) architecture tailored to T2DM risk estimation, which relies on the following resources.

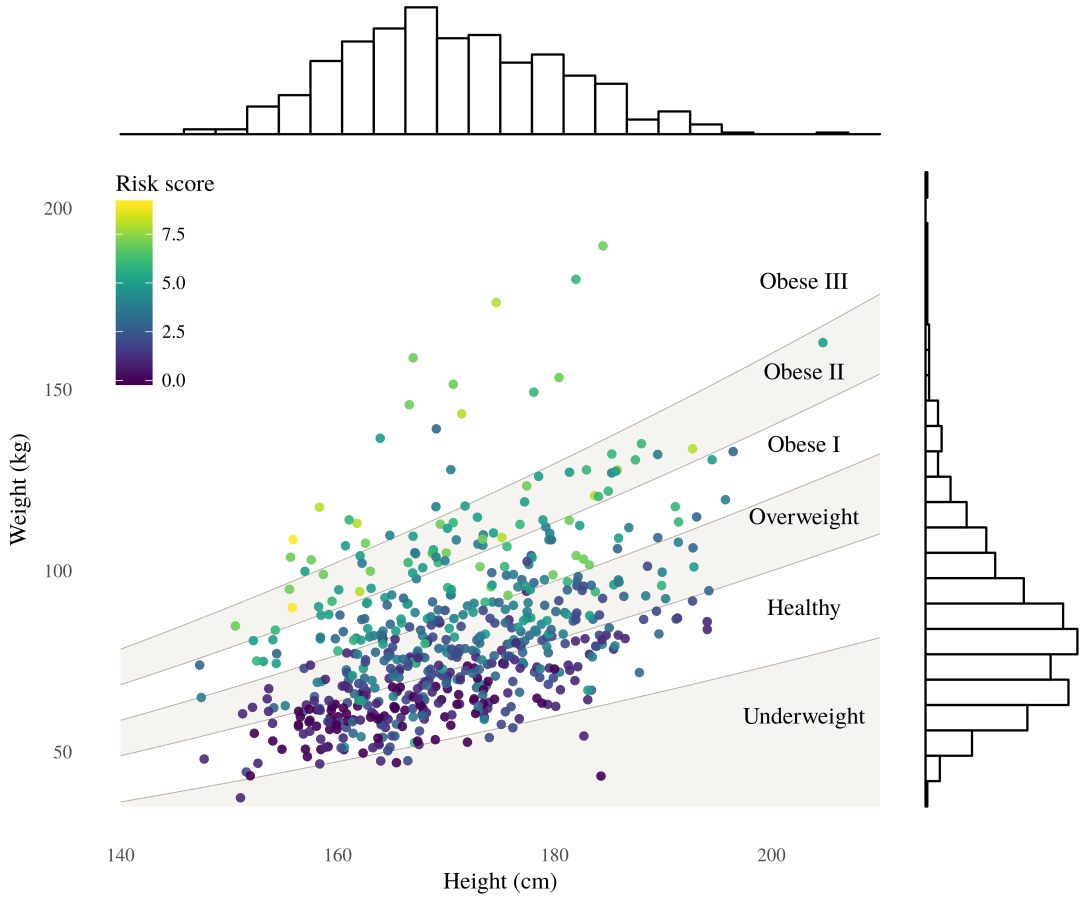

Figure 2: An illustration of the relationship between height, weight, BMI, and risk score for respondents who provided a valid Twitter handle. The BMI categories (underweight, healthy, overweight, etc.) are assigned according to boundaries set by the World Health Organization. The marginal histograms denote the distribution of height (top) and weight (right) in the sample. This figure is best viewed in color.

3.1 Resources

Custom dictionaries: In early experiments, we observed that no model that trained on the posts' entire content outperformed a simple baseline. We explain this result by the fact that indicators of risk factors (e.g., diet or activity words) are sparsely represented in this data, and the models cannot reliably identify them. To mitigate this problem, we created domain-specific dictionaries of words and hashtags indicating foods (*pizza*), exercise (*#5k*), chain restaurant names (*#mcdonalds*), and hashtags related to being overweight (*#fatguyproblems*). The food words were derived from a domain-specific Spanish-English glossary[4] and food vocabulary set[5], following Fried et al. (2014). Exercise words and restaurant names were adapted from Wikipedia lists of sports[6] and restaurants[7]. The smaller list of 13 overweight-related terms were hand-chosen based on Twitter searches.

To adapt the food dictionary to Twitter, we automatically expanded it using semantic vectors. We trained the `word2vec` algorithm (Mikolov et al., 2013) over an independent dataset of 12.3 M food-related tweets[8], creating 200-dimension vectors for each word. From each existing dictionary term, we found the 5 closest candidate words, as measured by cosine distance. Each candidate could appear in multiple lists (e.g. *#breakfastburrito* is similar to both *burrito* and *taco*), so we calculated the softmax of the distances for each candidate. We then expanded our dictionary with the top 500 candidates, which included words such as

[4] www.lingolex.com/spanishfood/a-b.htm
[5] www.enchantedlearning.com/wordlist/food.shtml
[6] en.wikipedia.org/wiki/List_of_sports
[7] en.wikipedia.org/wiki/List_of_the_largest_fast_food_restaurant_chains
[8] Collected automatically using a set of seven diet-related hashtags such as *#breakfast* and *#lunch*.

halloumi, *muesli*, and *sriracha*. After these additions, there was a total of 2,871 features.

Gender: It is well established that language use differs by gender (Rao et al., 2010; Burger et al., 2011; Volkova et al., 2013; Johannsen et al., 2015). On the hypothesis that conditioning classification on these secondary variables would maximize the informativeness of other features[9], we automatically annotated each account for gender. We predicted gender using a SVM model trained on a separate corpus of 1,000 Twitter accounts hand-annotated with gender information (man or woman).[10] This gender classifier used solely unigram features extracted from the account description and its tweets. The macro-averaged F_1 of this model is 75.79 on the T2DM dataset (c.f. human annotators, who averaged 71% accuracy on a similar task (Nguyen et al., 2014)).

3.2 Neural network architecture

We propose a feedforward neural network with one hidden layer, which captures both post recency (by weighing each input word by the recency of the corresponding post) and gender information (captured through domain adaptation). The proposed architecture is depicted and summarized in Fig. 3. This network uses pre-trained word embeddings of 200 dimensions generated using `word2vec` (Mikolov et al., 2013) on the above corpus of food-related tweets. The $tanh$ layer has 128 neurons, and was trained under a 40% dropout. Importantly, this network uses only account words that matched entries in the above custom dictionaries.[11]

Recency weighting: Our preliminary analysis indicated that more recent tweets are more relevant for classification. We attribute the effect of recency to transitions from high to low risk or vice versa due to lifestyle changes, in which case more recent tweets are more representative. To capture recency, we introduce a simple attention mechanism where each word is weighted by its recency, defined as normalized tweet position in the corresponding account. More formally, the recency weight (r_i) of a word w_i is defined as:

$$r_i = \frac{position\_of\_tweet\_containing\_w_i}{\#tweets\_in\_account}$$

where the newest tweet in an account has the highest position. The average embedding ($\overline{x}$) is calculated as:

$$x_i = w_i r_i, \qquad \overline{x} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} r_i}$$

Domain adaptation: We capture gender information using the domain adaptation method of Daumé (2007), adapted to neural networks. As shown in Figure 3, we replicate the output of the $tanh$ layer $\langle t \rangle$ to have a domain-independent version, and one version specific to each domain modeled. For example, the concatenated vector for a female account is $\langle t, t, 0 \rangle$, where 0 is the zero vector corresponding to the male-account domain. This routing process is automatically implemented using the gender classifier described in the previous subsection. All in all, this allows the top sigmoid layer to detect information that generalizes across all domains, in which case the domain-independent vector (t_g) receives a larger update during backpropagation, or is specific to a domain, in which case the corresponding domain-specific vector (t_{d1} or t_{d2}) is updated more.

3.3 Baselines

We implemented three baselines:

(1) All at risk: This baseline assumes all individuals are at risk, i.e., they have a score 5 or higher.

(2) Support vector machines (SVM): This baseline model uses a linear SVM with unigram features from words and hashtags that match our custom dictionaries.[12] Similarly, following the domain adaption method of Daumé (2007), we incorporate gender information by prepending each feature name with the account's gender annotation (in addition to keeping the original feature). For example, an account annotated as a woman who used the word *coffee* 16 times would yield an unigram feature `coffee` in all models, and additionally a feature `gender:woman_coffee`, both with a feature value of 16. (The accounts feature `gender:man_coffee` would have a value of 0.) This allows these models to discover the best generalization for this task, e.g., if *coffee* is an important classification word for women only, the models will put the greatest weight on the *gender:woman_coffee* feature; conversely, if coffee is

[9] We additionally tested this hypothesis, classifying participants' ages into 5 classes (0-20, 21-30, 31-40, 51-60, 61+). Although the age classifier itself performed better than chance, initial experiments showed that age provided no benefit in classifying diabetes risk, and so the influence of age was left to future work.

[10] Non-binary individuals represented $< 1\%$ of our dataset.

[11] Implemented in PyTorch: `http://pytorch.org/`.

[12] Other kernels, larger n-grams, and using all words did not improve performance.

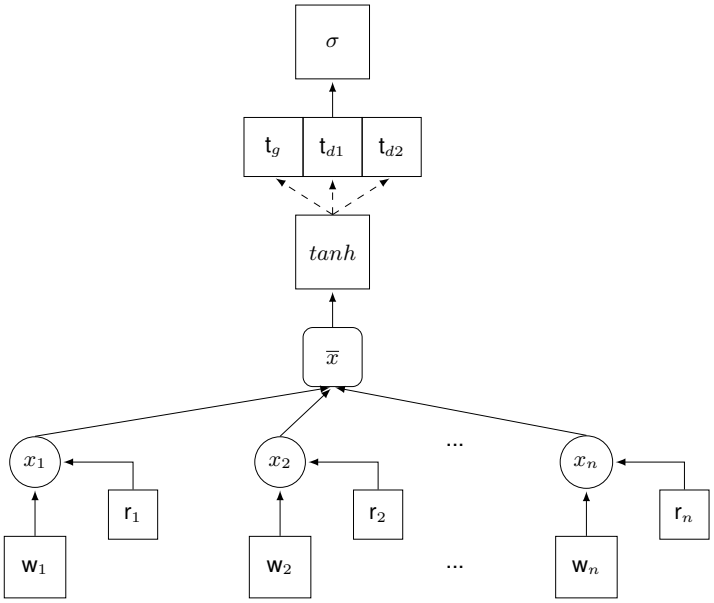

Figure 3: An illustration of the proprosed NN architecture. The bag of words from a user's account matching our custom dictionaries is translated into a set of word embeddings $w_1, w_2, ..., w_n$. The embeddings are multiplied by recency weights $r_1, r_2, ..., r_n$. The resulting vectors are averaged ($\overline{x}$) and passed to the $tanh$ hidden layer. The output of this layer is replicated, producing copies for the general domain, t_g, and for each of the domains, t_{d1} and t_{d2}, e.g., $d1$ = female, and $d2$ = male. If an account belongs to domain $d1$, the copy t_{d2} is set to the zero vector, and vice versa. The copies are then concatenated and fed to the top sigmoid (σ) layer.

always important, the generic unigram feature *coffee* will be assigned greater weight.

(3) Convolutional neural networks (CNN): For this baseline, we apply a CNN layer to the sequence of embeddings of dictionary words that occur in the corresponding account, followed by a rectified linear operator (ReLU). We implement domain adaptation for gender by augmenting the output of the ReLU layer, similarly to the $tanh$ layer in Figure 3. The resulting vector feeds a top $sigmoid$ layer that makes the prediction.[13]

3.4 Ceiling models

We also developed two ceiling models against which to compare our text-based approaches. The first model (Ceiling) is an SVM trained with all the risk assessment variables collected in the survey mentioned in Section 2. This dataset is maximally informative, because these are precisely the variables that determine the risk score (Bang et al., 2009). However, it is not realistic, because most of these features are not available in social media, neither directly nor through machine learning techniques. For this reason, we also implemented an alternative and more realistic version of the ceiling system (Realistic Ceiling) that incorporates only those features that have previously been predicted by automatic systems through social media text or images (see Section 5). The features are summarized in Table 2.

[13]We also experimented with gated recurrent units, and with using all words instead of just dictionary words/hashtags. None outperformed this CNN configuration.

Feature	*Type*	*Ceiling*	*Realistic Ceiling*
age	Integer	✓	✓
gender	Boolean	✓	✓
BMI	Float	✓	✓
diabetic relatives	Boolean	✓	
high blood pressure	Boolean	✓	
little physical activity	Boolean	✓	
gestational diabetes	Boolean	✓	

Table 2: Features available to each ceiling system.

4 Results

We used 10-fold cross-validation to train and evaluate each model on the binary classes *at-risk* and *less-risk* (see Section 2), using the same folds across all models. For each of the 10 runs, we reserve one fold for development, to tune hyperparameters such as classifier confidence cutoff, one fold for testing, and the rest for training. Table 3 summarizes the results of the proposed models, compared against the baselines described in Section 3. In the table, -R marks models that have ***r***ecency information (models without recency used uniform r_i weights), -GG marks

models that used the ***g***old ***g***ender information collected during the questionnaire, and -PG marks models that used ***p***redicted ***g***ender information. The SVM-U is an SVM model using all available words except a stoplist of closed-class words.

The table underlines several observations:

(1) The proposed NN models outperform all baselines, demonstrating that our NNs generalize better on this task dominated by sparse signals. Importantly, most of the strong baselines we include are below the performance of the simple "all at risk" baseline, highlighting again the difficulty of the task. The only baseline that outperformed "all at risk" is CNN-GG, which uses gold gender information, which would not be available in real-world deployments. Interestingly, our approach, which essentially relies on a (recency-weighted) bag-of-word model outperforms all the baselines that rely on sequence models. Similar observations about bag-of-word models outperforming sequence models on complex NLP tasks have been made in the past (Iyyer et al., 2015; Wang and Manning, 2012, inter alia).

(2) Both recency and gender information help. Our best model includes both, validating our original hypotheses. Surprisingly, models using predicted gender performed slightly better than models using gold gender information, but this difference was not statistically significant.

(3) This bag-of-word NN that uses only words/hashtags from relevant dictionaries outperforms considerably other complex NN sequence models that had access to the entire account texts (CNN-all). This highlights the importance of task-specific information (food and activity dictionaries in our case), which, in turn, emphasizes the need of collaboration between NLP researchers and domain (i.e., nutritional science and health care) experts.

(4) Even the Ceiling and Realistic Ceiling classifiers have considerably less than perfect performance at 68.1 and 62.7, respectively. Better performance would be likely with a larger dataset, which would likely also improve the performance of the proposed classifiers.

4.1 Feature analysis

To understand the influence of individual features (tokens) to the classification of an account by the best-performing neural net (using predicted gender and recency-weighted averaging), we adapted the Layerwise Relevance Propagation (Bach et al., 2015; Binder et al., 2016; Arras et al., 2016, 2017) technique. LRP has the advantage of maintaining both positive and negative relevances, representing in this case contribution to the at-risk and less-risk class scores, respectively. In contrast, the commonly used Sensitivity Analysis (Dimopoulos et al., 1995; Gevrey et al., 2003; Simonyan et al., 2013; Li et al., 2015) measures relevance to the decision, rather than to a given class's score, and is therefore always non-negative. LRP assigns relevance to each neuron (including input values) as a function of how much they contribute to the final layer's values, as a share of its layer's contribution. To accomplish this, the neuron's activation must be divided by the sum of whole layer's activation, which can lead to unbounded values when a layer's activations sum to near zero. For this reason, we employ Bach et al. (2015)'s equation 58, which applies a small smoothing constant to the layer's summed activation to the avoid this value explosion.

Examples of accounts' most recent words marked with their relevances according to the NN-PG-R model are shown in Table 4. As the table shows, the health value of words broadly aligns to relevance scores. However, because of the recency weighting of this model, making older tweets' words progressively less relevant, and because of variance in the training of different cross-validation folds, these relevance scores are highly variable. The result is that sometimes a given

Model	*P*	*R*	F_1
All at-risk baseline	22.59	100.00	36.86
Ceiling	67.14	69.12	68.12**
Realistic Ceiling	62.96	62.50	62.73**
SVM-U	**32.82**	31.62	32.21
SVM	28.90	36.77	32.36
SVM-GG	27.95	33.09	30.30
SVM-PG	30.91	37.50	33.89
CNN-GG	26.67	**76.47**	39.54
CNN-PG	24.17	75.00	36.56
NN	28.10	63.24	38.91
NN-R	30.94	60.29	40.90
NN-GG-R	29.39	71.32	41.63*
NN-PG-R	29.38	72.79	**41.86****

Table 3: The precision, recall, and F_1 score of each model in predicting the *at-risk* label. See Section 3 for a description of the models. The *s indicate that the difference in F_1 score between the corresponding model and the best baseline is statistically significant (* indicates $p < 0.05$, and ** indicates $p < 0.01$). All significance values were determined through a one-tailed bootstrap resampling test with 100,000 iterations.

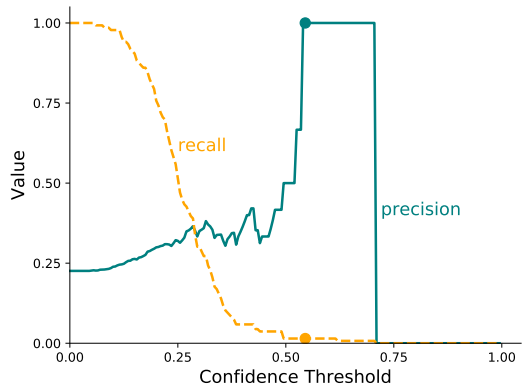

Figure 4: Precision and recall in the NN-PG-R ensemble model, as a function of classifier confidence. The dots mark a threshold of 0.55, at which precision is 100%, and recall is 1.47%. Note that no accounts meet the highest confidence thresholds, leading to a precision and recall of 0, which explains the steep drop of precision for high confidence thresholds.

token is counted as relevant to one classification (e.g., *at-risk*), and other times another (e.g., *less-risk*). This is likely due to both the modest dataset size and to the indirectness of the connection between language use and health.

4.2 Real-world deployment using a high-precision model

The practical application of this risk detection system would involve pointing high-risk individuals toward the Bang et al. (2009) survey, and, if at-risk, to further medical diabetes screening (Rains et al., 2018). To mitigate the drawbacks of false positives (i.e., unnecessary and stressful medical testing), it is likely that in real-world deployments of this technology a high-precision variant of the learned model would be used.

In Figure 4, we show the classification performance at different thresholds for the classifier confidence. In this experiment, in order to increase stability, we have built an ensemble of models through bagging (Breiman, 1996): we generated 50 different versions of the training set by resampling it with replacement, and we trained a different model of the NN-PG-R architecture on each sampled training set. The final predictions are obtained averaging the outputs of the resulting models. As shown in Fig. 4, a threshold of 0.55, for example, yields a precision of 100% and a recall of 1.47%.

Despite the modest recall of such a high-precision model, this classifier would detect a large number of individuals at risk, if applied to the all of Twitter. Assuming the 28.2% prediabetes rate of (Rowley et al., 2017), and the 11% prediabetes diagnosis rate of Li et al. (2013), there are approximately 62 million undiagnosed prediabetic individuals in the US. If we further assume a lower-than-average Twitter adoption of 15%—compared to (Pew Research Center, 2018)'s estimate of approximately 25%—there are roughly 9.2 million Americans who use Twitter and have undiagnosed prediabetes. A similar application to the estimated 7.2 million Americans with undiagnosed diabetes (National Center for Health Statistics, 2017) produces an estimate of 1.1 million unknowingly diabetic Americans on Twitter. Therefore, the successful application of this classfier would identify an estimated 16,000 diabetic and 140,000 prediabetic Americans. Of course, expanding to other English-using Twitter users, and other languages[14] further increases this estimate.

5 Related work

Analysis of social media content for health has been a topic of wide interest (Aramaki et al., 2011; Bian et al., 2012; Prier et al., 2011; Culotta, 2014; Nguyen et al., 2017). Similarly, the literature on detecting user attributes and the effects of those attributes on language use is extensive.

Rao et al. (2010) predict individuals' demographic characteristics of gender, age, and political affiliation based on their tweets. Burger et al. (2011) construct a multilingual dataset of over 100K Twitter accounts, and classify gender better than human annotators, based on account text. Johannsen et al. (2015) study cross-linguistic variation in syntax (part-of-speech and dependency patterns) according to age and gender in online reviews (chosen over tweets for ease of parsing and richer metadata).

Age and gender, while much studied, are not the only available latent characteristics. Mowery et al. (2016) and Vedula and Parthasarathy (2017), for example, predict depressive symptoms in tweet text, a long-term health-variable detection task similar to ours. Similarly, De Choudhury et al. (2013) predict postpartum depression from tweets and Twitter social network structure. Shuai et al. (2016) gather a rich, multi-network feature

[14] https://www.npr.org/sections/goatsandsoda/2017/04/05/522038318/how-diabetes-got-to-be-the-no-1-killer-in-mexico

Correct Label	*Predicted Label*	*Relevance*
less-risk	less-risk	chicken waffles tea reading...
less-risk	at-risk	cake food starving sit sit heart...
at-risk	less-risk	bacon run cup pack writing rolls parkour...
at-risk	at-risk	catfish peanut butter pie picnic bland pop...

Table 4: Examples of relevance displays for words used to assess accounts (most recent tweets first). The red words are relevant to the *less-risk* category, and the blue words to the *at-risk* category, with greater saturation indicating greater relevance.

set to detect social network mental disorders with symptoms such as excessive use of social network sites, measured against gold-data questionnaires. Likewise, Schwartz et al. (2013) predict not only age and gender from the text of Facebook messages, but also the Big Five personality traits (extraversion, emotional stability, agreeableness, conscientiousness, and openness to experience) (Digman, 1990). Moreover, these sometimes-latent user characteristics can inform other classification tasks. For example, Volkova et al. (2013) demonstrate an improvement in the sentiment classification of tweets in a language-independent rule-based model when sentiment vocabulary is adapted for gender-dependent language. Our work continues this direction: here we show that gender information, even when predicted automatically, considerably improves the accuracy of T2DM risk detection.

Much of the previous work on diabetes and weight detection on social media has been at the level of communities. Fried et al. (2014) predict population characteristics such as diabetes and overweight prevalence using location-tagged, food-related tweets. Abbar et al. (2015) analyze correlations between county-level obesity prevalence and food mentions. Again the focus is on predicting dietary choices on a large scale. Relatedly, Eichstaedt et al. (2015) detect heart disease mortality at the county level from tweet text.

There is no known work on detecting individual diabetes risk from social media text. However, Farseev and Chua (2017) capitalize on multiple social media inputs (e.g., a workout tracker) to predict individuals' Body Mass Index category. Wen and Guo (2013) and Kocabey et al. (2017) predict body mass index from images similar to profile pictures, the former from booking photographs and the latter from an internet forum for sharing fitness progress. Of these, only Farseev and Chua (2017) classify solely from text, which is often the only data available from a social media account. Their classification's F_1 is low (17.8) –understandable given the difficulty of this task– which limits its use for realistic T2DM risk prediction. In contrast, our approach obtains a F_1 score that is over 2 times higher, on a task that is arguably more complex.

6 Conclusions

We introduced an approach to the detection of individuals' diabetes risk from their Twitter posts. To this end, we collected a novel dataset linking Twitter activity to a validated, survey-based measure of T2DM risk (Bang et al., 2009). Using this dataset, we proposed the first machine learning approach to predict the T2DM risk of a Twitter account holder using only her tweets. This task is challenging because the data tends to be very sparse, and there are many latent contributing variables (such as genetic predisposition). Our analysis indicates that reducing noise with relevant dictionaries, modeling gender, and modeling posts' temporal recency are valuable in predicting T2DM risk. All in all, our best model achieves an F_1 of 41.9 (vs. the 36.9 "all at risk" baseline and 39.5 of a strong sequence model).

We estimate that if a high-precision variant of this approach were to be deployed at large, e.g., on the public posts of all American Twitter users, it would identify 16,000 diabetic and 140,000 prediabetic Americans that are currently not diagnosed.

Continuing this work, we envision a larger battery of social media-based tools for public-health intervention that focus on the early identification of multiple health risks such as heart disease and various cancers at scale.

7 Release

The system is available as open-source software at github.com/clulab/releases/tree/master/louhi2018-t2dmrisk.

References

Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 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*, CHI '15, pages 3197–3206, New York, NY, USA. ACM.

American Diabetes Association. 2013. Economic costs of diabetes in the US in 2012. *Diabetes care*, 36(4):1033–1046.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. *EMNLP 2017*, page 159.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Heejung Bang, Alison M Edwards, Andrew S Bomback, Christie M Ballantyne, David Brillon, Mark A Callahan, Steven M Teutsch, Alvin I Mushlin, and Lisa M Kern. 2009. Development and validation of a patient self-assessment score for diabetes risk. *Annals of internal medicine*, 151(11):775–783.

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

Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for deep neural network architectures. In *Information Science and Applications (ICISA) 2016*, pages 913–922. Springer.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Centers for Disease Control and Prevention. 2017. National diabetes statistics report, 2017.

A. Culotta. 2014. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1335–1344.

Hal Daumé. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Yannis Dimopoulos, Paul Bourret, and Sovan Lek. 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters*, 2(6):1–4.

Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Aleksandr Farseev and Tat-Seng Chua. 2017. Tweet-Fit: Fusing multiple social media and sensor data for wellness profile learning. In *AAAI*, pages 95–101.

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

Muriel Gevrey, Ioannis Dimopoulos, and Sovan Lek. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, 160(3):249–264.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *CoNLL*, pages 103–112.

Enes Kocabey, Mustafa Camurcu, Ferda Ofli, Yusuf Aytar, Javier Marin, Antonio Torralba, and Ingmar Weber. 2017. Face-to-BMI: Using computer vision to infer body mass index on social media. In *International AAAI Conference on Web and Social Media*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in NLP. *arXiv preprint arXiv:1506.01066*.

YanFeng Li, Linda S Geiss, Nilka R Burrows, Deborah B Rolka, and Ann Albright. 2013. Awareness of prediabetes — United States, 2005–2010. *Morbidity and Mortality Weekly Report*, 62(11):209–212.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 182–191.

National Center for Health Statistics. 2017. Health, united states, 2016: with chartbook on long-term trends in health.

Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961.

Thin Nguyen, Mark E Larsen, Bridianne O'Dea, Duc Thanh Nguyen, John Yearwood, Dinh Phung, Svetha Venkatesh, and Helen Christensen. 2017. Kernel-based features for predicting population health indices from geocoded social media data. *Decision Support Systems*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Pew Research Center. 2018. Social media fact sheet. http://www.pewinternet.org/fact-sheet/social-media/.

Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. 2011. Identifying health-related topics on Twitter: An exploration of tobacco-related tweets as a test topic. In *Proceedings of the 4th International Conference on Social Computing, Behavioral-cultural Modeling and Prediction*, SBP'11, pages 18–25, Berlin, Heidelberg. Springer-Verlag.

Stephen A. Rains, Melanie D. Hingle, Mihai Surdeanu, Dane Bell, and Stephen Kobourov. 2018. A test of the risk perception attitude framework as a message tailoring strategy to promote diabetes screening. *Health Communication*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

William R Rowley, Clement Bezold, Yasemin Arikan, Erin Byrne, and Shannon Krohe. 2017. Diabetes 2030: insights from yesterday, today, and future trends. *Population health management*, 20(1):6–12.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip S Yu, and Ming-Syan Chen. 2016. Mining online social data for detecting social network mental disorders. In *Proceedings of the 25th International Conference on World Wide Web*, pages 275–285. International World Wide Web Conferences Steering Committee.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Nikhita Vedula and Srinivasan Parthasarathy. 2017. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136. ACM.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *EMNLP*, pages 1815–1827.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Lingyun Wen and Guodong Guo. 2013. A computational approach to body mass index prediction from face images. *Image and Vision Computing*, 31(5):392–400.