What does the language of foods say about us?

Hoang Van[*], **Ahmad Musa**[*], **Hang Chen, Mihai Surdeanu, Stephen Kobourov**
Department of Computer Science, University of Arizona
{vnhh,ahmadmusa,hangchen,msurdeanu,kobourov}@email.arizona.edu

Abstract

In this work we investigate the signal contained in the language of food on social media. We experiment with a dataset of 24 million food-related tweets, and make several observations. First, the language of food has predictive power. We are able to predict if states in the United States (US) are above the median rates for type 2 diabetes mellitus (T2DM), income, poverty, and education – outperforming previous work by 4–18%. Second, we investigate the effect of socioeconomic factors (income, poverty, and education) on predicting state-level T2DM rates. Socioeconomic factors do improve T2DM prediction, with the greatest improvement coming from poverty information (6%), but, importantly, the language of food adds distinct information that is not captured by socioeconomics. Third, we analyze how the language of food has changed over a five-year period (2013 – 2017), which is indicative of the shift in eating habits in the US during that period. We find several food trends, and that the language of food is used differently by different groups such as different genders. Last, we provide an online visualization tool for real-time queries and semantic analysis.

1 Introduction

With an average of 6,000 new tweets posted every second, Twitter[1] has become a digital footprint of everyday life for a representative sample of the United States (US) population (Mislove et al., 2011). Previously, Fried et al. (2014) demonstrated that the language of food on Twitter can be used to predict health risks, political orientation, and geographic location. Here, we use predictive models to extend this analysis – exploring the ways in which the language of food can shed insight on health and the changing trends in both food culture and language use in different communities over time. We apply this methodology to the particular use case of predicting communities which are risk for type 2 diabetes mellitus (T2DM), a serious medical condition which affects over 30 million Americans and whose *diagnosis alone* costs $327 billion each year[2]. We refer to T2DM as diabetes in the rest of the paper. We show that by combining knowledge from tweets with other social characteristics (e.g., average income, level of education) we can better predict risk of T2DM. The contributions of this work are four-fold:

1. We use the same methods proposed by Fried et al. (2014) with a much larger (7 times) tweet corpus gathered from 2013 – 2017 to predict the risk of T2DM. We collected over 24 million tweets with meal-related hashtags (e.g., *#breakfast*, *#lunch*) and localized 5 million of them to states within the US. We show that more data helps, and that by training on this larger dataset the state-level T2DM risk prediction accuracy is improved by 4–18%, compared to the results in Fried et al. (2014). We also apply the same models to predict additional state-level indicators: income, poverty, and education levels in order to further investigate the predictive power of the language of food. On these prediction tasks, our model outperforms the majority baseline by 12–34%. We believe that this work may drive immediate policy decisions for the communities deemed at risk without awaiting for similar results from major health organizations, which take months or years to be generated and disseminated.[3] Equally as important, we believe that this state-level T2DM risk prediction task may improve predicting risks

[*]Equal contribution.

[1]https://twitter.com/

[2]http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

[3]https://www.cdc.gov/nchs/nhis/about_nhis.htm

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 87–96
Hong Kong, November 3, 2019.
https://doi.org/10.18653/v1/D19-62

for *individuals* from their social media activity, a task which often suffers from sparsity (Bell et al., 2018).

2. Unlike (Fried et al., 2014), we also investigate the effect of socioeconomic factors on the diabetes prediction task itself. We observe that aggregated US social demographic information from average income[4], poverty[5], and education[6] is complementary to the information gained from tweet language used for predicting diabetes risk. We add the correlation between each of these socioeconomic factors and the diabetes[7] rate in US states as additional features in the models in (1). We demonstrate that the T2DM prediction model strongly benefits from the additional information, as prediction accuracy further increases by 2–6%. However, importantly, the model that relies solely on these indicators performs considerably worse than the model that includes features from the language of food, which demonstrates that the language of food provides distinct signal from these indicators.

3. Furthermore, with a dataset that spans nearly five years, we also analyze language trends over time. Specifically, using pointwise mutual information (PMI) and a custom-built collection of healthy/unhealthy food words, we investigate the strength of healthy/unhealthy food references on Twitter, and observe a downward trend for healthy food references and an upward trend for unhealthy food words in the US.

4. Lastly, we provide a visualization tool to help understand and visualize semantic relations between words and various categories such as how different genders refer to vegetarian vs. low-carb diets.[8] Our tool is based on semantic axes plots (Heimerl and Gleicher, 2018).

2 Related Work

Many previous efforts have shown that social media can serve as a source of data to detect possible health risks. For example, Akbari et al. (2016) proposed a supervised learning approach that automatically extracts public wellness events from microblogs. The proposed method addresses several problems associated with social media such as insufficient data, noisiness and variance, and interrelations among social events. A second contribution of Akbari et al. (2016) is an automatically-constructed large-scale diabetes dataset that is extended with manually handcrafted ground-truth labels (positive, negative) for wellness events such as diet, exercise and health.

Bell et al. (2016) proposed a strategy that uses a game-like quiz with data and questions acquired semi-automatically from Twitter to acquire relevant training data necessary to detect individual T2DM risk. In following work, Bell et al. (2018) predicted individual T2DM risk using a neural approach, which incorporates tweet texts with gender information and information about the recency of posts.

Sadeque et al. (2018) discussed several approaches for predicting depression status from a user's social media posts. They proposed a new latency-based F1 metric to measure the quality and speed of the model. Further, they re-implemented some of the common approaches for this task, and analyzed their results using their proposed metric. Lastly, they introduced a window-based technique that trades off between latency and precision in predicting depression status.

Our work is closest to (Fried et al., 2014). Similar to us, Fried et al. (2014) predicted latent population characteristics from Twitter data such as overweight rate or T2DM risk in US states. Our work extends (Fried et al., 2014) in several ways. First, in addition of tweets, we incorporate state-level indicators such as poverty, education, and income in our risk classifier, and demonstrate that language provides distinct signal from these indicators. Second, we use the much larger tweet dataset to infer language-of-food trends over a five-year span. Third, we provide a visualization tool to explore food trends over time, as well as semantic relations between words and categories in this context.

3 Data

We collected tweets along with their meta data with Twitter's public streaming API[9]. Tweets have been filtered by a set of seven hashtags to make the dataset more relevant to food (see distribution in Table 1). We stored the tweets and their metadata

[4] https://www.census.gov/topics/income-poverty/income.html

[5] https://www.census.gov/topics/income-poverty/poverty.html

[6] https://talkpoverty.org/indicator/listing/higher_ed/2017

[7] https://www.kff.org/other/state-indicator/adults-with-diabetes

[8] http://t4f.cs.arizona.edu/

[9] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/connecting.html

Term	# of tweets	# of tweets localized in US
#dinner	5,455,890	1,367,745
#breakfast	5,125,014	1,183,462
#lunch	4,969,679	1,094,681
#brunch	1,910,950	681,978
#snack	797,676	220,697
#meal	495,073	101,976
#supper	124,979	22,154
Total	24,493,223	4,362,940

Table 1: Seven meal-related hashtags and their corresponding number of tweets filtered from Twitter. The right-most column indicates the number of tweets we could localize to a US state or Washington D.C.

into a Lucene-backed Solr instance.[10] This Solr instance is used to localize the tweets in the US and annotate them with topic models afterwards.

All in all, we collected over 24 million tweets from the period between October 2, 2013 to August 28, 2018, a dataset that is seven times larger than that of Fried et al. (2014). Both datasets contain tweets filtered using the same 7 meal-related hashtags. In order to localize the tweets in the US, we use self-reported location, timezone, and geotagging information (latitude and longitude). The geolocalization is performed in two steps. First, we use regular expressions to match a user's reported location data with the names or postal abbreviations of the 50 US states (e.g., Arizona or AZ) and Washington D.C., and also with city names or known abbreviations (e.g., New York City or NYC). Second, if we cannot find a match, then we use the latitude and longitude information (if provided in the metadata) to localize a tweet. This allowed us to successfully localize approximately 5 million out of the 24 million tweets. For the remaining tweets, latitude/longitude data is converted into city, state, or country using Geopy[11], successfully localizing an additional hundred thousand tweets[12]. Each tweet is preprocessed and filtered to remove punctuation marks, usernames, URLs, and non-alphanumeric characters (but not hashtags).

4 Approach

This work aims for four main goals: predicting state-level characteristics, evaluating the effect of socioeconomic factors in these prediction tasks, analyzing food trends, and using visualization tools to capture trends in the usage of the language of food by different population groups.

4.1 State-level prediction tasks

We investigate the predictive power of the language of food through four distinct prediction tasks: T2DM rate, income, poverty, and education level. We use the tweets from the above dataset as the only input for our prediction models.

T2DM rate prediction: We use the diabetes rate from the Kaiser Commission on Medicaid and Uninsured (KCMU)'s analysis of the Center for Disease Control's Behavioral Risk Factor Surveillance System (BRFSS) 2017 Survey (its most recent year)[7]. The state-level diabetes rate is defined as the percentage of adults in each state who have been told by a doctor that they have diabetes. The median diabetes rate for the US is 10.8%. For each state, we convert the diabetes rate into a binary variable with a value of 1 if the state diabetes rate is greater than or equal to the national median rate, and a value of 0 if it is below. For example, the state with highest diabetes rate, West Virginia (15.2%), is assigned a binary variable of 1 (high T2DM rates). On the other hand, states with below-national-median rate, like Arizona (10.4%), are assigned a label of 0 (low T2DM rates).

Income rate prediction: We collect income data from the United States Census Bureau (USCB)'s analysis of the American Community Survey (ACS)'s Income and Poverty in the United States: 2017[4]. The data shows that national median household income is $60,336. Similarly to above, we convert the household median income of the state into a binary variable with a value of 0 (low income) if its median household income is lower than national median, and a value of 1 (high income) if its median household income is equal or greater. For example, Alabama ($48,193) is labeled as low-income and Alaska ($74,058) is labeled as high-income.

Poverty rate prediction: To predict poverty rates, we also collect poverty data from the USCB's analysis of the ACS's Income and Poverty in the United States: 2017[5], which shows that national median poverty rate is 13.4%. Again, we assign each state a binary variable indicating whether its rate is above or below this national median.

Education rate prediction: For predicting education rate, we use the higher education attain-

[10] `https://lucene.apache.org/`. Solr is the open source NoSQL search platform from the Apache Lucene project.

[11] `https://pypi.org/project/geopy/`

[12] As our work is centered around state-level analysis, we do not use the remaining unlocalized tweets in this paper.

ment rate (HEAR) data from the Center of American Progress (CAP)[6]. The data shows that national median HEAR is 43.2%. Once again, the state-level HEAR is converted to a binary variable in the same manner as above.

Because each of these binary variables is at the state level, we group the tweets by state before feature extraction. We use leave-one-out cross-validation (LOOCV) as proposed by Fried et al. (2014). This approach is necessary because even though we have a large tweet corpus, we only have 51 aggregate data points (one for each state plus Washington, D.C.). For classification, we use Support Vector Machines (SVM) (Vapnik, 2013) for feature-based classification. To avoid overfitting, we tuned the classifier's hyper-parameters during training using the tweets from 2013 to 2016. We tested the tuned prediction models for each task using solely tweets from 2017.

We use two sets of features: lexical (words from tweets) and topical (sets of words appearing in similar contexts). For lexical features, we compare open (all unique tweet words or hashtags) and closed (800 food words) vocabularies, using the token counts as the tweet features. These experiments help us to determine the predictive power of the specific language of food versus the broader context in the full tweets (or socially compact hashtag). For topic model features, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003), to learn a set of topics from food tweets. Because tweets are very short in nature (up to 140 characters), this approach allows us to analyze correlations that could go beyond individual words. We chose 200 as the number of topics for LDA to learn. After LDA is trained using MALLET[13], we use it to create the set of topics for each tweet, and the topic with highest probability is then assigned to each tweet as an additional feature. Topics are counted across all tweets in a state in the same manner as the lexical features.

We also experimented with Deep Averaging Network (DAN) (Iyyer et al., 2015), a simple but robust bag-of-words model based on averaging word embeddings that has been shown to perform well in sentiment analysis and factoid question answering with little training data. In our case, we implemented DAN with embeddings generated using Word2Vec (Mikolov et al., 2013) trained over all 24 million tweets (including the ones that were not localized). We compute the embedding for each token in our dataset, and pass them to the network; see Figure 1. Again using LOOCV, in each pass we leave out one state, train the network on tweets from the 50 other states and predict the T2DM rate for the left out state.

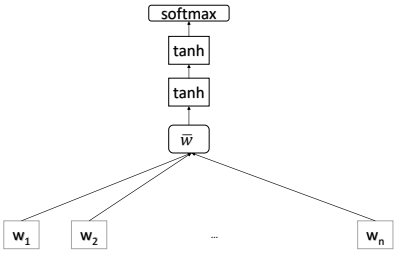

Figure 1: Deep Averaging Network for prediction tasks. The embeddings are averaged and passed to two non-linear layers (tanh).

4.2 Impact of socioeconomic factors

Previous work has shown that the T2DM rate can be predicted by socioeconomic factors such as poverty (Chih-Cheng et al., 2012), income (Yelena et al., 2015), and education (Ayyagari et al., 2011). Therefore, we incorporate these factors into our prediction models (Section 4.1) to assess their contribution. We represent each socioeconomic factor and its correlation with the T2DM rate in the corresponding state as a feature, and include these new features alongside the lexical and topic-based ones. Even though in general, the correlations are relatively low (see Table 2), we will show that the model strongly benefits from the additional information leading to accuracy increases of 2–6% (see Section 5). This indicates that the language of food captures different signal and reflects distinct information from these indicators. However, because these indicators are represented as single features, as opposed to the other features (e.g., there are tens of thousands of food word features, each of which is represented as an integer count), they tended to be ignored by the classifier. To account for this, we empirically explored a series of multipliers to increase the weights of the values of these indicator features[14]. For this task, we use the same SVM classifier from Section 4.1, as well as a Random Forest (RF) classifier (Breiman, 2001)[15].

[13] http://mallet.cs.umass.edu/

[14] For these correlation multipliers, we experimented with powers of 10, from 10^1 to 10^6.

[15] To avoid overfitting, we do not fine-tune the RF classifier's hyperparameters.

Socioeconomic factor	Correlation with T2DM
Education	-0.37
Income	-0.14
Poverty	0.18

Table 2: Correlation between socioeconomic factors (education, income, poverty) and type 2 diabetes mellitus (T2DM) in 2017. Each correlation is calculated from the binary data described in Section 4.1. For the correlation values we used Pearson correlation (Boslaugh, 2012).

4.3 Exploring food trends

We use pointwise mutual information (PMI) between food words/hashtags and years to analyze food trends over time. We divide our corpus of tweets into four parts, each containing a complete year's set of tweets (from 2014 to 2017) and then calculate PMI for pairs (food term t, year y) using the formula:

$$PMI(t, y) = \frac{C(t, y)}{C(t) * C(y)}, \quad (1)$$

where, $C(t, y)$ is the number of occurrences of term t in year y, $C(t)$ is the total number of occurrences of the term, and $C(y)$ is the number of tweets in year y. Intuitively, the higher the PMI value of a term in a given year, $PMI(t, y)$, the more that term is associated with tweets from that year in particular.

4.4 Semantic axes analysis

Word vector embeddings are a standard tool used to analyze text, as they capture similarity relationships between the different words. However, interpreting such embeddings and understanding the encoded grammatical and semantic relations between words can be difficult due to the high dimensionality of the embedding space (typically 50-300 dimensions).

Semantic axes visualizations allow us to view specific low dimensional projections where the new dimensions can be used to explore different semantic concepts (Heimerl and Gleicher, 2018). For our task, we generate several word embeddings from our dataset using the CBOW Word2Vec model of (Mikolov et al., 2013). Different than other visualization tools (e.g., t-SNE, PCA), when using semantic axes we need to define two semantic axes by two opposite concepts (e.g., *man* vs. *woman* and *breakfast* vs. *dinner*) and project a collection of vectors (words in embedding) based on the specific 2D space. The result is a 2D scatter plot with respect to two different concepts.

We first create a word embedding for all the tweets in our dataset. This allows us to explore the correlations between different concepts.

We further augment the semantic axes tool[16] provided by Heimerl and Gleicher (2018), to allow a concept axis to be defined by two sets of words (rather than exactly two words). For example, instead of having one axis defined by the pair (*vegetables*, *meat*) we can now use two sets of words (*vegetables, fruit, vegetarian, vegan*, etc., and *meat, fish, chicken, beef*, etc.). This allows us to capture more complex concepts such as "meat-eaters" that are not captured by individual words.

5 Results

We present the results for all prediction tasks of state level characteristics, as well as the evaluation of the contribution of socioeconomic factors alongside food language in predicting T2DM rate. We also investigate the shifts in eating habits over time (i.e., food trends), as well as the trends in different groups through our semantic axes experiments.

5.1 State-level characteristics prediction

In Table 3, we show the results for predicting state-level socioeconomic characteristics using various sets of features. We compare the results from our dataset with the results of Fried et al. (2014) for predicting T2DM rates. However, since Fried et al. (2014) do not experiment with predicting poverty, income, and education level, for these we compare against a majority baseline. As there are 51 states (including Washington D.C.), and each binary socioeconomic factor is based on the national median, this means that for each factor there will be 26 states either above or below (resulting in a majority baseline of 50.98%).

Comparing the effects of each type of lexical features and their combination with LDA topic features on these prediction tasks, we make several observations.

Performance comparison by feature set: First and foremost, the results demonstrate that the language of food can be used to predict health and social characteristics such as diabetes risk, income, poverty, and education level. The highest overall performance is achieved by using all tweet words (both with and without LDA). This suggests that

[16] http://embvis.flovis.net/

		Diabetes	Poverty	Income	Education	Average
#	Majority baseline	50.98	50.98	50.98	50.98	50.98
	All Words					
1	Fried et al. (2014)	64.71	–	–	–	–
2	Our dataset	**74.51**	64.71	80.39	74.51	73.53
	All Words + LDA					
3	Fried et al. (2014)	64.71	–	–	–	–
4	Our dataset	**70.59**	66.67	82.35	74.51	73.53
	Hashtags					
5	Fried et al. (2014)	68.63	–	–	–	–
6	Our dataset	**74.51**	64.71	80.39	66.67	71.57
	Hashtags + LDA					
7	Fried et al. (2014)	68.63	–	–	–	–
8	Our dataset	**72.55**	62.75	84.31	68.63	72.06
	Food					
9	Fried et al. (2014)	60.78	–	–	–	–
10	Our dataset	**72.55**	62.75	64.71	62.75	65.69
	Food + LDA					
11	Fried et al. (2014)	60.78	–	–	–	–
12	Our dataset	**78.43**	62.75	62.75	62.75	66.67
	Food+Hashtags					
13	Fried et al. (2014)	62.75	–	–	–	–
14	Our dataset	**72.55**	64.71	78.43	66.67	70.59
	Food+Hashtags+LDA					
15	Fried et al. (2014)	62.75	–	–	–	–
16	Our dataset	**74.51**	64.71	84.31	68.63	73.05

Table 3: Results from using various feature sets to predict state-level characteristics: whether a given state is above or below the national median for diabetes, poverty, income, and education. We also show the average performance across all characteristics. We compare against Fried et al. (2014) as well as the majority baseline. Note that Fried et al. do not predict poverty, income, or education level. The low number of data points (51 states) is responsible for the same accuracy value in multiple experiments.

we can capture significant predictive signal from tweets when capturing food words in context.

The highest prediction performance is seen when predicting the state-level income rate, demonstrating a high correlation between food-related words and income. When predicting state-level diabetes rate, we also see strong predictive power from the language of food – all models perform above 70%, up to 78.43%. This confirms our hypothesis that there is a strong correlation between food-related words (and presumably food behaviors) and diabetes rate, one indicator of public health.

Amount and recency of data: For diabetes prediction, with our larger dataset, we improve upon the results of Fried et al. (2014) (ranging from 4 to 18%). In particular, when we use the food-word features combined with LDA topics, we increase prediction accuracy by almost 18%. These results suggest that more data matters in this type of analysis, as evidenced by the learning curves shown in Figure 2, where we compare performance against amount of training data (by year).

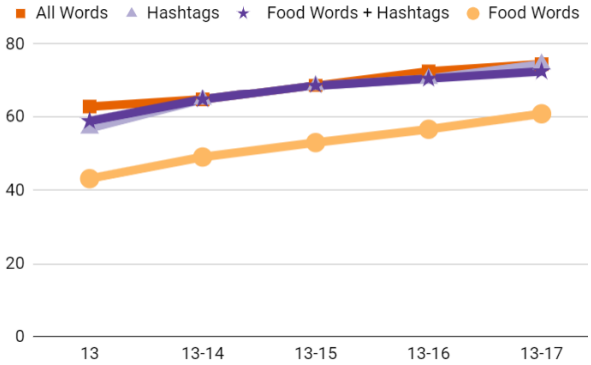

Figure 2: The learning curves for each lexical feature set in terms of predicting diabetes rate in 2017. The horizontal axis corresponds to the cumulative date range used, i.e., 13 only uses tweets from 2013, and 13-14 uses tweets from 2013 through 2014, etc. The y-axis is the state-level prediction accuracy.

We also created learning curves for prediction of T2DM, but from the opposite direction, i.e., starting from tweets from 2017 only, and then adding tweets from earlier years one year at a time. We observe that the more recent the data, the more useful it is for prediction. We hypothesize that in terms of the utility of increased data, the performance of food-word features is improved only as the amount of *relevant* data increases. For the first part of the curve (only from 17, combined 17–16, combined 17–15), the classifier's performance is improved with additional tweets. However, after this peak, additional older tweets decrease performance, suggesting that people change their eating behavior over a period spanning multiple years. The importance of recency of tweet data is also discussed in (Bell et al., 2016).

Comparison to previous work: The best performing model of Fried et al. (2014) relies on hash-tags (see Table 3, lines 5 and 7) and the worst performing model use food words (lines 9 and 11). However, with more data we find that we get the best performance with food words (line 12). We hypothesize that with smaller data, the concise semantics of hashtags are more informative, but with more data the model is able to learn the relative semantics of the food words themselves. Further, while LDA topics do not benefit any model of Fried et al. in terms of predicting diabetes, here we find that with additional data, LDA topics benefit the food words model (compare lines 10 and 12), and in fact contribute to our best performing model (line 12), perhaps because additional data leads to more representative LDA topics.

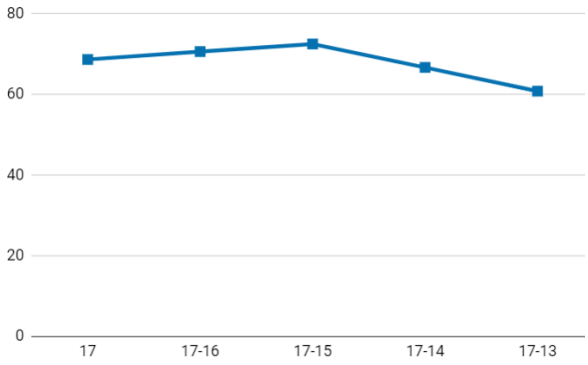

Figure 3: The learning curve using food word features, on the diabetes rate prediction task for 2017. In this figure, the data portion used for each point is in reverse order compared to Figure 2, that is, starting from most recent tweets and going back in time. The horizontal axis is labeled based on the year(s) from which the tweets used for prediction were used.

The Deep Averaging Network from Figure 1 consistently underperformed the results reported in Table 3. This approach obtained an accuracy of 60.78% for the T2DM rate prediction task, considerably lower than the 78% obtained by the best SVM configuration in Table 3. We hypothesize that the reason behind this low performance is the small number of data points (51), which is insufficient to train a neural network.

5.2 Impact of socioeconomic factors

In Table 5, we show the SVM results for predicting T2DM rate from extending the feature matrix from 5.1 with one additional feature based on the correlation between each socioeconomic factor (education, income, and poverty) and T2DM. For each factor, we compare several multipliers (see Section 4.2) to amplify the impact of the socioeconomic correlations. Consistently, we find that models with more features benefit from larger multipliers. For example, the extended food word models that have several hundred features perform best with a multiplier of 10^2, while the other extended models, which all have tens of thousands of features, perform best with a multiplier of 10^3. The best multiplier for each model, according to SVM performance, is used in our Random Forest models (Table 4).

From these extended models we see that using poverty information as an additional feature improves our SVM performance by a range of 2–8% and our RF performance by up to 6%. The other socioeconomic factors, i.e., income and education, do not help when using an SVM classifier (Table 5), but when using a RF classifier we see up to 6% improvement (Table 4). Overall, our highest T2DM prediction performance is obtained with SVM using Food + LDA + poverty. This performance surpasses 80% accuracy and is the highest value reported for this task. Further, to the best of our knowledge, the effect of using poverty information to improve T2DM rate prediction is novel and suggests a potential avenue for improving classifiers with socioeconomic correlation information.

#	Features	Results from best performing multiplier
	All Words+LDA with RF	62.75
	Fried et al. (2014)	64.71
1	+ Education	64.71
2	+ Income	64.71
3	+ Poverty	**64.71**
	Food+LDA with RF	70.59
	Fried et al. (2014)	60.78
4	+ Education	74.51
5	+ Income	72.55
6	+ Poverty	**76.47**
	Hashtags+LDA with RF	68.63
	Fried et al. (2014)	68.63
7	+ Education	64.71
8	+ Income	64.71
9	+ Poverty	**68.63**
	Food+Hashtags+LDA with RF	66.67
	Fried et al. (2014)	62.75
10	+ Education	**72.55**
11	+ Income	**72.55**
12	+ Poverty	70.59

Table 4: Results for predicting T2DM rate using a random forest classifier with our additional socioeconomic correlation features. For each feature set, we use the best performing multiplier, as determined in the previous experiment that used a SVM classifier (Table 5). That is, the best performing multiplier for food word features is 10^2, while other features' multipliers are 10^3.

Importantly, predicting the T2DM below/above median labels from the poverty indicator alone has an accuracy of 58.82%. This value is considerably lower than that of the classifier that uses poverty coupled with the extended word features from tweets, which obtained 80% accuracy. This demonstrates that the language of food provides signal that is distinct from this indicator, which suggests that there is value in social media mining for the monitoring of health risks.

5.3 Food trends

Given our dataset that spans nearly five years, we are also able to investigate whether changes in food habits over time can be detected in social media language. To this end, we explored a list of 800 food words and their change in PMI values in the different years. To understand which food words

#		10^1	10^2	10^3	10^4	10^5	10^6
	All Words + LDA	70.59	–	–	–	–	–
1	+ Education	70.59	70.59	70.59	70.59	70.59	66.67
2	+ Income	70.59	70.59	70.59	66.67	66.67	66.67
3	+ Poverty	66.67	72.55	**78.43**	74.51	70.59	70.59
	Food + LDA	78.43	–	–	–	–	–
4	+ Education	70.59	74.51	68.63	70.59	68.63	68.63
5	+ Income	70.59	74.51	68.63	70.59	66.67	62.75
6	+ Poverty	78.43	**80.39**	76.47	68.63	70.59	70.59
	Hashtags+LDA	72.55	–	–	–	–	–
7	+ Education	70.59	70.59	74.51	70.59	66.67	68.63
8	+ Income	66.67	68.63	70.59	66.67	62.75	66.67
9	+ Poverty	72.55	74.51	**76.47**	64.71	68.63	68.63
	Food+Hashtags+LDA	74.51	–	–	–	–	–
10	+ Education	70.59	70.59	72.55	68.63	68.63	66.67
11	+ Income	66.67	72.55	74.51	68.63	68.63	66.67
12	+ Poverty	72.55	74.51	**78.43**	72.55	68.63	68.63

Table 5: Results for predicting T2DM rate using our SVM classifier, which is similar to that of Fried et al. (2014), but with additional socioeconomic correlation features. Columns show results under different multipliers used to boost the importance of the indicator features (see Section 4.2).

1st annotator	2nd annotator	Score
annotator 1	annotator 2	0.72
annotator 1	annotator 3	0.39
annotator 2	annotator 3	0.58

Table 6: The Cohen's kappa inter-annotator agreement scores among the three annotators.

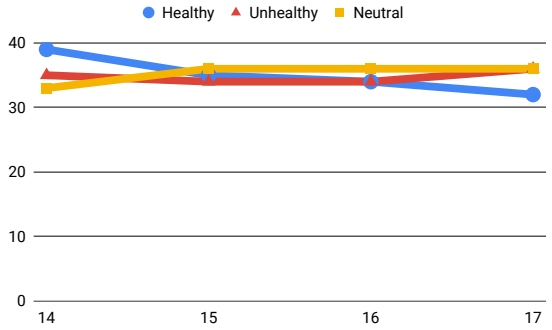

Figure 4: PMI values for each category foods annotated by the 1st annotator. The y-axis shows PMI values 10^6. The trends based on the other annotators are similar.

indicate healthy vs. unhealthy diets, we manually classified the 800 food words into three categories – healthy, unhealthy and neutral – using reliable online resources[17]. The annotations were independently performed by three annotators. The inter-annotator Kappa agreement scores[18] shown in Table 6 indicate fair to good agreement between the three annotators.

We computed PMI values for each of these 800 words and each year in our US dataset. We also computed the PMI values for the three categories and each year (here all words from each category are treated as one). The category trends in our US dataset indicate a slight increase of mentions of unhealthy food words and a slight decrease in mentions of healthy food words in US tweets; see Figure 4. These results suggest a continued decline in dietary patterns in the US, despite seemingly increased interest in health benefits from food[19].

5.4 Semantic axes visualization

As discussed in Section 4.4, visualizations can help discover correlations between different concepts, as well as look at trends over time. In Figure 5, we consider the two axes defined by *man* vs. *woman*, and *breakfast* vs. *dinner*. Exploring the four corners we can identify particular types of foods representative of those coordinates. In the top-left we see words associated with women and breakfast (*yogurt, cupcake, pastry*), whereas in the bottom-left we see words associated with men and breakfast (*sausage, bacon, ham*). Similarly, in the top-right we see words associated with women and dinner (*mussels, halibut, eggplant*) whereas the bottom right we see words associated with men and dinner (*lasagna, lamb, teriyaki*). This data confirms common stereotypes, e.g., (1) men tend to eat more meat, whereas women often prefer fish, and (2) women are more health-conscious compared to men.

We also consider topics (defined by a collection of words) as axes, as illustrated in Table 7. The two axes now are *man* vs. *woman*, and vegetarian words vs. low-carb diets. To represent the vegetarian topic we use the words *vegan*, *vegetarian*, *tofu*, and to represent the low-carb topic we use *keto*, *paleo*, and *atkins*. We then average the word embedding vectors for all words in the topic to create the 2D projection.

We list the 4 corners in the projection as 4 rows in Table 7, where the left column corresponds to the concepts and the right column contains the words. Several patterns emerge: vegetarian words associated with women tend to be soups, salads,

[17] http://www.diabetes.org/ and https://www.healthline.com/health/diabetes/

[18] We use the scikit learn library to calculate the score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[19] https://foodinsight.org/wp-content/uploads/2018/05/2018-FHS-Report-FINAL.pdf

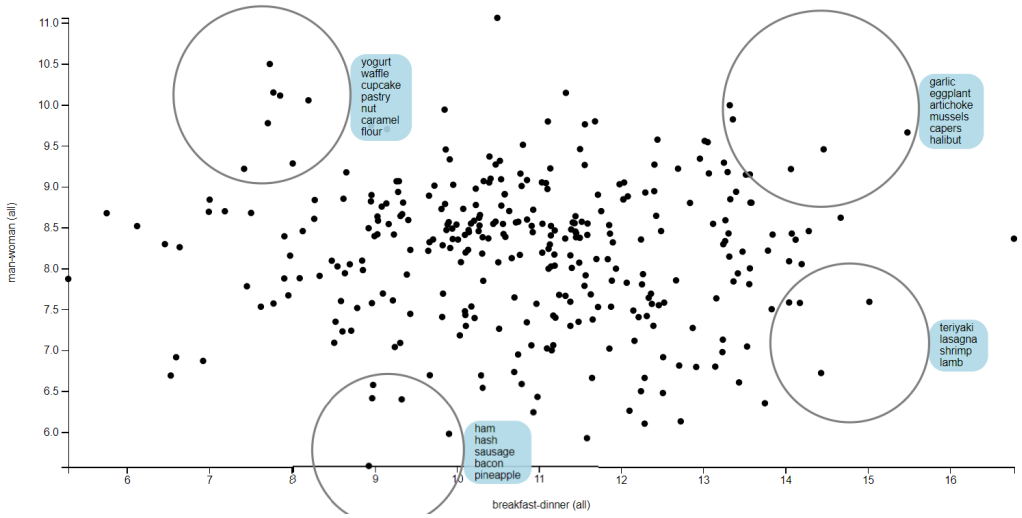

Figure 5: Semantic axes 2D plot using man vs. woman and breakfast vs. dinner as the two axes. We highlight the four corners where interesting patterns can be seen (e.g., the top-left corner is associated with with women and breakfast). Note that this image is a composite of four images, each highlighting one corner.

and gourmet types of foods (saffron, fennel). In contrast, the vegetarian words associated with men tend to be vegetables (spinach, kale, carrot). In the low-carb and women corner we find breakfast words and deserts (cupcake, pastry, caramel, wheat) whereas in the low-carb and men corner we see more hearty foods (spaghetti, hamburgers, buns).

man vs. woman, and vegetarian vs. low-carb diets

woman, vegetarian diet	mint, saffron, fennel, squash, soup, tomato, eggplant
man, vegetarian diet	beet, onion, coconut, spinach, kale, carrot
woman, low-carb diet	hazelnut, nut, cupcake, pastry, grain, caramel, wheat
man, low-carb diet	cereal, spaghetti, buns, hamburger, pepperoni, crunch

Table 7: The 4 corners in the man vs. woman and vegetarian words vs. low-carb diets plot. Each row represents one corner, The left column contains the pair of concepts; the right column contains the foods associated with those concepts.

6 Conclusion

We showed that the language of food has predictive power for non-trivial state-level health tasks such as predicting if a state has higher/lower diabetes risk than the median. When augmented with socio-economic data such as poverty indicators, performance improves further, but we demonstrate that the language of food captures different signal and reflect distinct information from these socio-economic data. We also provide visualization tools to analyze the underlying data and visualize patterns and trends. This work may have immediate use in public health, e.g., by driving rapid policy decisions for the communities deemed at health risk. Further, we hope that this work complements predicting health risk for individuals, a task that is plagued by sparsity, and which could potentially benefit from additional community-level information.

Acknowledgments

Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

M. Akbari, X. Hu, N. Liqiang, and T. Chua. 2016. From tweets to wellness: Wellness event detection from twitter streams. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 87–93. AAAI Press.

P. Ayyagari, D. Grossman, and F. Sloan. 2011. Education and health: evidence on adults with diabetes. volume 11, pages 35–54. Springer.

D. Bell, D. Fried, L. Huangfu, M. Surdeanu, and S. Kobourov. 2016. Towards using social media to identify individuals at risk for preventable chronic illness. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. LREC.

D. Bell, E. Laparra, A. Kousik, T. Ishihara, M. Surdeanu, and S. Kobourov. 2018. Detecting diabetes risk from social media activity. In *Ninth Interna-*

tional Workshop on Health Text Mining and Information Analysis (LOUHI).

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichelet allocation. In *The Journal of Machine Learning Research*, pages 993–1022. JMLR.

S. Boslaugh. 2012. *Statistics in a Nutshell*, volume 2. O'Reilly Media, Inc, Boston, MA.

L. Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

H. Chih-Cheng, L. Cheng-Hua, W. L Mark, H. Hsiao-Ling, C. Hsing-Yi, C. Likwang, S. Shu-Fang, S. Shyi-Jang, T. Wen-Chen, C. Ted, H. Chi-Ting, and C. Jur-Shan. 2012. Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. In *Diabetes Care Vol 35*, pages 2286–2292. ADS.

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.

F. Heimerl and M. Gleicher. 2018. Interactive analysis of word vector embeddings. In *Computer Graphics Forum*, volume 37, pages 253–265. Wiley Online Library.

M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. 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.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

A. Mislove, S. Lehmann, Y. Ahn, J. Onnela, and J. N. Rosenquist. 2011. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

F. Sadeque, D. Xu, and S. Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 495–503, New York, NY, USA. ACM.

V. Vapnik. 2013. *The nature of statistical learning theory*. Springer science & business media.

B. Yelena, L. Mark, R. Marla, and M. John. 2015. The relationship between socioeconomic status/income and prevalence of diabetes and associated conditions: A cross-sectional population-based study in saskatchewan, canada. In *2015 International Journal for Equity in Health*, pages 93–101. BMC.