TODO

Anonymous NAACL submission

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. The performance is further improved by 2–8% with best performance obtained using poverty information (6%), but, importantly, the language of food adds distinct information that is not captured by these indicators. 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. Lastly, we provide an online visualization tool for real-time queries and further analysis of discriminate power of food-related tweets over time.

1 Introduction

Twitter[1], a popular social media network, is a major part of everyday life. With an average of 6,000 new tweets posted every second, Twitter is used for everything from posting screenshots of users' daily lives, to launching political campaigns, and everything in between. This social media now has a digital footprint of our 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 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 collect over 24 million tweets with meal-related hashtags (e.g., #breakfast, #lunch, #snack, etc.) and localize them to states within the US. We show that more data helps, and that by training on this larger dataset the 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. The performance of the model on these prediction tasks outperforms the majority baseline by 12–34%.

2. We also investigate the effect of socioeconomic factors on the diabetes prediction task. We observe that aggregated US social demographic informa-

[1] https://twitter.com/

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

tion from average income[3], poverty[4], and education[5] 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[6] rate in US states as additional features in the models in (1) and demonstrate that the accuracy increases further 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 has distinct information from these indicators.

3. We provide an analysis of food trends over time. Using pointwise mutual information (PMI), we observe the changes in references to popular food items over five years, which provides insight into specific aspects of cultural trends over a given period of time. For example, we discover that healthy foods (e.g. *turmeric*, *jackfruit*) have an upward trend in the american food habits whereas unhealthy foods (e.g. *margarine*, *carbohydrates*) have a downward trend. TODO: Attempted. Ahmad: say that healthier foods trend upwards, and bad ones downwards.

4. We provide a visualization tool[7], to help understand and visualize semantic relations between words and various categories such as how different genders refer to vegetarian vs. low-carb diets. Our tool is based on semantic axes plots (Heimerl and Gleicher, 2018). TODO: For Stephen: One high level sentence to explain what semantic axes really is

2 Data

We collected tweets along with their meta data by querying a public streaming API[8] provided by Twitter. Tweets have been filtered by a set of seven hashtags to make the dataset more relevant to meals (see distributions in Table 1). We put the tweets and their metadata into a Lucene-backed Solr instance. The Solr instance is used to localize the tweets in the US and annotate them with topic models afterwards (Section 3).

We identified the gender of the users posting these tweets TODO: I DIDN'T put this line here. suprised to see it. We didn't experiment with gender in any way. ahmad: explain how you did this? Did you use the classifier built by Dane? If so, summarize it and cite.

All in all, we collected over 24 million tweets, a dataset that is seven times larger than that of Fried et al. (2014), from the period between October 2, 2013 to August 28, 2018. Both datasets have tweets filtered by the same hashtags. In order to localize the tweets in the US, we use the user's self reported location, timezone, and geotagging information (latitude-longitude). The geolocalization is performed in two steps. First, we use regular expressions to match 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. Location data from the metadata successfully localizes about 5 million out of the 24 million tweets. For the remaining tweets, latitude-longitude data is converted into city, state, or country using GeopyTODO: DONE. Ahmad: replace with the exact name[9], successfully localizing an additional 100 thousand tweets[10]. Each tweet is preprocessed and filtered to remove punctuation marks, usernames, URLs, and non-alphanumeric characters (but not hashtags).

3 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 categories such as genders. This section is structured along these four goals.

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

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

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

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

[7] Website anonymized for blind review.

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

[9] https://pypi.org/project/geopy/TODO: DONE Ahmad: Fill in the url

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

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.

3.1 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 Diabetes Control's Behavioral Risk Factor Surveillance System (BRFSS) 2017 Survey (its most recent year)[6]. 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[3]. 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 ACSs Income and Poverty in the United States: 2017[4], 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 attainment rate (HEAR) data from the Center of American Progress (CAP)[5]. 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). To avoid overfitting, we fine-tune the classifier's hyper-parameters during training with the tweets from 2016. Then we only test the fine-tuned prediction models once for each task in 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 correlation that could go beyond individual words. We chose 200 as the number of topics for LDA to learn. After LDA is trained using MALLET[11], 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. Top-

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

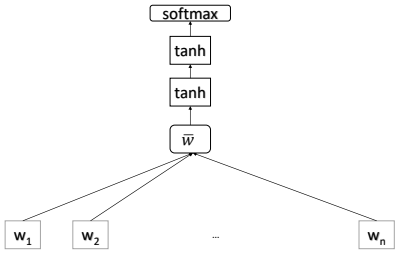


Figure 1: Deep Averaging Network for prediction tasks. The embeddings are generated from the tweet tokens, then they are averaged and passed to the tanh layer. After performing two non-linear operations, they are fed to a softmax layer for prediction.

ics are counted across all tweets in a state in the same manner as the lexical features.

We also experiment 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 proven 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) TODO: DONE. Ahmad: cite the Mikolov paper here not Goldberg trained over all words (preprocessed to filter out punctuation marks, usernames, URLs, and non-alphanumeric characters) in our tweet corpus.TODO: DONE. Ahmad: corpus detail. We compute the embedding for each token in our dataset, and pass them to the network (see Figure 1 for architecture). 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.

3.2 Effect of socioeconomic factors in predicting T2DM rate

Previous work has shown that T2DM rate can be predicted by socioeconomic factors such as poverty (Chih-Cheng et al., 2012), income (Yelena et al., 2015), and education (Padmaja et al., 2011). Therefore, we incorporate these factors into our prediction models (Section 3.1) 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. In general, the correlations are relatively low (see Table 2). Further, 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.[12].

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 3.1. For the correlation values we used Pearson[13] correlation that is available with Microsoft Excel.TODO: DONE Hoang: what correlation formula did you use?.

For this task, we use the same SVM classifier from Section 3.1, as well as a Random Forest (RF) classifier (Breiman, 2001). To avoid overfitting, we do not fine-tune the RF classifier's hyperparameters.

3.3 Exploring food trends with pointwise mutual information

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 (2014-2017) 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.

3.4 Using semantic axes to understand the usage of the language of food by different categories

TODO: Needs a full pass by Stephen.

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

As mentioned in Fried et al. (2014) paper, topic of vegan's words is highly correlated with democratic state. One interesting question is whether such predictive features still help us for prediction by using 50 month rather just 8 month tweets. Is there a way to find a good predictive feature candidate for any prediction tasks by using tweets?

One of the famous way to study word relation is using word embedding. However, those word embedding vectors are usually in 300D space, which makes it hard to interpret. Semantic axes is a 2D space visualization tool which allow use to study high dimensional space. For our task, we generate several word embeddings from our dataset using word2Vec (CBOW model)(Mikolov et al., 2013). Different than other visualization tools (e.g., t-SNE, PCA), one needs to define two semantic axes by two opposite concepts (e.g., man vs. woman) and project a collection of vectors(words in embedding) based on the specific 2D space, which allows them to view different aspects of the original embedding by modifying axes concepts.

Semantic axes is suitable for our task. We can study the relation between democrat and vegan by setting democrat vs. republican as one axis and vegan vs. meat as another axis. As our dataset is a collection of 5 years of tweets, we separate the dataset by year and train an embedding for each year of tweets. We also train one embedding for all tweets. We modify the semantic axes tool[14] provided by Heimerl and Gleicher (2018), to allow us to visualize a set of food words projected by a set of opposite concepts. For each set of concepts, we average the concept's embedding vector. By doing this, we can reduce the bias in the embedding by use a set of synonyms.(not sure is this true or not. what is the motivation for combine several word?)

4 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).

[14] 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

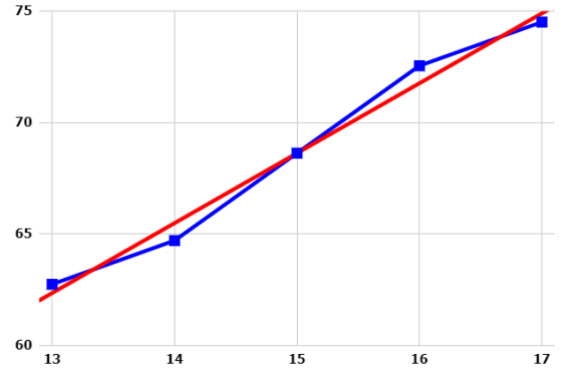
Table 3: Results from using various feature sets to predict state-level characteristics: whether a given state is above or below the national median for diabetes, poverty, income, and education. We also show the average performance across all characteristics. We compare against Fried et al. (2014) as well as the majority baseline. Note that Fried et al. do not predict poverty, income, or education level.

4.1 State-level characteristics prediction

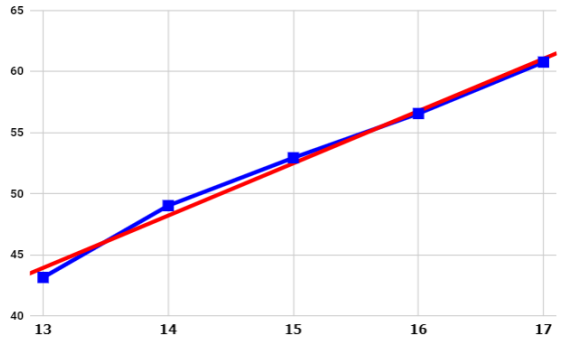
In Table 3, we show the results for predicting state-level socioeconomic characteristics using various sets of features. We compare the result from our dataset with results of Fried et al. (2014) for predicting T2DM rates. However, since Fried et al. (2014) do not experiment with predicting poverty, 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.

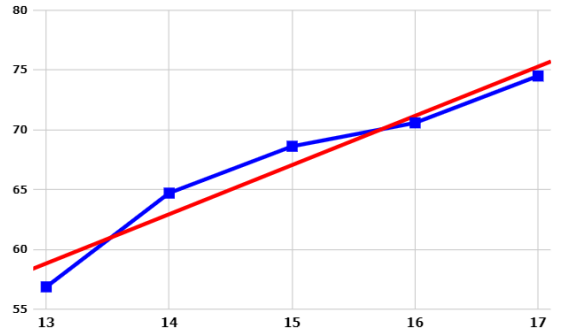
Performance comparison by feature set: First and most important, the results demonstrate that the language of food can be used to predict social characteristics such as diabetes (health), income, poverty, and education level. The highest overall performance is achieved by using all tweet words (both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



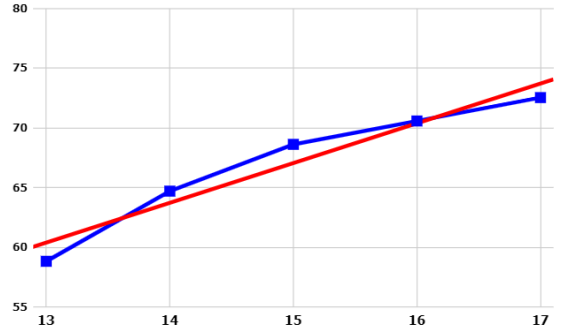
(a) All Words



(b) Food Words



(c) Hashtags



(d) Food Words + Hashtags

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 14 uses tweets from 2013 through 2014, etc. The y-axis is the state-level prediction accuracy. TODO: DONE Hoang: redo x-axis and rescale to fit

we can capture significant predictive signal from

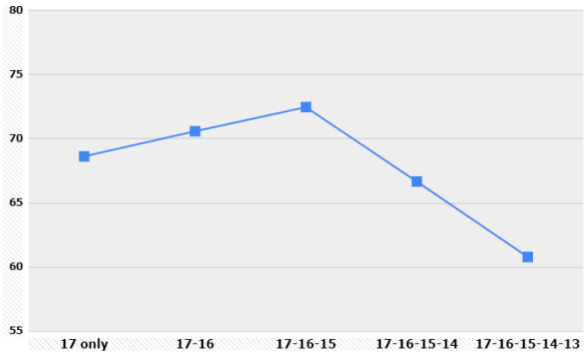


Figure 3: The learning curve using food words features, on the diabetes rate prediction task for 2017. In this figure, the data portion used for each point is in reverse order compared to 2b, that is, starting from most recent tweets and going back in time. The horizontal axis is labeled based on the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

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).

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 the recent data is more useful for prediction. We hypothesize that in terms of the utility of increased data, the performance of food-word features are improved only as the amount of *relevant* data increases. For the first part of the curve (only from 17, combined 17-16, combined 17-16-15), the model's performance is improved with additional tweets. However, after this peak, additional, older tweets decrease performance, suggesting that people change their eating

#		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 4: 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. The columns show results under different multipliers used to boost the importance of the indicator features (see Section 4.2).

#	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 5: 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 4). That is, the best performing multiplier for food word features is 10^2, while other features' multipliers are 10^3.

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: When we compare our results with those of Fried et al. (2014), we see that their best performing models relies on hashtags (see Table 3, lines 5 and 7) and their worst performing models 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.

4.2 Effect of socioeconomic factors on prediction

In Table 4, we show the SVM results for predicting T2DM rate from extending the feature matrix from 4.1 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 (RF) models (Table 5).

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 4), but when using a RF classifier we see up to 6% improvement (Table 5). Overall, our highest T2DM prediction performance is with the SVM with Food + LDA + poverty. This performance surpasses 80% accuracy, which, to our knowledge, 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.

Importantly, predicting the T2DM below/above median labels from the poverty indicator alone has an accuracy of 58.82%. Compared to that with ex-

tended word features from tweets (80% from our best performing model), this proves that the language of food provides signal that is distinct from this indicator.TODO: DONE Hoang: discuss this value, and say that this proves that the language of food provides signal that is distinct from this indicator!

4.3 Food trends

Another hypothesis investigated in this work is if food habits change over time, and if this change is reflected in social media. To this end, we explored a list of approximately 800 foods and their change in PMI values with different years.

We first inspect upward trends, i.e., food words that have low PMI value in 2014 but have high value of PMI in later years. For example, in Fig. 4 we observe that *jackfruit*, which is not native to North America, has recently gained popularity. It is a fruit with a meaty texture that has received recent popularity as a meat replacement. To investigate this, we computed the cosine similarity between a simple mean of the projection weight vectors of *jackfruit* and the vectors for each word found in the tweets containing *jackfruit*.[15] We report these results in Table 6. The table confirms our intuition that jackfruit tends to be used a meat replacement in vegan diets. Similarly, we find that *turmeric*, whose active ingredient is *curcumin*, has an upward trend among Twitter users. Turmeric has been commonly used as a spice in Asian, Indian, African dishes, and has several health benefits such as anti-cancer and anti-inflamatory activity. (Cemil et al., 2010).[16] This trend indicates that usage of turmeric in American cuisine has increased in the past five years.

We also find food words that have decreased in popularity over the same period of time. One such example is *margarine*, which is probably due to the fact that margarine, which was made as a replacement of butter, was found to have a high concentration of saturated fats. Also, margarine consumption is associated with allergic symptoms and diseases in children (Bolte et al., 2001). Also, *Carbohydrates* has been losing popularity because

[15] We used the Word2Vec implementation in Gensim to compute the embedding vectors: https://radimrehurek.com/gensim/models/keyedvectors.html.

[16] See also: https://www.mskcc.org/cancer-care/integrative-medicine/herbs/turmeric.

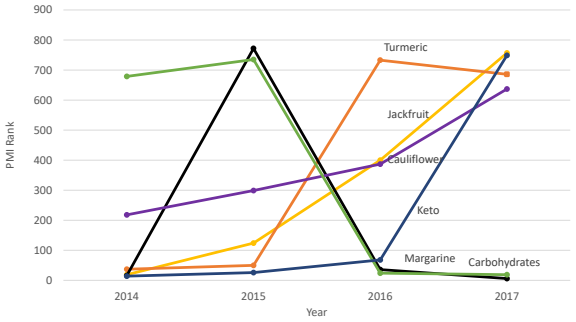


Figure 4: PMI value for each food item over the years. Increase in rank value denotes the upward trend of the food item. These trends highlight that healthy foods, e.g., jackfruit and cauliflower, are on an upward trend, while unhealthy ones, e.g., margarine, are on a downward trend.

more and more people started doing keto or paleo diet. These diets are mostly associated with eating low carbs and more fruits (Fig. 4). TODO: Made an attempt with keto and carbohydrates. Ahmad: can you find another downward trend

All in all, these trends suggest an encouraging discovery: American foods tend to incorporate more healthy foods, and reduce the usage of unhealthy ones.

Word	Words that have highest similarity in context
jackfruit	plantbased, breakfast, vegan, snack, recipe, pulled
margarine	butter, cup, breakfast, cinnamon, vegan, egg
fillet	seared, fryday, snails, grilled, omega, protein
turmeric	spicy, egg, green, yummy, sliders, antiinflammato
cauliflower	paleo, weightloss, seared, potato, mealprep, baked
carbohydrates	cereal, low, complex, infographic, ketogenic, carbs
keto	sauteed, banting, paleo, avacado, cauliflower, pickled

Table 6: Each word on the left is a trending food item in our PMI lookup table and the words on the right for each row are top six words in the matched corpus that have highest cosine similarity with the trending word in embedding space. *Jackfruit* is associated with a plant-based diet, and tastes like pulled pork, which is also evident in the tweets. *Fillet* is associated with frying fish and seafood, which contains a lot of omega and proteins. *Turmeric* has anti-inflammatory effects. *Cauliflower* is popular as a healthy alternative to white rice and gluten-filled grains. *Margarine* is an unhealthy alternative to butter. Diets (e.g. *Keto*) are associated with eating low carbs which is evident in the neighbor words around *Carbohydrates*.

4.4 Semantic axes visualization

TODO: Needs full pass by Hang and Stephen

As we discussed in Sec.3.4, semantic axes helps us to understand language of food. In Fig.5, We define man vs. woman, breakfast vs. dinner as our 2 axes. We divide the 2D space to 4 parts from middle by considering combination of 2 axes. e.g.

the top-left part is more associate with woman and breakfast, and bottom-right corner is associate with man and dinner. For easier to extract information, we put typical foods in a corner(e.g. top-left) into a table. For example In Fig.5, words in the circle is typical food in top-left corner, represent as first row in Table.7

From Table.7, we define woman vs. man, breakfast vs. dinner as 2 axes. woman usually post health or baking food in the morning as their breakfast. Man on the other hand, post about protein food. Woman usually post vegetables and seafood in dinner. Man post about meats.

In Table.8, we define man vs. woman, vegetarian words vs. low-carb diets as 2 axes. In order to represent vegetarian words vs. low-carb diets, We selected list of words for both concepts. We use "vegan", "vegetarian", "tofu" to represent vegetarian word; use "keto", "paleo", "atkins", "meat" to represent low-carb diets.TODO: observation

man vs. woman and breakfast vs. dinner

woman,breakfast	yogurt, waffle, cupcake, pastry, nut, caramel, flour
man, breakfast	ham, hash, sausage, bacon, pineapple
woman, dinner	garlic, eggplant, artichoke, mussels, capers, halibut
man, dinner	teriyaki, lasagna, shrimp, lamb

Table 7: Table of semantic axes by define man vs. woman, breakfast vs. dinner as 2 axes. each row represent one corner , e.g. first row is top-left corner in the axes. The left column is associated concepts, right column is typical food associated with those concept.

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

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

Table 8: Table of semantic axes by define man vs. woman, vegetarian words vs. low-carb diets as 2 axes. In order to represent vegetarian words vs. low-carb diets, We selected list of words for both concepts. We use "vegan", "vegetarian", "tofu" to represent vegetarian word; use "keto", "paleo", "atkins", "meat" to represent low-carb diets. each row represent one corner, e.g. first row is top-left corner in the 2d plot. The left column is associated concepts, right column is typical food associated with those concept.

TODO: Should we have an example about time

5 Conclusion

TODO: Stephen: take a shot?

References

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)*.

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

Gabriele Bolte, Christian Frye, Bernd Hoelscher, Ines Meyer, Matthias Wjst, and Joachim Heinrich. 2001. Margarine consumption and allergy in children. *American journal of respiratory and critical care medicine*, 163(1):277–279.

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

Berker Cemil, Kivanc Topuz, Mehmet Nusret Demircan, Gokhan Kurt, Kagan Tun, Murat Kutlay, Osman Ipcioglu, and Zafer Kucukodaci. 2010. Curcumin improves early functional results after experimental spinal cord injury. *Acta neurochirurgica*, 152(9):1583–1590.

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.

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.

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

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

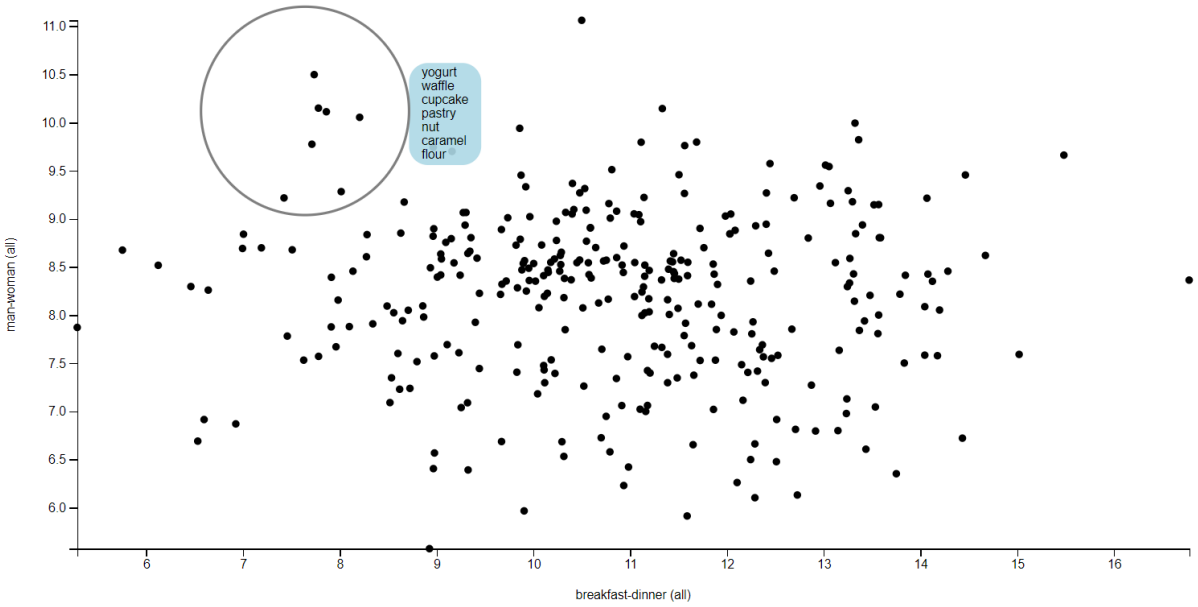


Figure 5: Semantic axes 2D visualization tool. In this specific example, we define man vs. woman, breakfast vs. dinner as 2 axes, the food words in circle are associate with woman and breakfast.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media*.

A. Padmaja, G. Daniel, and S. Frank. 2011. Education and health: Evidence on adults with diabetes. In *Int J Health Care Finance Econ*, pages 35–54.

Vladimir 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.