Analyzing the Language of Food on Social Media

Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, and Dane Bell
University of Arizona, Tucson, AZ, USA
Email: {dfried, msurdeanu, kobourov, hinglem, dane}@email.arizona.edu

***Abstract*—We investigate the predictive power behind the language of food on social media. We collect a corpus of over three million food-related posts from Twitter and demonstrate that many latent population characteristics can be directly predicted from this data: overweight rate, diabetes rate, political leaning, and home geographical location of authors. For all tasks, our language-based models significantly outperform the majority-class baselines. Performance is further improved with more complex natural language processing, such as topic modeling. We analyze which textual features have greatest predictive power for these datasets, providing insight into the connections between the language of food, geographic locale, and community characteristics. Lastly, we design and implement an online system for real-time query and visualization of the dataset. Visualization tools, such as geo-referenced heatmaps and temporal histograms, allow us to discover more complex, global patterns mirrored in the language of food.**

I. Introduction

Our diets reflect our identities. The food we eat is influenced by our lifestyles, habits, upbringing, cultural and family heritage. In addition to reflecting our current selves, our diets also shape who we will be, by impacting our health and well-being. The purpose of this work is to understand if information about individuals' diets, reflected in the language they use to describe their food, can convey latent information about a community, such as its location, likelihood of diabetes, and even political preferences. This information can be used for a variety of purposes, ranging from improving public health to better targeted marketing.

In this work we use Twitter as a source of language about food. The informal, colloquial nature of Twitter posts, as well as the ease of data access, make it possible to assemble a large corpus describing the type of food consumed and the context of the discussion. Over eight months, we collected such a corpus of meal-related tweets together with relevant meta data, such as geographic locations and time of posting. We construct a system for aggregating, annotating, and querying these tweets to create predictive models and interactive visualizations (Fig. 1). Building on this dataset and system, the contributions of this work are fourfold:

1. We analyze the predictive power of the language of food by predicting several latent population characteristics from the tweets alone (after filtering out location-related words to avoid learning trivial correlations). We demonstrate that this data can be used to predict multiple characteristics, which are conceivably connected with food: a state's percentage of overweight population, the rate of diagnosed diabetes, and even political voting history. Our results indicate that the language-based model yields statistically-significant improvements over the majority-class baseline in all configurations, and that more complex natural language processing (NLP), such as topic modeling, further improves results.

2. We demonstrate that the same data accurately predicts geographic home locale of the authors (from city-level, through state-level, to region-level), with our model significantly outperforming the random baseline (e.g., more than 10 times better for city prediction).

3. In addition to examining the effectiveness of our models on these predictive tasks, we analyze which textual features have most predictive power for these datasets, providing insight into the connections between the language of food, geographic locale, and community characteristics.

4. Lastly, we show that visualizations of the language of food over geographical or temporal dimensions can be used to infer additional information such as the importance of various daily meals in different regions, the distribution of different foods and drinks over the course of days, weeks and seasons, as well as some migration patterns in the United States and worldwide.

II. Data

Twitter provides an accessible source of data with broad demographic penetration across ethnicities, genders, and income levels[1], making it well-suited for examining the dietary habits of individuals on a large scale. To identify and collect tweets about food, we queried Twitter's public streaming API[2] for posts containing hashtags related to meals (Table I). We

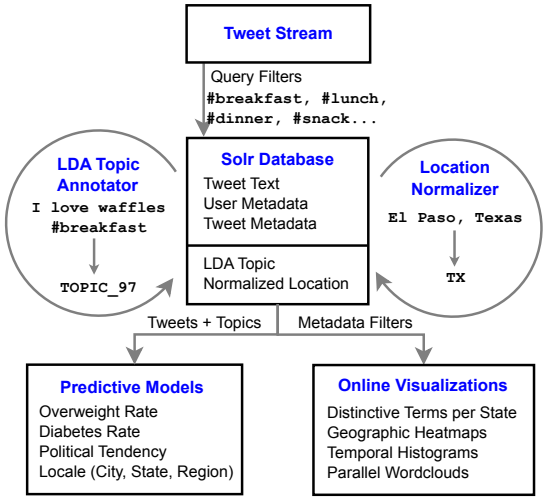

Fig. 1: The main steps of the system are: collecting tweets from Twitter using a set of meal-related filters, loading the tweets and their meta data into a Lucene-backed Solr instance, annotating the tweets with topic model labels (Section IV-B) and normalizing locations (Section II), and then querying the tweets for use in the predictive models (Section III) or visualization systems (Section VI).

[1] http://www.pewinternet.org/2014/01/08/social-media-update-2013/twitter-users/

[2] https://dev.twitter.com/docs/api/streaming. Note: Twitter caps the number of possible tweets returned by the streaming API to a fraction of the total number of tweets available at a given moment.

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

Term	# of Tweets	# with normalized US Location
#dinner	1,156,630	173,634
#breakfast	979,031	161,214
#lunch	931,633	129,853
#brunch	287,305	86,239
#snack	139,136	21,539
#meal	94,266	12,149
#supper	32,235	2,971
Total	3,498,749	562,547

TABLE I: Hashtags used to collect tweets, and number of tweets containing each hashtag. "Normalized US location" indicates that we could extract at least the user's state from the meta data. Since some tweets contain multiple meal hashtags, the total number of tweets (bottom row) is less than the column sum.

collected approximately 3.5 million tweets containing at least one of these hashtags from the period between October 2, 2013 and May 29, 2014.

Tweets are very short texts, limited to 140 characters. In our collection, the average length of a tweet is 8.7 words, after filtering out usernames, non-alphanumeric characters (hashtags excepted), and punctuation. After filtering, the tweet collection contains a total of about 30 million words and hashtags, with a total vocabulary size of 1.5 million words and hashtags.

Fig. 1 describes the system used to collect, annotate, and process the tweets for prediction and visualization. Along with the text of each tweet, we store the user's self-reported location, time zone, and geotagging information, whenever these fields are available. This meta data is used to group tweets by the home location of the author, e.g., specified as city and/or state for those users located within the United States (US). For most experiments in this paper, geolocation normalization is performed using regular expressions, matching state names or postal abbreviations of one of the 50 US states or Washington, D.C. (e.g., *Texas* or *TX*), followed by matching city names or known abbreviations (e.g., *New York City* or *NYC*) within the author's location field. In case of ambiguities (e.g., *LA* stands for both Los Angeles and Louisiana) we used the user's time zone to disambiguate. About 16% (562,547) of the collected tweets could be located within a state using this method (Table I). We chose to use the self-reported user location instead of the geotagging information because: (a) it is more common, (b) it tends to have a standard, easily parseable form for US addresses, and (c) to avoid potential biases introduced by travel. However, in Section VI, we extend our analysis to discover world-wide food-related patterns. In this context, because world addresses are considerably harder to parse than US addresses, we revert to geotagging information to identify the location of tweet authors.

Using this dataset, we can immediately see food-driven patterns. For example, Fig. 2 shows prominent food-related words that appeared in the tweets normalized to each state. Tweet text is filtered using a list of approximately 800 food-related words (see Sec. IV-A). Terms are ranked using term frequency–inverse document frequency (*tf-idf*) [1] to discount words that occur frequently across all states, and give priority to those words that are highly representative of a state. Each state's food word with the highest *tf-idf* ranking is displayed in the map. Regional trends can be seen, for example *grits*, a breakfast food made from ground corn, is a common dish in the southern states, and various types of seafood (*halibut*, *caviar*, *cod*, *clam*) are popular in the eastern and western coastal states.

Fig. 2: Prominent food word per state from the corpus of food-related tweets. Terms are filtered by a list containing approximately 800 food-related terms (Section IV-A) and ranked using *tf-idf*. Note that "Prune" is the name of a popular restaurant. Other word ranking criteria are offered on the website accompanying this paper: https://sites.google.com/site/twitter4food/.

III. Tasks

To understand the predictive power of the language of food, we implement several prediction tasks that use the tweets in the above dataset as their only input. We group these tasks into two categories: state-level characteristic prediction and locale prediction.

A. Predicting State-Level Characteristics

Here we predict three aggregate characteristics for US states, using features extracted from the tweets produced by individuals in each state:

1) Diabetes Rate: This is the percentage of adults in each state who have been told by a doctor that they have diabetes. Data in this set is taken from the Kaiser Commission on Medicaid and the Uninsured (KCMU)'s analysis of the Center for Disease Control's Behavioral Risk Factor Surveillance System (BRFSS) 2012 survey.[3] We convert this data into a binary dependent variable by considering whether a state's rate of diabetes is above or below the national median. The median diabetes rate is 9.7%, and the range is 6.0% (7.0% in Alaska to 13.0% in West Virginia). For example, Alabama has a diabetes rate of 12.3%, which is above the national median of 9.7%, so it is labelled as high-diabetes, while Alaska, with a rate of 7.0%, is labelled as low-diabetes.

2) Overweight Rate: This is the percentage of adults within each state who reported having a Body Mass Index (BMI) of at least 25.0 kilograms per meter squared, placing them within the "overweight" or "obese" categories defined by the National Institutes of Health.[4] As with the diabetes rate dataset, data is taken from KCMU's analysis of the BRFSS 2012 survey results[5]. Similarly, the corresponding binary dependent variable indicates if a state's overweight rate is above/below the national median. The median overweight rate is 64.2%, and the range is 17.7% (51.9% in Washington, D.C. to 69.6% in Louisiana).

3) Political Tendency: This dataset measures historical voting history over a 5-year period: whether a state is more

[3]The BRFSS is a random-digit-dialed telephone survey of adults age 18 and over. For more details see: http://kff.org/other/state-indicator/adults-with-diabetes/

[4]http://www.nhlbi.nih.gov/guidelines/obesity/BMI/bmi-m.htm

[5]http://kff.org/other/state-indicator/adult-overweightobesity-rate/

Democratic or Republican relative to the median US state, as measured by proportion of Democratic/Republican votes in general presidential, gubernatorial, and senatorial elections, in the interval from 2008 to 2013.[6]. For example, Alaska cast 554,565 total votes for Democratic candidates and 748,488 for Republican candidates in these three types of elections during the six-year period, for a fraction of 42.6% Democratic votes. This is below the median fraction of 51.6%, so Alaska is labelled as *Republican*. Votes are compared relative to the median because of a slight bias toward Democratic votes during this time period. The median fraction of Democratic votes is 51.6%, and the range is 65.4% (27.0% in Wyoming to 92.4% in Washington D.C.).

Because the above dependent variables are at state level, each state is treated as a single instance for these three tasks: all of the tweets produced within the state are aggregated into a single pool for feature extraction (detailed in the next section). We used Support Vector Machines (SVM) with a linear kernel [2] for classification.

Although such a prediction task has many features (from all tweets in a given state), it has a small number of data points (51, one for each state plus Washington, D.C.). For this reason, we use leave-one-out cross-validation to evaluate the accuracy of the model. For each of the three data sets (overweight, diabetes, and political), we use the following process: Each state is held out in turn. The SVM is trained on features of tweets taken from the remaining 50 states, using the labels of the current data set. The SVM is then used to predict the current dataset's label of the held-out state. The accuracy of the model on the label set is calculated as the number of correct predictions out of the total number of states. To avoid overfitting, we do not tune the classifier's hyper-parameters.

B. Predicting Locales

To examine the connection between the language of food and geographic location, we seek to predict the locale of a group of tweets, using only the text of the tweets as input. We investigated predicting the cities, states, and geographic regions of tweets, but only report results for city prediction because of space limitations. See the full paper [3] for the state and region prediction tasks. The locales in the city prediction task are the 15 most populous cities in the US.[7] The accuracy on the city prediction task is the number of cities correctly identified, divided by the total number of cities. To focus our analysis on the predictive power of the language of food, we remove as many state and city names as possible from the tweets to avoid learning trivial correlations (see Sec. IV-A).

IV. Feature Descriptions

We use two sets of features: *lexical* (from tweet words) and *topical* (sets of words appearing in similar contexts).

A. Lexical Features

We take the simple approach of representing each locale as a bag of words assembled from all the tweets in that group. Each word becomes a feature with value equal to the number of times it occurs across all tweets for that locale. We tokenize the tweets using the Stanford CoreNLP software.[8] An additional pre-processing step removes the following tokens: (a) tokens that do not contain alpha-numeric characters or punctuation (to reduce noise); (b) stopwords and words that occur a single time (to reduce data size); and, most importantly, (c) URLs, usernames (preceded by an @ symbol), and words and hashtags naming state and city locations[9] (to avoid learning trivial correlations, such as *#TX* indicating a tweet from Texas).

We also experiment with open versus closed vocabularies. For open vocabularies, we use two configurations: all words produced by the above pre-processing step, or only hashtags. For a closed vocabulary experiment, we use a set of 809 words related to food, meals, and eating, obtained from the English portion of a Spanish-English food glossary[10] and an online food vocabulary list[11]. These experiments help us understand how much predictive power is contained in food words alone versus the full text (or hashtags) of the tweets, which capture a much broader context.

B. Topic Model Features

Topic models provide a method to infer the themes present in tweets. Topics are clusters of words that tend to appear in similar contexts. For example, a topic learned by the model, which we refer to as the *American diet* topic, contains *chicken*, *baked*, *beans*, and *fried*, among other terms. We use topics as features for several reasons: (1) topics provide a method to address the sparsity resulting from having very short documents (tweets are limited to 140 characters) by treating groups of related words as a single feature; (2) topical features aid in post-hoc analysis by allowing us to detect correlations that go beyond individual words.

We use Latent Dirichlet Allocation (LDA)[4] to learn a set of topics from the food tweets in an unsupervised fashion. LDA treats each tweet as a mixture of topics, which are themselves probability distributions over clusters of words. The LDA model (topic distributions and mixtures) is trained from all available tweets in the corpus, using the MALLET software package.[12] We chose 200 as the number of topics for the model to learn. This number produced topics that seemed fine-grained enough to capture specific patterns in diet, language, or lifestyle – clusters of foods of various nationalities, or specific diets such as vegetarian. Once the LDA topic model is trained, we use it to infer the mixture of topics for each tweet in the prediction tasks. The topic most strongly associated with the tweet (the topic with highest probability given the model and the tweet) is used as an additional feature for the tweet, similarly to the lexical features generated from the words of the tweet. Topics are counted across all tweets in a state in the same manner as the lexical features.

When applied in combination with the configuration containing solely food word or hashtag vocabularies, the LDA topics are constructed using the corresponding filtered versions of the tweets, i.e., with all non-food words or non-hashtag words removed.

For clarity in our analysis, we have manually assigned subject labels, such as *American diet*, to some of these topics based on the words contained in the topic.[13] We use these

[6] http://uselectionatlas.org/

[7] http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

[8] http://nlp.stanford.edu/software/corenlp.shtml

[9] In addition of known state names and abbreviations we used a list of the 250 most populous cities in the US from http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population, together with common nicknames, such as "NYC" for New York City, "#sanfran" for San Francisco and "atl" for Atlanta. In total, we remove 892 distinct location words and hashtags.

[10] http://www.lingolex.com/spanishfood/a-b.htm

[11] http://www.enchantedlearning.com/wordlist/food.shtml

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

[13] But these topic labels are not visible to the classifier.

	overweight	diabetes	political	average
majority baseline	50.98	50.98	50.98	50.98
All Words	76.47‡	64.71	66.67‡	69.28‡
All Words + LDA	**80.39**‡	64.71	68.63‡	**71.24**‡
Hashtags	72.55‡	**68.63**†	60.78	67.32‡
Hashtags + LDA	74.51‡	**68.63**†	62.75	68.63‡
Food	70.59‡	60.78	68.63‡	66.67‡
Food + LDA	68.63†	60.78	**72.55**‡	67.32‡
Food + Hashtags	64.71†	62.75	64.71†	64.05‡
Food + Hashtags + LDA	74.51$^{\ddagger++}$	62.75	64.71†	67.32$^{\ddagger+}$

TABLE II: Using features of tweets to predict state-level characteristics: whether a given state is above or below the national median for overweight rate, above or below the median diagnosed diabetes rate, and the state's historical political voting trend (D or R). This table compares the effect of filtering the lexical features to: food words, hashtags, both, or keeping the entire text of the tweets; as well as the effect of adding LDA topics. Throughout the paper, we mark results as follows: ‡denotes a significant ($p <= 0.05$) and †a nearly-significant ($0.05 < p <= 0.10$) improvements over the majority baseline. Similarly, $^{++}$denotes that the LDA model has a statistically significant ($p <= 0.05$) and $^{+}$a nearly statistically significant ($0.05 < p <= 0.10$) improvement over the model without LDA. Statistical significance testing is implemented using one-tailed, non-parametric bootstrap resampling with 10,000 iterations.

assigned labels to refer to the topics in the remainder of this paper.

To account for the large differences in the number of tweets available for each state (for example, the state with the most normalized tweets, New York has 83,670 tweets, while the state with the fewest, Wyoming, has 339), we scale all the features collected for each state. Each feature's value within a state's feature set is divided by the number of tweets collected for the state.

V. Results

We present empirical results for both categories of tasks introduced in the previous section: predicting state-level characteristics and predicting locales. We also analyze the effectiveness of the language of food for these prediction tasks by examining the most important textual features in the classification models, and investigating the importance of open versus closed vocabularies.

A. State-Level Characteristics

Table II shows classification results on the state-level statistics prediction task (Section III-A) for varying feature sets. Since all three datasets are nearly evenly split between the binary classes (each dataset has either 25 or 26 states out of 51 in each of the two classes), a baseline that predicts the majority label achieves approximately 51% accuracy. We compare the performance of the tweet-based predictive models to this majority baseline, and evaluate how filtering the lexical content of the tweets and adding topical features affects accuracy on these prediction tasks. We draw several observations from this experiment:

(a) First and foremost, the language of food can indeed infer all the latent characteristics investigated: all configurations investigated statistically outperform the majority-class baseline. The best performance is obtained when the entire text of the tweets is used (All Words), which captures not only direct references to food, but also the context in which it is discussed. However, the performance of the closed vocabulary of food words (Food) is within 5% of the best performance, demonstrating that most of the predictive signal is captured by direct references to food.

(b) The classifiers achieve the highest accuracy on the overweight dataset. This is an intuitive result, which confirms that there is a strong correlation between food and likelihood of obesity. However, the fact that this correlation can be detected solely from social media posts is, to our knowledge, novel and suggests potential avenues for better and personalized public health. A similar correlation with political preferences is also interesting, indicating potential marketing applications in the political domain.

(c) More complex NLP (topic modeling in our case) is beneficial: the performance of the models that include LDA topics is, on average, better than that of the configurations without topics.[14] We plan to use more informative representations of text, e.g., based on deep learning [5], in future work.

In the full paper [3], we show the words and topical features assigned the greatest importance, i.e., largest magnitude weights, by the SVM training process, for each dataset and class. It is interesting to note that a dietary topic we have labeled as *American Diet*, containing terms such as *chicken*, *baked*, *beans* and *fried*, is an important feature for predicting both that a state has higher rates of overweight and diabetes than normal, whereas other diets, such as *#vegan* and *Paleo Diet* are important predictors for the opposite. Pronouns have high weights in the overweight prediction task: the first-person singular *I* and *my* are valuable for predicting that a state is overweight, while collective words such as the *You, We* topic cluster are valuable for predicting that a state is below the median. This is less surprising in view of prior work, such as Ranganath et al. [6], showing that the types of pronouns used by an individual are associated with a host of traits such as gender and intention. For the political affiliation task, we observe that features correlated with Republican states include those centered around work (the *Airport* topic) and home (the *After Work* topic, including words such as *home*, *after*, *work*). The most predictive feature for Democratic states is *#vegan*, and we also see topics associated with urban life and eating out, such as *Deli*, *#brunch*, promotions such as *Restaurant Advertising*, and *Eating Out*.

model	accuracy (%)
Random Baseline	6.67
All Words	66.67‡
All Words + LDA	80.00$^{\ddagger+}$
Food	40.00‡
Food + LDA	40.00‡
Hashtags	53.33‡
Hashtags + LDA	66.67‡
Food + Hashtags	53.33‡
Food + Hashtags + LDA	**86.67**$^{\ddagger++}$

TABLE III: City prediction accuracy (15 most populous US cities) for the various feature sets. Statistical significance testing is performed similarly to Table II.

B. City Prediction

Here we present results for the city prediction task. Results and feature analysis for the state and region locale prediction tasks are available in the full paper [3]. Table III shows the accuracies of the various feature sets for this task. The input

[14]The improvement is not statistically significant for most experiments, but this can be attributed to the small size of the dataset (51 data points).

for this task is 15 cities, so the random-prediction baseline accuracy is 6.67%. As in the previous task, every set of features improves significantly upon this baseline, ranging from 40% accuracy using only Food words to 86.67% accuracy using Food words, Hashtags, and LDA topics, demonstrating once again the predictive power of the language of food. The significant improvement of the closed food vocabulary alone (Food) over the baseline indicates that the diets in each of these 15 cities are distinct enough to have some predictive power. However, diets alone are not enough to completely identify the cities, and we see that for this task more context is beneficial: adding hashtags helps considerably (53.33% accuracy), and adding topical features to the food and hashtag filtered set of lexical features improves performance even further (86.67%).

In the full paper [3], we list the top five features for each city in this task. The table shows that variations in diet are clear: *tacos* are significant in Austin, *#vegetarian* food is indicative of San Francisco, *#brunch* is representative of New York, etc. Using the context around food is clearly important. We see that several cities in California are associated with *#foodie* (Los Angeles and San Francisco) or eating while on *Vacation* (San Diego and San Francisco). First-person pronouns are highly weighted in cities in Texas (*we* in Austin, *I* in Houston, and *my* and *I* in San Antonio).

VI. Visualization Tools

While the machine learning models described above are well-suited for prediction on predefined tasks, we also constructed several visualization tools to discover previously unknown trends in the Twitter dataset. These tools aim to allow aggregate analysis of tweet content in the context of geographic and temporal location. Although we normalize word and feature counts by the number of available tweets in our machine learning tasks (Sec. IV-B), we currently do not perform normalization in the various visualizations described below: all visualizations reflect the raw amount of tweets matching a given query.

A. Top Terms by State

The first of these tools, the term visualizer (Fig. 2), does a simple keyword analysis of the tweets available for each US state. We extract all terms that are contained within a list of around 800 food-related words (see Sec. IV-A) and rank them using *tf-idf*, treating all tweets normalized to a given state as a single document: each term's score is the number of times it occurred within a state, multiplied by the logarithm of the inverse proportion of the number of states it occurred in [1]. Ranking by *tf-idf* emphasizes words that are common in a particular state, but ensures that words used frequently in all states, such as *food* and *eat*, are not highly ranked. The term(s) with the highest ranking in each state are displayed on the state in the map. As discussed previously, this tool immediately highlights dietary patterns: *grits* in the Southern states, seafood on both coasts, etc.

B. Temporal Histograms

Temporal histograms allow us to visualize the changing popularity of terms over the course of a day, week, or year. About 71% of the collected tweets (2,503,351) are from users who have listed their time zone. For these tweets, we compute the time local to the user when the tweet was posted. The temporal visualization tool [3] allows querying these time-localized tweets by phrase and constructing histograms at varying time granularities: hour of day, day of the week, or month of the year.

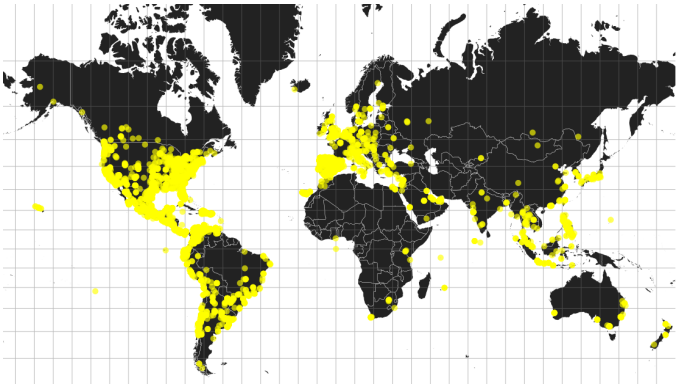

Fig. 3: Tweet geolocation plot showing migration patterns reflected in diet: yellow dots mark the locations of 11,827 tweets matching five *Spanish/Latin American food* topics (*tacos*, *burrito*, *salsa*, *pollo*, *arroz*, *paella*, etc.).

C. Tweet Location Maps

About 10% of the collected tweets (362,978) have associated geolocation information – the user's longitude and latitude at the moment the tweet was posted. We use this meta-information to build a system for querying and plotting worldwide geographic maps of tweets. The interface allows searching by phrase or LDA topic and displays geographic plots or heatmaps showing the locations of all tweets matching the query.

This system allows the discovery of broad geographic trends in the data, which are perhaps reflective of immigration patterns to the US or worldwide. For example, Fig. 3 shows the prevalence of Spanish and Latin-American influenced food throughout the Spanish-speaking world. We found similar trends with other topics: an *Italian food* topic has high intensity in Italy and New York City and a *Vietnamese food* topic has high intensity in Vietnam and in Southern California [3].

VII. Related Work

Previous work has used textual analysis of Twitter posts to study diverse and global populations, including investigating temporal changes in mood [7] and correlations between religious expression and sentiment [8]. Several other works predict latent characteristics of individuals and communities using social media posts and metadata, including predicting Twitter users' age, region, and political orientation [9] and gender [10], [11]. Jurafsky et al. [12] analyze a corpus of restaurant reviews and predict restaurant ratings using linguistic features such as sentiment, narrative, and self-portrayal.

Paul and Dredze [13] apply the Ailment Topic Aspect Model to 1.5 million health-related tweets and discover mentions of over a dozen ailments, including allergies and insomnia. Schwartz et al. [14] use Twitter to predict public health and well-being statistics on a state-wide level. As in our study, LDA topics improved accuracy. Hingle et al. [15] use Twitter together with analytical software to capture real-time food consumption and diet-related behavior. While this study identifies relationships between dietary and behavioral patterns the results were based on a small dataset (50 participants and 773 tweets). Nascimento et al. [16] evaluate self-reported migraine headache suffering using over 20,000 migrane-related tweets over a seven-day period, finding different peaking hours on weekdays and weekends. Yom-Tov et al. [17] show how Twitter can be used to discover possible outbreaks of communicable diseases at large public gatherings. Myslín et

al. [18] use machine classification of tobacco-related Twitter posts to detect tobacco-relevant posts and sentiment towards tobacco products.

Previous work has modelled linguistic variation on Twitter in terms of demographic and geographic variables. O'Connor et al. [19] create a generative model of word use from demographic traits, and show clusters of Twitter users with common lexicons. Eisenstein et al. [20] show that despite the global diffusion of social media, geographic regions have distinct word and topic use on Twitter. Additional work on aggregating, processing, and visualizing tweets includes systems for detecting newsworthy events and clustering tweets in real-time [21], and producing geographic visualizations of tweet sentiment [22].

Our work builds upon these previous results, and is, to the best of our knowledge, the first to provide a large-scale, empirical analysis of the predictive power of the language of food.

VIII. Conclusion and Future Work

This work empirically validates that food and food discussion are a major part of who we are. We develop a system for collecting a large corpus of food-related tweets and use these tweets to predict many latent population characteristics: overweight and diabetes rates, political leaning, and geographic location of authors. Furthermore, we integrate several visualization tools that summarize and query this data, allowing us to discover more complex geographical/temporal trends that are driven by the language of food, such as potential migration patterns. Although there are inherent biases in Twitter data caused by demographic imbalances [23] and the self-reported nature of the data [24], our analysis indicates that the language of food alone is nonetheless very powerful. For example, on most predictive tasks, a closed vocabulary of only 800 food words approaches the peak performance obtained when using the entire 1.5 million word vocabulary. Perhaps most importantly, our analysis of the learned predictive models provides big-data-driven insights into connections between the language of food and the investigated population characteristics.

We note that our choice of populations (e.g., cities, regions) for these tasks is purely practical (driven by the size of Twitter data at this granularity, and availability of dependent variables for the predictive tasks) and not a limitation of the proposed approach. In the future we would like to use our system to predict characteristics of individuals (e.g., propensity for diabetes), using the individuals' food information. Given sufficient amounts of available data, this can lead to non-trivial public health applications and, in a commercial and/or political space, to improved targeted marketing.

This paper is accompanied by a website, https://sites.google.com/site/twitter4food/, which includes a live version of all visualization tools presented, as well as a full version of this paper [3], containing additional tasks such as state and region prediction, and additional visualizations such as parallel word clouds.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

[2] V. N. Vapnik, *Statistical Learning Theory*. Wiley and Sons, NY, 1998.

[3] D. Fried, M. Surdeanu, S. Kobourov, M. Hingle, and D. Bell, "Analyzing the language of food on social media," *arXiv preprint 1409.2195v2*, 2014.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *EMNLP*. Association for Computational Linguistics, 2013.

[6] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you, it's me: Detecting flirting and its misperception in speed-dates," in *EMNLP*, 2009, pp. 334–342.

[7] S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[8] R. S. Ritter, J. L. Preston, and I. Hernandez, "Happy tweets: Christians are happier, more socially connected, and less analytical than atheists on Twitter," *Social Psychological and Personality Science*, pp. 243–249, 2013.

[9] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *2nd Intl. Workshop on Search and mining user-generated contents*. ACM, 2010, pp. 37–44.

[10] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *EMNLP*. Association for Computational Linguistics, 2011, pp. 1301–1309.

[11] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," *Journal of Sociolinguistics*, vol. 18, no. 2, pp. 135–160, 2014.

[12] D. Jurafsky, V. Chahuneau, B. Routledge, and N. Smith, "Narrative framing of consumer sentiment in online restaurant reviews," *First Monday*, vol. 19, no. 4, 2014. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/4944

[13] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health." in *ICWSM*, 2011.

[14] H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, M. Agrawal, G. Park, S. Lakshmikanth, S. Jha, M. Seligman, L. Ungar *et al.*, "Characterizing geographic variation in well-being using tweets," in *7th Intl. AAAI ICWSM*, 2013.

[15] M. Hingle, D. Yoon, J. Fowler, S. G. Kobourov, M. Schneider, D. Falk, and R. Burd, "Collection and visualization of dietary behavior and reasons for eating using a popular and free social media software application," *Journal of Medical Internet Research (JMIR)*, vol. 15, no. 6, pp. 125–145, 2013.

[16] D. T. Nascimento, F. M. DosSantos, T. Danciu, M. DeBoer, H. van Holsbeeck, R. S. Lucas, C. Aiello, L. Khatib, A. M. Bender, , J.-K. Zubieta, and F. A. DaSilva, "Real-time sharing and expression of migraine headache suffering on Twitter: A cross-sectional infodemiology study," *J Med Internet Res*, vol. 16, no. 4, p. e96, Apr 2014. [Online]. Available: http://www.jmir.org/2014/4/e96/

[17] E. Yom-Tov, D. Borsa, J. I. Cox, and A. R. McKendry, "Detecting disease outbreaks in mass gatherings using internet data," *J Med Internet Res*, vol. 16, no. 6, p. e154, Jun 2014. [Online]. Available: http://www.jmir.org/2014/6/e154/

[18] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway, "Using twitter to examine smoking behavior and perceptions of emerging tobacco products," *J Med Internet Res*, vol. 15, no. 8, p. e174, Aug 2013. [Online]. Available: http://www.jmir.org/2013/8/e174/

[19] B. OConnor, J. Eisenstein, E. P. Xing, and N. A. Smith, "A mixture model of demographic lexical variation," in *Proc. of NIPS workshop on machine learning in computational social science*, 2010, pp. 1–7.

[20] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "Mapping the geographical diffusion of new words," *arXiv preprint 1210.5268*, 2012.

[21] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic, "Scalable distributed event detection for twitter," in *Int. Conf. on Big Data, 2013*. IEEE, 2013, pp. 543–549.

[22] V. D. Nguyen, B. Varghese, and A. Barker, "The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter," in *Int. Conf. on Big Data, 2013*. IEEE, 2013, pp. 46–54.

[23] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users." *ICWSM*, vol. 11, p. 5th, 2011.

[24] E. Kıcıman, "OMG, I have to tweet that! A study of factors that influence tweet rates," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.