Estimating County Health Statistics with Twitter

Aron Culotta
Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616
culotta@cs.iit.edu

ABSTRACT

Understanding the relationships among environment, behavior, and health is a core concern of public health researchers. While a number of recent studies have investigated the use of social media to track infectious diseases such as influenza, little work has been done to determine if other health concerns can be inferred. In this paper, we present a large-scale study of 27 health-related statistics, including obesity, health insurance coverage, access to healthy foods, and teen birth rates. We perform a linguistic analysis of the Twitter activity in the top 100 most populous counties in the U.S., and find a significant correlation with 6 of the 27 health statistics. When compared to traditional models based on demographic variables alone, we find that augmenting models with Twitter-derived information improves predictive accuracy for 20 of 27 statistics, suggesting that this new methodology can complement existing approaches.

Author Keywords
social media; public health; natural language processing

ACM Classification Keywords
H.3.4 H.5.2: H.5.3

INTRODUCTION

Chronic diseases are the leading cause of death and disability in the U.S. and account for 75% of health care costs.[1] Understanding the interaction among environment, behaviors, and health outcomes is critical to developing informed intervention strategies. In response, the U.S. Centers for Disease Control and Prevention leads multiple community health data collection and intervention efforts such as the Behavioral Risk Factor Surveillance System, the National Health Interview Survey, and the Health Communities Program. A major goal of these initiatives is to identify vulnerable populations in order to better target intervention strategies. While these programs provide tremendous insight, they require considerable time and effort and are often limited in sample size, frequency, or geographic granularity.

In this paper, we investigate the use of social media as a complementary data source to identify at-risk communities. The popularity of websites like Twitter and Facebook continues to grow, making unprecedented amounts of information about attitudes and behaviors publicly available. Given the research in economics [2], socio-linguistics [18], and psychiatry [13] indicating the relationship between language and health, we examine whether linguistic patterns in Twitter correlate with health-related statistics.

For each of the 100 most populous counties in the U.S., we collect 27 health-related statistics from the County Health Rankings & Roadmaps project, including health outcomes, behaviors, socio-economic status, and environmental factors. We also collect over 1.4M user profiles and 4.3M posts from Twitter over a nine month span from the same 100 counties. We then perform a statistical analysis to identify how accurately these health outcomes can be predicted from the Twitter data and which linguistic markers are most predictive of each statistic.

Our experiments[2] investigate four research questions, the answers to which we summarize below:

RQ1. **Predictive accuracy.** Is Twitter activity predictive of county-level health statistics? We find a significant correlation on held-out data for 6 of 27 statistics, including obesity, diabetes, teen births, health insurance coverage, and access to healthy foods.

RQ2. **Representation.** How does the linguistic representation affect accuracy? We find that the LIWC lexicon [24] is more predictive than alternatives, and that normalizing linguistic vectors by the number of users in a county can greatly improve accuracy.

RQ3. **Beyond demographics.** Does Twitter activity provide more information than common demographic covariates? We find that models that augment demographic variables (race, age, gender, income) with linguistic variables (from Twitter) are more accurate than models using demographic variables alone for 20 of the 27 health statistics we consider. For two (limited access to health foods, prevalence of fast foods), the Twitter model in isolation is actually more accurate than the demographic variable model. These results suggest that the two sources of information are complementary.

RQ4. **Identifying linguistic indicators.** What are the linguistic indicators that are most predictive of each outcome? After controlling for five demographic vari-

[1] `http://www.cdc.gov/chronicdisease/overview`

[2] Code is available: `https://github.com/tapilab/twcounty`.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2473-1/14/04..$15.00.
http://dx.doi.org/10.1145/2556288.2557139

ables, we identify 33 linguistic categories that are significantly predictive of 6 different health-related statistics. For example, references to religion and certain pronouns ("we", "her") correlate with better socio-emotional support; references to money and inhibition correlate with lower unemployment; and references to family and love correlate with higher rates of teen births.

While this new methodology requires further experimentation, we believe it can aid public health researchers by providing (1) a more nuanced alternative to demographic profiles for identifying at-risk populations; (2) a low-cost method to measure risk across different subpopulations; (3) a process to help formulate new hypotheses about the relationship between environment, behaviors, and health outcomes, which can then be tested in a more controlled setting.

The remainder of the paper is organized as follows: we first review related work, then we describe the data and its collection. We next present our method for representing the linguistic activity of each county and the experimental framework for measuring accuracy and identifying significant linguistic variables. Finally, we present the results and discuss their implications.

RELATED WORK

We first briefly review related work in the study of language, health, and social media.

Language and Heath

Language has long been investigated as an indicator of health. For example, Gottschalk [13] performed a content analysis of patients to determine psychological state, such as anxiety, hostility, and alienation. Pennebaker [18] provides an excellent review of research connecting linguistic patterns to demographics, personality, psychology, mental health.

While many studies support the connection between mental health and language, the connection between physical health and language is less well-established. Some studies have reported correlations between "Type A" language and heart diseases[14] and positive emotional language with longevity [7]. Given growing evidence supporting the link between emotional well-being and health [17], estimating psychological health may serve as a predictive surrogate for physical health.

The emerging study of the economics of language has also investigated how language relates to decision-making, which in turn can affect health. For example, in a study of 76 countries, Chen [2] found that certain grammatical properties correlate with higher rates of savings and lower rates of smoking and obesity, concluding that some linguistic constructs may foster future-oriented behavior. Chiswick [3] investigates how language proficiency of immigrants can impact employment and other socio-economic factors.

Social Media and Health

There is a growing body of work investigating social media to track health concerns such as influenza [20, 5, 22, 30, 27], E. coli [32], alcohol consumption [6], Adderall use [15], insomnia [19] and depression [8]. See Dredze [9] for an overview. Most of these focus on detecting explicit mentions of a symptom of interest (e.g., "Staying home from work today with a sore throat"). In contrast, the present work investigates more nuanced linguistic cues that correlate with the overall health of a population.

Ghosh & Guha [12] identified geo-spatial patterns in specific obesity-related tweets (e.g. "fast food"), using topic models to qualitatively characterize discussions of obesity on Twitter. While some ancillary data is used for comparison (e.g., location of fast food restaurants), no correlation analysis is performed with obesity statistics. Additionally, Paul & Dredze [22] use a topic model to discover obesity-related tweets, finding a .28 correlation with state obesity statistics.

Our methodology is most similar to that of Schwartz et al. [28], who find tweets to be predictive of county-level surveys of life satisfaction. Here, we also use LIWC and PERMA lexicons as features in a regression model of county statistics.

In the context of this related work, the primary contributions of this paper are as follows: (1) we present the first large-scale social media analysis across a diverse set of 27 measures of community health; (2) we provide an empirical comparison of several important methodological decisions, such as linguistic lexicons, vector normalization, and source of linguistic content; (3) we provide a rigorous statistical treatment that identifies linguistic indicators from social media that are significant predictors of health outcomes even after controlling for demographic variables.

DATA

Here we describe how we collected the health and Twitter data and provide descriptive statistics of their contents.

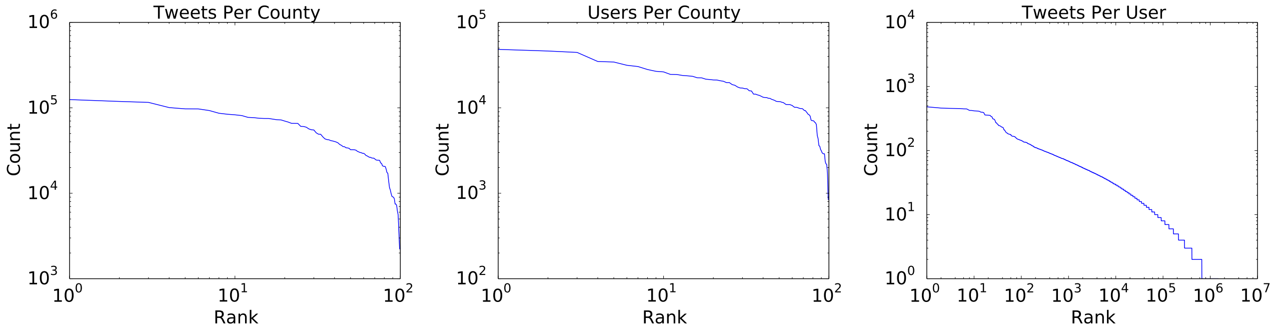
County Health Data

Using data from the U.S. Census' State-Based Counties Gazetteer,[3] we collected the top 100 most populous counties in the U.S. along with their geographical coordinates. Each county is assigned a Federal Information Processing Standards (FIPS) code as a unique identifier. The County Health Rankings & Roadmaps,[4] a partnership between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute, aggregates county-level health factors from a wide range of sources, including the Behavioral Risk Factor Surveillance System, American Community Survey, and the National Center for Health Statistics, collected over the past three years.[5] These publicly available data contain county statistics on 30 measures of mortality, morbidity, health behaviors, clinical care, socio-economic factors, and physical environment.

[3] `http://www.census.gov/geo/maps-data/data/docs/gazetteer/Gaz_counties_national.zip`

[4] `http://www.countyhealthrankings.org/`

[5] While the Twitter was collected more recently, most county-level statistics, and particularly their relative differences, are slow to change.



(a) (b) (c)

Figure 1: Distributions over the 4.31M tweets, 1.46M users, and 100 counties in the dataset.

For each of the top 100 most populous counties, we collected 27 health statistics (3 were removed because of missing values for some counties). These are listed in Table 3. As space precludes a precise definition of how each statistic was computed, we refer the reader to the County Health Rankings website for more information. We describe some of these in more detail in the Discussion section.

Twitter Data

We next constructed a set of 100 Twitter queries consisting of one geographical bounding box for each county, approximated by a 50 square mile area centered at the county coordinates obtained from the U.S. Census.[6] We then submitted these queries continuously to Twitter's search API from December 5, 2012 to August 31, 2013 (with intermittent stoppages for technical difficulties). These queries return tweets that have been geolocated, typically tweets issued from a mobile device. This resulted in 4.31M tweets from 1.46M unique users. For each tweet, we retain the tweet content as well as the user description field, a short, user-provided summary (e.g., "motivated law student"). Figure 1 shows distributions of tweets per county, users per county, and tweets per user. While the demographic distributions of Twitter users are thought to skew young and urban [10], it is worth noting that these 1.46M users represent over 1% of the total population of these 100 counties (130M). As expected, Twitter usage varies significantly by county size. On average, we collect 14.5K users per county, with 66 counties containing at least 10K users. Hudson County (part of the New York metropolitan area) has the most with 52K users, Honolulu County the least with 845. The tweets per user graph exhibits a typical long tail — a few users tweet very often, but most tweet infrequently.

We note that this data collection methodology differs from that of Schwartz et al. [28], who collect the 10% "garden hose" sample of the entire Twitter stream, then use heuristics to filter by location using the user's profile information. This can yield more tweets (since only a small percentage of tweets are geocoded), but can introduce additional geolocation noise due to the unreliability of the location field [16].

LINGUISTIC REPRESENTATION

Given a collection of tweets categorized by county, we next must distill them into a set of variables to correlate with the health statistics. Due to the small number of validation points (100 counties) and the large number of potential variables (hundreds of thousands of unique words), rather than considering words as variables, we instead consider word categories. We build on prior work that considers two lexicons:

- **LIWC:** The 2001 Linguistic Inquiry and Word Count lexicon [24] contains 74 categories and 2,300 word patterns (which includes exact matches as well as prefixes like *awake**). Each word pattern may belong to multiple categories (e.g., *Physical*, *Sleep*). This lexicon was developed over a number of years to identify categories that capture emotional and cognitive cues of interest to health, sociology, and psychology. It has been used in numerous studies [18], including Twitter studies [25, 28, 8].
- **PERMA:** The PERMA lexicon [29] contains 10 categories and 1,522 words. The categories reflect the five dimensions of positive psychology (Positive emotion, Engagement, Relationships, Meaning, Achievement) — each category is either positive or negative. For example, *R+* indicates positive relationships and *P-* indicates negative emotions. Only exact matches are considered, and each word belongs to exactly one category.

We select these lexicons based on their use in prior work [28] and the fact that they were designed to represent categories of relevance to health and personality.

For each county, then, we record the frequency with which each lexical category is used. To do this, we use a simple tokenizer to process each tweet that removes punctuation and then splits by whitespace to return a list of tokens. Additionally, we remove all mentions and URLs. The remaining tokens are matched against the above lexicons, resulting in a vector of category frequencies for each county.

We distinguish between tokens appearing in the tweet text and tokens appearing in the user description, denoted by the prefixes (*d*=) and (*t*=). For example, [*d=Sleep: 2*, *t=R+: 1*]

[6] This introduces a small amount of noise – 957 tweets came from overlapping bounding boxes. This can be eliminated by using the county polygon data from the Census. We thank the anonymous reviewer for this suggestion.

indicates that two tokens in the description field map to the *Sleep* category and that one token in the tweet text maps to the positive relationship category.

We found that only 70 of the LIWC categories appear in our data, along with all 10 of the PERMA categories, yielding a total of 80 linguistic categories.

For each county, we create a vector of 160 values reflecting the frequency of each category (80 categories each for description and text tokens). Since the magnitude of these values will vary greatly based on the number of tweets collected from each county, we consider several normalization strategies to make the vectors comparable across counties:

- **None:** No normalization used; each vector contains the raw frequency of each category.
- **Log:** We take the natural log of one plus the value (as advocated by Schwartz et al. [28]). This dampens large values.
- **Word:** We divide each value by the sum of all values in that county's frequency vector. This represents the relative prevalence of a category as compared to overall usage in that county.
- **User:** We store the proportion of users from the county who use a word from this category. Note that if one user tweets the same word category many times, this will only increase the numerator by one; the denominator is the total number of users from that county.

EXPERIMENTS

To address our four research questions from the Introduction, we perform regression to predict each of the 27 health-related statistics using the 180 linguistic variables described above. Given the large number of independent variables (180) relative to the number of validation points (100 counties), we use ridge regression to reduce overfitting.[7]

To estimate generalization accuracy, we use five fold cross-validation — each fold fits the model on 80 counties and predicts on the remaining 20. The splits are created uniformly at random, except that we additionally ensure that counties from the same state do not appear in both the training and test split in one fold. This is to confirm that the model is learning more than simply the state identity of each county.[8]

We report two measures of accuracy:

- **Pearson's r:** We collect all the predicted values from the held-out data in each fold (100 counties total) and compute the correlation with the true values; $r \in [-1, 1]$; larger is better.
- **SMAPE:** Symmetric Mean Absolute Percentage Error [11] measures the relative error between the predicted and true value. This is a useful alternative to the more common mean-squared error as it can compare outcome variables that have different ranges. If y_i is the true value and $\hat{y}_i$ is the predicted value, then SMAPE $= \frac{\sum_i |y_i - \hat{y}_i|}{\sum_i y_i + \hat{y}_i}$. For non-negative y values, SMAPE $\in [0, 1]$; smaller is better.

In addition to LIWC and PERMA, some experiments also include five demographic control variables:

- $<$ **18**: the proportion of people under the age of 18.
- **65 and over**: the proportion of people at least 65 years old.
- **Female:** the proportion of people who are female.
- **Afro-Hispanic:** the proportion of people who are African-American or Hispanic.
- **Med_income:** the log of the median household income.

We select these variables because of they are used in prior Twitter work [28], they are prevalent in governmental data collection for health studies (e.g., the Behavioral Risk Factor Surveillance System), and they have been linked to health outcomes in epidemiological studies [31, 21, 26]. We collect these variables from the County Health Rankings Roadmap data.

RESULTS

Below, we first briefly summarize the main results, then discuss them in more detail.

RQ1: Predictive accuracy

Our first research question asks whether Twitter-derived linguistic variables are predictive of a county's health statistics. Columns labeled **T** (Twitter) in Table 3 display the results for our two evaluation metrics across 27 statistics for the model containing all 160 linguistic variables (**LIWC+PERMA**) with **User** normalization (we will revisit these choices in the next section). To compute statistical significance of each correlation value, we use a Bonferroni correction to adjust for multiple comparisons. Additionally, we replace the traditional p-value calculation with the Clifford & Richardson correction [4], which computes an effective sample size based on spatial autocorrelation[9], as measured by Moran's I (using the R SpatialPack[10] library.)

We find that for nine statistics the prediction of the linguistic model is significantly correlated with the health statistic. The strongest correlations are for No Insurance (percent of population under the age of 65 without health insurance) ($r = .59$), Vehicle Mortality (motor vehicle crash deaths per 100k) ($r = .52$), Limited Healthy Food (percent of population who live in poverty and are 1 mile from a supermarket in urban areas or 10 miles in rural areas) ($r = .51$), teen birth rate (per 1k females age 15-19) ($r = .50$), Dentist Access (ratio of population to dentists) ($r = .44$), and Obesity (percent of adults that report a BMI ≥ 30) ($r = .43$).

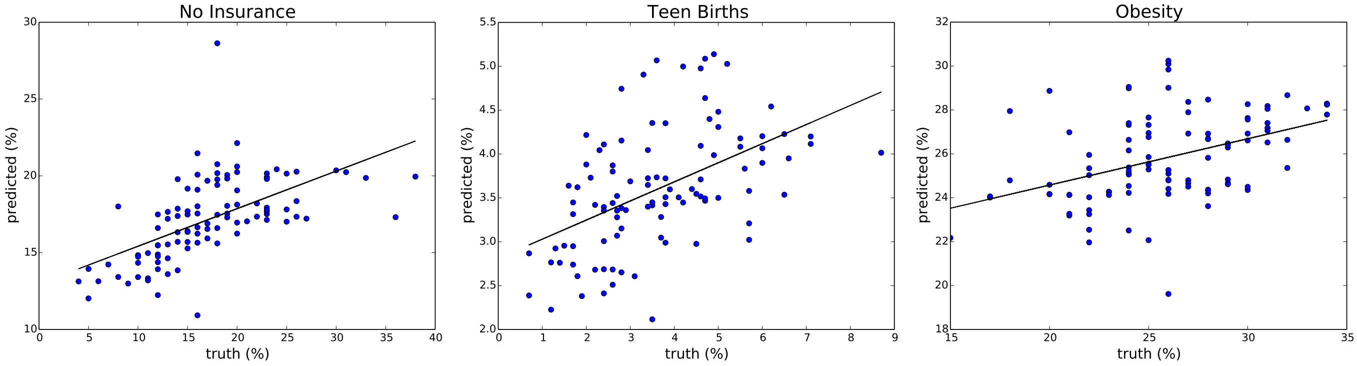
Figure 2 shows scatter plots of the true and predicted values on held-out data for three of the significantly correlated predictions using the **LIWC+PERMA** model. The largest errors generally appear at extreme values. For example, Hidalgo County in Texas has the highest Teen Birth rate of the 100 counties (8.7%, about 1.5% higher than the next highest

[7] We use the implementation in `scikit-learn` [23] with smoothing parameter $\alpha = 0.1$.

[8] Indeed, we find that splitting at random instead of by state increases the overall average correlation for the LIWC model from .25 to .29.

[9] We thank the anonymous reviewer for this suggestion.

[10] **`http://spatialpack.mat.utfsm.cl`**



(a) (b) (c)

Figure 2: Scatter plots of true versus predicted values on held-out data using the user-normalized LIWC+PERMA model (no demographic variables used) for No Insurance ($r = .59$), Teen Births ($r = .50$), and Obesity rates ($r = .43$).

Variables	r	SMAPE
PERMA	-0.07	10.45%
LIWC	0.25	9.67%
LIWC+PERMA	0.25	9.67%
Controls	0.59	7.65%
Controls+PERMA	0.60	7.57%
Controls+LIWC	0.63	7.37%
Controls+LIWC+PERMA	0.63	7.37%

Table 1: Held-out correlation and SMAPE averaged across all 27 output variables using various combinations of input variables. All models use **User** normalization.

county), though the linguistic model predicts only 4%. Similarly, Miami-Dade County in Florida has a high uninsured rate of 36%, while the model predicts only 17%. The outlier in the No Insurance plot (predicted=28.6%, truth=18%) is Pima County, AZ — this may in part be explained by the limited Twitter data from that county (4k users).

RQ2: Representation

Our second research question asks how the choices of representation and normalization affects accuracy. Table 1 displays the evaluation metrics averaged across all 27 outcomes for all combinations of lexicon choice and inclusion of the demographic control variables. Somewhat surprisingly, the PERMA lexicon does not appear to add much value — this may in part be due to the fact that it only contains 10 categories (versus 74 for LIWC). As it does not hurt performance, we retain it in other experiments — we show below that for certain health statistics it does produce statistically significant predictors. We delay discussion of the demographic control variables until the next section.

Table 2 evaluates the different normalization strategies using the **Controls+LIWC+PERMA** model. It is clear that using no normalization at all leads to poor results. This is not unexpected, since the variables will have very different ranges across different counties. We do find it informative that user normalization outperforms the alternative, more common normalization strategies. For example, Schwartz et al. [28] use **Log** normalization, and many Twitter influenza models use **Word** normalization [30]. We speculate that the superiority of **User** normalization here is mostly due to the inclusion of user description variables, which should only be counted once per user.

Norm	r	SMAPE
None	0.07	25.23%
Log	0.47	11.34%
Word	0.59	7.63%
User	0.63	7.37%

Table 2: Held-out correlation and SMAPE for the Controls+LIWC+PERMA model averaged across all 27 output variables using various normalization strategies.

RQ3: Beyond demographics

Our third research question asks what if any predictive value these linguistic variables provide beyond that of commonly used demographic covariates. It is possible that the correlations found in the linguistic variables in RQ1 are simply surrogates for demographic variables. Given the strong predictive accuracy of the control variables (c.f., Table 1), it is important to quantify the additional value added by Twitter. Table 1 provides a partial answer to this question — averaged across all 27 health statistics, including linguistic variables leads to an absolute 3% improvement in held-out correlation and a .28% improvement in SMAPE.

Table 3 provides a more detailed answer for each health statistic. By comparing values for **C** and **T+C**, we can see the change in held-out accuracy obtained by including linguistic variables in the model. We see that higher correlation is obtained by including linguistic variables for 19 of the 27 statistics and lower SMAPE is obtained for 20. Performing a Wilcoxon signed-rank test on the SMAPE values, we find three statistics that are significantly more accurately predicted (no socio-emotional support, no insurance, low birth weight) versus one that is significantly less accurately predicted (high school graduation rate). Moreover, for two statistics (Limited Healthy Food, Fast Food) the model using linguistic variables

	Pearson's r				SMAPE			
Outcome	**T**	**C**	**T+C**	**Δ**	**T**	**C**	**T+C**	**Δ**
Ambulatory Care	0.08	0.15	0.30	105%	9.1% ± 1.8	9.1% ± 2.4	8.5% ± 2.1	7%
Limited Healthy Food	0.51**	0.31	0.48**	53%	21.1% ± 4.6	26.2% ± 4.0	23.4% ± 4.3	11%○
Fast Food	0.31	0.24	0.31○	30%	4.3% ± 1.5	4.5% ± 0.9	4.4% ± 1.1	4%
No socio-emotional support	0.32○	0.59***	0.72***	22%	7.0% ± 1.3	6.2% ± 1.0	5.2% ± 0.8	16%*
Vehicle Mortality	0.52	0.52*	0.62○	20%	12.5% ± 2.5	11.8% ± 3.0	10.5% ± 1.8	10%
Unemployment	-0.11	0.36*	0.43***	19%	10.1% ± 1.4	9.1% ± 3.5	8.9% ± 2.3	2%
Diabetes	0.35**	0.45**	0.53***	17%	1.4% ± 0.2	1.3% ± 0.2	1.3% ± 0.1	4%
No Insurance	0.59	0.70**	0.80**	15%	11.6% ± 1.9	10.5% ± 1.6	8.4% ± 1.3	20%*
Low Birth Weight	0.41***	0.68***	0.77***	13%	5.8% ± 0.9	4.5% ± 0.8	3.9% ± 0.8	12%*
Obesity	0.43***	0.57***	0.64***	13%	6.0% ± 1.2	4.9% ± 1.2	4.7% ± 1.2	3%
Poor Health	0.23	0.71***	0.76***	6%	9.6% ± 1.7	6.8% ± 1.4	6.5% ± 1.2	4%
Unhealthy Days	0.01	0.63***	0.66***	6%	6.0% ± 0.6	4.3% ± 0.3	4.3% ± 0.6	1%
Inactivity	0.11	0.55***	0.58***	6%	7.6% ± 1.2	6.0% ± 1.0	6.0% ± 1.0	1%
Mentally Unhealthy	-0.22	0.46***	0.49***	6%	5.8% ± 0.9	5.1% ± 1.2	5.0% ± 1.3	1%
Drinking	0.12	0.18	0.19	5%	5.6% ± 1.2	5.8% ± 1.2	5.7% ± 1.5	1%
Smokers	0.15	0.65***	0.67***	3%	8.7% ± 1.3	6.7% ± 0.5	6.3% ± 0.8	5%
College	0.15	0.85***	0.87***	2%	5.0% ± 1.1	2.8% ± 0.4	2.6% ± 0.4	6%
Dentist Access	0.44	0.66***	0.67***	1%	10.1% ± 0.5	9.0% ± 2.3	8.9% ± 2.7	1%
Teen Births	0.50	0.86***	0.87***	1%	15.1% ± 2.4	8.6% ± 2.0	8.4% ± 2.5	3%
Child Poverty	0.29	0.93***	0.93***	0%	15.0% ± 1.2	5.5% ± 1.2	5.4% ± 1.2	1%
Single Parent	0.20	0.88***	0.88***	-0%	11.0% ± 1.9	5.1% ± 0.4	5.1% ± 0.6	0%
Chlamydia	0.20	0.74***	0.74***	-1%	19.3% ± 4.5	13.7% ± 2.0	13.8% ± 1.6	-1%
Mammography	0.25	0.57***	0.56***	-1%	2.9% ± 0.6	2.5% ± 0.6	2.5% ± 0.5	-1%
Violent Crime	0.32*	0.73***	0.72***	-2%	19.7% ± 3.5	14.0% ± 1.9	15.2% ± 0.9	-9%
Primary Care	0.30	0.68***	0.65***	-4%	11.4% ± 1.5	8.7% ± 1.4	9.0% ± 1.7	-4%
Rec Facilities	0.34	0.70***	0.67***	-4%	14.3% ± 3.1	10.7% ± 2.9	11.2% ± 3.0	-5%○
HS Grad Rate	-0.10	0.66***	0.58***	-12%	5.2% ± 1.1	3.4% ± 1.0	3.9% ± 1.3	-13%*

Table 3: Held-out correlation and mean SMAPE (with standard deviation) for each outcome under three models — *T*: Twitter model using LIWC and PERMA lexicons; *C*: control variables (age, gender, race, income); *T+C*: Twitter and controls. All models use **User** normalization. Δ is the percent relative improvement (either correlation or SMAPE) from model *C* to *T+C*, an estimate of how complementary the two models are. Pearson's r significance is indicated by $\circ = 0.1$, $* = 0.05$, $** = 0.01$, $*** = 0.001$ (degress of freedom = 98). The thresholds have been Bonferroni-corrected (using the 27 outcomes).

in isolation (**T**) is actually more accurate than the model that uses demographic variables in isolation (**C**).

These results appear to support the hypothesis that Twitter-derived variables complement demographic variables.

One note regarding the relatively poor performance for Violent crime and Chlamydia — according to the County Health Rankings project, the data collection methodology for these two statistics tends to vary widely across states, making inter-state comparisons difficult.[11]

RQ4: Identifying linguistic indicators

Our fourth research question seeks to identify linguistic categories that are significantly correlated with each health statistic. In addition to providing an additional validation, this may help health researchers formulate new hypotheses about the connection between behavior, personality, and health.

We begin by computing the correlation between each independent and dependent variable in isolation, across all 160 linguistic variables, 5 control variables, and 27 health statistics. Figure 3 plots the top 10 most correlated variables for the top 12 statistics from Table 3. Significance values are again computed using spatially-adjusted p-values.

To disentangle those linguistic variables that are acting as surrogates for demographic variables, we perform an additional analysis which controls for these factors. For each linguistic variable, we perform regression in which the independent variables consist of one linguistic variable and the five demographic control variables and the dependent variable is one of the 27 health statistics. We then compute the statistical significance of the coefficient estimated for each linguistic variable, again using a Bonferroni correction. To the best of our knowledge, no previous work has explicitly controlled for these demographic variables when identifying significant linguistic categories. This is important to determine which variables are simply recovering demographics, and which are providing additional information.

Because of the spatial autocorrelation inherent in this geographical data, rather than using ordinary least squares regression [1], we use two stage least squares spatial regres-

[11] **http://www.countyhealthrankings.org/resources/chr-2013-data-comparability-across-states**

sion[12] (using the `pysal`[13] Python package). We use a kernel weight matrix with the default values.

This analysis yielded 33 linguistic categories that were significant predictors of 6 different health statistics after controlling for demographics. Table 4 displays the 15 categories that were found to be significant predictors of at least two different statistics. For each, we display the top five most common words found in the category and the list of health statistics for which they are predictive. Overall, we find that the user description is more predictive than the tweet itself — 80% of the significant variables come from the user description.

Caution must be taken when interpreting these results — the true context of word usage on Twitter often differs from intuition. Below, we highlight a few significant categories, including examples of the most common phrases to provide missing context. As this is a purely correlational analysis, we make no claims as to the causal mechanisms underlying these findings.

- The Family category is correlated with several negative health outcomes (limited healthy foods, lack of health insurance, teen birth rate). This category contains words such as "family" (e.g., "I love my family"), "mom" (e.g., "stay at home mom" or "single mom"), and "daddy" (e.g., "daddy's girl", "r.i.p. daddy").
- The PosFeel (positive feelings) category is correlated with increased socio-emotional support. This category contains words such as "love", "happy", and "smile"; e.g. "Seeing people smile makes me happy", and "Do what makes you happy."
- Inhibition words (e.g., "stop") appear to correlate with positive health outcomes (lower unemployment, fewer physically and mentally unhealthy days). These words often appear with ambition-oriented quotes people enter in their description field (e.g., "don't stop when you are tired; stop when you are done"; "set your goals high and don't stop until you get there").
- Job-related terms are correlated with lower unemployment — the most common word in this category is "work", often used in phrases like "going to work" or "do I have to go to work today?".
- The word "god" is the most common term in both the *Relig*ious and *Metaph*ysical categories — it appears in user descriptions as in "I love God, my family, and friends." Such profiles tends to be correlated with limited access to healthy foods, lack of health insurance, and more vehicle mortalities. This may in part be explained by increased church attendance rates for counties in the deep South, which tend to be ranked lower by many health outcomes.

ERROR ANALYSIS

To better understand how the linguistic variables can improve accuracy, we identified counties that are similar demographically but different linguistically. To do this, we selected counties whose statistics were more accurately predicted using linguistic and control variables than using control variables alone, then identified the linguistic differences that contributed to that improved accuracy. We used the following process: For each health statistic, we compared the true value to the held-out value predicted by the controls only model (**C**) and to the Twitter plus controls model (**T + C**). We identified the county whose prediction was most improved by **T + C**. We then chose a second county that had a similar value predicted by the controls model, but a different true value (specifically, we sorted by the difference in the predicted values minus the difference in true value). Finally, we compared the linguistic variables from each of the two counties and identified those that differed the most, weighted by importance (specifically, we multiply the relative difference in the values multiplied by the absolute value of the correlation between that variable and the health statistic). In this way, we identified counties that appear similar when considering demographic variables, but exhibit different linguistic properties on Twitter.

	Kings	**Wayne**
Obesity	25.0	34.0
C	30.2	28.8
T+C	**26.8**	**30.2**
65 and over	11.5	12.8
d=Sports	.086	.179
E-	.017	.033
d=School	.133	.232
Sleep	.037	.067
Swear	.063	.118

Table 5: Comparison of obesity rates, predictions, and linguistic variables for Kings County, NY and Wayne County, MI, along with the most important linguistic categories. Augmenting the control variable model (**C**) with Twitter variables (**T+C**) improves accuracy.

	LA	**Jefferson**
Limited Healthy Food	1.0	11.0
C	5.3	5.9
T+C	**4.0**	**8.6**
med_income	52k	42k
d=Relig	.056	.127
d=Metaph	.068	.139
d=TV	.038	.016
d=Family	.036	.052

Table 6: Comparison of rates of limited healthy food access, predictions, and linguistic variables for Los Angeles County, CA and Jefferson County, AL, along with the most important linguistic categories. Augmenting the control variable model (**C**) with Twitter variables (**T+C**) improves accuracy.

We highlight two examples from this analysis. For obesity, we compared Kings County, NY (Brooklyn) and Wayne County, MI (which includes Detroit). (See Table 5.) Both are highly urbanized counties in the northern United States with similar demographics. However, Wayne County has a much

[12] We thank the anonymous reviewer for this suggestion.

[13] **`http://pysal.org`**

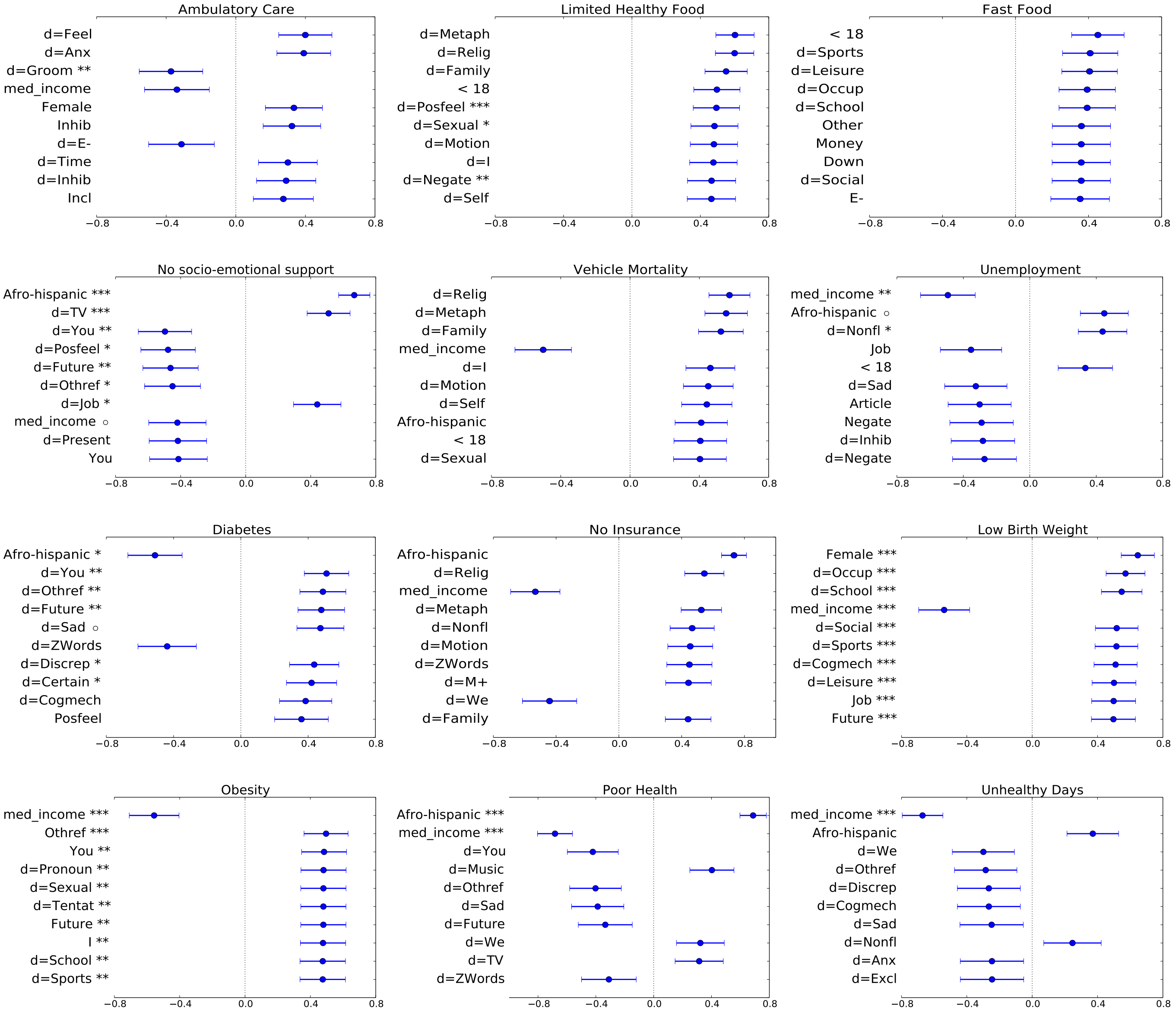


Figure 3: For the top 12 outcomes in Table 3, we plot the 10 variables with the highest correlation (error bars denote the 95% confidence interval.) Statistical significance is indicated by $\circ = 0.1$, $* = 0.05$, $** = 0.01$, $*** = 0.001$ (degress of freedom = 98). The thresholds have been Bonferroni-corrected using the total number of variables (160) times the number of outcomes (27). The prefix *d=* denotes lexical categories from the description field of a user's Twitter profile. Otherwise, the categories are derived from the tweet text. For comparison, the control variables are also included.

Cat	Examples	Outcomes
Affect	love, good, best, beautiful, happy	Low Birth Weight:d+ No socio-emotional support:d-
Family	family, mom, son, daddy, ex	No socio-emotional support:d- Teen Births:d+
Future	be, will, may, might, shall	Low Birth Weight:t+/d+ No socio-emotional support:t-
Metaph	god, die, jesus, blessed, christ	No socio-emotional support:t-/d- Teen Births:d+
Motion	follow, go, take, going, dance	Low Birth Weight:t+ No socio-emotional support:d-
Negate	not, no, nothing, without, nobody	Mentally Unhealthy:d- No socio-emotional support:t-/d- Poor Health:d- Unemployment:t-/d-
Other	they, she, her, he, them	Low Birth Weight:d+ No socio-emotional support:d-
Posemo	love, good, best, beautiful, happy	Low Birth Weight:d+ No socio-emotional support:d-
Present	is, follow, love, like, live	Low Birth Weight:d+ No socio-emotional support:d-
Pronoun	i, my, you, me, it	Low Birth Weight:d+ No socio-emotional support:t-/d-
Relig	god, jesus, blessed, christ, soul	No socio-emotional support:t-/d- Teen Births:d+
Sexual	love, loves, fu**ing, huge, gay	Low Birth Weight:d+ No socio-emotional support:d-
Social	you, we, who, girl, family	Low Birth Weight:d+ No socio-emotional support:d-
Sports	sports, football, basketball, play, teamfollowback	Low Birth Weight:t+/d+ No socio-emotional support:d-
TV	show, tv, movies, comedian, drama	No socio-emotional support:d+ Poor Health:d+

Table 4: A summary of 15 of the 80 lexical categories. These were selected by collecting all categories that are significantly correlated with at least two outcomes after controlling for demographics variables ($p < 0.05$, Bonferroni-corrected). We list the significantly correlated outcomes, the sign of correlation, and the field where the word was found: *t* for text and *d* for user description. E.g., the second row indicates that the presence of a word from the Family category in a user description is positively correlated with teen birth rates.

higher obesity rate (Wayne = 34, Kings = 25). In part because Wayne County has a higher proportion of people over 65 (1.3% higher), and this correlates with lower obesity rates, the controls-only model erroneously predicts a *lower* rate of obesity for Wayne County (Kings = 30.2, Wayne = 28.8). Including the linguistic variables improves accuracy considerably for both counties. We display the top five linguistic variables that influenced the score for **T+C**. As we can see, user descriptions from Wayne County are more likely to contain references to school and sports — the most common references are to football and basketball teams. Also, tweets from Wayne county exhibit a higher rate of negative engagement words (**E-**) (most common examples: “tired”, “bored,” “sleepy”), references to sleep (“bed”, “tired”, “wake”), and swear words (“ass”, “fu**ing”, “hell”). All of these lexical categories correlate with higher obesity rates. In this case, the linguistic variables provide a more nuanced distinction between two highly urbanized areas in the northern U.S.

Table 6 repeats this analysis for the Limited Healthy Food statistic (the percent of the population who live in poverty and are 1 mile from a supermarket in urban areas or 10 miles from a supermarket in rural areas). Here, we compare Los Angeles County, CA and Jefferson County, AL (which includes the major city of Birmingham). Los Angeles has much greater access to healthy foods, as predicted by both models. Income is a strong predictor of this statistic, so the smaller median income in Jefferson County has influenced the controls-only model. However, adding the Twitter variables results in a much larger (and more accurate) difference between the two counties. Jefferson County is more likely to have user descriptions containing religious and metaphysical words (e.g., “god,” “jesus,” “blessed”), family words (“family”, “mom”, “son”), and less likely to contain TV references (“show”, “tv”, “movies”). In this case, the linguistic categories appear to be distinguishing between two very different types of urban environments (West Coast versus Deep South). We find similar patterns with these linguistic categories for No Insurance, Vehicle Mortality, and Teen Birth Rate.

CONCLUSION AND FUTURE WORK

The main conclusion of our analysis is that Twitter activity provides a more fine-grained representation of a community’s health than demographics alone — for example, the health of counties with similar demographics can be distinguished by the prevalence of words indicating negative engagement (“tired”, “bored”), television habits (“show”, “tv”, “movies”), and religious observance (“jesus”, “blessed”). The reason for this appears to come from the insights Twitter provides into personality, attitudes, and behavior, which in turn correlate with health outcomes. We have provided a methodology to discover such predictive patterns from county-aligned Twitter data. Given the large number of variables explored, we have used very conservative estimates of significance to reduce the chance of Type 1 errors and adjust for spatial autocorrelation.

In the future, we plan to consider automatically-learned word categories (e.g., using topic models), as well as extra-linguistic attributes (e.g., graph connectivity, posting frequency).

REFERENCES

1. Anselin, L. *Spatial econometrics: methods and models*. Kluwer Academic Publishers, Dordrecht; Boston, 1988.
2. Chen, M. K. The effect of language on economic behavior: Evidence from savings rates, health behaviors,

and retirement assets. *American Economic Review 103*, 2 (Apr. 2013), 690–731.

3. Chiswick, B. R., and Miller, P. W. *The economics of language international analyses*. Routledge, London; New York, 2007.
4. Clifford, P., Richardson, S., and Hmon, D. Assessing the significance of the correlation between two spatial processes. *Biometrics 45*, 1 (Mar. 1989), 123–134. PMID: 2720048.
5. Culotta, A. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, ACM (New York, NY, USA, 2010), 115–122.
6. Culotta, A. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Lang. Resour. Eval. 47*, 1 (Mar. 2013), 217238.
7. Danner, D. D., Snowdon, D. A., and Friesen, W. V. Positive emotions in early life and longevity: findings from the nun study. *Journal of personality and social psychology 80*, 5 (May 2001), 804–813. PMID: 11374751.
8. De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In *ICWSM* (2013).
9. Dredze, M. How social media will change public health. *IEEE Intelligent Systems 27*, 4 (2012), 81–84.
10. Duggan, M., and Brenner, J. The demographics of social media users – 2012. Pew Internet & American Life Project, Feb 2013.
11. Flores, B. E. A pragmatic view of accuracy measurement in forecasting. *Omega 14*, 2 (1986), 93–98.
12. Ghosh, D. D., and Guha, R. What are we tweeting about obesity? Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science 40*, 2 (2013), 90–102.
13. Gottschalk, L. A., and Gleser, G. C. *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. University of California Press, Jan. 1979.
14. Graham, L E, n., Scherwitz, L., and Brand, R. Self-reference and coronary heart disease incidence in the western collaborative group study. *Psychosomatic medicine 51*, 2 (Apr. 1989), 137–144. PMID: 2710908.
15. Hanson, C. L., Burton, S. H., Giraud-Carrier, C., West, J. H., Barnes, M. D., and Hansen, B. Tweaking and tweeting: Exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of Medical Internet Research 15*, 4 (Apr. 2013), e62.
16. Hecht, B., Hong, L., Suh, B., and Chi, E. H. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *CHI* (New York, NY, USA, 2011), 237–246.
17. Howell, R. T., Kern, M. L., and Lyubomirsky, S. Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review 1*, 1 (2007), 83–136.
18. James W Pennebaker, M. R. M. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology 54* (2003), 547–77.
19. Jamison-Powell, S., Linehan, C., Daley, L., Garbett, A., and Lawson, S. "I can't get no sleep": discussing #insomnia on Twitter. In *CHI*, ACM (New York, NY, USA, 2012), 1501–1510.
20. Lampos, V., De Bie, T., and Cristianini, N. Flu detector: tracking epidemics on twitter. In *ECML/PKDD* (2010), 599–602.
21. Messer, L. C. Neighborhood-level characteristics as predictors of preterm birth: Examples from wake county, north carolina. Tech. rep., North Carolina Department of Health and Human Services, 2005.
22. Paul, M. J., and Dredze, M. You are what you tweet: Analyzing Twitter for public health. In *ICWSM* (2011).
23. Pedregosa, F., et al. Scikit-learn: Machine learning in python. *Machine Learning Research 12* (2011), 28252830.
24. Pennebaker, J., Francis, J., and Booth, R. Linguistic inquiry and word count: LIWC 2001. *World Journal of the International Linguistic Association* (2001).
25. Qiu, L., Lin, H., Ramsay, J., and Yang, F. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality 46*, 6 (Dec. 2012), 710–718.
26. Rabi, D. M., et al. Association of socio-economic status with diabetes prevalence and utilization of diabetes care services. *BMC Health Services Research 6*, 1 (Oct. 2006), 124. PMID: 17018153.
27. Sadilek, A., Kautz, H., and Silenzio, V. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI* (Dec. 2012).
28. Schwartz, H. A., et al. Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)* (2013).
29. Seligman, M. E. P. *Flourish: a visionary new understanding of happiness and well-being*. Free Press, New York, 2011.
30. Signorini, A., Segre, A. M., and Polgreen, P. M. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic. *PLoS ONE 6*, 5 (May 2011), e19467.
31. Sobal, J., and Stunkard, A. J. Socioeconomic status and obesity: A review of the literature. *Psychological Bulletin 105*, 2 (1989), 260–275.
32. Stewart, A., and Diaz, E. Epidemic intelligence: For the crowd, by the crowd. In *Web Engineering*, M. Brambilla, T. Tokuda, and R. Tolksdorf, Eds., no. 7387 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Jan. 2012, 504–505.