Prediction of Population Health Indices from Social Media using Kernel-based Textual and Temporal Features

Thin Nguyen
Deakin University, Australia
first.last@deakin.edu.au

Duc Thanh Nguyen
Deakin University, Australia
first.last@deakin.edu.au

Mark E. Larsen
Black Dog Institute, UNSW Australia
first.last@blackdog.org.au

Bridianne O'Dea
Black Dog Institute, UNSW Australia
b.odea@blackdog.org.au

John Yearwood
Deakin University, Australia
first.last@deakin.edu.au

Dinh Phung
Deakin University, Australia
first.last@deakin.edu.au

Svetha Venkatesh
Deakin University, Australia
first.last@deakin.edu.au

Helen Christensen
Black Dog Institute, UNSW Australia
h.christensen@blackdog.org.au

ABSTRACT

From 1984, the US has annually conducted the Behavioral Risk Factor Surveillance System (BRFSS) surveys to capture either health behaviors, such as drinking or smoking, or health outcomes, including mental, physical, and generic health, of the population. Although this kind of information at a population level, such as US counties, is important for local governments to identify local needs, traditional datasets may take years to collate and to become publicly available. Geocoded social media data can provide an alternative reflection of local health trends. In this work, to predict the percentage of adults in a county reporting "insufficient sleep", a health behavior, and, at the same time, their health outcomes, novel textual and temporal features are proposed. The proposed textual features are defined at mid-level and can be applied on top of various low-level textual features. They are computed via kernel functions on underlying features and encode the relationships between individual underlying features over a population. To further enrich the predictive ability of the health indices, the textual features are augmented with temporal information. We evaluated the proposed features and compared them with existing features using a dataset collected from the BRFSS. Experimental results show that the combination of kernel-based textual features and temporal information predict well both the health behavior (with best performance at rho=0.82) and health outcomes (with best performance at rho=0.78), demonstrating the capability of social media data in prediction of population health indices. The results also show that our proposed features gained higher correlation coefficients than did the existing ones, increasing the correlation coefficient by up to 0.16, suggesting the potential of the approach in a wide spectrum of applications on data analytics at population levels.

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054136

Keywords

cognitive computing; feature engineering; prediction; population health indices; textual features; temporal information; kernel-based features; geo-referenced tweets; online texts

1. INTRODUCTION

Medical and health practices in the twenty-first century are anticipated to closely integrate with 'digital footprints' left by individuals – the concept of a digital phenotype as an 'extended phenotype' to human biology [16]. The integration of the World Wide Web (WWW) with cognitive computing has shown a trend in contemporary studies on WWW which enable learning systems using data mining, pattern recognition and natural language processing to analyze and understand human behaviors. The main objective of this work is to advance behavioral medicine [1] and population health using massive scales of digital data. In particular, the traces are used to estimate the degree of population health indices, including heath behavior and health outcomes.

In 2014, "sleep pattern" was first included in a US national survey, showing an emerging interest in the health behavior, in addition to health outcomes, including mental, physical, and generic health, by the government. Local health data are crucial for providing indicators of health indices and identifying local needs. However, traditional datasets take years to become publicly available. For example, data regarding health-related risk behaviors of US residents in 2014 were available to the public from early 2016, more than one year lagged [2]. Geocoded social media data can provide a complementary, real-time, reflection of local health trends. Social media has improved health care quality with better communication between patients and clinicians. It offers patients an online platform to record and share health data about themselves through which people might learn efficient

ways to deal with their own disease. Through Facebook or Twitter, for example, social media provides a novel channel that quickly disseminates information to a large number of people virtually at no cost [4].

This paper aims to assess whether Twitter data can be used to predict population health behaviors and outcomes. These health indices are measured based on responses to health-related questions in the Behavioral Risk Factor Surveillance System (BRFSS), annually conducted by the United States Centers for Disease Control and Prevention (CDC). For health behaviors, "sleep pattern", firstly introduced in BRFSS 2014, is considered. The question is "On average, how many hours of sleep do you get in a 24-hour period?" and tweets are utilized to predict an "inadequate sleep" index – the percentage of adults that report sleeping less than 7 hours per night in a county. For health outcomes, the questions, for year 2014, include i) "Would you say that in general your health is excellent, very good, good, fair, or poor?", ii) "Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?", and iii) "Thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?" [2].

When using tweets to predict the health behavior and health outcomes of a population, textual features are extracted from the tweets and are used in predictive models. The Linguistic Inquiry and Word Count (LIWC) [21], a set of handcrafted features capturing the psychological meaning of words, can be used. Alternatively, latent topics can also be used as the features. The topics are not directly observed from corpus data but can be extracted automatically using latent Dirichlet allocation (LDA) [3].

Due to the scale of data, often in billions of tweets, aggregating the tweets at a population level is common practice [8, 22]. Textual features are then extracted on the aggregated tweets and the prediction of health indices can be performed using a regression model. This approach alleviates the computational cost of big data analysis. However, at the same time, the aggregation operation loses information on the distribution of the textual data over the population, and such information may be important for identifying the health behavior and health outcomes of the population.

Since textual features are often in high dimensions, direct analysis of those high- and multi-variate distributions of billions of data points would require exponential computational complexity. Also, there could be relationships between features and those relationships could convey predictive information of the health behavior and health outcomes. In this paper, we propose kernel-based features as mid-level features to encode textual information. The kernel-based features are constructed on other low-level textual features, e.g., latent topics learned using the LDA method [3] or LIWC features [21], via kernel functions and take into account the distributions of textual features and their relationships over populations. Thus, they capture both the characteristics of the textual information and the correlation between individual textual features at population level.

In this work, we implemented the kernel-based features on two different textual features including the latent topics extracted using the LDA method [3] and LIWC features [21]. We investigated three different kernel types including polynomial, Gaussian radial basis function (RBF), and sigmoidal kernel. The kernel-based features are then augmented by the temporal information extracted from the time-stamps of tweets to form richer features. We applied the proposed features for prediction of the health behavior and health outcomes for counties in the CDC dataset for 2014 [2]. On this dataset, counties were ranked according to one health behavior: inadequate sleep and three health outcomes: i) self-reported poor health, ii) days of poor physical health, and iii) days of poor mental health. A contemporaneous dataset of 1.96 billion tweets containing latitude and longitude coordinates was collected, and linked to the health data via the Federal Information Processing Standard (FIPS, e.g., Los Angeles County: FIPS code 06037) county code, mapped from latitude and longitude information using the cartographic boundary shapefiles provided by the US Census Bureau[1]. The data were then partitioned such that 70% of the data were randomly selected for training and the remaining 30% were used for testing. The Spearman rank correlation coefficient (rho) between the actual health ranking (from the BRFSS) and estimated health ranking was used to evaluate prediction performance. Experimental results favorably show the improvement gained by our approach using the kernel-based textual features in comparison to conventional approach [8, 22]. The improvement was substantial (up to 16%) and consistent over different prediction tasks. In addition, the prediction performance was also improved when the temporal information was employed.

The remainder of the paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the low-level features used in the estimation of the health indices. Section 4 presents our proposed features. The experiments are presented in Section 5, with the results reported in Section 6. Section 7 provides remarks and potential future work and Section 8 concludes the paper.

2. RELATED WORK

Tweets have been utilized as a sensor in several health care applications. For example, in public health surveillance, Signorini et al. [23] illustrated that the information extracted from Twitter can be a real-time, effective indicator of public sentiment and influenza disease activity. In [7], Chunara et al. found that the trend of the 2010 Haitian cholera outbreak derived from tweets was significantly correlated with that of official case data. Other than that, the reports from these studies were available up to two weeks earlier than conventional publications. In other areas, the features extracted from tweets were found to be powerful predictors of heart disease mortality across counties [11], or county-level HIV prevalence [15].

Features for population health.

To predict health indices for a population, its demographics is often used as the predictors in baseline models, such as in [8], where health statistics for US counties were estimated.

When social media data are used as an external source supporting the prediction, textual features of the data, such as topics and language styles, are often chosen as the discriminators. For example, to estimate health-related statistics, including obesity, health insurance coverage, access to healthy foods, and teen birth rates, linguistic features of

[1]http://bit.ly/2aF7hC5, downloaded 15 May 2016.

tweets were employed [8]. The linguistic features were also used to predict the level of well-being of US counties [22].

Topics, another type of textual features, was also used as the predictors of population health. For example, in [22], topics were found to be better than linguistic features in predicting well-being for US counties. In addition, when both of the textual features were used as an augmented set of features in the prediction, along with demographic and socio-economic characters (age, gender, ethnicity, income, and education), the accuracy was 10% improved.

In these work, to characterize the content, topics and language styles have been extracted [8, 22]. Online behavior has also been used as features in the health care domain, such as in [9], where an insomnia index is used to predict depression. In the work, the insomnia index is defined using the pattern of posting during the course of a 24 hour cycle. The authors found that users with depression tend to be active during the "night" window of 9PM and 6AM.

Aggregation of tweets.

Due to the large-scale of data in these work, extracting features for each tweet would had been expensive. Therefore, the popular practice has been to aggregate tweets by county to reduce the number of documents to be processed from millions to thousands, and hence reducing computational cost [8, 22]. However, much information could be lost by the aggregation operation. In addition, the distribution of the textual data within a county captures the characteristics of that county and thus may be useful for identifying the health indices.

Population health indices.

For the survey in 2014, CDC introduced "sleep duration" in the questionnaire [18]. This behavior has also been characterized using Twitter [20]. Tweeting time was also used to build an insomnia index, which was in turn was used to predict the prevalence of depression [9].

Tweets have also been employed as a proxy to estimate population health statistics. For example, topics and language styles extracted from tweets by county were claimed to be strong markers of life satisfaction at county level [22]. Language styles of tweets were also found to estimate well several county health statistics, such as obesity, health insurance coverage, or teen birth rates [8].

Large-scale computing.

To deal with the large-scale of social media data, such as 82 million county-mapped tweets in [22], advances in computing platform has been employed. For example, in [22], a Hadoop (an implementation of MapReduce [10]) cluster was employed to aggregate words in tweets by counties.

3. FEATURES OF HEALTH INDICES

To estimate health indices for a county, either the content or temporal information from tweets made in the county are extracted and used as the features for the estimation.

3.1 Textual features

Give a tweet, the content is extracted for constructing textual features. Two types of textual features commonly used in text analysis are language style and latent topics.

Language style.

To extract the language style of tweets, we used the LIWC 2015 package [21]. For an input tweet, the package returns 78 psycho-linguistic categories, such as linguistic, social, affective, cognitive, perceptual, biological, relativity, time orientations, drives, personal concerns, and informal language.

Latent topics.

To extract latent topics, latent Dirichlet allocation (LDA) [3], a Bayesian probabilistic modeling framework, is often used. In our implementation of the LDA framework, Gibbs inference was adopted [13]. We set the number of topics to 80, comparable to the number of LIWC features, and ran the Gibbs sampling for 5,000 samples. The last Gibbs sample was then used to estimate $P(\text{topic} \mid \text{word})$ - the probability of a topic given a word. Finally, each tweet was encoded by probabilities$P(\text{topic} \mid \text{tweet})$ computed as:

$$P_{(topic|tweet)} = \sum_{word \in topic} P_{(topic|word)} \times P_{(word|tweet)} \quad (1)$$

3.2 Temporal features

The posting time of a tweet is expected to capture the relevant health behavior – the prevalence of insufficient sleep in a county. The assumption is that the percentage of people online late would be a powerful predictor of insufficient sleep. While this health behavior and the health outcomes are strongly related [6, 14], the posting time feature is also expected to be a good indicator of the health outcomes.

For this feature, each county is represented as a 24-dimension vector where each element $i = 0..23$ is the portion of tweets made on hour i in the county in 2014. As the time-stamp of tweets is in UTC (Coordinated Universal Time), a conversion into local time is needed to make the texting time convey daily cycles of the posters. Firstly, the location (the county) of a tweet is determined by converting the latitude and longitude information tagged to the tweet into a US county using the shape-files provided by US Census Bureau[2]. Secondly, the timezone for the tweet is identified through county-timezone tuples provided by the National Weather Service[3]. Finally, the UTC time of the tweet is converted into local time using the timezone determined.

4. PROPOSED FEATURES

4.1 Kernel-based textual features

As presented in the introductory section, the distribution of textual features of a county could be informative and predictive for the health behavior and health outcomes of that county. Moreover, the relationships between textual features may also be important. In this paper, we propose mid-level features encoding the correlation between individual low-level textual features via a kernel function. Specifically, let C be a county on which a set of N_C tweets $\{t_1, t_2, ..., t_{N_C}\}$ are collected. Suppose that d different feature types (e.g., latent topics) are used to describe a tweet, i.e., each tweet t_i is encoded by a set of textual features $\mathbf{v}_i = [v_{i,1}, v_{i,2}, ..., v_{i,d}] \in \mathbb{R}^d$. We can construct a matrix $M_C \in \mathbb{R}^d \times \mathbb{R}^{N_C}$ as follow,

[2]http://bit.ly/2aF7hC5, downloaded 15 May 2016.
[3]http://www.nws.noaa.gov/geodata/catalog/county/html/county.htm

$$M_C = \begin{pmatrix} v_{1,1} & v_{2,1} & ... & v_{N_C,1} \\ v_{1,2} & v_{2,2} & ..... & v_{N_C,2} \\ ... & ... & ... & ... \\ v_{1,d} & v_{2,d} & ... & v_{N_C,d} \end{pmatrix} \quad (2)$$

The matrix M_C represents the distribution of textual features $\mathbf{v}_i$ over the county C. Based on M_C, we define a vector $\bar{\mathbf{v}} \in \mathbb{R}^d$ which contains the mean values of feature types j,$j = 1, 2, ..., d$ over N_C tweets as follow,

$$\bar{\mathbf{v}} = [\bar{v}_1, \bar{v}_2, ..., \bar{v}_d] = \left[\frac{\sum_{i=1}^{N_C} v_{i,1}}{N_C}, \frac{\sum_{i=1}^{N_C} v_{i,2}}{N_C}, ..., \frac{\sum_{i=1}^{N_C} v_{i,d}}{N_C} \right] \quad (3)$$

We define a centric-normalized matrix $\hat{M}_C$ which can be obtained by translating M_C by $\bar{\mathbf{v}}$ and normalizing it by $\sqrt{N_C}$. In particular, we compute

$$\hat{M}_C = \frac{1}{\sqrt{N_C}} \left(M_C - \bar{\mathbf{v}}^\top \mathbf{1} \right) \quad (4)$$

$$= \begin{pmatrix} \frac{v_{1,1}-\bar{v}_1}{\sqrt{N_C}} & \frac{v_{2,1}-\bar{v}_1}{\sqrt{N_C}} & ... & \frac{v_{N_C,1}-\bar{v}_1}{\sqrt{N_C}} \\ \frac{v_{1,2}-\bar{v}_2}{\sqrt{N_C}} & \frac{v_{2,2}-\bar{v}_2}{\sqrt{N_C}} & ..... & \frac{v_{N_C,2}-\bar{v}_2}{\sqrt{N_C}} \\ ... & ... & ... & ... \\ \frac{v_{1,d}-\bar{v}_d}{\sqrt{N_C}} & \frac{v_{2,d}-\bar{v}_d}{\sqrt{N_C}} & ... & \frac{v_{N_C,d}-\bar{v}_d}{\sqrt{N_C}} \end{pmatrix} \quad (5)$$

where $\mathbf{1} = [1, 1, ..., 1] \in \mathbb{R}^{N_C}$.

Finally, we define a set of kernel-based features $\mathbf{D}_C = [D_{C,1,2}, D_{C,1,3}, ..., D_{C,d-1,d}] \in \mathbb{R}^{\frac{d(d-1)}{2}}$ in which each element $D_{C,j,k}$ is the result of a kernel function K applied on two rows j and k of $\hat{M}_C$. In particular, let $\hat{M}_C(j)$ and $\hat{M}_C(k)$ respectively denote the j-th and k-th row of $\hat{M}_C$, $D_{C,j,k}$ is calculated as,

$$D_{C,j,k} = K\left(\hat{M}_C(j), \hat{M}_C(k)\right) = \left\langle \Phi\left(\hat{M}_C(j)\right), \Phi\left(\hat{M}_C(k)\right) \right\rangle \quad (6)$$

where $\Phi(\mathbf{x})$ is an implicit function that maps a vector $\mathbf{x}$ in a low dimensional space $\mathcal{L}$ (e.g., of L dimensions) to a higher (possible infinite) dimensional space $\mathcal{H}$ and $\langle \cdot, \cdot \rangle$ is the inner product of two vectors.

Note that the kernel-based features $\mathbf{D}_C$ are extracted on $\hat{M}_C$ instead of M_C because $\hat{M}_C$ has been aligned by $\bar{\mathbf{v}}$ and thus captures the variation of features which could be important to encode the characteristics of the textual information at population level. We also note that the normalized factor $\sqrt{N_C}$ is used in $\hat{M}_C$ to compensate the variation of numbers of tweets acrossing counties.

Using kernels for feature construction holds several advantages. First, if the kernel function K can be represented in the form the inner product of $\Phi\left(\hat{M}_C(j)\right)$ and $\Phi\left(\hat{M}_C(k)\right)$, as shown in Eq (6), the function Φ is not necessary to be known. Second, instead of working on a high dimensional space $\mathcal{H}$, all the computations can be done in a lower dimensional space $\mathcal{L}$. As defined in Eq (6), the feature set $\mathbf{D}_C$ captures both the characteristics of the textual information and the correlation between individual textual features j and k of county C.

In this paper, we investigate three kernel types which have been commonly used. The kernels are defined as follows,

Polynomial kernel

$$K\left(\hat{M}_C(j), \hat{M}_C(k)\right) = \left\langle \hat{M}_C(j), \hat{M}_C(k) \right\rangle^p \quad (7)$$

where p is the degree of the kernel. In our experiments, we set p to 3 as often used in the literature.

Gaussian radial basis function (RBF) kernel

$$K\left(\hat{M}_C(j), \hat{M}_C(k)\right) = \exp\left(-\frac{\left\| \hat{M}_C(j) - \hat{M}_C(k) \right\|_2^2}{2\sigma^2} \right) \quad (8)$$

where $\|\cdot\|_2$ is the L_2- norm and σ is a user parameter set to 0.1 in our experiments.

Sigmoidal kernel

$$K\left(\hat{M}_C(j), \hat{M}_C(k)\right) = \tanh\left(\left\langle \hat{M}_C(j), \hat{M}_C(k) \right\rangle \right) \quad (9)$$

where $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$.

Although it is not necessary to obtain an explicit definition of the mapping function Φ, it could be also determined in several specific cases. For example, with the polynomial kernel presented in Eq (7), it can be shown that Φ has the form as

$$\Phi\left(\hat{M}_C(j)\right) = \sqrt{\left(\frac{p!}{r_1! r_2! ... r_{N_C}!}\right)} \times \left[\hat{M}_C(j,1)\right]^{r_1} \left[\hat{M}_C(j,2)\right]^{r_2} ... \left[\hat{M}_C(j,N_C)\right]^{r_{N_C}}$$

where $r_l \geq 0, \forall l = 1, 2, ..., N_C$ and $\sum_{l=1}^{N_C} r_l = p$, and $\hat{M}_C(j,l)$ is the element at row j and column l of matrix $\hat{M}_C$.

In addition, as shown in [5], the dimension of the mapping space $\mathcal{H}$ of a p-degree polynomial kernel working on an N_C-dimensional space (as each row in $\hat{M}_C$ has N_C elements) will be $\begin{pmatrix} N_C + p - 1 \\ p \end{pmatrix}$. For example, if we are to process $N_C = 200$ tweets using a polynomial kernel of $p = 3$ degree, then $\mathcal{H} = \Phi(\mathcal{L})$ will have 1,353,400 dimensions. It is also noticeable that the mapping space $\mathcal{H}$ can also have infinite dimensions. For example, as proved in [5], for the RBF kernel, since $K\left(\hat{M}_C(j), \hat{M}_C(k)\right) \to 0$ as $\left\| \hat{M}_C(j) - \hat{M}_C(k) \right\|_2 \to \infty$, $\mathcal{H}$ will be an infinite dimensional space. However, irrespective of the high dimensionality of the mapping space $\mathcal{H}$, we are still able to perform all the operations, e.g., calculating inner products or distances, on the lower dimensional space $\mathcal{L}$ while getting the same effect on $\mathcal{H}$ without increasing any computations. For example, there would be no computational cost incurred when the degree of the polynomial kernel increases as the main operation with this kernel is the inner product which does not depend on the degree p.

4.2 Fusion of textual and temporal features

The time-stamps associated with tweets provide the temporal information which could be another important cue for prediction of health behaviors and health outcomes. We consider the 24-dimensional frequency extracted from a county as the temporal feature vector of that county. In addition

to using the temporal information as the features in the prediction of population health indices, augmenting the information to the conventional textual features is also considered. The textual features (for example, $\mathbf{D}_C$ proposed in Section 4.1) and the temporal feature can be combined by a concatenation operator to form a richer feature vector capturing both textual and temporal information.

5. EXPERIMENTAL SETUP

5.1 Dataset

Tweets.

From June 2013 to July 2016, tweets geo-tagged within a bounding box of the USA (-170.0,18.0,-60.0,72.0) were streamed using the Twitter API[4]. This resulted in a corpus of 1,961,536,285 *non-re-tweeted* tweets, posted by 15,635,491 users, and archived in 5,414 gigabytes of storage. The latitude and longitude information of the tweets was mapped to US counties using the cartographic boundary shapefiles provided by US Census Bureau[5]. As CDC data for all the health indices is only available for year 2014, only tweets for 2014 were used in this study, consisting of 768,791,808 tweets mapped to 3,221 US counties.

County health indices.

All data for county-level health indices, including the health behavior and health outcomes, were drawn from the BRFSS [2]. The BRFSS is an ongoing telephone-based population health survey, conducted by the CDC. Survey results are published online annually by the CDC.

For the health behaviors, participants in the surveys were asked about smoking, drinking, and, for the first time, sleep patterns ("On average, how many hours of sleep do you get in a 24-hour period?") [2]. As tweeting time appears to be a powerful predictor of sleep pattern, this health behavior index was chosen into this study. In particular, "inadequate sleep" index (or insomnia index) is defined as the percentage of adults that report sleeping less than 7 hours per night.

For health outcomes, the indices include i) poor mental health days (referred to as "mental health") – the average number of reported mentally unhealthy days per month, ii) poor physical health days (referred to as "physical health") – the average number of reported physically unhealthy days per month, and iii) self-reported poor or fair health (referred to as "general health") – the percentage of adults that report fair or poor health.

At the time of writing, the most recent data available were for year 2014. As tweet data were available from June 2013, county health indices for 2014 were downloaded.

5.2 Prediction model

The experiment setup was originally designed to use data from previous years to predict health indices for a given year. However, as the inadequate sleep data from CDC is available for 2014 only, we turned to conduct across-county prediction of the health indices. This could be applied to the scenario where the health indices by CDC are available for some counties, but not for all. Specifically, the data from year 2014 was used for both model training and testing. In this case study, the 3,221 counties were randomly sampled into a training group of 2,255 counties (70%) and a validation group of 966 counties (30%).

As defined in [2], we predicted three health outcome indices: "mental health", "physical health", and "general health"; and one health behavior index: inadequate sleep. As often used in population health prediction, e.g.,[8, 22], linear regression was adopted in our experiment. We note that by using the same prediction model, comparison of existing methods can be done straightforwardly. In particular, the model maps input variables, which consists of features, to an output variable, that is a particular CDC health index. Let y_i denote the value of a health index variable Y for an observation unit i (an individual county). Let x_i be the input (i.e., features) of the unit i. The relationship between y_i and x_i is modeled as follow,

$$y_i = \beta_0 + \beta^T x_i + e_i \tag{10}$$

where e_i is an error term.

The regression model can be learned by fitting the coefficient vector β. This task is equivalent to minimizing the prediction error and hence can be regularized by a Lasso constraint $\|\beta\|_1 < c$, where the threshold c was determined via 5-fold cross validation [12].

To evaluate and compare feature types, the Spearman rank correlation coefficient between the actual health ranking and estimated health ranking was adopted as the measure of prediction performance.

5.3 Computing environment

To compute the proposed kernel-based features, the latent topics and the LIWC for each tweet has to be extracted and the number of tweets to be handled is up to several hundred millions. To deal with this scale, we made use of Apache Spark, an emerging cluster computing platform [25]. Spark is shown to perform better than Hadoop [10] in both optimization using gradient descent and interactive analytics, e.g., querying large corpora [25, 24]. The key difference is that while MapReduce of Hadoop is a disk-based system, Spark is an in-memory one. For example, in performing gradient descent, while MapReduce reads the same data from disks repeatedly for every iteration, Spark loads the data once and keeps it in memory for following iterations [25, 24]. In addition, Spark enables distributed and parallel computations and thus makes the computations efficiently.

In our experiments, a Spark cluster of eight worker nodes was employed. Each node was featured by a dual eight-core Intel® Xeon® E5-2670@2.60GHz processors, 128 gigabytes of main memory, and CentOS 7.2 operating system.

6. EVALUATION AND COMPARISON

6.1 Evaluation of kernel-based textual features

We first evaluated the kernel-based textual features on the prediction of health indices (see Table 1). In this experiment, the result of the conventional approach [8, 22] is also included. Recall that in [8, 22], all tweets in a county are aggregated to a so-call aggregated tweet on which textual features, e.g., latent topics/LIWC are extracted. We call this feature type the Aggregated (Agg) feature. We also highlight the best results for each health index in Table 1. As shown in our experiments, for prediction of inadequate sleep

[4]https://stream.twitter.com/1.1/statuses

[5]http://bit.ly/2aF7hC5, downloaded 15 May 2016.

	LIWC				Latent topics			
	Inadequate sleep	Mental health	Physical health	General health	Inadequate sleep	Mental health	Physical health	General health
Agg	0.66	0.56	0.60	0.69	0.65	0.64	0.56	0.58
Polynomial	0.71	0.61	0.63	0.70	0.74	0.73	0.68	0.72
Sigmoidal	0.72	0.64	0.65	0.72	0.74	0.73	0.68	0.72
RBF	0.70	0.61	0.63	0.70	**0.78**	**0.76**	**0.72**	**0.74**

Table 1: Prediction performance (in term of Spearman's rho between CDC values and estimated values) of kernel-based features on various health indices.

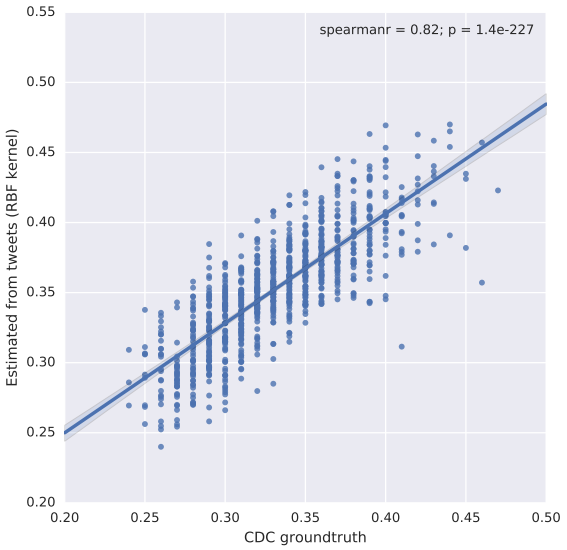

Figure 1: Prediction performance of RBF kernel on the topics as features to predict the insomnia health index. With temporal information augmented, the model gains the best result in the prediction of the health behavior.

(the 2nd and 6th column in Table 1), there was a slight difference in prediction performance between the kernel-based features applied on the LIWC; yet sigmoidal kernel achieved the best performance (rho=0.72). In contrast, on the latent topics, RBF kernel showed a significant improvement (about 0.04) compared with polynomial and sigmoidal kernel. In addition, this is the best performance (rho=0.78) for prediction of the health behavior.

For prediction of health outcomes, sigmoidal kernel applied on the LIWC (see the 3rd, 4th, and 5th column in Table 1) performed best on all the health outcomes. On the other hand, using the latent topics (see the 7th, 8th, and 9th column in Table 1), the RBF kernel obtained the highest prediction accuracy, also on all the health outcomes. In all, the kernel-based features showed a better performance when applied on the latent topics than on the LIWC and RBF kernel consistently remained the best performance on prediction of all health indices.

6.2 Evaluation of temporal features

We verified the impact of the temporal information in prediction of the health behavior and health outcomes. Table 2 reports the prediction performance of the temporal information. Experimental results show that the temporal information can be an indicator of inadequate sleep. In addition, the temporal information also made the greatest impact on prediction of inadequate sleep (rho=0.60) in comparison to other health indices. Note that in this experiment, the same prediction model (presented in Section 5.2) was used.

Health index	Correlation
Inadequate sleep	0.60
Poor mental health	0.43
Poor physical health	0.42
Poor or fair health	0.52

Table 2: Prediction performance of temporal information on various health indices.

We also verified the role of the temporal information when it was combined with other textual features (see Table 3). Our experiments show that, when the temporal information was augmented with other textual features, the prediction of health indices could be improved. Generally speaking, the improvement on prediction of the health behavior was more prominent compared with prediction of the health outcomes. In addition, the combination of the temporal information and RBF kernel applied latent topics performed best on prediction of all the health indices. Figure 1 shows an example of the approach in the prediction of the insomnia health index.

We note that the use of the temporal information to augment the conventional Agg feature could also enhance the prediction ability on all health indices and the improvement was up to 10% on prediction of inadequate sleep using topics.

6.3 Kernel-based versus aggregated features

In addition to evaluation of the proposed kernel-based textual and temporal features, we also compared the kernel-based features with the Agg feature used in existing works [8, 22]. In those methods, the Agg feature was the LIWC or latent topics extracted on aggregated tweets.

Table 3 shows the comparison between our kernel-based features and the Agg feature. As shown in this table, in overall, our kernel-based features significantly outperform the Agg feature in prediction of all health indices. Experimental results also show that the improvement gained by the proposed kernel-based features over the Agg feature was up to 0.16 in the prediction of both physical health (rho=0.72 by RBF kernel and 0.56 by Agg) and general health indices (0.74 by RBF kernel and 0.58 by Agg).

6.4 Demography versus Twitter activity

Demographic and socio-economic variables, such as age, gender, and ethnicity, have been found to be correlated with

	LIWC		Latent topics	
	Without temporal	With temporal	Without temporal	With temporal
Agg	0.66	0.70	0.65	0.75
Polynomial	0.71	0.75	0.74	0.79
Sigmoidal	0.72	0.76	0.74	0.79
RBF	0.70	0.76	0.78	**0.82**

(a) Prediction of inadequate sleep index.

	LIWC		Latent topics	
	Without temporal	With temporal	Without temporal	With temporal
Agg	0.56	0.59	0.64	0.68
Polynomial	0.61	0.63	0.73	0.76
Sigmoidal	0.64	0.64	0.73	0.76
RBF	0.61	0.65	0.76	0.78

(b) Prediction of mental health index.

	LIWC		Latent topics	
	Without temporal	With temporal	Without temporal	With temporal
Agg	0.60	0.63	0.56	0.62
Polynomial	0.63	0.65	0.68	0.72
Sigmoidal	0.65	0.66	0.68	0.72
RBF	0.63	0.65	0.72	0.75

(c) Prediction of physical health index.

	LIWC		Latent topics	
	Without temporal	With temporal	Without temporal	With temporal
Agg	0.69	0.72	0.58	0.66
Polynomial	0.70	0.74	0.72	0.74
Sigmoidal	0.72	0.75	0.72	0.74
RBF	0.70	0.74	0.74	0.77

(d) Prediction of generic health index.

Table 3: Performance (rho) of aggregated and proposed kernel-based features (in correlation with CDC data), with and without augmenting temporal information, in predicting of the population health behavior and health outcomes.

Heath index	Demo.	LIWC+Time		Topics+Time	
		No demo.	With demo.	No demo.	With demo.
Inadequate sleep	0.60	0.76	0.80	0.82	**0.85**
Mental health	0.49	0.65	0.67	0.78	0.80
Physical health	0.50	0.65	0.69	0.75	0.78
Generic health	0.57	0.74	0.79	0.77	0.81

Table 4: Performance (rho) of demographics and Twitter activity features, with and without augmenting demographic information, in predicting of the population health behavior and health outcomes.

county-level health outcomes [8, 22]. At state-level, these variables alone could achieve good predictions of population non-communicable disease outcomes (median correlation of 88%) [19]. We examined whether these factors could be potential confounders for the predictions of the population health indices. In particular, we ran the predictions using either demographic variables alone or include these variables into textual and temporal features (generated by the best kernel RBF). The demographic variables at county level for 2014 were provided by the County Health Rankings & Roadmaps[6]. They included ten variables for gender, age, race, and education. The result is shown in Table 4.

For all the health indices, the predictions using demography along are significantly correlated with the health statistic provided by CDC, with Pearson's rho ranges from 0.49 to 0.60. This performance is worse than those achieved with Twitter activity, including the textual content represented in either LIWC or topics augmented with temporal information. However, when being included into the best models for Twitter activity, demographic variables always improved the accuracy. The improvement was up to 0.05 in the case of predicting generic health index using LIWC and temporal information from tweets.

7. DISCUSSION

In general, while the Agg feature performs similarly on both the latent topics and LIWC, the kernel-based features have a clear benefit over the latent topics. This shows that, compared with the language style, the content covered by the latent topics, has higher inter-correlation and this relationship is important to prediction of population health indices. This could be probably because tweets are often presented in short forms and thus would be difficult to express the language style. However, extracting latent topics requires an inference operation on every tweet and thus is time consuming. In addition, the number of topics is crucial to the prediction performance yet unknown a priori. Literature has shown advances in topic modeling, for example, discriminative LDA [17] or maximum entropy discrimination LDA [26]. We consider applying those techniques to learn latent topics on which the proposed kernel-based textual features can be extracted as our future work. For linguistic representation, rather than LIWC, other lexicons could also be used as the low level features of health indices, such as PERMA, which is said to be relevant to health and personality [8].

In this paper, we investigated the capability of the kernel-based textual features in prediction of overall health outcomes (mental, physical, and generic health) and health behavior (inadequate sleep). However, our proposed kernel-based features are general and thus can be used to predict other health-related statistics, such as obesity, health insurance coverage, well-being, or teen birth rates, as in [8, 22]. We are extending our features in other prediction tasks.

8. CONCLUSION

This paper proposes kernel-based textual and temporal features for prediction of population health indices. The kernel-based features are formed by considering the distributions of textual features over the population tweets and encode the relationships between individual textual features

[6]http://www.countyhealthrankings.org/rankings/data

using kernel functions. The kernel-based features can be applied on top of other low-level textual features to obtain mid-level textual features that are able to capture the characteristics in the corpus data of populations. The kernel-based textual features are then augmented by temporal information to enhance the predictive ability of health indices.

We evaluated the proposed features with various kernel types and on a big dataset of hundred millions geospatially coded tweets. Experimental results show that the use of kernel-based textual features significantly outperformed (up to 16%) the conventional approach of extracting features by aggregating tweets at population level. It is also shown that the use of temporal information further improved the prediction performance. This suggests the potential and applicability of the proposed features as well as the temporal information in a wide spectrum of applications requiring data analytics at population levels.

Acknowledgment

This work is partially supported by the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning.

9. REFERENCES

[1] Ayers, J. W., Althouse, B. M., and Dredze, M. Could behavioral medicine lead the web data revolution? *JAMA 311*, 14 (2014), 1399–1400.

[2] Behavioral Risk Factor Surveillance System. *2014 Behavioral Risk Factor Surveillance System Questionnaire*, December 2013. http://bit.ly/2aJOXIl, retrieved May 2016.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[4] Bull, S. S., Breslin, L. T., Wright, E. E., Black, S. R., Levine, D., and Santelli, J. S. Case study: An ethics case study of HIV prevention research on Facebook: The just/us study. *Journal of Pediatric Psychology 36*, 10 (2011), 1082–1092.

[5] Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*, 2 (1998), 121–167.

[6] Buysse, D. J., Grunstein, R., Horne, J., and Lavie, P. Can an improvement in sleep positively impact on health? *Sleep Medicine Reviews 14*, 6 (2010), 405–410.

[7] Chunara, R., Andrews, J. R., and Brownstein, J. S. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene 86*, 1 (2012), 39–45.

[8] Culotta, A. Estimating county health statistics with Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), pp. 1335–1344.

[9] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2013), pp. 128–137.

[10] Dean, J., and Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Communications of the ACM 51*, 1 (2008), 107–113.

[11] Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., and Sap, M. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science 26*, 2 (2015), 159–169.

[12] Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1 (2010), 1.

[13] Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences 101*, 90001 (2004), 5228–5235.

[14] Harris, M., Glozier, N., Ratnavadivel, R., and Grunstein, R. R. Obstructive sleep apnea and depression. *Sleep Medicine Reviews 13*, 6 (2009), 437–444.

[15] Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L. H., and Albarracín, D. Future-oriented tweets predict lower county-level HIV prevalence in the United States. *Health Psychology 34*, S (2015), 1252.

[16] Jain, S. H., Powers, B. W., Hawkins, J. B., and Brownstein, J. S. The digital phenotype. *Nature Biotechnology 33*, 5 (2015), 462–463.

[17] Lacoste-Julien, S., Sha, F., and Jordan, M. I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems* (2009), pp. 897–904.

[18] Liu, Y., Wheaton, A., Chapman, D., Cunningham, T., Lu, H., and Croft, J. Prevalence of healthy sleep duration among adults – United States, 2014. *MMWR Morbidity and Mortality Weekly Report 65*, 6 (2015), 137–141.

[19] Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., Phung, D., Venkatesh, S., and Allender, S. Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. *PLOS ONE 10*, 5 (05 2015), 1–13.

[20] McIver, D. J., Hawkins, J. B., Chunara, R., Chatterjee, A. K., Bhandari, A., Fitzgerald, T. P., Jain, S. H., and Brownstein, J. S. Characterizing sleep issues using Twitter. *Journal of Medical Internet Research 17*, 6 (2015), e140.

[21] Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. *Linguistic Inquiry and Word Count: LIWC 2015 [Computer software]*. Pennebaker Conglomerates, Inc., 2015.

[22] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E., and Ungar, L. Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2013), pp. 583–591.

[23] Signorini, A., Segre, A. M., and Polgreen, P. M. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS ONE 6*, 5 (2011), e19467.

[24] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., Mccauley, M., Franklin, M., Shenker,

S., and Stoica, I. Fast and interactive analytics over Hadoop data with Spark. *;login: 37*, 4 (2012), 45–51.

[25] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. Spark: Cluster computing with working sets. In *Proceedings of the USENIX Conference on Hot Topics in Cloud Computing* (2010), p. 10.

[26] Zhu, J., Ahmed, A., and Xing, E. P. MedLDA: Maximum margin supervised topic models. *Journal of Machine Learning Research 13*, Aug (2012), 2237–2278.