Multimedia Tools and Applications (2021) 80:7187–7204
https://doi.org/10.1007/s11042-020-10034-0

A graph-based approach for population health analysis using Geo-tagged tweets

Hung Nguyen[1] · Thin Nguyen[2] · Duc Thanh Nguyen[3]

Received: 7 May 2020 / Revised: 13 August 2020 / Accepted: 6 October 2020 /
Published online: 26 October 2020

Abstract
We propose in this work a graph-based approach for automatic public health analysis using social media. In our approach, graphs are created to model the interactions between features and between tweets in social media. We investigated different graph properties and methods in constructing graph-based representations for population health analysis. The proposed approach is applied in two case studies: (1) estimating health indices, and (2) classifying health situation of counties in the US. We evaluate our approach on a dataset including more than one billion tweets collected in three years 2014, 2015, and 2016, and the health surveys from the Behavioral Risk Factor Surveillance System. We conducted realistic and large-scale experiments on various textual features and graph-based representations. Experimental results verified the robustness of the proposed approach and its superiority over existing ones in both case studies, confirming the potential of graph-based approach for modeling interactions in social networks for population health analysis.

Keywords Graphs · Large-scale computing · Health on the web · Population health · Geo-tagged tweets

1 Introduction

Population health measurement reflects the dynamic state of physical, mental, and social well-being of a community [18, 43]. Understanding population health is thus essential for

✉ Duc Thanh Nguyen
duc.nguyen@deakin.edu.au

Hung Nguyen
hungnd@ntu.edu.vn

Thin Nguyen
thin.nguyen@deakin.edu.au

[1] Faculty of IT, Nha Trang University, Nha Trang, Vietnam

[2] Applied Artificial Intelligence Institute, Deakin University, Geelong, VIC 3220, Australia

[3] School of Information Technology, Deakin University, Geelong, VIC 3220, Australia

governments to identify health-related concerns and develop strategic healthcare programs for communities.

Traditionally, population health data is collected via telephone interviews or postal questionnaires. The benefits of this approach include the simplicity of data collection and the reliability of responses. This is mostly because the questionnaires have been designed by professionals, and the population of interest have been actively and intentionally targeted. Despite these advantages, traditional health surveys exhibit two major limitations: expensive cost and time-consuming. For instance, the budget spent for the Behavioral Risk Factor Surveillance System (BRFSS) survey in Florida, US over 5 years 2011 - 2015 was more than 3.5 million USD,[1] and the BRFSS reports in 2017 were typically based on the data collected in or before 2015,[2] which, in turn, could lead to delayed public health policy decisions.

Social behaviors of a population provide cues for the health status of that population. The challenge here is how to obtain large-scale and diversified datasets of such behaviors in an automatic and low-cost manner. Fortunately, with the advent of the social networks that allow billions of people to easily connect and communicate, social media has become an abundant and diversified source of information for many healthcare studies [26]. Examples include tracking influenza-like illness in populations from Google search queries [23], localizing illnesses by region [44], or measuring life satisfaction of populations using Twitter data [50]. Importantly, several studies have shown that data collected from social networks (e.g., Twitter) highly correlates to the results achieved by phone-based surveys [19, 35]. Moreover, community behaviors can be analyzed through social media in real-time and at extremely low cost. Therefore, social media, if exploited properly, can provide important insights into understanding people's health behaviors at both individual and population level. Our work is also motivated by this trend, i.e., exploiting social media as an information source for automatic population health analysis. More specifically, we aim at using geo-tagged tweets to predict and classify population health behaviors and outcomes.

Conventional social media-based health analysis methods extract health-related information from the content of the social media data, e.g., from textual features [14, 15] or built via relationships between the features [35, 36]. As shown in many studies in psychology and sociology [5, 53], interactions in social networks are important factors to understanding behaviors of communities [2]. However, this sort of information has not been explored in social media-based healthcare research. In this paper, we propose a graph-based approach taking into account these interactions for population health analysis. Specifically, our contributions are four-fold as follows,

- We propose to model interactions in social media data using graph theory. Graphs offer a natural way to capture relationships in data and is often used to represent interactions in social networks. However, existing approaches build graphs from social networks' users and hence require expensive computation to handle large-scale networks. In contrast, our graphs are constructed in a more manageable and scalable manner. Specifically, we propose two graph structures called inter-feature and inter-tweet graphs built via the coincidence of features and interactions between groups of tweets. To the best of our knowledge, our graph construction methods are novel and our work is the first taking into account interactions in social media data using graphs for population health analysis.
- We investigate various graph-based representations defined on the two proposed graph structures for representing social media data at population scale.

[1] https://bit.ly/2JjWqgn
[2] https://www.cdc.gov/brfss/

- We apply our model in two case studies: (1) estimating health indices, and (2) classifying health situation of counties in the US.
- We conduct extensive experiments on a large-scale dataset consisting of more than one billion geo-tagged tweets over three years 2014, 2015 and 2016. Experimental results show that our proposed approach outperformed existing ones in both case studies.

Parts of this paper have been published in our recent work [38]. Compared with the previous version, this work makes several extensions. First, in this version, we propose inter-tweet graphs for modelling response behaviour, e.g. like/reply, in social networks. Second, we extend our experiments with the case study of health status classification and provide detailed insights with in-depth discussion.

2 Related work

2.1 Public health analytics with social media

Social media has been serving as a rapid vehicle for public health analytics. Social media data has also been proven to be superior to traditional means due to time- and cost-effectiveness [11]. Applications of social media in public healthcare can be found in detection and monitoring of health issues including social well-being, positive mental health, and self-rated health [6], in forecasting public health trends [33], and in developing prevention programs [31].

The correlations between social media and clinical datasets haven been demonstrated in various ways [4, 14, 19, 44]. For instance, in [25] Facebook likes were used to predict mortality, diseases, and lifestyle behaviors of 214 counties across the US and the prediction results were shown to be comparable with those obtained from the BRFSS. In [22], tweets were used to build sentiment scores which were found to highly relate to self-rated mental health, sleep quality, and heart disease at census tract level for the city of San Diego over the period of 2014-12-06 to 2017-05-24. In [13], natural language processing tools were applied on social media data, e.g., Twitter, Reddit, and Facebook, to answer public health research questions.

Social media offers an effective means for tracking public health attitudes and behaviors. Applications include tracking disease-relevant behaviors and sentiments [49], understanding patient experiences and healthcare quality [48], building disease surveillance systems, supporting public health tracking and prevention [1]. For instance, in [44], health-related tweets with geo-tags were used with the Ailment Topic Aspect Model for tracking influenza over time. The tracking results were benchmarked against the influenza database from the Centers for Disease Control and Prevention (CDC). In [42], health-related tweets were combined with Wikipedia articles for identifying public health concerns in populations. Similarly, in [28], tweets were used to monitor the rate of alcohol consumption across regions in the UK. In [15], linguistic analysis was applied on Twitter's data and provided a finer-grained representation of population health. In the same manner, tweets have been found useful in prediction of depression in populations [16]. Tweets also have been incorporated with prior knowledge for tracking illnesses, measuring behavioral risk factors, or localizing illnesses by geographic regions [44].

Social media also plays an important role in preventive healthcare. For example, launched in 2019, Facebook's Preventive Health [31] provided preventive health recommendations customized to users' age and sex. This demonstrates social media's capacity

in encouraging healthy behaviors to a wide population, including vulnerable and isolated ones. Most recently, social media has been integrated to support managing the COVID-19 pandemic in both preparedness and emergency response [32]. As shown in [27], in difficult situations, e.g., during the MERS outbreak in South Korea where the information from public health officials was untrustworthy, social media could be considered as an alternative source.

Besides the supportive role, social media has also been found to have detrimental influences on public health [41], e.g., on suicidal behavior [51]. Likewise, social media has been considered to be a main source spreading health misinformation, such as the COVID-19 conspiracy theories [3]. Such information could make people reluctant to engage in health-protective behaviors [3].

2.2 Social media data types

Literature has shown the role of various social media data types in population health analysis [37]. Analysis of textual data can provide user demographics, personality, psychological state, and mental health situation. To encode the textual data, textual features are extracted. Those features capture both the content and emotion from the language used in social media [15, 36, 46].

In prediction of population health indices through social media, linguistic style is often employed. Linguistic style is an indicator of emotion in the language and has been discovered to highly relate to health outcomes [15, 36]. For instance, when reading depressing stories, judges tend to get depressed accordingly. Based on these findings, a software package, namely Linguistic Inquiry and Word Count (LIWC), was developed to extract psycho-linguistic features from textual data, e.g., documents or tweets [45]. Basically, given a text to be analyzed, the LIWC software first goes through every word of the text, makes comparison between each word with a pre-built dictionary [45], then calculates the percentage of each LIWC category occurring in the given text, and finally results in a list of categories with their rates.

Various studies have adopted LIWC features to problems of social media-based health analysis [15, 50]. For instance, Culotta [15] performed linguistic analysis of activities on Twitter to estimate health indices from County Health Rankings and Roadmaps. Experimental results showed significant correlations (with 6 of the 27 indices) between the language that people used and their health situation. This study also indicated that tweets better captured the health status of a community than demographics. In addition, the linguistic style features were found to be predictive of well-being of the US counties [50].

Another popular type of textual features in analysis of health-related concerns is latent topics. Topics capture the content of the textual data and can be learned using topic modeling techniques. A commonly used topic modeling method is called latent Dirichlet allocation (LDA) [7]. LDA is an unsupervised technique using Bayesian probabilistic framework to learn latent topics from a corpus. As shown in [50], LDA topics derived from tweets were more useful than LIWC in predicting life satisfaction in the US counties. Moreover, when combined with LIWC, demographic and social-economic controls (age, sex, ethnicity, income, and education), prediction performance was significantly improved.

Spatio/temporal-referenced data have also been utilized in public healthcare. In particular, accumulated geo-tagged data can be harnessed to determine health issues, monitor the spread of infectious diseases, and analyze the effects of clinical concepts on public health. For instance, geo-tagged data was employed to estimate geographic densities of clinical concepts in regions of interest [20], cluster groups of data having similar location

characteristics [47], and build recommendation systems to advise locations of interest [55]. In [36], temporal information encoded in tweets was augmented with textual features to predict sleep patterns of populations.

2.3 Graph-based representation of interactions

Interactions among users are an important aspect of social networks and are often modeled as graphs. For instance, Chun et al. [12] constructed a graph to model the interactions in Cyworld social network. In this graph, nodes were users and edges between two users represented communications (writing/responding) between them. Edge weights were computed based on the frequency of communications. Similarly, Leskovec and Horvitz [30] created a communication network for the Microsoft Messenger instant-messaging system in which each user was represented by a node and an edge was formed between nodes if the corresponding users exchanged at least one message during the month of observation. In [53], Wilson et al. argued that, as observed from Facebook data, not all social links represented active social relationships. They then recommended to build interaction graphs with a constraint on the minimum number of interaction events (e.g., respond, like) within a stipulated window of time.

The graph models in the existing works, e.g., [30, 53] cannot be applied to our problem for two reasons. First, modeling tweets as nodes and pairwise interactions between tweets as edges is not scalable, especially when dealing with large-scale datasets, e.g., our dataset contains billions of tweets. Second, tweets themselves do not have identities while nodes in a graph require such information.

3 Proposed approach

As presented in the introductory section, interactions in social media data could be an important implication of behaviors of communities and thus may play a role in predicting health status of populations. In this section, we first describe how to model interactions in social media data using graph theory (Section 3.1). We then introduce graph-based representations of social media data at population scale (Section 3.2). Fig. 1 illustrates the flowchart of our approach.

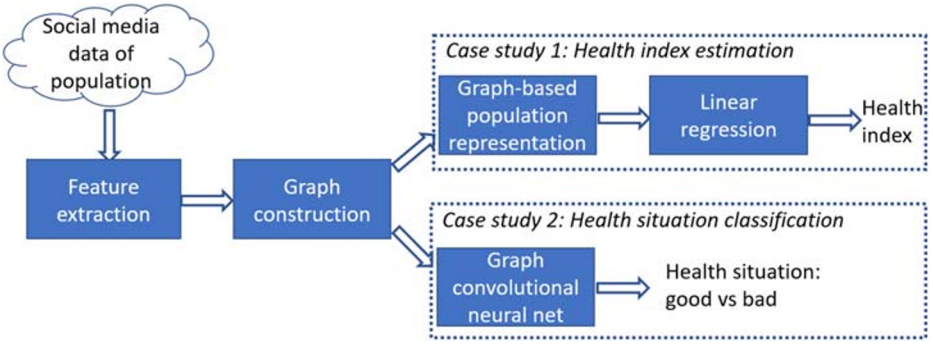


Fig. 1 Flowchart of our approach

3.1 Graph construction

We define interactions in social media data in two ways: via coincidences of features and responses/likes between groups of tweets. We then model these interactions in so-called inter-feature and inter-tweet graphs.

Specifically, let $T^P = \left\{t_1^P, ..., t_N^P\right\}$ denote a set consisting of N tweets collected from a population P. Suppose that each tweet $t_i^P \in T^P$ can be described by a feature vector $\mathbf{f}_i^P = \left[f_{i,1}^P, ..., f_{i,d}^P\right] \in \mathbb{R}^d$, e.g., $d = 78$ psycho-linguistic features in LIWC.

3.1.1 Inter-feature graph

Having the low-level feature vectors $\left\{\mathbf{f}_1^P, ..., \mathbf{f}_N^P\right\}$, we define the interaction $I_{j,k}^P$, where $j, k \in \{1, ..., d\}$ between two arbitrary features j and k using the radial basis function (RBF) as,

$$I_{j,k}^P = \exp\left[-\left(\frac{1}{N}\sum_{i=1}^{N} f_{i,j}^P - \frac{1}{N}\sum_{i=1}^{N} f_{i,k}^P\right)^2 / 2\sigma^2\right] \tag{1}$$

where σ is a free parameter which controls the width of the RBF and is used to normalize feature distances into probabilistic metrics. In our implementation, σ is set to 0.1. We empirically found inter-feature graphs achieved the similar yet best overall performance for $\sigma \in [0.1, 1.0]$. Note that our features were also normalized into [0,1].

As shown in (1), the representative value for each feature j is accumulated over all the tweets in P. Therefore, by using the RBF, $I_{j,k}^P$ capture the coincidence of features j and k, and thus represent the inter-feature relationships within the population P. The larger $I_{j,k}^P$ is, the more correlated feature j to feature k is.

We represent the interactions between features in P via a graph $G^P\left(V^P, E^P\right)$ where $V^P = \left\{v_1^P, ..., v_d^P\right\}$ is the set of vertices, each vertex corresponds to a feature and E^P is the set of undirected edges defined as,

$$E^P = \left\{\left(v_j^P, v_k^P\right) \in V^P \times V^P | I_{j,k}^P > \theta\right\} \tag{2}$$

where θ is a user-defined threshold. In our experiment, θ was set so that edges with the top 20% of $I_{j,k}^P$ were maintained in the graph, i.e., only top 20% of highly correlated features were considered. We found graphs whose number of edges take 20-30% of the total number of connections performed best on our dataset.

3.1.2 Inter-tweet graph

A straightforward approach to model pairwise interactions between tweets is to consider each tweet as a node in a graph and interactions as edges. However, this approach is not scalable. In addition, tweets do not have identities to be nodes. To overcome these issues, we cluster a training tweet set using a K-means algorithm wherein each tweet is encoded by its feature vector and the dissimilarity between two feature vectors is measured by Euclidean distance. This step results in a set of K centroids that are then used in both training and testing. Given a tweet set T^P of a population P (T^P can be either a training or test set), we cluster T^P into K subsets $T_1^P, ..., T_K^P$, i.e., $T^P = \bigcup_{j=1}^K T_j^P$. The partition is done by

assigning each tweet in T^P to its nearest centroid from the K centroids. In our implementation, we empirically set $K = 50$. We observed that there were subtle changes in the performances of inter-tweet graphs while this setting achieved the best overall performance across all years in our dataset.

We define the interaction $S^P_{j,k}$ between two subsets T^P_j and T^P_k as the proportion of the interactions between tweets in these two subsets. Specifically,

$$S^P_{j,k} = \frac{\sum_{t^P_m \in T^P_j} \sum_{t^P_n \in T^P_k} r\left(t^P_m, t^P_n\right)}{|T^P_j||T^P_k|} \tag{3}$$

where $r\left(t^P_m, t^P_n\right) = 1$ if t^P_m is a response/like to t^P_n or vice versa, and $r\left(t^P_m, t^P_n\right) = 0$, otherwise; $|T^P_j|$ and $|T^P_k|$ is the cardinality of T^P_j and T^P_k respectively.

We then construct a graph $G^P\left(V^P, E^P\right)$ in which each node $v^P_j \in V^P$ corresponds to a subset T^P_j. Like inter-feature graphs, two nodes v^P_j and v^P_k are connected (by an undirected edge) if their interaction $S^P_{j,k} > \theta$. We note that the subsets T^P_j are deterministic in their feature space and thus they imply identities.

3.2 Graph-based population representation

Given the population P, a graph $G^P\left(V^P, E^P\right)$ is constructed from either features or tweets as above. The graph-based representation of P is denoted as $\mathbf{h}^P$. In the following subsections, we present different ways to define $\mathbf{h}^P$.

3.2.1 Graph properties

By using graph properties, $\mathbf{h}^P$ can be represented as a vector of V^P dimensions, i.e., $\mathbf{h}^P = \left[h^P_1, ..., h^P_{|V^P|}\right]$ where h^P_j are computed from properties of vertices v^P_j. In this work, we investigate commonly used graph properties including Closeness Centrality, Betweenness Centrality, and PageRank [39].

Closeness Centrality (CC) The closeness centrality of a node $v^P_j \in V^P$ is defined as the reciprocal of the sum of the shortest path distances from v^P_j to all other nodes [21],

$$h^P_j = CC\left(v^P_j\right) = \frac{|V^P| - 1}{\sum_{v^P_k \in V^P - \left\{v^P_j\right\}} l\left(v^P_j, v^P_k\right)} \tag{4}$$

where $l\left(v^P_j, v^P_k\right)$ is the length of the shortest-path from node v^P_j to node v^P_k.

Intuitively, h^P_j represents the proximity of v^P_j to other nodes in the graph G^P. For instance, if the graph G^P is built based on tweets, $l\left(v^P_j, v^P_k\right)$ represents how often tweets in cluster j interact with tweets in cluster k. If G^P is constructed from features, $l\left(v^P_j, v^P_k\right)$ is calculated from the similarity between feature j and feature k. The shorter $l\left(v^P_j, v^P_k\right)$ is, the more direct v^P_j can be linked to v^P_k, e.g., more interactions exist between cluster j and cluster k. In other words, h^P_j captures the centrality (or sparsity) of the graph G^P.

Betweenness Centrality (BC) The betweenness centrality of a node $v_j^P \in V^P$ is the sum of the fraction of all-pairs shortest paths that pass through v_j^P [10],

$$h_j^P = BC\left(v_j^P\right) = \sum_{v_k^P, v_l^P \in V^P} \frac{\beta\left(v_k^P, v_l^P | v_j^P\right)}{\beta\left(v_k^P, v_l^P\right)} \tag{5}$$

where $\beta\left(v_k^P, v_l^P\right)$ is the number of shortest paths from v_k^P to v_l^P and $\beta\left(v_k^P, v_l^P | v_j^P\right)$ is the number of those paths passing through v_j^P other than v_k^P and v_l^P. If $v_k^P = v_l^P$, $\beta\left(v_k^P, v_l^P\right) = \beta\left(v_k^P, v_k^P\right) = 1$, and if $v_j^P \in \left\{v_k^P, v_l^P\right\}$, $\beta\left(v_k^P, v_l^P | v_j^P\right) = 0$.

As shown in (5), in contrast to closeness centrality, betweeness centrality takes into account indirect connections.

PageRank (PR) PageRank [40] was developed for measuring the importance of websites on the Internet. This method makes use of an underlying assumption that more important websites are likely to receive more links from others. In our case, we define the PageRank property of a node $v_j^P \in V^P$ as:

$$h_j^P = PR\left(v_j^P\right) = \sum_{v_k^P \in \mathcal{N}\left(v_j^P\right)} \frac{PR\left(v_k^P\right)}{L\left(v_k^P\right)} \tag{6}$$

where $\mathcal{N}(v_j^P)$ is the set of all nodes linking to node v_j^P and $L(v_k^P)$ is the number of links from v_k^P.

3.2.2 Graph kernels

Graph-based representations $\mathbf{h}^P$ defined in Section 3.2.1 are created from the graph G^P of the population P. Alternatively, one may consider the similarity among different graphs in a training dataset in creating graph-based representations. Graph kernels have been proven an effective tool to calculate the similarity among graph structures [9]. The core idea of graph kernels is to decompose a graph into sub-graphs, then applies a kernel to measure the similarity between these sub-graphs.

In general, let $\mathbf{P} = \{P_j\}$ be the set of populations in a training dataset and $\mathbf{G} = \left\{G^{P_j}\right\}$ be the set of graphs constructed using the methods presented in Section 3.1. The graph-based representation $\mathbf{h}^P$ of a population P is a vector of $|\mathbf{P}|$ dimensions, $h^P = \left[h_1^P, ..., h_{|\mathbf{P}|}^P\right]$ where

$$h_j^P = \mathcal{K}\left(G^P, G^{P_j}\right) = \langle \phi\left(G^P\right), \phi\left(G^{P_j}\right) \rangle \tag{7}$$

where $G^{P_j} \in \mathbf{G}$ and ϕ is a function that maps a graph G into the Hilbert space $\mathcal{H}$ that supports the structure of inner products $\langle ., . \rangle$.

The advantage of using kernel methods is that the mapping function ϕ is not necessary to be determined explicitly. This is because the kernel function $\mathcal{K}$ can be conveniently computed using inner products. Intuitively, $\mathcal{K}$ measures the similarity between G^P and G^{P_j}. Note that if $G^P = G^{P_j}$, $\mathcal{K}\left(G^P, G^{P_j}\right) = 1$ (i.e., G^P and G^{P_j} are isomorphic).

Different kernels make use of different decomposition techniques and similarity measures [9]. For instance, shortest path kernel [8], computing the shortest path lengths between all pairs of nodes in two graphs G and G' is defined as follows,

$$\mathcal{K}\left(G, G'\right) = \Sigma_{v_i, v_j \in G} \Sigma_{v'_m, v'_n \in G'} \kappa \left(l\left(v_i, v_j\right), l\left(v'_m, v'_n\right)\right) \tag{8}$$

where $l\left(v_i, v_j\right)$ is the length of the shortest path between node v_i and v_j, and κ is calculated as,

$$\kappa \left(l\left(v_i, v_j\right), l\left(v'_m, v'_n\right)\right) = \begin{cases} 1 & \text{if } l\left(v_i, v_j\right) = l\left(v'_m, v'_n\right) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We note that computing the shortest paths between all pairs of nodes in a graph of n nodes can be done efficiently in $O(n^3)$ using the Floyd-Warshall algorithm. In this work, we investigate two kernels: shortest path (SP) [8] and Weisfeiler-Lehman (WL) subtree [52].

4 Experiments

4.1 Case studies

We applied our approach in two case studies: 1) estimating health indices, and 2) classifying health situation of counties in the US.

4.1.1 Case study 1. Population health index estimation

We conducted an across-county prediction task to estimate health indices. Technically, this is a regression task where, given a county P, input is a feature vector extracted from that county and output is a population health index. In this case study, three primary health indices in BRFSS: "*generic health*", "*physical health*", and "*mental health*" were estimated.

We employed a linear regression model for the estimation task. Specifically, the health index y^P of a population P can be estimated as follows,

$$y^P = \mathbf{w}^\top \mathbf{h}^P + e \tag{10}$$

where $\mathbf{h}^P$ is defined in Section 3.2, $e \sim N\left(0, \epsilon^2\right)$ is a Gaussian error term, and $\mathbf{w}$ is the weight vector that can be learned directly from training data.

4.1.2 Case study 2. Population health situation classification

This case study aims to classify the health status of a given population into two classes: *good* or *bad* [54]. Like case study 1, input of each county is the feature vector extracted from that county and output is a health status (good vs bad).

We adopted the deep graph convolutional neural network (DGCNN) proposed in [57] for classifying population health status. DGCNN allows end-to-end learning on original graphs without preprocessing while demonstrating state-of-the-art performance on many tasks. To apply DGCNN on our graphs, we used a single network structure consisting of four graph convolutional layers, two 1-D convolutional layers followed by a dense layer, and a softmax layer as output. Activation functions include the hyperbolic tangent function (tanh) in graph convolutional layers and rectified linear units (ReLU) in other layers. Stochastic gradient descent with the Adam updating rule [29] was employed in training the network.

4.2 Dataset

We crawled 1,129,928,183 tweets in years 2014, 2015, and 2016, with associated US geo-codes. We also collected 152,853,038 tweets made in 2013 for learning latent topics.

The collected tweets were associated with the US counties by mapping their geo-codes to the Federal Information Processing System (FIPS) codes using the cartographic boundary files provided by the US Census Bureau in 2013. There were 3,221 different geo-codes (i.e., 3,221 counties) in the US. Note that we used only tweets with associated latitude/longitude coordinates, those with self-reported location information but without coordinates were not considered in our study.

We used BRFSS survey reports as the ground truth. The surveys were conducted by the CDC via telephone interviews of the US residents regarding to their health-related risk behaviors, chronic health conditions, and health outcomes. BRFSS contains more than 400,000 interviews conducted each year and is currently the largest health survey system, not only in the US but also in the world. The questionnaires in BRFSS surveys are categorized into core sections including current health status, number of healthy days, inadequate sleep, chronic health conditions, and optional modules such as healthcare access or social context.

For estimating health indices (case study 1), we used the annual health ranking data of counties in BRFSS surveys including i) poor or fair health - percent of adults that report fair or poor health, ii) poor physical health days - the average number of reported physically unhealthy days per month, and iii) poor mental health days - the average number of reported mentally unhealthy days per month. The ranges of the health indices in the ground-truth are as follows: [4, 51] for poor or fair health, [1, 10] for poor physical health, and [1, 10] for poor mental health.

For classifying health situation (case study 2), for each health index, top 500 counties with highest scores were assigned to "good", and top 500 counties with lowest scores were assigned to "bad".[3]

4.3 Computational resources

To process large-scale data in data aggregation, county mapping, feature extraction (from billions of tweets), and graph construction, we employed Spark on top of Hadoop [56]. Spark is a computing platform which enables distributed and parallel computations on a cluster scaled up to 8,000 nodes. Furthermore, Spark is an in-memory based system which is convenient to keep data in memory for subsequent processing, thus allows much faster computations than disk-based systems like Hadoop MapReduce [17]. Specifically, Spark Hadoop cluster comprises 8 CentOS 7.2 physical machines, each of which is equipped with Intel Xeon E5-26700 (8 cores, 16 threads) CPU, 128 GB RAM, Intel Xeon Phi Coprocessor (60 cores), and 24TB HDD.

5 Results

There are several technical contributions proposed in the paper, including two graph construction methods (Section 3.1), different graph-based representations (Section 3.2), and

[3] https://www.usnews.com/news/healthiest-communities/rankings

two textual feature types: LIWC and latent topics. Different combinations of these proposals result in different performances. In the following subsections, we investigate such combinations in each case study and compare our proposed approach with existing works.

5.1 Case study 1. Population health index estimation

We used 70% of the counties (2,255 counties) in every year for training the regression model in (10) and the remainder (966 counties) for testing it. To measure the performance of health index estimation, we used Spearman's rank correlation coefficient (or Spearman's rho [34]): $rho = 1 - \frac{6\Sigma_{i=1}^{n} d_i^2}{n(n^2-1)}$ where n is the number of counties and d_i is the difference between the estimated and actual health index of the i-th county.

We first evaluated inter-feature graphs in population heath index estimation. Specifically, we constructed inter-feature graphs using LIWC and latent topics respectively. For LIWC, all 78 features were used. For latent topics, we varied the number of latent topics from 10 to 100. We observed that the performance of health index estimation slightly changed when the number of latent topics was within 50-80 and reached the highest performance when the number of topics was 80. The performance then dropped when the number of latent topics was outside this range. Therefore, 80 topics were used in all following experiments.

Table 1 shows the performances in health index estimation of inter-feature graphs built on LIWC and latent topics respectively. For each feature type (LIWC/Topics), we evaluate all graph properties presented in Section 3.2. Experimental results show that, among graph properties, Betweeness Centrality (BC) performed best on both LIWC and latent topics. In addition, WL subtree kernel outperformed SP kernel. Therefore, to make Table 1 easy to follow, only the results of BC property and WL subtree kernel (i.e., the best representations of graph properties and graph kernels) are included. As shown in Table 1, BC slightly but consistently outperforms WL subtree kernel in all cases. In addition, BC achieves the highest performance across all health indices, years, and on both LIWC and latent topics.

We then combined graph-based representations with features used to build the graphs, e.g., BC property + LIWC, and found that these combinations made no improvements on the use of LIWC. In contrast, the combination of BC property and latent topics gained significant advance and also achieved the best overall performance across all years and on all health indices.

We compared our graph-based approach with other non-graph-based ones. Culotta in [15] established a seminal in the field of population health analysis through social media. In [15], tweets within the same county were gathered into an aggregated tweet on which features (LIWC and latent topics) were extracted. This approach is referred to as non-graph approach and has been widely adopted in following studies such as [24, 37, 50]. We re-implemented the non-graph approach and evaluated it on our collected data. Comparison results, presented in Table 1, show that graph-based methods, including graph properties and graph kernels, significantly outperform the non-graph ones. In particular, compared with the non-graph approach, the graph-based representations built on inter-feature graphs of latent topics and BC property improved the health index estimation performance on all health indices and in all years, and the improvement was up to 14% in estimation of mental health in year 2014.

We evaluated the performances in health index estimation of inter-tweet graphs on LIWC and latent topics and reported results in Table 2. In general, inter-tweet graphs show lower performances in comparison with inter-feature graphs on all health indices, years, and on both LIWC and latent topics. Unlike inter-feature graphs, BC property and WL subtree

Table 1 Case study 1: health index estimation performance (Spearman's *rho*) of *inter-feature* graphs vs existing work

Features	Methods	2014			2015			2016		
		Generic	Physical	Mental	Generic	Physical	Mental	Generic	Physical	Mental
LIWC	Non-graph [15]	0.67	0.59	0.55	0.58	0.52	0.52	0.57	0.50	0.49
	WL subtree [52]	0.69	0.60	0.55	0.62	0.56	0.52	0.57	0.50	0.49
	BC property [10]	0.69	0.62	0.58	0.62	0.57	0.56	0.58	0.52	0.51
	WL subtree [52] + LIWC	0.69	0.60	0.55	0.62	0.56	0.52	0.57	0.50	0.49
	BC property [10] + LIWC	0.69	0.62	0.58	0.62	0.57	0.56	0.58	0.52	0.51
Topics	Non-graph [15]	0.69	0.65	0.68	0.62	0.59	0.62	0.57	0.55	0.59
Latent	WL subtree [52]	0.33	0.27	0.27	0.28	0.25	0.22	0.24	0.18	0.18
	BC property [10]	0.64	0.53	0.51	0.61	0.53	0.48	0.55	0.45	0.43
	WL subtree [52] + Latent topics	0.67	0.61	0.64	0.62	0.59	0.62	0.54	0.49	0.52
	BC property [10] + Latent topics	**0.70**	**0.66**	**0.69**	**0.67**	**0.65**	**0.65**	**0.63**	**0.58**	**0.61**

For graph-based methods, only BC property's results and WL subtree kernel's results are presented as BC property and WL subtree kernel respectively perform best among other graph properties and kernels. The index ranges are as follows: [4, 51] for generic health, [1, 10] for physical health, and [1, 10] for mental health. In each year and on each health index, best performances are highlighted

kernel with inter-tweet graphs obtained comparable performances. For instance, as shown in Table 2, WL subtree kernel is more dominant than BC property in year 2014 but less favor in years 2015 and 2016. Compared with the non-graph approach, both WL kernel and BC property defined on latent topics showed superior performance. Like inter-feature graphs,

Table 2 Case study 1: health index estimation performance (Spearman's *rho*) of *inter-tweet* graphs vs existing work

Features	Methods	2014			2015			2016		
		Generic	Physical	Mental	Generic	Physical	Mental	Generic	Physical	Mental
LIWC	Non-graph [15]	0.46	0.37	0.31	0.30	0.22	0.13	0.33	0.15	0.11
	WL subtree [52]	0.46	0.37	0.33	0.33	0.24	0.17	0.34	0.17	0.15
	BC property [10]	0.46	0.36	0.33	0.34	0.25	0.17	0.34	0.20	0.17
Latent topics	Non-graph [15]	0.49	0.36	0.31	0.36	0.26	0.22	0.35	0.26	0.23
	WL subtree [52]	**0.54**	**0.43**	**0.38**	**0.37**	**0.27**	0.23	**0.38**	0.26	0.27
	BC property [10]	0.53	**0.43**	0.37	**0.37**	0.26	**0.24**	0.37	**0.27**	**0.29**

For graph-based methods, only BC property's results and WL subtree kernel's results are presented as BC property and WL subtree kernel respectively perform best among other graph properties and kernels. The index ranges are as follows: [4, 51] for generic health, [1, 10] for physical health, and [1, 10] for mental health. In each year and on each health index, best performances are highlighted

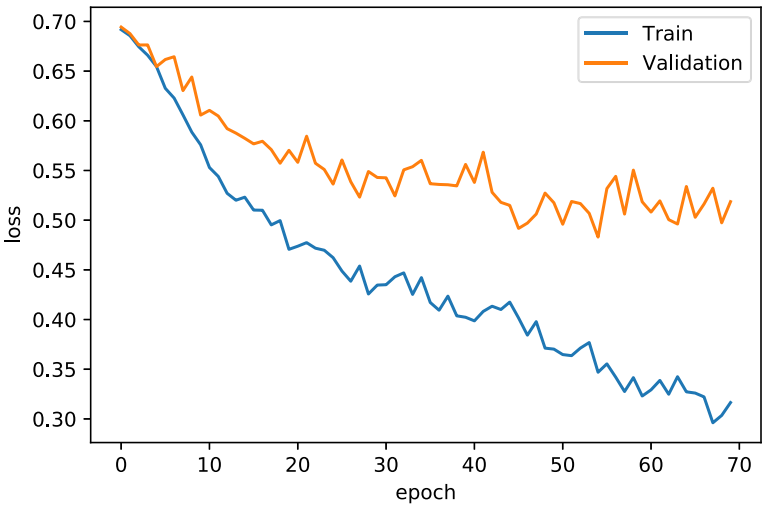


Fig. 2 Learning curves of our graph convolution model on the training and validation set. We used the model trained at 50 epochs in evaluations and comparisons

the graph-based representations built on latent topics and BC property improved 18% in estimation of mental health in year 2016, in comparison with the non-graph approach.

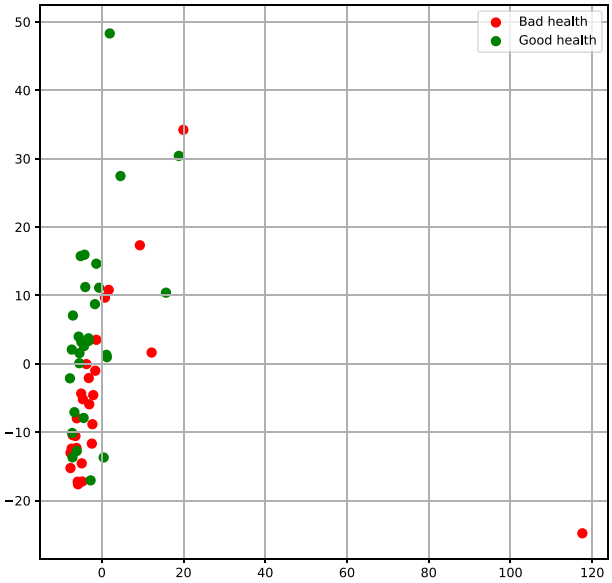
5.2 Case study 2: population health situation classification

As shown in our experiments, inter-feature graphs significantly outperformed inter-tweet graphs on both LIWC and latent topics. Therefore, in this case study, we focused on

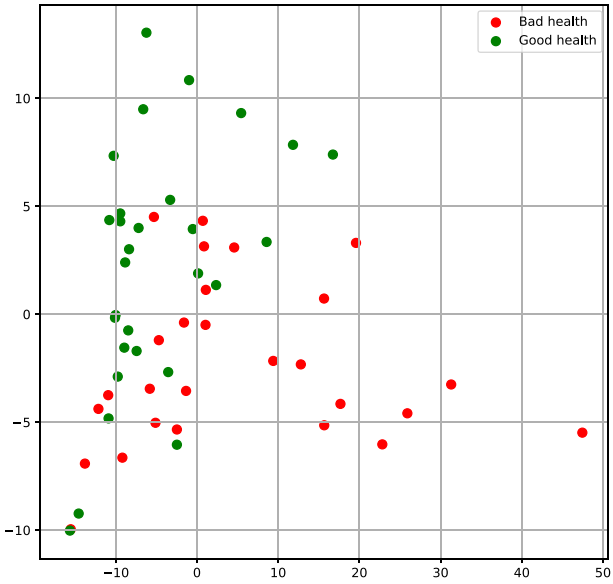
Table 3 Case study 2: classification performance of *inter-feature* graphs on three health indices: generic, physical, and mental

Year	Features	Generic					Physical					Mental				
		Acc	Sen	Spe	Mean AUC	Std AUC	Acc	Sen	Spe	Mean AUC	Std AUC	Acc	Sen	Spe	Mean AUC	Std AUC
2014	LIWC	0.87	0.74	1.00	0.92	0.02	0.83	0.67	1.00	**0.91**	0.04	0.80	0.61	1.00	**0.91**	0.05
	Latent topics	0.87	0.75	1.00	**0.94**	0.03	0.82	0.65	0.99	0.90	0.05	0.82	0.65	1.00	0.88	0.06
2015	LIWC	0.85	0.71	1.00	0.91	0.04	0.82	0.64	1.00	**0.90**	0.04	0.80	0.61	1.00	**0.88**	0.05
	Latent topics	0.84	0.69	1.00	**0.93**	0.04	0.81	0.62	1.00	0.89	0.05	0.77	0.54	1.00	**0.88**	0.06
2016	LIWC	0.84	0.68	1.00	**0.91**	0.05	0.78	0.57	1.00	**0.88**	0.06	0.77	0.55	1.00	**0.89**	0.06
	Latent topics	0.80	0.60	1.00	0.88	0.04	0.79	0.59	1.00	0.87	0.06	0.77	0.54	1.00	0.86	0.05

We report the classification performance of our inter-feature graphs built upon LIWC and latent topics in years 2014, 2015, and 2016. In each year and on each health index, best mean AUC performances are highlighted



(a)



(b)

Fig. 3 Visualization of learning features in graph convolution. Each data point corresponds to a county whose the graph is built on LIWC: **a** input graph features, **b** features learned after 50 epochs. Features are created by concatenating node features in graphs and presented in 2D using Principal Component Analysis. The most two prominent components are selected for this visualization. As shown, compared with input features, learned features are better separated

inter-feature graphs only. Similarly to case study 1, we also experimented inter-feature graphs on both LIWC and latent topics. For evaluation, we performed 10-fold validation. For each health index, based on the ground truth status scores, the top/bottom 500 counties were considered as true positives (i.e., *good*) and true negatives (i.e., *bad*). We show the learning curve of our graph convolution model in Fig. 2.

Since this problem is a binary classification problem, we measured the classification performance using,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specificity = \frac{TN}{TN + FP}$$

where TP/TN is the number of cases correctly classified as *good*/*bad* and FP/FN is the number of cases incorrectly classified as *good*/*bad*.

In addition, for each health index, we generated Receiver Operating Characteristic (ROC) curves, representing the trade-off between the sensitivity and specificity, in different training/test splits of the 10-fold setting. We then calculated the mean and standard deviation of Area Under the Curve (AUC) of the ROC curves across all training/test splits.

We report the performances of inter-feature graphs built on LIWC and latent topics in case study 2 in Table 3. Experimental results show that, compared with physical and mental health, generic health was always classified at the highest accuracy in all years and on both LIWC and latent topics. The classification accuracy of generic health reached its highest performance in year 2014 at 94% of AUC on latent topics. Physical health took the second place and got its highest position at 91% of AUC on LIWC in 2014. Unlike case study 1, both LIWC and latent topics performed similarly in most cases. Table 3 also shows that the classification was performed consistently across all experimental settings, e.g., the standard deviation of AUC, denoted as "std AUC", $\leq 6\%$. We further validate the potential of graph convolution method by illustrating the distributions of features learned by the method in Fig. 3.

Like case study 1, we compared our graph-based approach (i.e., combination of inter-feature graphs and DGCNN) with non-graph ones. For the non-graph methods, e.g., [15], we created features for a population by aggregating features from individual tweets of that

Table 4 Case study 2: comparison of our method with existing ones using mean AUC in three years 2014, 2015, and 2016

Features	Methods	2014			2015			2016		
		Generic	Physical	Mental	Generic	Physical	Mental	Generic	Physical	Mental
LIWC	Non-graph [15]	0.80	0.76	0.75	0.74	0.71	0.70	0.63	0.65	0.61
	Ours	**0.92**	**0.91**	**0.91**	**0.91**	**0.90**	**0.88**	**0.91**	**0.88**	**0.89**
Latent topics	Non-graph [15]	0.85	0.80	0.81	0.82	**0.82**	0.76	0.75	0.76	0.76
	Ours	**0.94**	**0.90**	**0.88**	**0.93**	**0.89**	**0.88**	**0.88**	**0.87**	**0.86**

We report the performance of our method and the work in [15] on both LIWC and latent topics. In each year and on each health index, best performances are highlighted

population. These features were then fed to a logistic classifier for classifying the population health status. The same evaluation protocol (i.e., 10-fold validation with the same training/test split per fold) was applied. We present the comparison results in Table 4. As shown in our results, our graph-based approach significantly outperforms the non-graph ones (up to 18% on LIWC and 13% on latent topics).

6 Conclusion

This paper proposes a novel approach for population health analysis through social media. In our approach, interactions in social media data are modeled in graphs and defined via the coincidences of features and responses/likes between groups of tweets. We investigated various graph-based representations. We applied the proposed approach in two tasks: health index estimation and health situation classification of counties in the US, and conducted extensive experiments on a large-scale dataset benchmarked by the Behavioral Risk Factor Surveillance System of the US. Experimental results verified the importance of interactions in social media for health analysis at population scale. Specifically, our approach achieved state-of-the-art performance on both the tasks while inter-feature graphs built on latent topics performed best on health index estimation. Studying the interactions between different feature types and combination of various social media data types, e.g., text, images, etc., will be our future work.

References

1. Aiello AE, Renson A, Zivich PN (2020) Social media–and internet-based disease surveillance for public health. Annu Rev Public Health 41:101–118
2. Akbari M, Relia K, Elghafari A, Chunara R (2018) From the user to the medium: neural profiling across web communities. In: Proceedings of the international conference on web and social media, pp 552–555
3. Allington D, Duffy B, Wessely S, Dhavan N, Rubin J (2020) Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency. Psychol Med, pp 1–7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7298098/
4. Bai H, Chunara R, Varshney LR (2015) Social capital deserts: obesity surveillance using a location-based social network. In: Proceedings of the data for good exchange, pp 1–7
5. Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
6. Bekalu MA, McCloud RF, Viswanath K (2019) Association of social media use with social well-being, positive mental health, and self-rated health: disentangling routine use from emotional connection to use. Health Education & Behavior 46(2_suppl):69S–80S
7. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
8. Borgwardt KM, Kriegel H-P (2005) Shortest-path kernels on graphs. In: Proceedings of the IEEE international conference on data mining, pp 74–81
9. Borgwardt KM, Kriegel H-P, Vishwanathan S, Schraudolph NN (2007) Graph kernels for disease outcome prediction from protein-protein interaction networks. In: Proceedings of the pacific symposium on biocomputing, pp 4–15
10. Brandes U (2008) On variants of shortest-path betweenness centrality and their generic computation. Soc Networks 30(2):136–145
11. Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, Buckeridge DL (2010) Information technology and global surveillance of cases of 2009 H1N1 influenza. N Engl J Med 362(18):1731–1735
12. Chun H, Kwak H, Eom Y-H, Ahn Y-Y, Moon S, Jeong H (2008) Comparison of online social relations in volume vs interaction: a case study of Cyworld. In: Proceedings of the ACM SIGCOMM conference on internet measurement, pp 57–70

13. Conway M, Hu M, Chapman WW (2019) Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearbook of Medical Informatics 28(1):208–217
14. Culotta A (2010) Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the workshop on social media analytics, pp 115–122
15. Culotta A (2014) Estimating county health statistics with Twitter. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1335–1344
16. De Choudhury M, Counts S, Horvitz E (2013) Social media as a measurement tool of depression in populations. In: Proceedings of the annual ACM web science conference, pp 47–56
17. Dittrich J, Quiané-Ruiz J-A (2012) Efficient big data processing in Hadoop MapReduce. Proceedings of the VLDB Endowment 5(12):2014–2015
18. Dredze M (2012) How social media will change public health. IEEE Intell Syst 27(4):81–84
19. Farhadloo M, Winneg K, Chan M-PS, Jamieson KH, Albarracin D (2018) Associations of topics of discussion on Twitter with survey measures of attitudes, knowledge, and behaviors related to zika: probabilistic study in the United States. JMIR Public Health and Surveillance 4(1):e16
20. França U, Sayama H, McSwiggen C, Daneshvar R, Bar-Yam Y (2016) Visualizing the 'heartbeat' of a city with tweets. Complexity 21(6):280–287
21. Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Networks 1(3):215–239
22. Gibbons J, Malouf R, Spitzberg B, Martinez L, Appleyard B, Thompson C, Nara A, Tsou M-H (2019) Twitter-based measures of neighborhood sentiment as predictors of residential population health. PLoS ONE 14 (7):e0219550
23. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
24. Giorgi S, Preoţiuc-Pietro D, Buffone A, Rieman D, Ungar L, Schwartz HA (2018) The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In: Proceedings of the conference on empirical methods in natural language processing, pp 1167–1172
25. Gittelman S, Lange V, Crawford CAG, Okoro CA, Lieb E, Dhingra SS, Trimarchi E (2015) A new source of data for public health surveillance: Facebook likes. Journal of Medical Internet Research 17(4):e98
26. House JS, Landis KR, Umberson D (1988) Social relationships and health. Science 241(4865): 540–545
27. Jang K, Baek YM (2019) When information from public health officials is untrustworthy: the use of online news, interpersonal networks, and social media during the MERS outbreak in South Korea. Health Commun 34(9):991–998
28. Kershaw D, Rowe M, Stacey P (2014) Towards tracking and analysing regional alcohol consumption patterns in the UK through the use of social media. In: Proceedings of the ACM conference on web science, pp 220–228
29. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. arXiv:1412.6980
30. Leskovec J, Horvitz E (2008) Planetary-scale views on a large instant-messaging network. In: Proceedings of the international world wide web conference, pp 915–924
31. Merchant RM (2020) Evaluating the potential role of social media in preventive health care. JAMA 323(5):411–412
32. Merchant RM, Lurie N (2020) Social media and emergency preparedness in response to novel coronavirus. JAMA 323(20):2011–2012
33. Müller MM, Salathé M (2019) Crowdbreaks: tracking health trends using public social media data and crowdsourcing. Frontiers in Public Health 7(81):1–6
34. Myers JL, Well AD, Lorch RF Jr (2010). In: 3 (ed) Research design and statistical analysis. Routledge, London
35. Nguyen T, Larsen M, O'Dea B, Nguyen H, Nguyen DT, Yearwood J, Phung D, Venkatesh S, Christensen H (2020) Using spatiotemporal distribution of geocoded Twitter data to predict US county-level health indices. Futur Gener Comput Syst 110:620–628
36. Nguyen T, Nguyen DT, Larsen ME, O'Dea B, Yearwood J, Phung D, Venkatesh S, Christensen H (2017) Prediction of population health indices from social media using kernel-based textual and temporal features. In: Proceedings of the international conference on world wide web companion, pp 99–107
37. Nguyen H, Nguyen T, Nguyen DT (2019) An empirical study on prediction of population health through social media. J Biomed Inform 99(103277):1–9
38. Nguyen H, Nguyen T, Nguyen DT (2019) Estimating county health indices using graph neural networks. In: Australasian conference on data mining, pp 64–76
39. Opsahl T, Agneessens F, Skvoretz J (2010) Node centrality in weighted networks: generalizing degree and shortest paths. Soc Networks 32(3):245–251

40. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Tech. rep., Stanford InfoLab
41. Pagoto S, Waring ME, Xu R (2019) A call for a public health agenda for social media research. Journal of Medical Internet Research 21(12):e16661
42. Parker J, Wei Y, Yates A, Frieder O, Goharian N (2013) A framework for detecting public health trends with Twitter. In: Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining, pp 556–563
43. Parrish RG (2010) Peer reviewed: measuring population health outcomes. Preventing Chronic Disease 7(4):1–11
44. Paul MJ, Dredze M (2011) You are what you tweet: analysing Twitter for public health. In: Proceedings of the international AAAI conference on weblogs and social media, pp 265–272
45. Pennebaker JW, Booth RJ, Boyd RL, Francis ME (2015) Linguistic inquiry and word count: LIWC 2015 [Computer software], Pennebaker Conglomerates, Inc
46. Pennebaker JW, Mehl MR, Niederhoffer KG (2003) Psychological aspects of natural language use: our words, our selves. Annu Rev Psychol 54(1):547–577
47. Quercia D, Capra L, Crowcroft J (2012) The social world of Twitter: topics, geography, and emotions. In: Proceedings of the international AAAI conference on weblogs and social media, pp 298–305
48. Rozenblum R, Bates DW (2013) Patient-centred healthcare, social media and the internet: the perfect storm? BMJ Quality & Safety 22(3):183–186
49. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS (2013) Influenza a (H7N9) and the importance of digital epidemiology. N Engl J Med 369(5):401–404
50. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Lucas RE, Agrawal M, Park GJ, Lakshmikanth SK, Jha S, Seligman ME, Ungar L (2013) Characterizing geographic variation in well-being using tweets. In: Proceedings of the international AAAI conference on weblogs and social media, pp 583–591
51. Sedgwick R, Epstein S, Dutta R, Ougrin D (2019) Social media, internet use and suicide attempts in adolescents. Current Opinion in Psychiatry 32(6):534–541
52. Shervashidze N, Schweitzer P, Leeuwen EJV, Mehlhorn K, Borgwardt KM (2011) Weisfeiler-Lehman graph kernels. J Mach Learn Res 12:2539–2561
53. Wilson C, Boe B, Sala A, Puttaswamy KPN, Zhao BY (2009) User interactions in social networks and their implications. In: Proceedings of the EuroSys conference, pp 205–218
54. Wu S, Wang R, Zhao Y, Ma X, Wu M, Yan X, He J (2013) The relationship between self-rated health and objective health status: a population-based study. BMC Public Health 13(320):1–9
55. Ye M, Yin P, Lee W-C (2010) Location recommendation for location-based social networks. In: Proceedings of the SIGSPATIAL international conference on advances in geographic information systems, pp 458–461
56. Zaharia M, Chowdhury M, Das T, Dave A, Ma J, Mccauley M, Franklin M, Shenker S, Stoica I (2012) Fast and interactive analytics over Hadoop data with Spark. Usenix Login 37(4):45–51
57. Zhang M, Cui Z, Neumann M, Chen Y (2018) An end-to-end deep learning architecture for graph classification. In: Proceedings of the AAAI conference on artificial inteligence, pp 4438–4445