Applied Geography 63 (2015) 184–191

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Effect of climate and seasonality on depressed mood among twitter users

Wei Yang [a], Lan Mu [a,*], Ye Shen [b]

[a] Department of Geography, University of Georgia, United States
[b] Epidemiology and Biostatistics, College of Public Health, University of Georgia, United States

ARTICLE INFO

Article history:
Received 6 March 2015
Received in revised form
27 June 2015
Accepted 28 June 2015
Available online xxx

Keywords:
Climate
Depression
GIS
Seasonality
Social media
Twitter

ABSTRACT

Location-based social media provide an enormous stream of data about humans' life and behavior. With geospatial methods, those data can offer rich insights into public health. In this research, we study the effect of climate and seasonality on the prevalence of depression in Twitter users in the U.S. Text mining and geospatial methods are used to detect tweets related to depression and their spatiotemporal patterns at the scale of Metropolitan Statistical Area. We find the relationship between depression rates, climate risk factors and seasonality are varied and geographically localized. The same climate measure may have opposite association with depression rates at different places. Relative humidity, temperature, sea level pressure, precipitation, snowfall, weed speed, globe solar radiation, and length of day all contribute to the geographic variations of depression rates. A conceptual compact map is designed to visualize scattered geographic phenomena in a large area. We also propose a three-stage framework that semi-automatically detects and analyzes geographically distributed health issues using location-based social media data.

1. Introduction

In recent years, social media have received considerable attention as a new data source for health research. With location-based techniques and wireless devices, social media have provided an enormous stream of data about human's life and behavior. Those petabytes data have high spatio-temporal resolution, thus represent huge potential for geographical analysis. Combining GIS methods with social media analytics can offer rich insights to human's perceptions of space and its significance to public health.

Former research has shown that it is feasible and inexpensive to use social media for health study. Lee, Agrawal, and Choudhary (2013) examine influenza spread and compare cancer types at U.S. states level using Twitter. Ghosh and Guha (2013) compared spatial clusters of obesity-related tweets with the distribution of McDonald's restaurant. Park, Cha, and Cha (2012), De Choudhury, Counts, and Horvitz (2013), Yang and Mu (2015) highlighted the potential of using Twitter to detect depression and its geographical pattern. Harman (2014) pointed out that although Twitter users are not a representative sample of the entire population, individual and population level analysis can still be made because of the diverse set of quantifiable signals relevant to mental health observable in Twitter.

Whilst social media offer many opportunities in geographic analysis of health issues, they also poses a number of challenges. First is the need for new automated methods of handling and analyzing big data that are being generated at a very high speed (Batty et al., 2012; Kitchin, 2013). Geographic (usually 3D), temporal and unstructured textual attributes construct a five-dimension space. However, the originally designed computational underpinning of GIS cannot afford to integrate those information both at a very large volume and at a high speed. (Goodchild, 2013; Gorman, 2013). The second challenge is how to extract and distill useful information considering the volume and the depth of data. Researchers and general public are interested in the deeper complex patterns implied by the data (González-Bailón, 2013; Manovich, 2011; Ruppert, 2013). Third, big data often lack rigorous sampling, documentation, and quality assurance (Goodchild, 2013; Kitchin, 2013; Sui & Goodchild, 2011). This brings difficulty in confirming the validity and accuracy of crowdsourcing (Gorman, 2013). Fourth, although social media have better data

* Corresponding author. Department of Geography, University of Georgia, 210 Field St., GG Building, Rm 103, Athens, GA 30602, United States.
E-mail addresses: weiyang@uga.edu (W. Yang), mulan@uga.edu (L. Mu), yeshen@uga.edu (Y. Shen).

http://dx.doi.org/10.1016/j.apgeog.2015.06.017
0143-6228/

granularity that improves the level of details in observations, deciding how to find the right scale for analysis for both temporal and spatial resolution is still a primary issue (González-Bailón, 2013). In addition, new methods are needed to synthesize geo-visual analytics and social media analytics. On the one hand, time often interacts with space, and therefore maps need to be more dynamic. On the other hand, considering the discontinuity and broad coverage of study areas, maps need to be merged in a creative and meaningful way to enable humans to visualize the large amount of information intuitively (González-Bailón, 2013; Sui & Goodchild, 2011).

In this research, we tackle the above challenges. We use text mining and spatial analysis to explore the relationship between climates, seasonality and depression rate in the U.S. using Twitter data. Data are downloaded in a one-year time period with more than 600 million tweets. We then create a conceptual compact map to visualize those large amount of data. Additionally, we propose and demonstrate the feasibility of a three-stage framework which can semi-automatically detect health-related topics and their spatiotemporal patterns using location-based social media data.

2. Effect of climate and seasonality on depressed mood

A research conducted by American Psychological Association shows that climate has an impact on the Americans' psychological well-being (Clayton, Manning & Hodge, 2014). Climate impacts are related to stress-related problems or negative emotions, such as anger and depression (Cunsolo Willox et al., 2012; Neria & Shultz, 2012). These latest research direct us to an argumentative geography topic — Environmental Determinism, which has received much prominence in geographic history in the early 1900s and declined in the 1920s (Peet, 1985). The main argument of Environmental Determinism is that the physical environment determines the patterns of human culture and societal development. This theory has been developed and replaced as "Environmental Possibilism" by the 1950s as one central theory in geography (Human Geography, 2015). Environmental Possibilism holds the opinion that although the environment sets limitations for cultural development, it does not completely define culture. In this paper, we only discuss the possible relative relationship between physical environment and depressed feelings, not the causal relationship between them.

Depression is a common chronic disorder and has a high prevalence in the U.S. (MMWR, 2010). Studies on the relationship between climate and depression rates have yielded mixed results due to some major limitations using traditional data collecting methods.

Radua, Pertusa, and Cardoner (2010) found that high accumulated solar radiation, high temperature and low barometric pressure are related to high depression rate. The limitation is that data collected by questionnaires or from local hospitals are within a small-scale area, such as neighborhood or community.

Molin, Mellerup, Bolwig, Scheike, and Dam (1996) showed that increases in sun time, day length and temperature were associated with lower depression scores, while rainfall was not significantly associated. Their limitation is that they used a small data sample of only 126 patients. Research using self-report data usually suffer from low response rates, and may introduce selection bias.

Zung and Green (1974) found a significant correlation between number of depressed patients admitted to hospital and length of day. Lee, Tsai, and Lin (2007) found that admission rate for depression are positively correlated with temperature. Their limitation is that the time of hospital admission is usually monthly delayed than the onset of depression.

Furthermore, many people are unaware of the symptoms when they have depression. Even people realize they have mental health issues, few of them would go to see a clinical phycologist. Thus, only use data from hospitals may result in spurious associations between depression rate and its risk factors.

Seasonal depression is a form of recurrent depressive disorder, in which people who have normal mental health throughout most of the year experience depressive symptoms in winter or summer (Partonen & Lönnqvist, 1998). The relationship between seasonality and the prevalence of depressive problems has been explored (Huibers, de Graaf, Peeters, & Arntz, 2010; Magnusson, 2000; Nillni, Rohan, Rettew, & Achenbach, 2009). There is also evidence that seasonal mood variations are even recognized in healthy people (Okawa et al., 1996; Schlager, Schwartz, & Bromet, 1993).

Radua et al. (2010) and Winkler et al. (2002) indicated that the distribution of depression varies depending on the geographical location. Mersch, Middendorp, Bouhuys, Beersma, and van den Hoofdakker (1999) found a significant positive correlation between the prevalence of depression and latitude in North America.

In summary, based on the literature, we start with detecting the interaction between the rate of tweets expressing depressed feelings, climate, seasonality, and geographical locations by exploring Twitter data from textual, spatial and temporal aspects in the U.S.

3. Methods

3.1. Data acquisition

We downloaded Twitter data of an entire year from September 5th, 2013 to September 5th, 2014 using Twitter Streaming Application Program Interfaces (APIs). We used the entire U.S. as one geographical bounding-box to filter the tweets. Tweets acquired in this way are the complete set of public geo-tagged tweets and represent the unbiased geographical distribution of Twitter users' activities (Morstatter, Pfeffer, Liu, & Carley, 2013).

Due to the restriction in Twitter Streaming APIs, we are not able to set up a keyword filter of depression together with a geographical filter. Thus, the tweet collection we acquired contains both relevant and irrelevant tweets for studying depressed mood on Twitter. In the next step, we built a customized filter to select tweets relevant to depression.

3.2. Data reduction and text mining

The large volume of tweets resulting from data acquisition presented a significant challenge to extract useful semantic information from tweet texts for studying the geographical patterns of tweets related to depression. Popular text mining and topic modeling methods in the field of computer science and machine learning such as latent Dirichlet allocation are often not scalable to such large data sets with 600 million documents or tweets (Kuang & Park, 2013).

In our research, we accurately identified tweets related to depression in two steps. First, we significantly reduced the size of data set by selecting only the tweets with the keyword "depress" or its variations. This strategy served our purpose well because the feeling of depression is very different from that of other moods and commonly expressed by the word "depress" (De Choudhury et al., 2013; Kim, Li, Lebanon, & Essa, 2012; Park et al., 2012). Second, the selected subset contained a wide range of word contexts such as true depressed feelings, "Great Depression", "tropic depression", "pet depression", and more. We differentiated the word context associated with true depressed feelings by employing an advanced text clustering method called nonnegative matrix factorization (NMF) (Lee & Seung, 1999; Yang & Mu, 2015). The basic idea of matrix factorization is that each tweet can be represented as a high-