GIS ANALYSIS OF DEPRESSION USING LOCATION-BASED SOCIAL MEDIA DATA

by

WEI YANG

(Under the Direction of Lan Mu)

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 apply GIS methods and text mining techniques to social media data to provide new perspectives for public health research.

Depression is a common chronic disorder and has a high prevalence in the U.S. It often goes undetected due to limited diagnosis methods and bring serious results to public and personal health. Former research detected geographic pattern for depression using questionnaires or self-reported measures of mental health, this may induce same-source bias. Recent studies use social media for depression detection but none of them examines the geographic patterns.

We design a procedure to automatically detect depressed users in Twitter and analyze their spatial patterns using GIS technology. We use socioeconomic risk factors to explain the pattern at county level and find that race, education and income have an impact on depression rate. Our method can improve diagnosis techniques for depression, and it can be expanded to detect other major events in near real-time, such as disease outbreaks and earthquakes. It is faster at collecting data and more promptly at analyzing and providing results. We then study the effect of climate and seasonality on the prevalence of depression in Twitter users in the U.S. by

examining the spatiotemporal patterns of depression rate at Metropolitan Statistical Areas level, which has never before been conducted in depression related research. We conduct a stepwise regression and find that the relationship between depression, climate and seasonality is different and geographically localized. 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 rate. We also propose a three-stage framework that semi-automatically detects and analyzes geographically distributed health issues using location-based social media data. This framework can help us understand how social and behavioral interventions influence humans' health and illness. We further detect the temporal pattern of depression during holidays and non-holidays, weekdays and weekends, and its daily trends by exploring Twitter data.

INDEX WORDS: GIS, Tweet, Depression, Cluster, Seasonality, Climate, Social media

GIS ANALYSIS OF DEPRESSION USING LOCATION-BASED SOCIAL MEDIA DATA

by

WEI YANG

B.E., Wuhan University, P.R.China, June 30, 2009

M.S., Eastern Michigan University, United State, April 19, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

GIS ANALYSIS OF DEPRESSION USING LOCATION-BASED SOCIAL MEDIA DATA

by

WEI YANG

Major Professor:	Lan Mu
Committee:	Marguerite Madden
	Ye Shen
	Xiaobai Yao

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2015

DEDICATION

I lovingly dedicate this dissertation to my parents, who supported me each step of the way. They always gave me best education since I was young, and encouraged me to pursue my Ph.D. dream. They set a good example for me by working hard at their position, and gave me constant love and endless support.

ACKNOWLEDGEMENTS

I cannot express enough thanks to everyone who has contributed to the completion of my dissertation in any way, shape or form. First and foremost, I would like to thank my academic advisor, Professor Lan Mu, for her conscientious input, supportive comments and continuous patience throughout every step of this process. Research life is never easy. Every time I felt distressed and confused, she was always here gave me guidance, feedbacks and inspiration, kindly and wisely. I really appreciate for her time and generous help. I would like to thank my committee members for their support and help. I would also like to thank University of Georgia and the department of Geography for a well-rounded education. Finally, I would like to thank my family members for their endless love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Background

The geographic context of places and the connectedness between places play a major role in shaping environmental risks as well as many other health effects. For example, detecting disease outbreaks, targeting public health strategies, locating healthcare facilities, monitoring infectious disease trends, improving access to healthcare and lowering overall costs all have a geographic context (Little, Wicks, Vaughan, & Pentland, 2013) (Ginsberg, et al., 2009). Understanding how disease risk factors interact with the natural and social environments is direct and crucial to public health. GIS with its computer-based technology provides a digital lens for analyzing the dynamic associations between health issues and environments. It helps to interpret and address health issues for different neighborhoods or communities with their own particular demographic structures and social economic status (SES).

In recent years, social media have received considerable attention as a new data source for public health. With location-based techniques and wireless devices, social media have provided an enormous stream of fine-grained data on humans' life and behavior. Those petabytes data are highly resolute with time stamp and spatial attributes. With geospatial methods, location-based social media data can offer rich insights to humans' perceptions of space and its significance to public health. On another thought, health is not just about disease, but the result of interaction of physical, social, and emotion at individual, regional, or global scales, thus health has the geographical context of that person. In order to better understand the pathological mechanism

and risk factors for health issues, we need to analysis data from textual, spatial and temporal aspects. Thus, social media opens a bridge for communication between public health and GIS (Figure1.1).

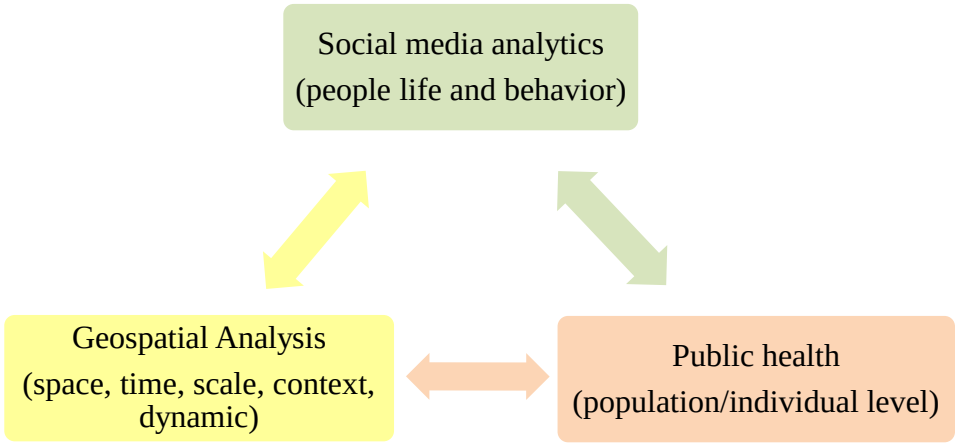


Figure1.1 Relationship between social media, GIS and public health

The fusion of GIS with social media provides new ways for public health research to extract and distill data into information implied by the data, which also emphasize the need for tackling challenges arising from handling the large amount of geographic data. First is the need for new automated methods of handling and analyzing big data that are being generated at very high speed (Kitchin, 2013) (Batty, et al., 2012). 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 very large volume and at high speeds. (Gorman, 2013) (Goodchild, 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 not content with knowing the surface phenomena about the data, but also want to know the deeper complex patterns implied by the data (González-Bailón, 2013) (Ruppert, 2013) (Manovich, 2011). Third, big data often lack rigorous sampling,

documentation, and quality assurance (Kitchin, 2013) (Goodchild, 2013) (Sui & Goodchild, 2011). This brings difficulty in confirming the validity and accuracy of crowdsourcing (Gorman, 2013). Fourth, although social media have better data 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 geovisual analytics and social media analytics. On 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 creative and meaningful ways to enable humans to visualize the large volume of information easily (González-Bailón, 2013) (Sui & Goodchild, 2011).

In this dissertation, we tackle those problems one by one by exploring the spatiotemporal pattern of depression among Twitter users in the U.S. We use text mining techniques, Geospatial methods and statistics methods to study the interaction between depression, environment socioeconomic statues, climate and seasonality.

Depression is a common chronic disorder with adverse effects for well-being and daily functioning, and is associated with high suicide rates (Barlow & Durand, 2011). Of all depression types, major depression disorder (MDD) is the most common one. The centers for disease control and prevention has reported that an estimated 3.4 percent U.S. adults have MDD (MMWR, 2010). Usually, patients with MDD are characterized by depressive mood accompanied by low self-esteem, apathy, anhedonia or lack of pleasure. Also, people with MDD suffer from lethargy, poor concentration, and fatigue. They always feel excessive guilt and hopelessness, and even have suicidal ideas or death wish.

Because depression is common, it often goes undetected. There is only a fraction of depression patients receiving adequate treatment, and the result is that nearly half of the undetected patients developed MDD in their later life (Houston, et al., 2001). Looking at the global environment of provisions and services for identifying, supporting, and treating mental illness, we have to admit it is insufficient: although 87% of the world's governments offer some primary care health services to tackle mental illness, 30% of them do not have programs, and 28% have no budget specifically identified for mental health (Detels, 2009). In fact, we have no reliable laboratory test for diagnosing most forms of mental illness (De Choudhury, Counts, & Horvitz, 2013b). As for MDD, diagnosis is based on the patient's self-reported experiences, behaviors reported by relatives or friends, and a mental status examination.

Depression can be treated if diagnosed earlier. How to detect and diagnosis depression quickly is the most important problem we need to solve. Researchers show that the population of Internet users searching for health-related information is up to 8 million per day (NLM/NIH, 2006). 23% of the online query is about mental health, such as anxiety, stress and depression (Rudd, et al., 2006). Social media such as Twitter, Myspace and Facebook provide a platform for people to share activities and a chance to express their feelings (Moreno, et al., 2011). The population that use social media as a source for health information is growing rapidly, particularly for young adults (Vance, Howe, & Dellavalle, 2009). Thus, social media with its open data online may indicate some opportunity for diagnosing and preventing depression if we can analysis the data in a proper way.

Twitter is a popular social media which allows users post tweets at a maximum length of 140 characters. In this dissertation, I chose Twitter as our data source because it provides many ready-to-use Application Programing Interfaces (APIs) for developers and it has a large

population base. As of 2012, there were about 500 million registered users, and now there are more than 400 million tweets posted within a single day. Another important reason is that it provides geotagged tweets with time stamp, which are two critical parts to understand health issues.

1.2 Research Objectives

The overarching research question of this dissertation asks how GIS analysis, big data and social media can be combined to work for public health studies. In this technological era, extracting and distilling useful information, and analyzing information at near-real time from the data deluge are the main trend of development. This research set out to analyze health issues using location-based social media data from textual, spatial and temporal aspects. Particularly, we studies depression using Twitter data. Our research are all from the technique view, and the methodology is suitable for any other disease, such as flu, postpartum depression and suicides. With open geo-data, the whole workflow can be used for disease outbreak detection, early diagnosis of disease, monitor and control the spread of disease, thus promote public health.

This dissertation research includes three main objectives. Each of the objectives corresponds to one chapter and addresses one technique contributions in GIS, text mining or statistics areas. More specifically, these three objects are:

1) Identify depressive users portrayed in Twitter; detect spatial patterns for auto-detected depressive users in Twitter; and find the association between depression rate and environment socioeconomic status.

2) Detect spatiotemporal patterns for depression rate among Twitter users; and probe into the effect of climate and seasonality on depression.

3) Detect temporal patterns for depression rate among Twitter users; and investigate the effect of U.S. holidays and weekdays on depression.

In general, Figure.1.2 shows the logical structure of the dissertation research.

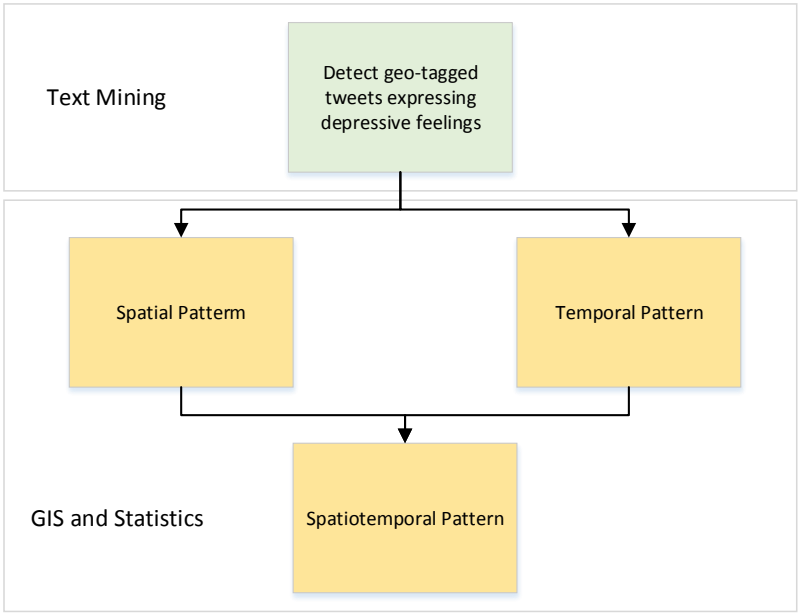


Figure 1.2 Logical structure of dissertation

1.3 Literature Review

1.3.1 Depression mood detection

There are some former researches about using Twitter to detect MDD patients. Park, Cha, & Cha (2012) and De Choudhury, et al. (2013b) shown that it is feasible and cheap to use Twitter for MDD detection. Coppersmith, Dredze, & Harman (2014) shown that although Twitter users are not a representative sample of the entirely population suffering from MDD, individual and population level analysis can still be made because of the diverse quantifiable signals relevant to depression observed in Twitter.

Moreno, et al. (2011) explored the depression moods within universities in Facebook. They found that about 25% of profiles on college students' Facebook expressing one or more depressive symptoms on status updates. The limitation is that the data set is from only one college, and they didn't prove the validity and accuracy of the identified MDD Facebook user.

Youn, et al. (2013) did a research to show that it is a feasible and cheap way to use social media (such as Facebook) for MDD detection and education. They listed some advantages for using social media. For example, it can reduce the staff needed to screen MDD and increase customization of interventions for a particular group of population. It is easy in data capture, since researchers are not restricting to geographic, time, mobility constraints. It is easy for data output because no more data entry was needed. It may reduce the stigma in traditional questionnaires, since social media has some confidentiality protocols and anonymity features.

Another research showed that signs identified by The American Association of Suicidology (AAS) were frequently included on websites (Rudd, et al., 2006). Among all the signs, the largest individual category for psychological symptoms is about depression, anxiety, psychosis, and maladaptive emotional states. Additionally, they found that young adults like to text to share their feelings. This was reported to be the second most common method of reaching out to others.

Some researchers are studying the depressive language in written text (Rudd, et al., 2006). Bollen, Gonçalves, Ruan, & Mao (2011) pointed out that investigation using Twitter should consider linguistic, cultural and geographic factors together.

1.3.2 Spatial cluster analysis

There are some other studies about neighborhood effects and mental health. Early in 1939, a research showed that different types of psychoses have a tightly relationship with certain conditions of communities (Faris & Dunham, 1939). A study in Midtown Manhattan also showed that sociocultural features of urban living can potentially influence mental health (Srole, Langner, Michael, Opler, & Rennie, 1962).

Former research showed that depression exhibited geographic clusters. Neighborhood racial composition (Mair, Diez Roux, Osypuk, et al., 2010), household income and education (Akhtar-Danesh & Landeen, 2007), and neighborhood family structures (Mair, Diez Roux, & Morenoff, 2010) are all related to depression. Their limitation lies in traditional data collection that conducts an interview or uses self-reported measures of mental health. Most people are unaware of the symptoms when they have depression; people may also conceal facts for privacy reasons. This may result in the same-source bias and spurious associations (Mair, et al., 2009).

Most of the SES factors used in these studies are defined by rational choice social theory. Oakes & Rossi (2003) gave a definition for SES with three different aspects. The first part is called material standard of living, which is measured by income, deprivation and poverty. The second part is called skills, which is defined as the number of years for education received and unemployment. The last part is social relationships composed of civic participation (e.g. member of a social organization or a volunteer worker) and living arrangements (e.g. people living alone).

1.3.3 Temporal cluster analysis

Climatic factors have relationship with the prevalence of depression such as temperature (H.-C. Lee, Tsai, & Lin, 2007), barometric pressure (Radua, Pertusa, & Cardoner, 2010), minutes of sunshine, global radiation and length of daylight (Molin, Mellerup, Bolwig, Scheike, & Dam, 1996).

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 (Nillni, Rohan, Rettew, & Achenbach, 2009; Magnusson, 2000; Huibers, de Graaf, Peeters, & Arntz, 2010). There is also evidence that

seasonal mood variations are even recognized in healthy people (Okawa, et al., 1996; Schlager, Schwartz, & Bromet, 1993).

Depression may occur at any time of the year, but the stress and anxiety of the holiday season may cause even those who are usually content to experience loneliness and a lack of fulfillment, especially during the months of November and December, and period before Valentine's day (Hillard, Holland, & Ramm, 1981; Sansone & Sansone, 2011; Velamoor, Voruganti, & Nadkarni, 1999) . The weekday-to-weekend effect was seen across age, gender and partner status, though retired people felt a less drastic mood shift than those still working, adding to the wealth of research that show work contributes to low moods (Stone, A. A., Schneider, S. & Harter, J. K., 2012; Golder & Macy, 2011).

1.3.4 Limitation of previous studies

There are some limitations for previous studies. First shortcoming is the same-source bias. A common method for detecting relationship between neighborhood effects and built environment is to collect neighborhood measures from participants' perceptions in the study area. Participants may have depression so their information may not accurately reflect the true situation. The same problem also existed when self-report measures are used in mental health study. This can bring error in the relationship. Considering all these, we believe that using social media to collect patients' data is better than traditional questionaries' or telephone surveys' data collection. Data are more accurate to reflect the real situation because website can provide some kind of privacy contract for the patients. Also, the process of collecting data using social media is a random selection in all kinds of people (gender, age, race, occupation, education, etc.). What's more, there are no time and space constraints for collecting data using social media.

Second, many research focused on adults for detecting the relationship between neighborhood effects and depression. This is not suitable. Compared to adults, young people are more susceptible to their neighborhood environment and have more space constrains near their home. Furthermore, adolescence is a time when depression often emerges for the first time. If we can focus on adolescent populations, we may get some fresh idea about depression and thus prevent MDD onset.

Third, many studies selected multilevel regression models to examine the relationship between neighborhood effects and depression. They didn't consider about the spatial effects. Although traditional multilevel regression model may explain for spatial heterogeneity at some degree, it cannot explain spatial autocorrelation due to spatial arrangement of data.

Fourth, when analyzing the relationship between climate and depression, former research often ignore the geographic location effect and only use a unified model to estimate the relationship through the whole study area. (Mersch, Middendorp, Bouhuys, Beersma, & van den Hoofdakker, 1999) found a significant positive correlation between the prevalence of depression and latitude in North America. They suggested the influence of latitude on prevalence is small and other factors such as climate, genetic vulnerability and social-cultural context may play a more important role.

1.4 Dissertation Structure

The proposal structure is organized into 3 chapters. Figure.1.3 shows the organization of this dissertation research. Chapter 1 is a brief introduction of the background and objectives of the dissertation research and literature review of the topics covered in this dissertation, including the detection of MDD users in Twitter, spatial clusters of depression, the relationship between depression and the environment social economic status, and the effect of climate and seasonality

on depression, as well as the temporal change of depression patterns during holidays and weekdays. The following three chapters are separate papers published in, submitted to, or prepared to be submitted to journals. In Chapter 2, topics were extracted for detecting MDD users in Twitter using text mining techniques. We use data in NY MSA to detect the spatial patterns of those auto-detected depressive users and detect their relationship with environment social economic status at county level. Chapter 3 explores the effect of climate and seasonality on depressed mood among Twitter users by examine the spatiotemporal patterns of depression. We use data in the top 20 MSA and 34th MSA to detect tweets expressing depressed feelings. We then detect their spatiotemporal patterns and the relationship between depression rate, climate and seasonality at MSA level. Chapter 4 uses different statistical models to study the temporal change of depression patterns during holiday season and non-holiday season, weekdays and weekends, and during different time in a day. Chapter 5 summaries our major results and contributions, and lists several limitations.

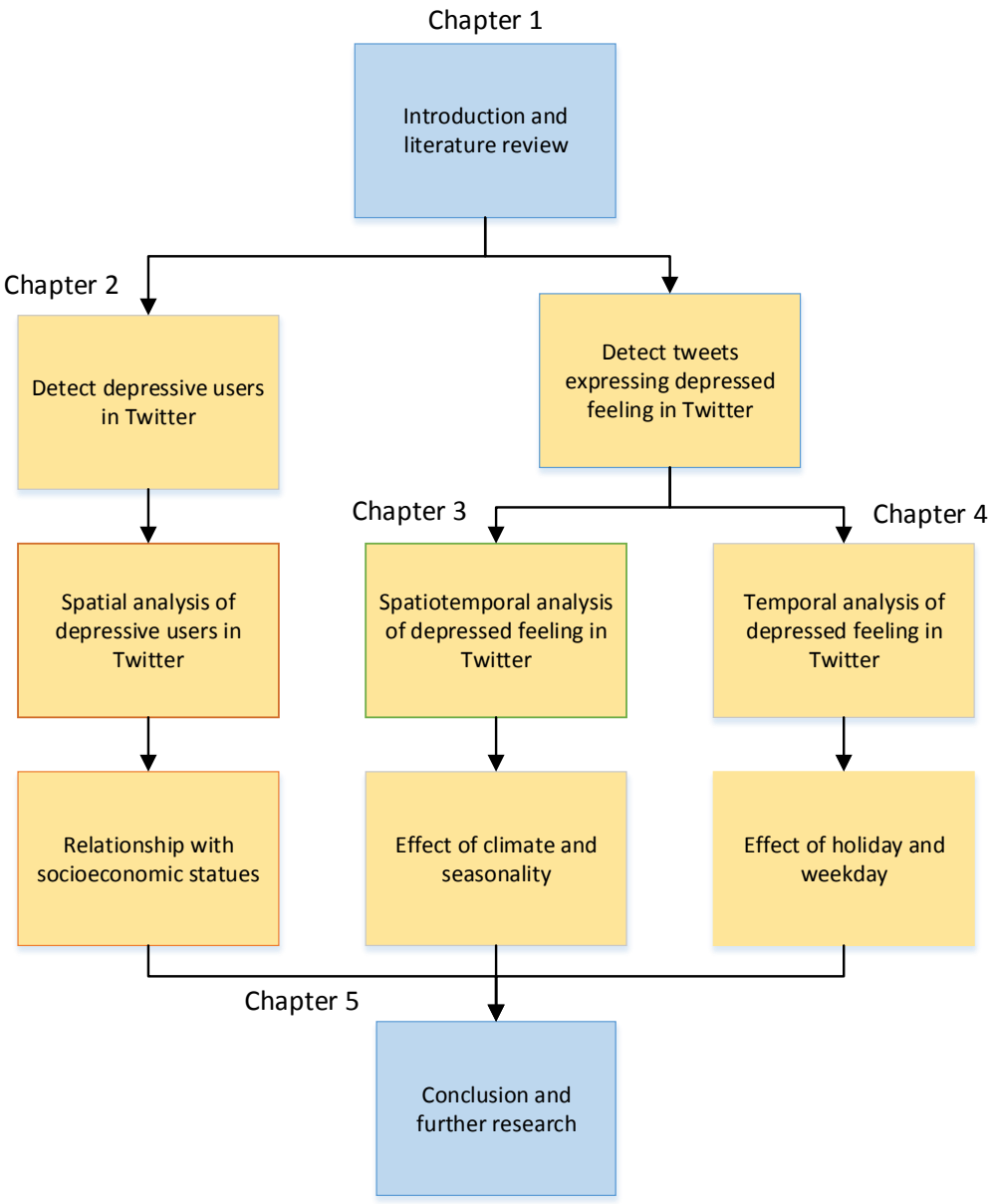


Figure.1.3 Organization of dissertation

References

Akhtar-Danesh, N., & Landeen, J. (2007). International Journal of Mental Health Systems. *International Journal of Mental Health Systems, 1*, 4.

Barlow, D. H., & Durand, V. M. (2011). *Abnormal Psychology: An Integrative Approach: An Integrative Approach*. Cengage Learning.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., & Portugali, Y. (2012). Smart cities of the future. *European Physical Journal-Special Topics*, *214* (1), 481.

Bollen, J., Gonçalves, B., Ruan, G., & Mao, H. (2011). Happiness is assortative in online social networks. *Artificial life*, *17* (3), 237-251.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3267-3276). ACM.

Detels, R. (2009). The scope and concerns of public health. *Oxford textbook of public health*, *2*, 3-19.

Faris, R. E. L., & Dunham, H. W. (1939). Mental disorders in urban areas: an ecological study of schizophrenia and other psychoses.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457* (7232), 1012-1014.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333* (6051), 1878-1881.

González-Bailón, S. (2013). Big data and the fabric of human geography. *Dialogues in Human Geography*, *3* (3), 292-296.

Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in Human Geography, 3* (3), 280-284.

Gorman, S. P. (2013). The danger of a big data episteme and the need to evolve geographic information systems. *Dialogues in Human Geography, 3* (3), 285-291.

Hillard, J. R., Holland, J. M., & Ramm, D. (1981). Christmas and psychopathology: Data from a psychiatric emergency room population. *Archives of general psychiatry, 38* (12), 1377-1381.

Houston, T. K., Cooper, L. A., Vu, H. T., Kahn, J., Toser, J., & Ford, D. E. (2001). Screening the public for depression through the Internet. *Psychiatric services, 52* (3), 362-367.

Huibers, M. J., de Graaf, L. E., Peeters, F. P., & Arntz, A. (2010). Does the weather make us sad? Meteorological determinants of mood and depression in the general population. *Psychiatry research, 180* (2), 143-146.

Kitchin, R. (2013). Big data and human geography Opportunities, challenges and risks. *Dialogues in Human Geography, 3* (3), 262-267.

Lee, H.-C., Tsai, S.-Y., & Lin, H.-C. (2007). Seasonal variations in bipolar disorder admissions and the association with climate: a population-based study. *Journal of affective disorders, 97* (1), 61-69.

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time disease surveillance using Twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1474-1477). ACM.

Little, M., Wicks, P., Vaughan, T., & Pentland, A. (2013). Quantifying short-term dynamics of parkinson's disease using self-reported symptom data from an internet social network. *Journal of medical Internet research, 15* (1).

Magnusson, A. (2000). An overview of epidemiological studies on seasonal affective disorder. *Acta Psychiatrica Scandinavica, 101* (3), 176-184.

Mair, C., Diez Roux, A. V., & Morenoff, J. D. (2010). Neighborhood stressors and social support as predictors of depressive symptoms in the Chicago Community Adult Health Study. *Health & place, 16* (5), 811-819.

Mair, C., Diez Roux, A. V., Osypuk, T. L., Rapp, S. R., Seeman, T., & Watson, K. E. (2010). Is neighborhood racial/ethnic composition associated with depressive symptoms? The multi-ethnic study of atherosclerosis. *Social science & medicine, 71* (3), 541-550.

Mair, C., Roux, A. V. D., Shen, M., Shea, S., Seeman, T., Echeverria, S., & O'meara, E. S. (2009). Cross-sectional and longitudinal associations of neighborhood cohesion and stressors with depressive symptoms in the multiethnic study of atherosclerosis. *Annals of epidemiology, 19* (1), 49-57.

Manovich, L. (2011). Trending: the promises and the challenges of big social data.

Mersch, P. P. A., Middendorp, H. M., Bouhuys, A. L., Beersma, D. G., & van den Hoofdakker, R. H. (1999). Seasonal affective disorder and latitude: a review of the literature. *Journal of affective disorders, 53* (1), 35-48.

Morbidity and Mortality Weekly Report (MMWR). (2010). Current Depression Among Adults - United States, 2006 and 2008. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5938a2.htm

Molin, J., Mellerup, E., Bolwig, T., Scheike, T., & Dam, H. (1996). The influence of climate on development of winter depression. *Journal of affective disorders, 37* (2), 151-155.

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., & Becker, T. (2011). Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety, 28* (6), 447-455.

National Library of Medicine (NLM)/National Institutes of Health. NLM Technical Bulletin: MLA 2006, NLM online users' meeting remarks. 2006. Available: http://www.nlm.nih.gov/pubs/techbull/ja06/ja06_mla_dg.html. Accessed 2008 April 25.

Nillni, Y. I., Rohan, K. J., Rettew, D., & Achenbach, T. M. (2009). Seasonal trends in depressive problems among United States children and adolescents: a representative population survey. *Psychiatry research, 170* (2), 224-228.

Oakes, J. M., & Rossi, P. H. (2003). The measurement of SES in health research: current practice and steps toward a new approach. *Social science & medicine, 56* (4), 769-784.

Okawa, M., Shirakawa, S., Uchiyama, M., Oguri, M., Kohsaka, M., Mishima, K., Sakamoto, K., Inoue, H., Kamei, K., & Takahashi, K. (1996). Seasonal variation of mood and behaviour in a healthy middle-aged population in Japan. *Acta Psychiatrica Scandinavica, 94* (4), 211-216.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)* (pp. 1-8).

Partonen, T., & Lönnqvist, J. (1998). Seasonal affective disorder. *The Lancet, 352* (9137), 1369-1374.

Radua, J., Pertusa, A., & Cardoner, N. (2010). Climatic relationships with specific clinical subtypes of depression. *Psychiatry research, 175* (3), 217-220.

Rudd, M. D., Berman, A. L., Joiner, T. E., Nock, M. K., Silverman, M. M., Mandrusiak, M., Van Orden, K., & Witte, T. (2006). Warning signs for suicide: Theory, research, and clinical applications. *Suicide and Life-Threatening Behavior, 36* (3), 255-262.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography, 3* (3), 268-273.

Sansone, R. A., & Sansone, L. A. (2011). The christmas effect on psychopathology. *Innovations in clinical neuroscience, 8* (12), 10.

Schlager, D., Schwartz, J. E., & Bromet, E. J. (1993). Seasonal variations of current symptoms in a healthy population. *The British Journal of Psychiatry, 163* (3), 322-326.

Srole, L., Langner, T. S., Michael, S. T., Opler, M. K., & Rennie, T. A. (1962). Mental health in the metropolis: The midtown Manhattan study.

Stone, A.A., S. Schneider, and J.K. Harter. (2012). Day-of-week mood patterns in the United States: On the existence of "blue Monday," "thank God it's Friday," and weekend effects. *Journal of Positive Psychology* 7(12):306-314.

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science, 25* (11), 1737-1748.

Vance, K., Howe, W., & Dellavalle, R. P. (2009). Social internet sites as a source of public health information. *Dermatologic clinics, 27* (2), 133-136.

Velamoor, V. R., Voruganti, L. P., & Nadkarni, N. K. (1999). Feelings about Christmas, as reported by psychiatric emergency patients. *Social Behavior and Personality: an international journal, 27* (3), 303-308.

Youn, S. J., Trinh, N.-H., Shyu, I., Chang, T., Fava, M., Kvedar, J., & Yeung, A. (2013). Using online social media, Facebook, in screening for major depressive disorder among college students. *International Journal of Clinical and Health Psychology, 13* (1), 74-80.

CHAPTER 2

GIS ANALYSIS OF DEPRESSION AMONG TWITTER USERS[1]

[1] Yang, W. and L. Mu., 2015, Applied Geography. Volume 60, Pages 217-223, Reprinted here with permission of the publisher.

Abstract

Depression is a common chronic disorder. It often goes undetected due to limited diagnosis methods and brings serious results to public and personal health. Former research detected geographic pattern for depression using questionnaires or self-reported measures of mental health, this may induce same-source bias. Recent studies use social media for depression detection but none of them examines the geographic patterns. In this paper, we apply GIS methods to social media data to provide new perspectives for public health research. We design a procedure to automatically detect depressed users in Twitter and analyze their spatial patterns using GIS technology. This method can improve diagnosis techniques for depression. It is faster at collecting data and more promptly at analyzing and providing results. Also, this method can be expanded to detect other major events in real-time, such as disease outbreaks and earthquakes.

Key words: depression, tweets, clustering, GIS, social media

2.1 Introduction

Depression is a common chronic disorder with adverse effects for well-being and daily functioning, and is associated with high suicide rates (Barlow & Durand, 2011). Major depressive disorder (MDD)[2] (Barlow & Durand, 2005) is the most common type. The centers for disease control and prevention has reported that an estimated 3.4 percent U.S. adults report MDD (MMWR, 2010).

[2] Abbreviations in this paper:
GIS Geographic Information System
MDD Major Depressive Disorder
NMF Non-negative Matrix Factorization
SES Socioeconomic Status
API Application Programming Interface
DSM-IV Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
TP True Positive
FP False Positive
FN False Negative
MSA Metropolitan Statistical Areas

Depression often goes undetected due to the absence of reliable laboratory test, and therefore new methodology for its diagnosis is needed. Social media provides a platform for people to share their activities and feelings. Paul and Dredze (2011) highlighted a slogan "You are what your tweet" and suggested that Twitter had broad applicability for public health research. Twitter is a popular social media website that allows users to post tweets with a maximum length of 140 characters. It has a large user base and provides geo-tagged tweets to researchers, thus the collected data are not restricted to geographic, time, and mobility constraints. Recent work has shown that it is feasible to use Twitter for MDD detection (Park, et al., 2012) (De Choudhury, Counts, & Horvitz, 2013a).

Former research showed that depression exhibited geographic clusters. Neighborhood racial composition (Mair, Diez Roux, Osypuk, et al., 2010), household income and education (Akhtar-Danesh & Landeen, 2007), and neighborhood family structures (Mair, Diez Roux, & Morenoff, 2010) are all related to depression. Their limitation lies in traditional data collection that conducts an interview or uses self-reported measures of mental health. Most people are unaware of the symptoms when they have depression; people may also conceal facts for privacy reasons. This may result in the same-source bias and spurious associations (Mair, Diez Roux, Shen, et al., 2009).

We contribute to the study of depression using social media in two aspects. First, we design a procedure to automatically detect MDD users before their estimated onset using Twitter. Because tweets are short and noisy texts and a term can have different meanings, we not only use the key word "depress" and its variations to filter tweets, but also leverage an advanced text mining algorithm, namely *non-negative matrix factorization* (NMF) (Lee & Seung, 1999). We evaluate our procedure on real tweets. This procedure offers a novel and automated way to

diagnose depression. It reduces human labor for MDD screening and facilitates personalized interventions for particular groups of population. Second, we further detect the spatial distribution of these MDD users on Twitter and examine the association with socioeconomic status (SES) (Krieger, Willians & Moss, 1997) at county level.

2.1.1 Using Twitter for depression study

There were some papers using Twitter to predict depression. Park, et al. (2012) applied sentiment analysis on tweets and showed that Twitter provided meaningful data for clinical studies on depression. De Choudhury, et al. (2013a) from Microsoft Research compared tweet text of depressed Twitter users to those of the normal users and highlighted the potential of Twitter as a tool for predicting MDD. Harman, Coppersmith, and Dredze (2014) pointed out that although Twitter users are not a representative sample of the entire population suffering from MDD, individual level and population level analysis can still be made because of the diverse set of quantifiable signals related to MDD in Twitter. A recent survey revealed that 26 percent of the online U.S. adults discussed their health information online, and 42 percent of them use social media to post or seek information about health conditions (GE Healthcare, 2012). Rudd, Berman, Joiner, et al., (2006) showed that signs identified by the American Association of Suicidology were frequently included on websites. Among all the signs, the largest individual category for psychological symptoms is about depression and anxiety states. Additionally, they found that young adults like to text to share their feelings.

2.1.2 Geographic analysis of health issues using Twitter

There were very few papers on geographic analysis of public health issues using social media data. K. Lee, Agrawal, and Choudhary (2013) examined influenza spread in the U.S. by measuring the weighted percentage of tweet volumes mentioning “flu” at state level. The

limitations were that they only used the key word “flu” to extract tweets and only considered user profiles that have valid U.S. state information in their home location field.

Ghosh and Guha (2013) compared spatial clusters of obesity-related tweets with the distribution of McDonald’s restaurants. They didn’t consider that the frequency of posting tweets is different for different people. Using tweets as units instead of Twitter users to explain obesity distribution is not meaningful. For example, someone who posting many tweets with obesity theme at one location doesn’t represent there is a higher incident there.

Morstatter, Pfeffer, Liu, and Carley (2013) showed that when geographic bounding boxes were used for Twitter data collection, the collected data were almost the complete set of geo-tagged tweets and thus can be trusted for analytical purposes. Also, text analysis was most accurate when data were downloaded from the Twitter streaming application programming interfaces (APIs) (Customer Information Manager, 2013).

2.2 Material and Method

2.2.1 Pilot study

Because a term can have different meanings, we conduct a pilot study to explore the different expressions in tweets related to depression. We use two kinds of Twitter APIs for downloading data: Twitter Streaming APIs give low latency access to the global stream of Twitter data. Twitter REST APIs provide interfaces for most of Twitter’s online functionality. We only keep tweets written in English and posted in the U.S. from 2013/09/05 to 2014/03/05, and remove all the others. First, we filter tweets by the key word “depress” or its variations to select tweets related to depression. Then, we write a computer program to randomly select a uniform sample with 0.1% rate (402 tweets) for content analysis.

We follow the classification scheme of tweets related to depression proposed by Park, et al. (2012), and manually label the tweets according to their meaning and purpose. Similar labels are placed into the same category in a hierarchical way (Table 2.1). Each label has a code. For example, "*don't want to be in this hospital anymore, I'm depressed.*" is labeled as "reason for depression" in category A - about my depressed feeling.

Table 2.1 Categories A-E, labels and codes 1-23 for tweets.

A. about my depressed feelings		13	seek depression info
1	depressed feeling	**C. sharing thoughts related to depression**	
2	reason for depression	14	my perception of depression
3	my own remedy for depression	15	attitude towards depression
4	treatment on one's depression	16	comments to encourage others with depression
5	change of depression feelings	**D. other topics**	
6	pattern of depression	17	various usage of the word depress
7	meeting doctors	18	investigating depression
8	take medication	19	pet depression
9	confess diagnosis	20	will be depressed if …
10	not depressed	21	meaning is not clear
B. delivery of depression info		**E. about other's depressed feelings**	
11	my symptoms	22	stories of others on TV, article..
12	Info w/o URL for depression	23	tweets about friends' depressed feelings

Figure 2.1 shows the number of tweets under different categories and labels. Among the 402 randomly selected tweets, the word "depress" is most frequently used to express one's own depressed feelings (70 percent of the sample). Among them, 146 tweets express depression directly, 119 tweets even give detail reasons for being depressed, and three tweets describe the change of depressed feelings. Some of them even post private information about their own remedy or diagnosis for depression, and pattern of depression. Eight tweets in category B deliver information regarding depression. Among them, seven tweets contain links leading to a song or a video for fighting against depression, or some medical articles and causes for depression. One tweet mentions the user's own depression symptoms.

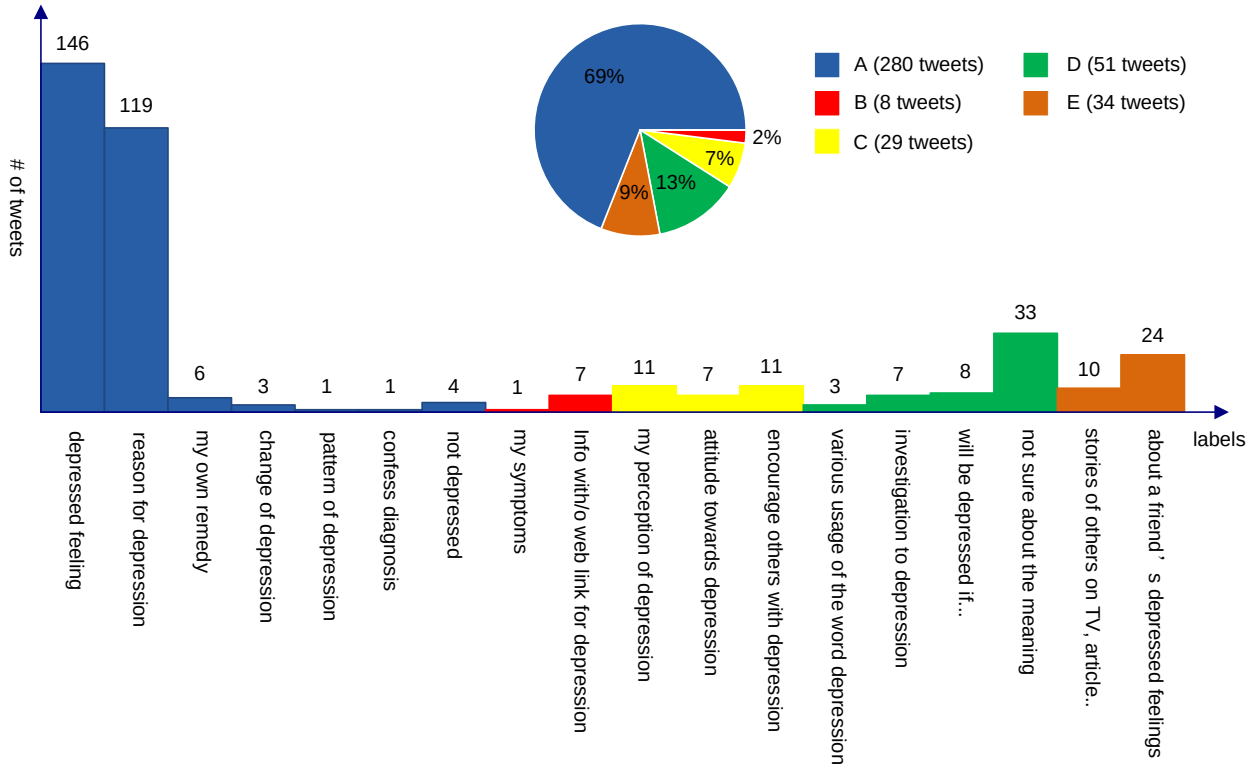


Figure 2.1 Tweets counts for different categories and labels.

2.2.2 Text mining for depression

In this section, we introduce a text mining technique we employed to select MDD candidates more accurately. Because a term can usually have different meanings, it is not sufficient to use the keyword "depress" or its variations to select tweets related to depression. Therefore, we apply a data mining technique, called *non-negative matrix factorization* (NMF), to differentiate the word context associated with depression from those not related to depression.

The basic idea of matrix factorization is that each tweet can be represented as a high-dimensional vector in the space of words, and such high-dimensional data can often be described approximately in a latent subspace with much lower dimensions (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). We can interpret the dimensions of the latent subspace as a recurring pattern of word contexts. NMF is a special type of matrix factorization that applies to

nonnegative data (D. D. Lee & Seung, 1999). It often drives better interpretation of the latent dimensions and is first applied to document clustering by Xu, Liu, & Gong (2003).

First, we transform the tweets into a word-document matrix (Manning, Raghavan, & Schütze, 2008). A common assumption is that each tweet (document) is composed of a bag of words and the sequence of words in a document can be ignored. Thus, we can represent a document as a vector of word counts for each word in a vocabulary V. We consider a fixed vocabulary containing *m* unique words, that is, $|V|=m$. We arrange *n* documents represented as vectors into an $m \times n$ matrix A, as shown in Figure 2.2.

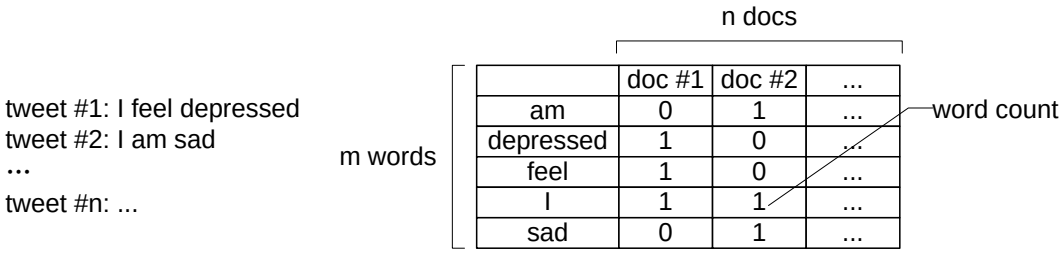


Figure 2.2 Example of a word-document matrix

In NMF, given a word-document matrix *A* as above and a positive integer *k*, the goal is to approximate *A* by the product of two nonnegative matrices, *W* and *H*:

$$A \approx WH. \qquad \text{Equation [2.1.]}$$

Here *W* is an $m \times k$ matrix and *H* is a $k \times n$ matrix. We show an intuitive illustration of NMF in Figure 2.3.

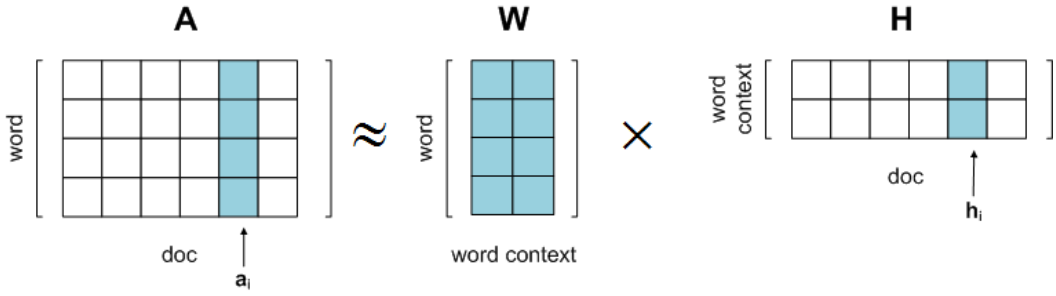


Figure 2.3 Illustration of NMF

In Equation [2.1.], each column of *A*, denoted as $\mathbf{a}_i$, is approximated by a weighted sum of the columns of *W*, that is,

$$\mathbf{a}_i \approx W \times \mathbf{h}_i \qquad \text{Equation [2.2]}$$

The weights are contained in the *i*-th column of *H*, denoted as $\mathbf{h}_i$.

We can interpret each column of *W* as a vector of frequencies for all the words in *V*, which we call a word context. The *k* word contexts define a *k*-dimensional latent subspace where we can describe the documents, and $\mathbf{h}_i$ contains the coordinates for the *i*-th document a_i in the latent subspace. We can interpret $\mathbf{h}_i$ as the proportions corresponding to the *k* word contexts that constitute $\mathbf{a}_i$. Hence, $\mathbf{a}_i$ can be assigned to the word context with the largest proportion value.

Choosing an appropriate number of word contexts *k* is difficult but important. Because tweets are very noisy, we manually choose the number of word contexts, and examine the word contexts found by NMF under different choices. Then we pick the word context directly related to the depression mood, and collect all the documents assigned to that context.

For illustration purposes, for each word context (a column of *W*), we pick the words corresponding to the ten largest frequencies in order to display the word contexts.

2.2.3 Auto-detected MDD users

Using text clustering to identify the word context for the depressed mood is not sufficient to detect the MDD candidates. Additionally, we use the diagnostic and statistical manual of mental disorders (DSM-IV) criteria (American Psychological Association, 2014) for MDD (Figure 2.4).

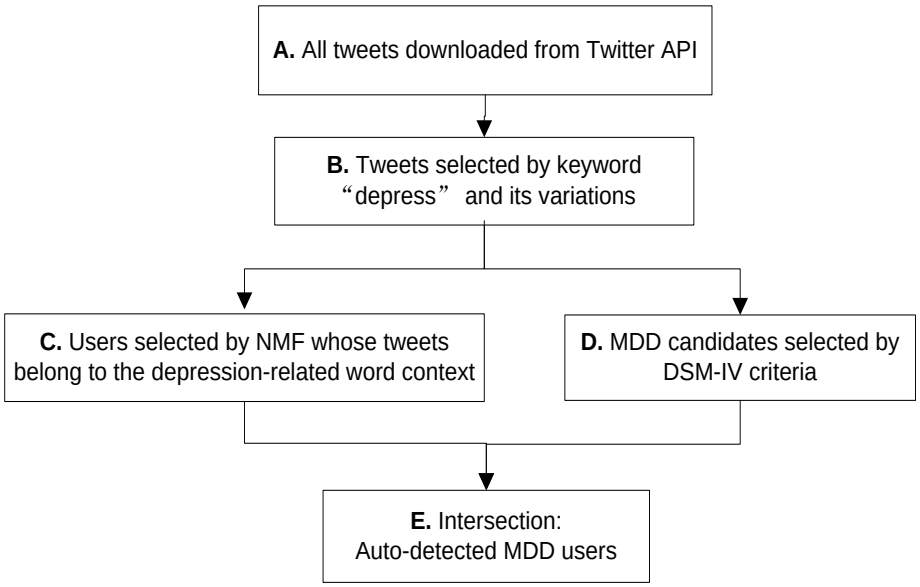


Figure 2.4 Workflow of detecting MDD users

The DSM-IV criteria require that MDD patients must suffer from at least five of the nine typical depression symptoms for more than two weeks (Skinner, Whiteley, & Ratner, 1990). These specific symptoms include depressed mood, decreased interest in daily activities, significant weight change or change in appetite, change in sleep or activity, fatigue, feelings of guilt or worthlessness, loss of concentration and having suicide plan. However, when analyzing Twitter text, we only focus on the depressed mood of the users. As a result, in order to apply DSM criteria to identify MDD candidates, we make the following change to the criteria: We require five or more tweets that are associated with depressed mood within a two-week period.

2.2.4 Procedure evaluation

We use the evaluation measures arising from information retrieval to assess the accuracy of our procedure (Rijsbergen, 1997): True positive (TP) is the number of times that our procedure correctly indicates that a user has MDD. False positive (FP) is the number of times that our procedure wrongly indicates that a user has MDD, while in fact the user does not have MDD.

False negative (FN) is the number of times that our procedure indicates that a user does not have MDD, while the user actually has MDD.

To find the true MDD users, we focused again on those MDD candidates identified by both key word and DSM-IV criteria. See part D in Figure 2.4. We manually go through their tweets related to depression and exclude users whose tweets do not express depressed feelings, such as those regarding tropical depression.

Recall measures the fraction of auto-detected MDD users out of all the true MDD users, and precision measures the fraction of true MDD users out of all the auto-detected MDD users (Figure 2.5). We use the F-score as a summary statistic to evaluate our procedure, which is the harmonic mean of precision and recall. A larger F-score measure corresponds to a more accurate procedure.

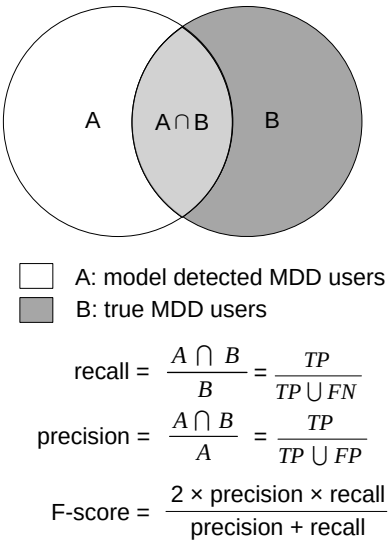


Figure 2.5 Illustration of calculation for recall, precision and f-score

2.2.5 Spatial analysis for MDD users

In this research, we want to detect the spatial patterns for MDD users at county level in New York-Newark-Jersey City Metropolitan Statistical Area (abridged as "NY MSA" in the following) (Nussle, 2008). Figure 2.6 shows the study area.

New York-Newark-Jersey City Metropolitan Statistical Area (NY MSA)

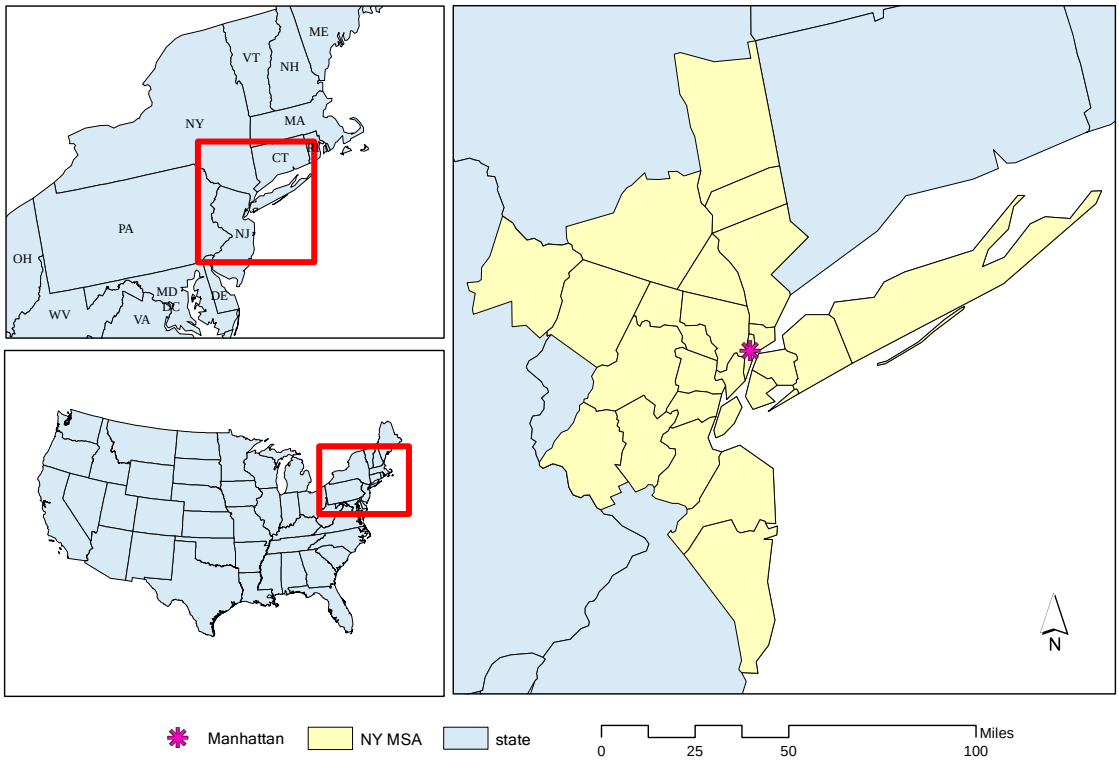


Figure 2.6 Study area of NY MSA

The time period is from 2013/09/05 to 2014/03/05. The null and alternative hypotheses are as follows:

H_0: The MDD users are geographically randomly distributed.

H_1: The distribution of MDD users are spatially clustered or dispersed.

To infer the location for an MDD user from their tweets related to depression, we create a user-by-county matrix. Each row represents an MDD user and each column represents a county. Each entry is the number of tweets by a user geo-tagged within a county. The sum of each row is the total number of tweets posted by that user, and the probability of user #1 living in county #1 is estimated as the proportion of tweets posted in county #1 by that user. The sum of each column is the number of MDD users located in that corresponding county.

After determining the number of MDD users in each county, we calculate Moran's I and use hot spot analysis to examine the spatial pattern of MDD users.

We then examine the association between depression rate at county level and local SES. It is reasonable to assume that the number of Twitter users is positively correlated with the size of population. Since over 88 percent of Twitter users are between 15 and 35 years old, we normalize the number of MDD users in each county by the population between 15 and 35 years old. Population and socioeconomic data such as education, income, and race are downloaded from the U.S. Census Bureau.

2.3. Results and Discussion

2.3.1 Word context and auto-detected MDD users

To find the word context related to depression, we use all the data downloaded from 2013/09/05 to 2014/03/05 in the U.S. When using the keyword "depress" and its variations to select tweets related to depression, we choose not to remove the tweets whose subject is not the user, since the pilot study shows that only eight percent of our sample of tweets are talking about other people's depressed feeling. Additionally, for detecting MDD users, we require the occurrence of five or more tweets related to depression within two weeks, which can exclude most of the tweets whose subjects are not the Twitter user.

For NMF, we experiment with different numbers of word contexts k ranging from two to six, and find that the word context related to depression emerges when $k=5$. This word context remains when k gets larger. Table 2.2. shows the NMF results: The word contexts 2, 3 and 4 are related more to music or weather; the word context 5 contains the key words "depressed, mood, stressed" and thus is identified as the word context related to depression. A more objective way to choose the depression-related word context is to determine the set of tweets assigned to each

word context via NMF, and find which set of tweets has the largest intersection with the users selected by the DSM criteria for MDD. In this way, the word context 5 is chosen as the depression-related word context again.

Table 2.2 Top ten words for each of the five auto-detected word contexts

1	feel, today, makes, make, sad, reason, bad, hate, idk, depresso
2	lol, *, tweets, im, music, ass, listening, twitter, songs, drake
3	depression, post, concert, deep, real, *, anxiety, great, seasonal, back
4	depressing, tweets, day, music, weather, today, *, news, song, movie
5	depressed, im, feeling, makes, stressed, dressed, making, *, happy, mood

Note: * represents cursed words or swear words

We create the word cloud using the top 50 words selected by frequency in the depression-related word context (Figure 2.7). The larger size of a word corresponds to a higher frequency. Then, we collect users who have tweets posted in this word context for further research.

Figure 2.7 Word cloud of depression-related word context

To focus on NY MSA, Table 2.3 lists the statistics of the group of users in our experiment corresponding to Figure 2.8. In order to evaluate the accuracy of our method, we identify 404 out of the 448 users are true MDD users. We then calculate TP (242 users), FP (44 users), and FN (162 users). The precision is 0.85, recall is 0.60, and F-score is 0.70.

Table 2.3 Actual number of users corresponding to the workflow

Procedure referring to the workflow	**# of users**
C. Users selected by NMF whose tweets belong to depression-related word context	1,225
D. MDD candidates selected by DSM-IV criteria	448
E. Intersection: auto-detected MDD users	286

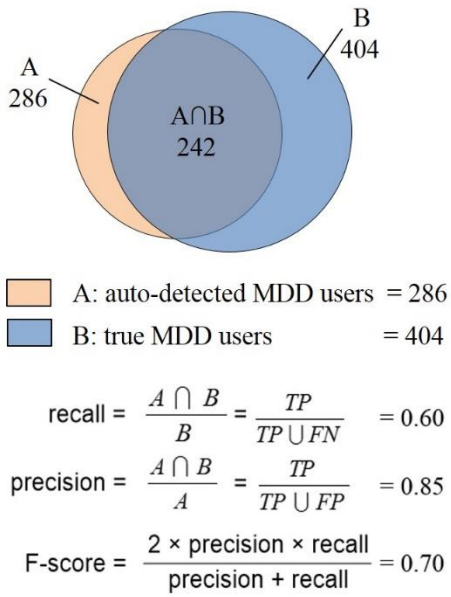


Figure 2.8 Evaluation of procedure

2.3.2 Spatial analysis of MDD users in NY MSA

There are 25 counties in NY MSA. 249 out of 286 MDD users have shared location information. Dutchess, Pike, Union, Richmond, Monmouth and Ocean counties have the highest rate of MDD users (Figure 2.9). Moran's I results (z-score = 3.54, p-value = 0.0004) also shows the clusters are significant. Hot spot analysis reveals that the hot spots clustering are concentrated around Ocean and Dutchess Counties (Figure 2.10).

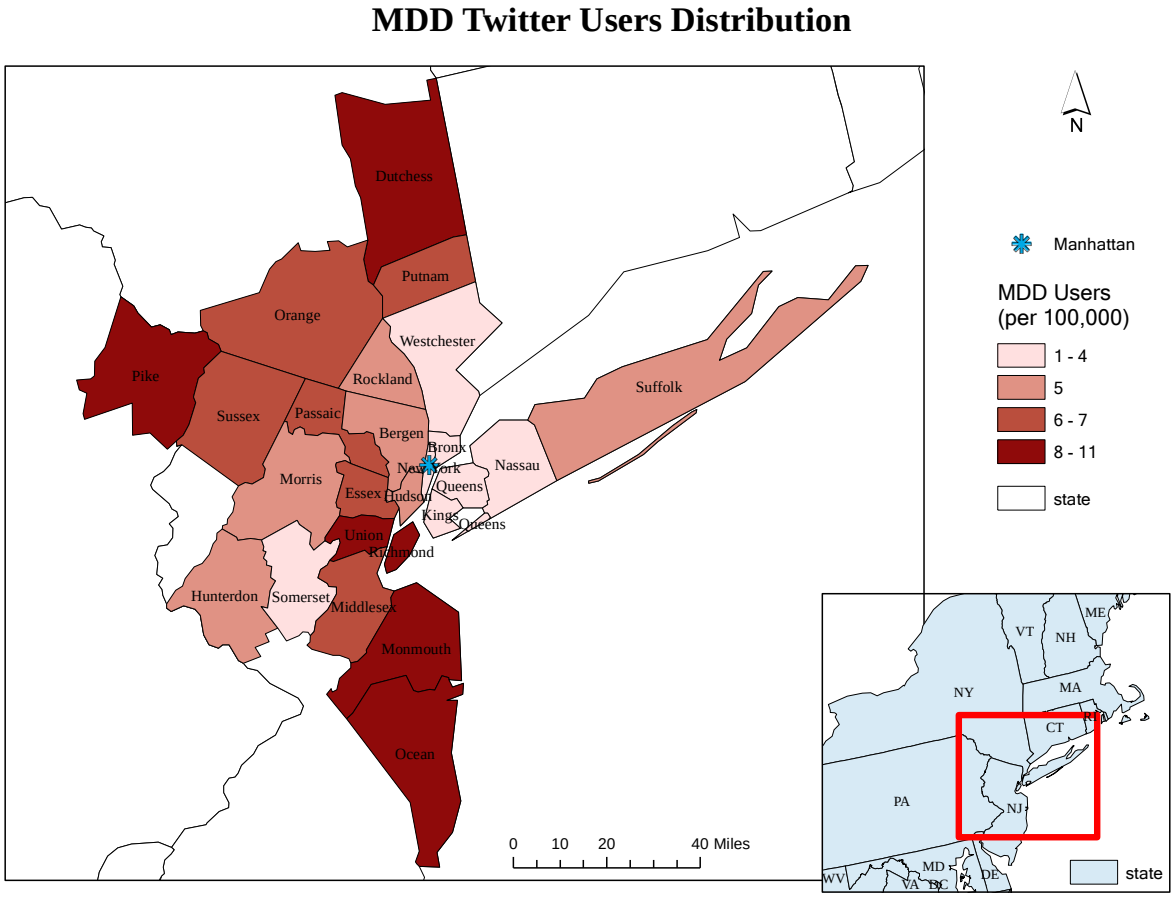


Figure 2.9 Distribution of MDD Twitter users in NY MSA

Hot Spot Analysis for MDD Twitter Users

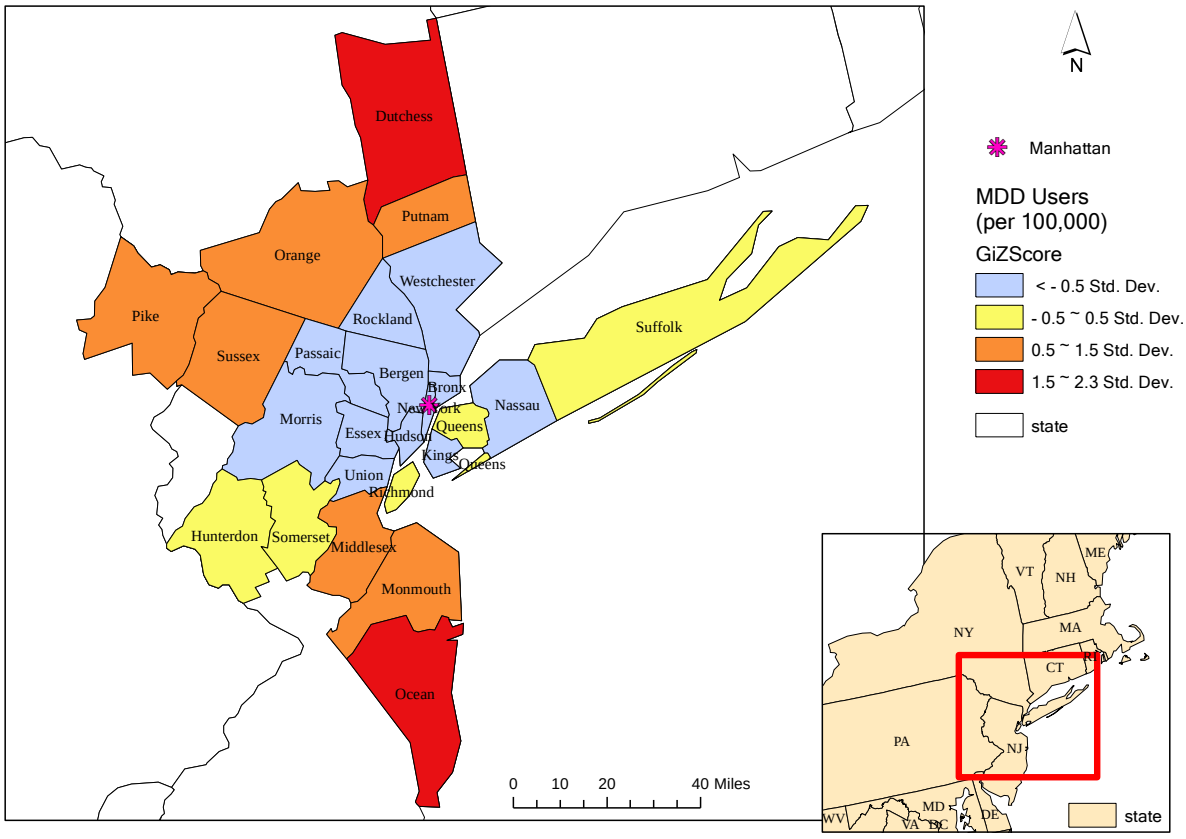


Figure 2.10 Hot spot analysis for MDD Twitter users in NY MSA

Bringing in SES variables to understand the auto-detected MDD results, we can see that the proportion of White population by county has a positive correlation with the total number of MDD users by county (Figure 2.11). Counties with higher proportion of population with college education have a lower MDD rate. Counties with 8 to15 percent of population with less than high school education have a higher MDD rate (Figure 2.12). Middle-class people with a median household income at around 60,000 dollars have a higher prevalence of depression (Figure 2.13), and this echoes some of the findings in the literature (Akhtar-Danesh & Landeen, 2007).

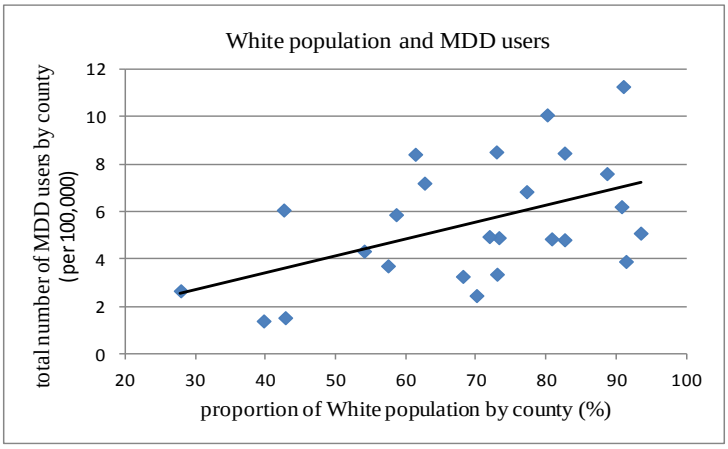


Figure 2.11 Relationship between rates of White population and MDD users

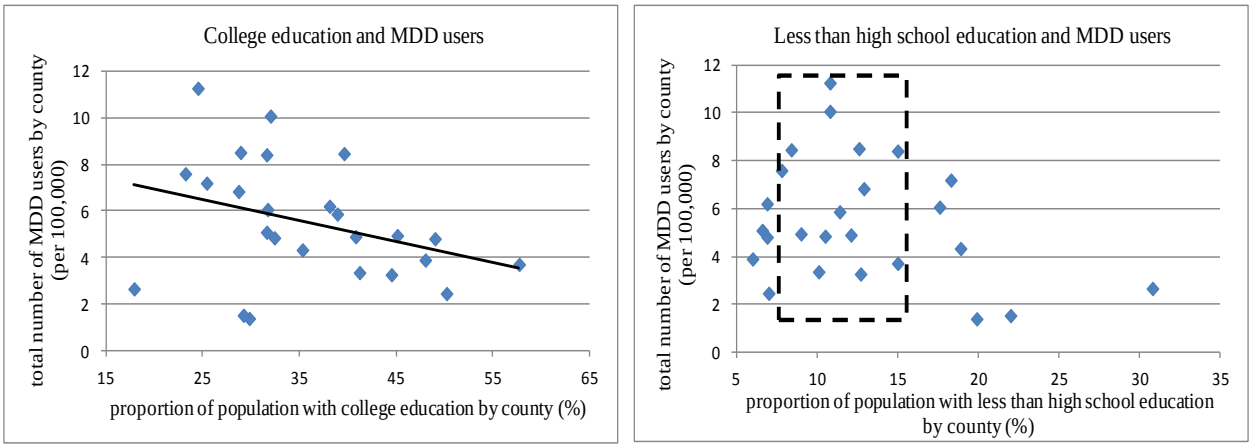


Figure 2.12 Relationship between education and rate of MDD users

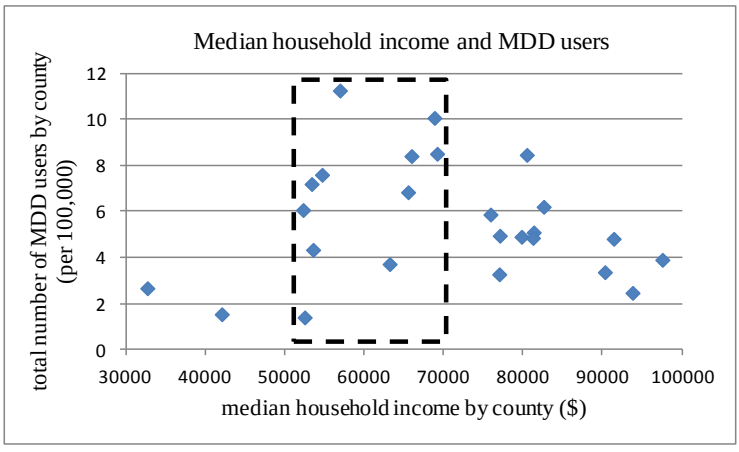


Figure 2.13 Relationship between income and the rate of MDD users

2.4 Conclusion and Limitation

In this paper, we present a new method for public health research combining GIS with social media. Compared with traditional data collection method, our automated method for detecting MDD users is faster and cheaper for analysis and diagnosis. The system can be applied to some online forum for detecting depression topic and forwarding related questions to psychiatrists. Our GIS results also provide novel knowledge about this disorder by examining the geographic clustering of MDD users and relationship with SES.

In this research, we didn't say that our method for MDD diagnosis can replace the work of a clinical psychologist. Our method can improve diagnosis techniques for depression. Further detailed clinical contexts are needed to make a formal diagnosis. Future study should probe into the difference between depression detected online and self-reported depression reported by a professional clinical scale table.

Secondly, the Twitter APIs only allows free access to a one percent convenience sampling of tweets. Data acquired are restricted to users with public profiles. These may bring some bias to our result.

Additionally, we only include tweets written in English and users who identified themselves living in the U.S. Changing any of those may affect our results.

References

Akhtar-Danesh, N., & Landeen, J. (2007). Relation between depression and sociodemographic factors. *International Journal of Mental Health Systems, 1*, 4.

American Psychiatric Association. (2014). DSM. http://www.psychiatry.org/practice/dsm.

Barlow,D. H., & Durand, V. M. (2005). *Abnormal psychology: An integrative approach (5th ed.)*. Belmont, CA, USA: Thomson Wadsworth, ISBN 0-534-63356-0.

Barlow, D. H., & Durand, V. M. (2011). *Abnormal Psychology: An Integrative Approach: An Integrative Approach*. Cengage Learning.

Customer Information Manager. *SOAP API Documentation*. 2013. http://www.authorize.net/support/CIM_SOAP_guide.pdf.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems* (pp. 3267-3276). ACM.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS, 41* (6), 391-407.

GE Healthcare. (2012). Twenty six percent of online adults discuss health information online; privacy cited as the biggest barrier to entry. http://www.businesswire.com/news/home/20121120005872/en

Ghosh, D., & Guha, R. (2013). What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science, 40* (2), 90-102.

Harman, C., Coppersmith, G., & Dredze, M. (2014). Quantifying Mental Health Signals in Twitter.

Krieger, N., Williams, D. R., & Moss, N. E. (1997). Measuring social class in US public health research: concepts, methodologies, and guidelines. *Annual review of public health, 18* (1), 341-378.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401* (6755), 788-791.

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time disease surveillance using Twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1474-1477). ACM.

Mair, C., Diez Roux, A. V., & Morenoff, J. D. (2010). Neighborhood stressors and social support as predictors of depressive symptoms in the Chicago Community Adult Health Study. *Health & place, 16* (5), 811-819.

Mair, C., Diez Roux, A. V., Osypuk, T. L., Rapp, S. R., Seeman, T., & Watson, K. E. (2010). Is neighborhood racial/ethnic composition associated with depressive symptoms? The multi-ethnic study of atherosclerosis. *Social science & medicine, 71* (3), 541-550.

Mair, C., Diez Roux, A. V., Shen, M., Shea, S., Seeman, T., Echeverria, S., & O'meara, E. S. (2009). Cross-sectional and longitudinal associations of neighborhood cohesion and stressors with depressive symptoms in the multiethnic study of atherosclerosis. *Annals of epidemiology, 19* (1), 49-57.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge university press Cambridge.

Morbidity and Mortality Weekly Report (MMWR). (2010). Current Depression Among Adults - United States, 2006 and 2008. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5938a2.htm

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*.

Nussle, J. (2008). Update of Statistical Area Definitions and Guidance on Their Uses. pp. 1-2. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/bulletins/fy2009/09-01.pdf

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proc. of the ACM SIGKDD Workshop on Healthcare Informatics, HI-KDD*.

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*.

Rijsbergen, C. V. (1997). Information Retrieval. 1979. *Btterworths, London*.

Rudd, M. D., A. L. Berman, T. E. Joiner, M. K. Nock, M. M. Silverman, M. Mandrusiak, K. Van Orden, and T. Witte. (2006). Warning signs for suicide: Theory, research, and clinical applications. Suicide and Life-Threatening Behavior 36 (3):255-262.

Skinner, B. F., Whiteley, J. M., & Ratner, H. (1990). *BF Skinner*. American Psychological Association.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 267-273). ACM.

CHAPTER 3

THE EFFECT OF CLIMATE AND SEASONALITY ON DEPRESSED MOOD AMONG TWITTER USERS[3]

[3] Yang, W., L. Mu, and Y. Shen. Accepted by Applied Geography. Reprinted here with permission of publisher.

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. Our results show that the relationship between depression, climate and seasonality is geographically localized. A two-direction stepwise regression is then conducted in each climate zone. We find that 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 rate. We also propose a three-stage framework that semi-automatically detects and analyzes geographically distributed health issues using location-based social media data.

Key words: climate, depression, GIS, seasonality, social media, Twitter

3.1 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 show that it is feasible and cheap to use social media for health study. Lee, et al. (2013) examine influenza spread and compare cancer types at U.S. states level using Twitter. Ghosh, et al. (2013) compared spatial clusters of obesity-related tweets with the

distribution of McDonald's restaurant. Park et al. (2012), De Choudhury, et al. (2013), Yang and Mu (2014) highlighted the potential of using Twitter to detect depression and its geographical pattern. Harman, et al. (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 (Kitchin, 2013; Batty, et al., 2012). 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. (Gorman, 2013; Goodchild, 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; Ruppert, 2013; Manovich, 2011). Third, big data often lack rigorous sampling, documentation, and quality assurance (Kitchin, 2013; Goodchild, 2013; Sui & Goodchild, 2011). This brings difficulty in confirming the validity and accuracy of crowdsourcing (Gorman, 2013). Fourth, although social media have better data 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 geovisual 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.

3.2 The 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, S., Manning, C.M., & Hodge, C., 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, et al. (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, et al. (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 (Nillni, et al., 2009; Magnusson, 2000; Huibers, et al., 2010). 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, et al. (1999) found a significant positive correlation between the prevalence of depression and latitude in North America.

In summary, based on the literatures, 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.

3.3 Methods

3.3.1 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, et al., 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.

3.2.2 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" (Kim, Li, Lebanon, & Essa, 2012; Park, Cha, & Cha, 2012; De Choudhury, Counts, & Horvitz, 2013). 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) (D. D. Lee & Seung, 1999; Yang & Mu, 2015). The basic idea of matrix factorization is that each tweet can be represented as a high-dimensional vector in the space of words, and such high-dimensional data can often be described approximately in a latent subspace with much lower dimensions (Deerwester, et al., 1990). We can interpret the dimensions of the latent subspace as a recurring pattern of word contexts. NMF is a special type of matrix factorization that applies to nonnegative data (Lee & Seung, 1999). It often drives better interpretation of the latent dimensions and is first applied to document clustering by Xu, Liu and Gong (2003).

We collected about 600 million tweets in a one-year time period in the U.S. About one in 1,000 tweets has the keyword "depress" or its variations. And among those one third are expressing depressed feelings. We didn't include re-tweets because the useful tweets are those expressing one's own depressed feelings. With the data collected and filtered, our major goal is to test the hypothesis that the prevalence of depression is comparable in all seasons and climate zones in the U.S.

3.3.3 Spatiotemporal analysis of depression

The study areas are the top 20 and the 34th metropolitan statistical areas (MSAs) in the U.S. measured by 2013 population (Table 3.1). The 34th MSA, San Jose is included because it is a fast growing MSA at the heart of Silicon Valley, the headquarter of hi-tech industry in the U.S.

Considering the special features of geographic information such as mountains and deserts, in a specific season, the climate can vary at a large degree through the whole U.S. Furthermore, seasonality effect on a specific location should also be considered. In order to capture the temporal variation of climate in different U.S. locations, we divided the 21 MSAs into different groups by climate zones. We adopted the Koppen-Geiger climate classification system, which depicts a detailed classification of the climate in the U.S. at county level (Kottek et al., 2006). In this way, we only have Twitter data covering seven climate zones listed in Table 3.2. None of the 21 MSAs is located in Semiarid steppe or Highland (alpine) climate zone.

Table 3.1 Top 20 MSAs and the 34th MSA in the U.S.

Population rank	MSA	Short name
1	New York-Newark-Jersey City, NY-NJ-PA	New York
2	Los Angeles-Long Beach-Anaheim, CA	LA
3	Chicago-Naperville-Elgin, IL-IN-WI	Chicago
4	Dallas-Fort Worth-Arlington, TX	Dallas
5	Houston-The Woodlands-Sugar Land, TX	Houston
6	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	Philadelphia
7	Washington-Arlington-Alexandria, DC-VA-MD-WV	DC
8	Miami-Fort Lauderdale-West Palm Beach, FL	Miami
9	Atlanta-Sandy Springs-Roswell, GA	Atlanta
10	Boston-Cambridge-Newton, MA-NH	Boston
11	San Francisco–Oakland–Hayward, CA	SF
12	Phoenix-Mesa-Scottsdale, AZ	Phoenix
13	Riverside-San Bernardino-Ontario, CA	Riverside
14	Detroit-Warren-Dearborn, MI	Detroit
15	Seattle-Tacoma-Bellevue, WA	Seattle
16	Minneapolis-St. Paul-Bloomington, MN-WI	Minneapolis
17	San Diego-Carlsbad, CA	San Diego
18	Tampa-St. Petersburg-Clearwater, FL	Tampa
19	St. Louis, MO-IL	St. Louis
20	Baltimore-Columbia-Towson, MD	Baltimore
34	San Jose-Sunnyvale-Santa Clara, CA	San Jose

Table 3.2 Climate zones of the continental U.S.

Climate	MSA short name
Humid continental (cool summer)	Minneapolis
Humid continental (warm summer)	St.Louis, Philadelphia, NewYork, Chicago, Detroit, Boston
Humid subtropical	DC, Baltimore, Dallas, Atlanta, Tampa, Houston
Marine westcoast	Seattle
Mediterranean	LA, San Jose, SF
Midlatitude desert	Riverside, San Diago, Phoenix
Tropical (wet/dry) season	Miami

For the analysis of seasonality effect on depression, we grouped Twitter data by different seasons according to the meteorological season calendar (National Climatic Data Center, 2013): spring begins on March 1, summer on June 1, autumn on September 1 and winter on December 1.

We aimed to test the hypothesis that the prevalence of depression is comparable in all seasons and climate zones in the U.S. We calculated the ratio of tweets related to depression in four seasons separately for each of the seven climate zone. In order to further understand how the depression rate changes by season and what local climatic variables are most salient in

explaining the spatiotemporal variations in depression rate in each climate zone, we conducted a stepwise regression (Hocking, 1976) in R (R Development Core Team, 2008) within each climate zone by comparing the Akaike Information Criterion (AIC) (Akaike, 1974) of different models. The dependent variable is the ratio of tweets related to depression. The independent variables (refer to previous sections for the inclusion of these variables) are relative humidity, temperature, sea level pressure, precipitation, snowfall, wind speed, globe solar radiation, and length of day. To mitigate the small number problem, we only select climate zones with at least three MSAs for regression analysis.

Climate data were downloaded from National Oceanic and Atmospheric Administration (NOAA). The temporal scale was month (monthly interval). We used the monthly average values of selected climatic variables from the major city in a MSA as the monthly average climatic measures for that MSA. Such climatic variables include temperature, precipitation, length of day, wind speed, relative humidity, number of rainy days, snow fall, cloudy days, sea level pressure, and daily solar radiation at global. As discussed early, these variables were chosen based on our literature review and data availability (Radua, et al., 2010; Molin, et al., 1996; Partonen & Lönnqvist, 1998; Shapira, et al., 2004; Terman & Terman, 2005; H.-C. Lee, et al., 2007).

3.4. Results

3.4.1 Spatiotemporal patterns of tweets related to depression

We used different colors to render the results (Figure 3.1). Figure 3.1A shows the exact location for each MSA. Figure 3.1C shows the classification of climate zones in the U.S. For better visualization, we created a conceptual compact map (Figure 3.1B). Each grey boundary represents a climate zone. To optimize visualization, we used varied map scales for the MSAs

and only kept the shapes and relative positions within each climate zone. Table 3.2 shows the abbreviation of MSAs in each climate zone.

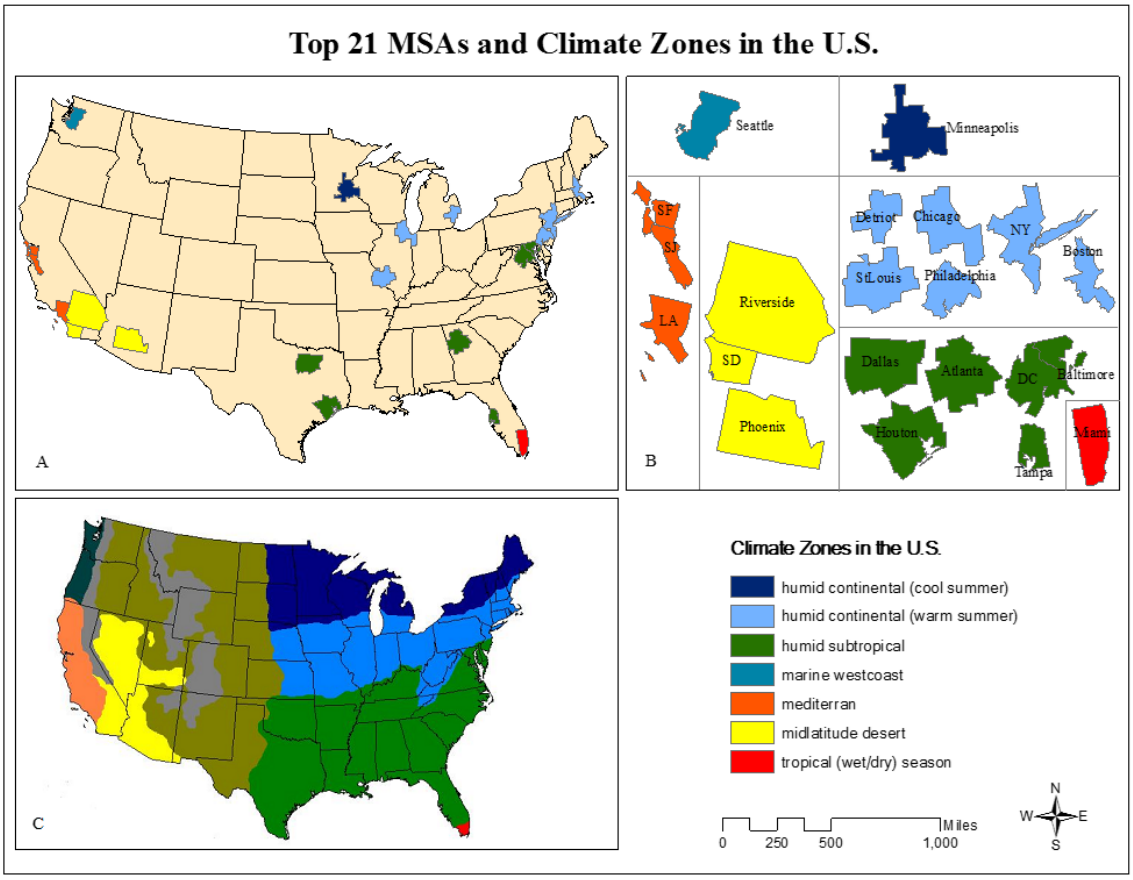


Figure 3.1 Climate zones of the top 21 MSAs in the U.S.

To detect the seasonality effect on the spatiotemporal pattern of depression, we calculated the depression rates (number of tweets expressing depressed feelings per 10,000 tweets) of the four seasons separately for each of the climate zones (Table 3.3). Figure 3.2 shows the variations of depression rates in a one-year time period in the seven climate zones of the U.S. The gradient from red into yellow represents depression rate from higher to lower. We observed that in summer, on the east half of the U.S., the depression rates of different climate zones from the highest to the lowest are: Humid subtropical, Humid continental (warm summer), Humid continental (cool summer). But the sequence is reversed in fall. Nationwide, we found that the

Midlatitude desert climate zone always has the highest depression rate among all climate zones all year round.

Table 3.3 Average rate of tweets related to depression

average rate of depressed tweets (per 10,000 tweets)	spring	summer	fall	winter
Humid subtropical climate	31.16	33.64	24.04	24.45
Humid continental (warm summer) climate	28.07	29.34	26.88	27.85
Mediterranean climate	27.17	28.46	26.97	26.73
Tropical (wet/dry) season climate	26.96	27.34	23.72	28.93
Humid continental (cool summer) climate	30.89	27.67	28.83	25.93
Midlatitude desert climate	36.44	35.04	33.39	32.16
Marine westcoast climate	26.29	24.41	27.00	26.99

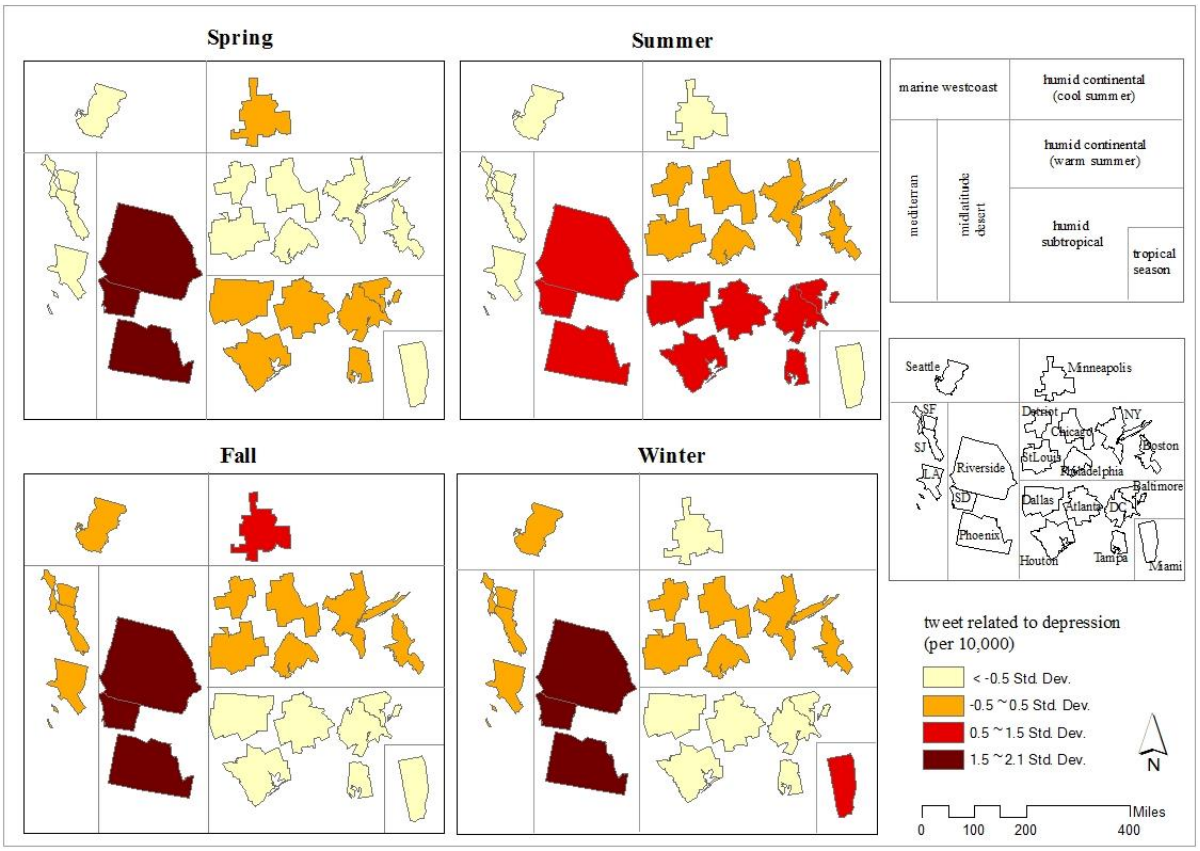


Figure 3.2 Rate of tweets related to depression in different climate zones and seasons

3.4.2 Relationship between climatic factors and depression rate

These diverse results observed in Figure 3.2 shows that it is not appropriate to use a unified model to predict the effect of climate on depression for the whole U.S. In order to detect the geographical localized relationship, we divided the Twitter data and climate data into different climate zones and conducted a stepwise regression analysis. To mitigate the small number problem, we only select four climate zones with at least three MSAs for regression analysis.

We found that the climatic risk factors for depression were different and localized. We summarized the monthly average values (raw data) of selected climatic factors for each climate zone by different seasons in Table 3.4. We listed the typical MSAs in each climate zone with the coefficient correlation and p-value in table 3.5 (data were normalized before conducting the regression analysis). Here, only significant relationships at 0.1 level or stronger were included. A negative signal represents a negative association. A positive signal represents a positive association.

Table 3.4 Average climatic factors in different seasons for each climate zone

	Humid continental (warm summer)		Humid subtropical		Mediterranean		Midlatitude desert	
Relative humidity (%)	63.24	68.27	66.55	69.93	52.90	53.44	54.52	52.71
	68.72	67.05	70.73	68.73	54.63	61.73	53.34	58.42
Temperature (C)	10.73	23.33	16.16	26.19	15.20	19.38	18.58	26.17
	13.31	-0.13	17.74	6.70	17.53	12.13	20.88	13.33
Sea level pressure (mbar)	1015.6	1015.3	1015.7	1015.5	1015.2	1013.1	1013.3	1010.8
	1018.0	1018.5	1017.7	1019.6	1014.5	1018.1	1013.3	1017.7
Precipitation (mm)	89.04	92.86	87.07	107.81	35.74	1.17	19.67	9.72
	82.58	68.91	90.73	69.00	24.18	85.27	14.54	40.11
Snowfall (cm)	3.77	0	1.43	0	0	0	0	0
	0.94	19.62	0.32	8.55	0	0	0	0
Wind speed (km/h)	17.80	14.30	15.56	12.21	13.14	13.77	11.55	11.45
	15.96	18.27	13.11	14.82	10.53	10.12	9.52	8.92
Globe solar radiation (Mj/m^2)	16.31	20.32	18.04	20.81	22.33	26.62	23.39	25.13
	11.15	7.17	13.17	9.27	15.42	10.00	16.36	12.07
Length of day (hours)	13.78	15.07	13.51	14.56	13.56	14.63	13.42	14.38
	11.60	10.37	11.74	10.73	11.70	10.68	11.78	10.87

Note: For each 2x2 small matrix, the legend of season is as follows:

Spring	Summer
Fall	Winter

Table 3.5 Relationship between climate and depression rate

coefficient p-value	Humid continental (warm summer)	Humid subtropical	Mediterranean	Midlatitude desert
Example MSA	Chicago	Atlanta	SF	Phoenix
Relative humidity	0.5234		-0.4433	-0.6543
	1.74e-05***		0.00586**	1.29e-05***
Temperature		-1.194	0.4822	
		5.46e-05***	0.07012#	
Sea level pressure	-0.6229			
	0.011507*			
Precipitation	-0.4571		0.5827	
	0.000731***		0.03800*	
Snowfall	-0.2789	-0.3138		
	0.058290#	0.02697*		
Wind speed	0.6757	-0.3687	-0.462	0.3103
	2.37e-05***	0.00306**	0.01518*	0.0349*
Globe solar radiation	1.777	1.310	0.6126	
	0.017343*	1.01e-08***	0.03330*	
Length of day	-1.810			
	0.018838*			

Note: significant codes 0 to 0.001: '***'; 0.001 to 0.01: '**'; 0.01 to 0.05: '*'; 0.05 to 0.1: '#'. The blanks represent none significant at $\alpha = 0.1$ level.

3.5. Discussion

Relative humidity is calculated using temperature and pressure. MSAs within Humid continental (warm summer) climate zone around the Great lakes are very humid in the whole year and humidity has a positive relationship with depression rate. However, MSAs within Mediterranean or Midlatitude desert climate zone have a much lower humidity level than those within Humid continental (warm summer) climate zone, and humidity has a negative relationship with depression rate. Average relative humidity is significant related to depression rate in these three climate zones just mentioned. To further understand this, take the example MSAs, higher humidity is related to higher depression rate in Chicago, but lower depression rate in San Francisco and Phoenix.

In MSAs within Mediterranean climate, the monthly average value of temperature is around 16^{o}C with small variance. In these MSAs, temperature has a positive relationship with the depression rate. However, in MSAs within humid subtropical climate zone, it is mild to hot in most of the days in a year, and cold in late October until early March. In cities such as Atlanta,

many residents do not have heating systems at home, and may get uncomfortable with freezing weather, thus cause a higher depression rate.

Sea level pressure is significantly (at $\alpha = 0.05$ level) negatively related with depression rate in Humid continental (warm summer) climate zone.

In MSAs within Humid continental warm summer climate zone, precipitation has a negative relationship with depression rate. The temperature is much higher in summer than other seasons, so precipitation can help to lower the temperature at some degree. However, the relationship between precipitation and depression in MSAs within the Mediterranean climate zone are negative.

According to our sample data analysis, snowing has a negative relationship with depression rate in MSAs within Humid continental (warm summer) and Humid subtropical climate zones.

MSAs within Humid continental (warm summer) climate zone are cold and windy in winter. MSAs within Midlatitude desert climate zone may be dusty when it is windy. Thus, people in these two climate zones may feel depressed due to bad weather. Compared with MSAs within Humid continental climate, MSAs within Humid subtropical and Mediterranean climate zones are not that cold in winter. Also, the weed speed in both Humid subtropical and Mediterranean climate zones is much lower. Thus people may feel more comfortable during windy days in those areas.

Globe solar radiation has a significant positive correlation with depression rate in MSA within Humid continental (warm summer) or Humid subtropical or Mediterranean climate zones.

Length of day is measured by hours, and is significant negative related to depression rate in MSAs within Humid continental (warm summer) climate zone. In those MSAs, winter has a much shorter length of day compared with that of summer.

To summarize our results in a nutshell, we have the following three conclusions. Among all eight climatic factors, wind speed plays a significant role in all zones. Sea level pressure and length of the day only matter in a single zone – Humid continental (warm summer). Among those significant climatic variables, snowfall has a negative relationship with depression. Globe solar radiation has a positive relationship with depression. Other six climatic factors show different signs of correlation with depression. In Humid continental (warm summer) climate zone, relative humidity, precipitation and wind speed explain the most variation in depression rate. In Humid subtropical climate zone, temperature and globe solar radiation explain the most. In Mediterranean and Midlatitude desert climate zones, relative humidity plays the most important role.

3.6. Conclusion and Limitation

Intrigued by the facts and discussion from literatures that climate and seasonality may play an important but previously overlooked role in the prevalence of depression, we set out to perform the analysis focus on their relationships using Twitter data. Our results show that the relationship between climate, seasonality and the prevalence of depression are localized and different. Our major contributions are using social media data for depression study to avoid or solve the shortcomings of using traditional data collecting methods for health study discussed in the literature reviews. Besides, social media can protect users from exposing their identities to a large extent, thus our research are better in terms of confidentiality and anonymity. Also, data acquired from social media can be seen as near-real time. Researchers do not have temporal or geographical constrains for collecting data. Thus, it is faster and more cost-effective for analyzing health data. Another important improvement of using location-based social media data is that researchers do not need to input the address of subjects manually into a database and then

geo-reference the location onto a digital map. This can help to avoid unnecessary errors. In addition, the relationship between depression and climate are analyzed at the aggregation level of MSA, which has never before been conducted in depression related research.

As we move further into the era of mobile, social media become a phenomenon and provide a lasting resource for social science. We can collect data from social media, and analyze them. However, social media data are not necessarily social science data (Wilson, 2014). Not all data downloaded from social media are complete, acceptable or accurate. Also, the data sample established on social media are collected in a passive way, thus was very noisy.

In this research, we proposed a framework to conceptualize the procedure of studying geographically distributed health issues using location-based social media data (Figure 3.3). The three-stage framework is a bridge between social media analytics and GIS analysis. Former research using traditional data collecting methods for health study limited in a small-scale area, thus didn't consider the geographical influence on the prevalence of disease. Other research using social media for health study also lack the spatiotemporal analysis of disease.

The first stage of the framework is geographical data acquisition. In this research, we use automated methods to collect geo-tagged tweets in the U.S. The 'First law of GIS management' says you get something for nothing by bringing together geographic information from different sources and using it in combination (Longley, 2005). The number of data combinations and potential uses rise rapidly as we collect more Twitter data. We also add map layers of the study area as base maps.

The second stage is data reduction and text mining. Knowledge discovery, data mining, and semantic analysis are most helpful in this procedure. We use keyword to remove noise, and apply an advanced text mining algorithm to extract useful information and raise accuracy.

The third stage is spatiotemporal analysis of health issues. We visualize the data at different spatiotemporal scales. Based on the first law of geography, we apply GIS representation methods, spatial interpolation and statistics to estimate and visualize values at areas of interest. We also need to consider the spatial heterogeneity when mapping a spatial relationship between two variables.

This framework can help us understand how social and behavioral interventions influence humans' health and illness. This implied that social media may have the potential to transform clinical practice considering some particular disease conditions (Parr, 2004). In addition, this framework can be used to detect major event outbreaks, such as flu and earthquakes.

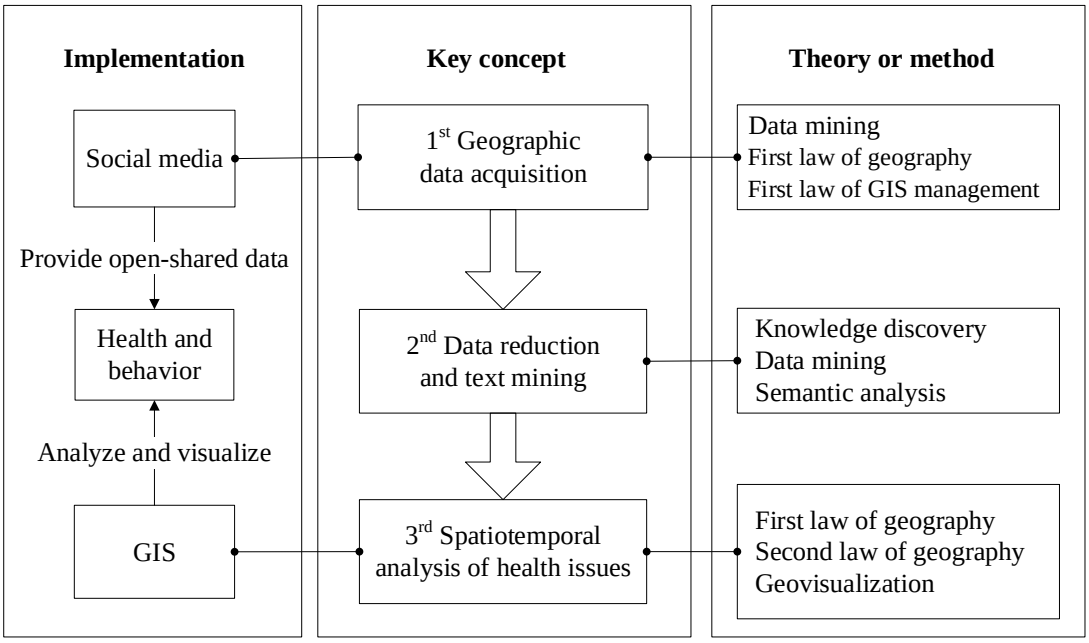


Figure 3.3 Conceptual framework

In our research, the Twitter APIs only allows free access to one percent convenience sampling of tweets. Data acquired are restricted to users with public profiles. We only include tweets written in English and users who identified themselves living in the U.S. Modifying any of those may change the results. In addition, this study only includes tweets posted in the top

MSAs in the U.S. We could not explore any relationship between tweets, seasonality and climate measures in the zones of humid continental (cool summer) climate and marine west coast climate due to limited data. In the future, we will compare Twitter data in multiple years for both urban and rural area. We will collaborate with hospitals and clinics and get approval from real patients to participate in the study.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19* (6), 716-723.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., & Portugali, Y. (2012). Smart cities of the future. *European Physical Journal-Special Topics, 214* (1), 481.

Clayton, S., Manning, C.M., & Hodge, C. (2014) Beyond Storms & Droughts: The Psychological Impacts of Climate Change. American Psychological Association. http://ecoamerica.org/wpcontent/uploads/2014/06/eA_Beyond_Storms_and_Droughts_Psych_Impacts_of_Climate_Change.pdf

Cunsolo Willox, A., Harper, S. L., Ford, J. D., Landman, K., Houle, K., & Edge, V. L. (2012). "From this place and of this place:" Climate change, sense of place, and health in Nunatsiavut, Canada. *Social science & medicine, 75* (3), 538-547.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems* (pp. 3267-3276). ACM.

Ghosh, D., and R. Guha. 2013. What are we 'tweeting'about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science* 40 (2):90-102.

González-Bailón, S. (2013). Big data and the fabric of human geography. *Dialogues in Human Geography, 3* (3), 292-296.

Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in Human Geography, 3* (3), 280-284.

Gorman, S. P. (2013). The danger of a big data episteme and the need to evolve geographic information systems. *Dialogues in Human Geography, 3* (3), 285-291.

Harman, G. C. M. D. C. 2014. Quantifying Mental Health Signals in Twitter. http://www.cs.jhu.edu/~mdredze/publications/2014_acl_mental_health.pdf

Hocking, R. R. (1976). A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 1-49.

Huibers, M. J., de Graaf, L. E., Peeters, F. P., & Arntz, A. (2010). Does the weather make us sad? Meteorological determinants of mood and depression in the general population. *Psychiatry research, 180* (2), 143-146.

Human Geography. (2015). "Theories of Geography." https://humangeography.wikispaces.com/Theories+of+Geography

Kim, S., Li, F., Lebanon, G., & Essa, I. (2012). Beyond sentiment: The manifold of human emotions. *arXiv preprint arXiv:1202.1568*.

Kitchin, R. (2013). Big data and human geography Opportunities, challenges and risks. *Dialogues in Human Geography, 3* (3), 262-267.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift,15(3), 259-263.

Kuang, D., & Park, H. (2013). Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 739-747). ACM, Chicago, Illinois, USA.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401* (6755), 788-791.

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time disease surveillance using Twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1474-1477). ACM.

Lee, H. C., Tsai, S. Y., & Lin, H. C. (2007). Seasonal variations in bipolar disorder admissions and the association with climate: a population-based study.Journal of affective disorders, 97(1), 61-69.

Longley, P. (2005). *Geographic information systems and science*. John Wiley & Sons.

Magnusson, A. (2000). An overview of epidemiological studies on seasonal affective disorder. *Acta Psychiatrica Scandinavica, 101* (3), 176-184.

Manovich, L. (2011). Trending: the promises and the challenges of big social data.

Mersch, P. P. A., Middendorp, H. M., Bouhuys, A. L., Beersma, D. G., & van den Hoofdakker, R. H. (1999). Seasonal affective disorder and latitude: a review of the literature. *Journal of affective disorders, 53* (1), 35-48.

Molin, J., Mellerup, E., Bolwig, T., Scheike, T., & Dam, H. (1996). The influence of climate on development of winter depression. *Journal of affective disorders, 37* (2), 151-155.

Morbidity and Mortality Weekly Report (MMWR). (2010). QuickStats: Prevalence of Current Depression Among Persons Aged ≥12 Years, by Age Group and Sex — United States, National Health and Nutrition Examination Survey, 2007–2010.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*.

National Climatic Data Center. (2013). Meteorological Versus Atronomical Summer – What's the difference？ http://www.ncdc.noaa.gov/news/meteorological-versus-astronomical-summer

Neria, Y., & Shultz, J. M. (2012). Mental health effects of Hurricane Sandy: Characteristics, potential aftermath, and response. *JAMA, 308* (24), 2571-2572.

Nillni, Y. I., Rohan, K. J., Rettew, D., & Achenbach, T. M. (2009). Seasonal trends in depressive problems among United States children and adolescents: a representative population survey. *Psychiatry research, 170* (2), 224-228.

Okawa, M., Shirakawa, S., Uchiyama, M., Oguri, M., Kohsaka, M., Mishima, K., Sakamoto, K., Inoue, H., Kamei, K., & Takahashi, K. (1996). Seasonal variation of mood and behaviour in a healthy middle-aged population in Japan. *Acta Psychiatrica Scandinavica, 94* (4), 211-216.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)* (pp. 1-8).

Parr, H. (2004). Medical geography: critical medical and health geography? *Progress in Human Geography, 28* (2), 246-257.

Partonen, T., & Lönnqvist, J. (1998). Seasonal affective disorder. *The Lancet, 352* (9137), 1369-1374.

Peet, R. (1985). The social origins of environmental determinism. *Annals of the Association of American Geographers, 75* (3), 309-333.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Radua, J., Pertusa, A., & Cardoner, N. (2010). Climatic relationships with specific clinical subtypes of depression. *Psychiatry research, 175* (3), 217-220.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography, 3* (3), 268-273.

Schlager, D., Schwartz, J. E., & Bromet, E. J. (1993). Seasonal variations of current symptoms in a healthy population. *The British Journal of Psychiatry, 163* (3), 322-326.

Shapira, A., Shiloh, R., Potchter, O., Hermesh, H., Popper, M., & Weizman, A. (2004). Admission rates of bipolar depressed patients increase during spring/summer and correlate with maximal environmental temperature. *Bipolar disorders, 6* (1), 90-93.

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science, 25* (11), 1737-1748.

Terman, M., & Terman, J. S. (2005). Light therapy for seasonal and nonseasonal depression: efficacy, protocol, safety and side effects. *CNS spectrums, 10* (8), 647.

Wilson, M. W. (2015). Morgan Freeman is dead and other big data stories.cultural geographies, 22(2), 345-349.

Winkler, D., Willeit, M., Praschak-Rieder, N., Lucht, M., Hilger, E., Konstantinidis, A., Stastny, J., Thierry, N., Pjrek, E., & Neumeister, A. (2002). Changes of clinical pattern in seasonal affective disorder (SAD) over time in a German-speaking sample. *European archives of psychiatry and clinical neuroscience, 252* (2), 54-62.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (pp. 267e273). ACM.

Yang, W., & Mu, L. (2015). GIS analysis of depression among Twitter users. *Applied Geography*. Volume 60, June 2015, Pages 217-223.

Zung, W. W., & Green, R. L. (1974). Seasonal variation of suicide and depression. *Archives of general psychiatry, 30* (1), 89-91.

CHAPTER 4

TEMPORAL ANLAYSIS OF DEPRESSION AMONGST TWITTER USERS: FROM THE VIEW OF HOLIDAY, WEEKDAY AND DAILY PATTERN[4]

[4] Yang, W., L. Mu, and Y. Shen. To be submitted to American Journal of Public Health.

Abstract

Depression is a common chronic disorder and it has a high prevalence in the U.S. Few studies have examined relationships between traditional holidays, weekdays and their effects on depression. In this paper, we apply text mining techniques to social media data to provide new perspectives for public health research. We design a procedure to automatically detect depressed mood in Twitter and analyze their temporal patterns using statistical methods. We compare the depression rates during holidays to those on non-holidays in the U.S. using Twitter data. We then set out to compare the daily depression rates from tweets by days in a week. We also investigate the time differences of posting tweets in a day for both depressive users and normal users on Twitter. Our method can improve diagnosis techniques for depression. It is faster at collecting data and more promptly at analyzing and providing results. Also, this method can be expanded to detect other major events in real-time, such as flu outbreaks and earthquakes.

Key words: depression, tweets, temporal pattern, holiday, weekday

4.1. Introduction

It is believed that Christmas and other holidays are associated with an increased incidence of psychopathology (Hillard, Holland, & Ramm, 1981). Urban legend suggests that psychopathology tends to increase around the holidays. Some research suggested although fewer people utilize emergency services or attempt suicide during holiday season, there is an increase in certain other kinds of psychopathology, including mood disorders such as dysphoria, depression and substance abuse (Sansone & Sansone, 2011).

Depression is a common chronic disorder and it has a high prevalence in the U.S. Few studies have examined relationships between traditional holidays, weekdays and their effects on depression. Depression may occur at any time of the year, but the stress and anxiety of the

holiday season may cause even those who are usually content to experience loneliness and a lack of fulfillment, especially during the months of November and December, and the period before Valentine's day (Domingo-Salvany, 2008). One of the most relevant factors in holiday depressions is that those who suffer from such kind of depression tend to believe that everyone else is having a good time and engaged in loving family relationships, which are not necessarily a fact. Velamoor, Voruganti, & Nadkarni (1999) found that the most common stressors reported by depression patients in a psychiatric emergency service during the Christmas season were loneliness and being without a family. In describing their feelings about the holiday, most participants used the phrase, "depressed".

A study published in 2011 in the Journal of Positive Psychology (Stone, A. A., Schneider, S. & Harter, J. K., 2012), used a Gallup poll of 340,000 adults who were asked to describe their mood from the day before. The weekday-to-weekend effect shown that retired people felt a less drastic mood shift than those still working. These results add to the wealth of research that show work contributes to low moods.

In a previous study that analyzed more than 500 million tweets sent worldwide, researchers (Golder & Macy, 2011) found people expressed the most negative emotions on Tuesdays and the most positive on Sunday mornings. On any given day, positivity peaked in the mornings, waned during the day, then both positivity and negativity rose again after about 6 p.m.

Intrigued by the fact and discussion on those literatures, we have three research questions. First, we compare the depression rates during holidays to those on non-holidays in the U.S. using Twitter data. Second, we set out to compare the daily depression rates from tweets by days in a week. Third, we investigate the time differences of posting tweets in a day for both depressive users and normal users on Twitter.

4.2 Methods

4.2.1 Geographic data acquisition

In Chapter 3, we proposed a three-step framework for geographical analysis of health issues using location-based social media data. In this study, we followed the framework using Twitter data to check the temporal pattern of depressed mood.

In this research, we downloaded more than 300 million tweets in the U.S. from 2013/09/05 to 2014/03/05 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, et al., 2013). We then extracted tweets posted in New York Metropolitan Statistical Area (NY MSA) for this study (Figure 4.1).

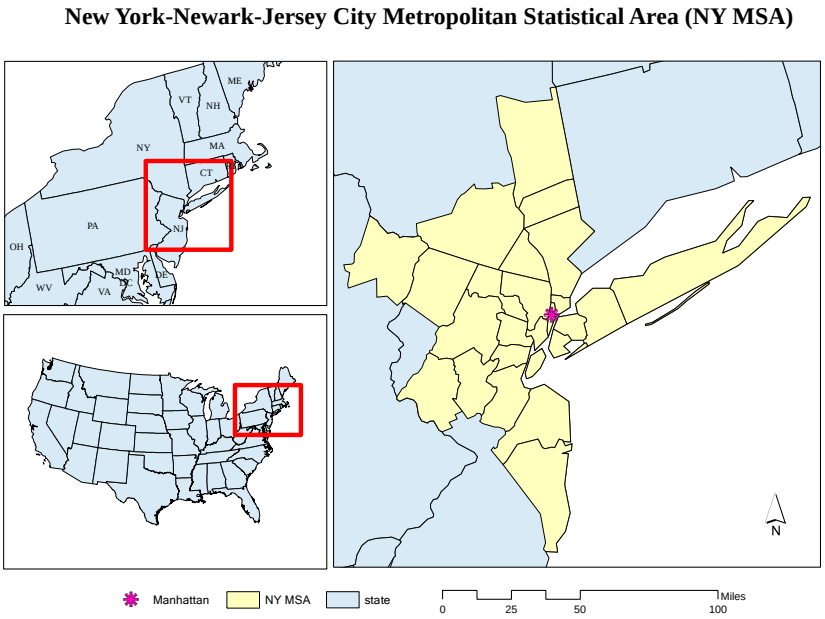


Figure 4.1 The study area – NY MSA

4.2.2 Data reduction and text mining

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” (Kim, Li, Lebanon, & Essa, 2012; Park, Cha, & Cha, 2012; De Choudhury, Counts, & Horvitz, 2013). 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) (D. D. Lee & Seung, 1999; Yang & Mu, 2015). The basic idea of matrix factorization is that each tweet can be represented as a high-dimensional vector in the space of words, and such high-dimensional data can often be described approximately in a latent subspace with much lower dimensions (Deerwester, et al., 1990). We can interpret the dimensions of the latent subspace as a recurring pattern of word contexts. NMF is a special type of matrix factorization that applies to nonnegative data (Lee & Seung, 1999). It often drives better interpretation of the latent dimensions and is first applied to document clustering by Xu, Liu and Gong (2003).

Using text clustering to identify the word context for the depressed mood is not sufficient to detect the depressive Twitter users. Thus, we added the diagnostic and statistical manual of mental disorders (DSM-IV) criteria (American Psychological Association, 2014) for MDD as a time criteria to detect depressive Twitter users (Figure 4.2).

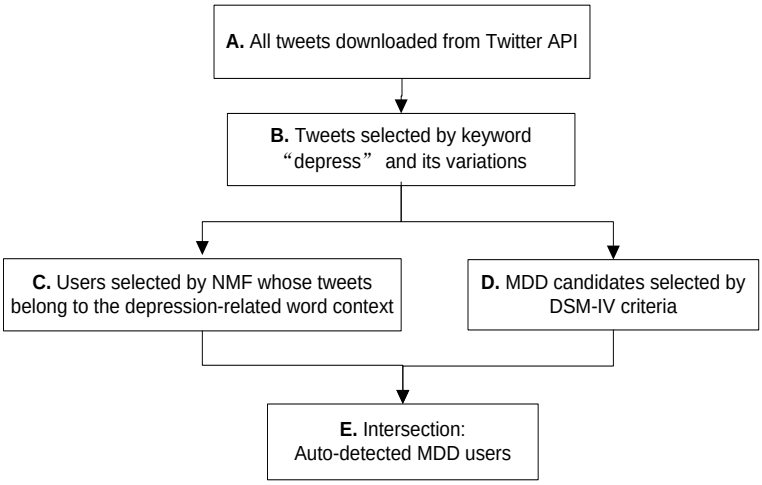


Figure 4.2 Workflow of detecting depressive users

The DSM-IV criteria require that MDD patients must suffer from at least five of the nine typical depression symptoms for more than two weeks (Skinner, Whiteley, & Ratner, 1990). These specific symptoms include depressed mood, decreased interest in daily activities, significant weight change or change in appetite, change in sleep or activity, fatigue, feelings of guilt or worthlessness, loss of concentration and having suicide plan. However, when analyzing Twitter text, we only focused on the depressed mood of the users. As a result, in order to apply DSM criteria to identify MDD candidates, we made the following change to the criteria: We required five or more tweets that are associated with depressed mood within a two-week period.

4.2.3 Temporal analysis of depression

To compare the depression rate in both holidays and non-holidays, we selected the traditional U.S. holidays according to the U.S. Holidays Calendar for the year 2013 to 2014 and only used Twitter data from 2013/09/05 to 2014/03/05 since most U.S. national and traditional holidays are concentrated in this half year time period (Table 4.1). However, two major holidays, Memory Day and Independent Day were not included in this research.

To define holidays and non-holidays time periods in this research, a flexible temporal buffer was made around the exact holiday depending on the weekday and the length of the holiday. For Thanksgiving, Christmas and Valentine's Day, we set a week-long buffer before the holiday. For holidays on Monday or Friday, we added the connective weekends into the holiday. Specifically, we included the first day for work after each holiday into the holidays to examine the post-holiday mood. In this way, 43 days out of 180 days were within the holidays time period we defined from 2013/09/05 to 2014/03/05. We set up the null hypothesis that the daily depression rates during holidays and non-holidays have equal mean values and comparable distributions.

Table 4.1 Definition of holidays

Date	**Weekday**	**Holiday Name**	**Holiday Type**	**Holidays**
2013/10/14	Monday	Columbus Day	National holiday	2013/10/12-2013/10-15
2013/10/31	Thursday	Halloween	Observance	2013/10/31
2013/11/11	Monday	Veterans Day	National holiday	2013/11/09-2013/11/12
2013/11/28	Thursday	Thanksgiving Day	National holiday	2013/11/21-2013/12/02
2013/12/25	Wednesday	Christmas Day	National holiday	2013/12/16-2013/12/26
2014/01/01	Wednesday	New Year's Day	National holiday	2014/01/01
2014/01/20	Monday	Martin Luther King Day	National holiday	2014/01/18-2014/01/21
2014/02/14	Friday	Valentine's Day	Observance	2014/02/10-2014/02/17
2014/05/26	Monday	Memorial Day	National holiday	Not included
2014/07/04	Friday	Independence Day	National holiday	Not included

To compare the daily depression rate during holidays and non-holidays, we divided data into holidays and non-holidays group as shown in Table 4.1. We then followed the workflow in Figure 4.3 to test our data. We first examined the mean and variances values and the distribution of depression rates separately in both groups. We further adopted Shapiro-Wilk normality test (Shapiro & Wilk, 1965) to check whether the distributions of daily depression rates during holidays and non-holidays follow a normal distribution. If the distribution follows a normal

distribution, we will use Welch Two Sample t-test (Equation 4.1) (Welch, 1947) to test whether the depression rates of the two time periods have comparable means:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{s_1}^2}{N_1} + \frac{{s_2}^2}{N_2}}} \qquad \text{Equation 4.1}$$

Where $\bar{X}_1$, ${s_1}^2$, and N_1 are the 1st sample mean, sample variance and sample size, respectively. If the data does not follow a normal distribution, we will adopt Mann-Whitney U test (Equation 4.2) (Mann & Whitney, 1947) to test whether the depression rates of the two time periods have comparable medians.

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad \text{Equation 4.2}$$

Where n_1 is the sample size for sample1, and R_1 is the sum of the ranks in sample1. To calculate R_1, we first assigned numeric ranks to all the observations, beginning with 1 for the smallest value. Where there are groups of tied values, assign a rank equal to the midpoint of unadjusted ranking [e.g., the ranks of (1, 3, 3, 5) are (1, 2.5, 2.5,4)]. We then added up the ranks for the observations which came from sample 1. The sum of ranks in sample 2 is now determinated, since the sum of all the ranks equals N(N+1)/2 where N is the total number of observations. It does not matter which of the two samples is considered sample 1. The smaller value of U_1 and U_2 is the one used when consulting significance tables.

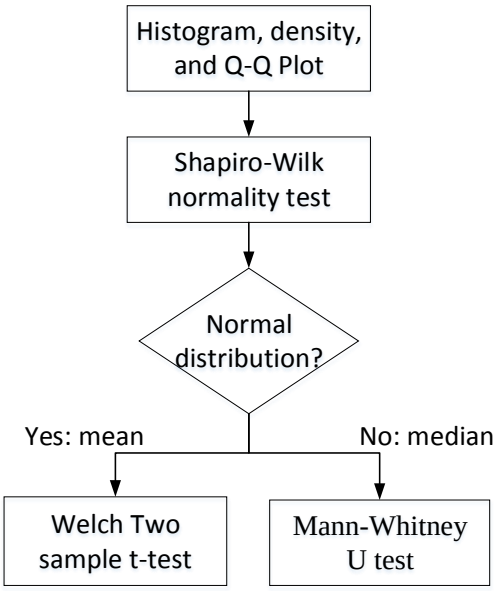


Figure 4.3 Workflow of using statistical methods to test null hypothesis

To examine the variation of depression rate in a week, we divided the dataset into seven groups by day in the week, from Monday to Sunday. We set up the null hypothesis that the mean and variation values of depression rate in each day of a week are comparable. We first calculated the mean and variation values of each groups, and then used box plot to visualize the distribution of depression rate in days of a week. Analysis of variance (ANOVA) test (Anscombe, 1948) was used to analyze the differences of daily depression rates between the mean and variation values of the seven groups. Welch Two sample t-test was further used to compare the significant differences in mean values of depression rate between weekdays and weekends, and between Saturday and Sunday.

To compare the daily temporal pattern difference of posting tweets for both depressive users and normal users on Twitter, we first divided the whole dataset of users into two groups: major depressive disorder (MDD) group and normal group using Non-negative Matrix Factorization methods (Yang and Mu, 2014). We divided the time period in a day into 24 time slots, each slot is one hour, started from 0am to 1am. For each group, we summed the number of tweets posted

in each time slot for the half a year time period. We then plotted the rate of posting tweets in each time slot for both depression group and normal group separately, which equal to the number of tweets posted in a time slot divided by the total number of tweets posted in the whole time period for that group.

4.3 Results and Discussion

4.3.1 Holidays and non-holidays temporal pattern

Our results help us know better about the temporal pattern of depression during holidays and non-holidays. The basic statistics are summarized as Table 4.2. The mean value of depression rate during holidays is 29.33 (number of tweets related to depression per 10,000 tweets), while the mean value during non-holidays is 26.77. The variance value of depression rate during holidays is 30.02, while the variance value during non-holidays is 28.94. The workflow of using statistical methods to test the hypothesis is as Figure 4.2.

Table 4.2 Summary of depression rates

	Mean	Median	Var.	Std. Dev.	Pop.	Confidence Interval	Shapiro-Wilk p-value
holiday	29.33	29.16	30.02	5.48	43	27.24 – 31.44	0.644
non-holiday	26.77	26.04	28.94	5.38	137	24.70 – 28.84	0.030

The Shapiro-Wilk normality test results show that the distribution of daily depression rate in holidays follows a normal distribution (W = 0.9799, p-value = 0.644), while the distribution of daily depression rate in non-holidays doesn't follow a normal distribution (W = 0.9787, p-value = 0.030). We can see this point directly from the histogram, density and Q-Q Plot of daily depression rate in holidays and non-holidays (Figure 4.4).

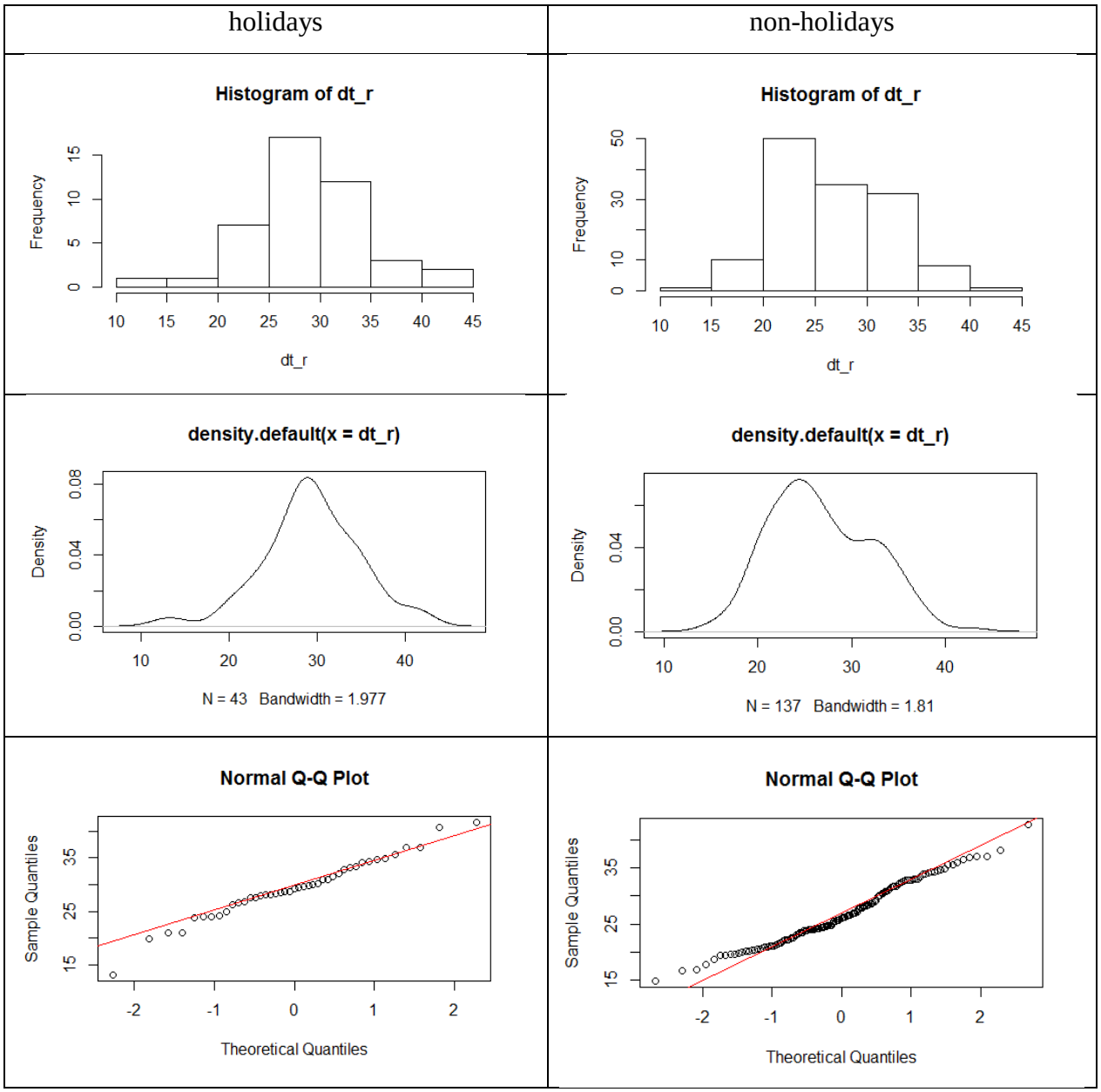


Figure 4.4 Histogram, density and Q-Q Plot of depression rates

Both the means and variances of the two data sample are unequal. The results of Welch Two sample t-test (t = -2.6879, p = 0.009) shows that the true difference in the mean values of depression rates during holidays and non-holidays is not equal to 0, and this difference is significant.

We find that the median value of daily depression rates during holidays is about 29.16, while the median value of daily depression rates during non-holidays is 26.04. The results of

Mann-Whitney U test (W = 2093, p = 0.004) shows that the true location shift of daily depression rate of the holidays and non-holidays is not equal to 0, and this difference is significant. So our results show that the depression rate during holidays is significant higher than that during non-holidays.

Maybe part of the problem is the bombardment of media during the holidays showing images of smiling families and friends on Twitter? People who feel lonely and being without a family may start to question the quality of their own relationships. A number of other factors, such as unrealistic expectations, financial pressures, and too many commitments, can also cause stress and anxiety at holiday time. Headaches, excessive drinking, overeating, and insomnia are some of the possible consequences of poorly managed holiday stress too.

4.3.2 Weekly temporal pattern

We then used ANOVA test to see whether the depression rate is comparable during weekday and weekends. Table 4.3 shows the mean, median, and standard deviation of tweets related to depression (per 10,000 tweets). We also summarized the number of population, confidence intervals and the p-values of shapiro-Wilk test from Monday to Sunday in this half an year time period. In Figure 4.5, we observed that Sunday has the highest depression rate, followed by Monday and Tuesday. The depression rate attempts tend to drop off Started from Tuesday till Saturday. Saturday has the lowest depression rate.

Table 4.3 Statistics of depression rates in a week

Weekday	Mean	Median	Std. Dev.	Pop.	Confidence Interval	Shapiro-Wilk p-value
Monday	28.72	29.00	4.13	25	27.10 – 30.34	0.248
Tuesday	28.24	28.00	4.76	25	26.37 – 30.11	0.572
Wednesday	25.96	25.00	5.43	25	23.83 – 28.09	0.080
Thursday	26.77	25.50	5.64	26	24.60 – 28.94	0.019
Friday	26.96	25.50	5.53	26	24.84 – 29.09	0.032
Saturday	25.23	25.00	5.72	26	23.03 – 27.43	0.594
Sunday	29.54	28.50	5.92	26	27.26 – 31.81	0.988

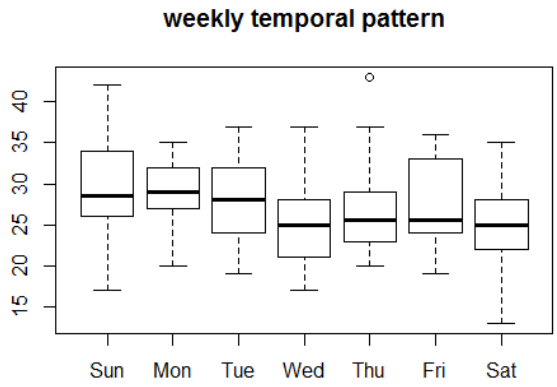


Figure 4.5 Box plot for each day of a week

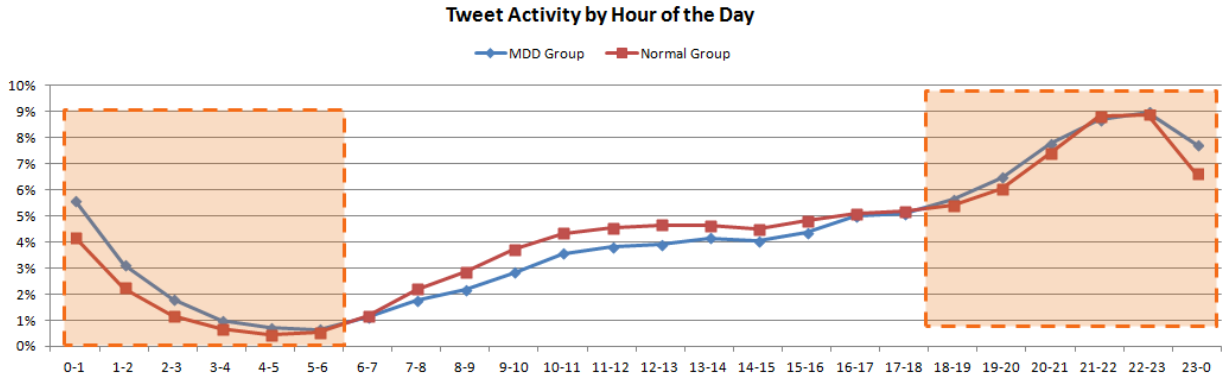
The result of ANOVA test (F-value = 0.671) shows the mean and variation values of depression rate in each day of a week is significant different. With the extreme high depression rate on Sundays and low depression rate on Saturdays, we further examined the difference in mean values of Saturday and Sunday using Welch Two Sample t-test. The result (p-value = 0.0103) shows the true difference in means is not equal to 0. This results show the depression rate on Sundays is significant higher than that on Saturdays.

Our research find that weekly temporal pattern of the depression rate can be divided as two major groups: people were no more glum on Sunday than they were on Monday, or Tuesday. They were significantly more chipper on Wednesday, Thursday, Friday, and Saturday, with significant difference in depression levels between the weekend days. Monday gets a bad rap as being the week's most miserable day, which may be the major reason of the extreme high

depression rate on Sundays. The anticipation of a bad day is sometimes the worst part about it. Also, the highest depression rate on Sundays could have something to do with drinking too much the night before than anything else, since Saturday has the lowest depression rate. Usually, people are busiest on Mondays and Tuesdays at work, so that may be a cause of the higher depression rate on those two days. Wednesdays and Saturdays seem to be appealing to most people. The depression rate on Friday is a little higher than those on Tuesdays and Saturdays. Maybe people need to finalize their work for the entire week, as well as plan for the activities during the weekends.

4.3.3 Daily temporal pattern

Our results show that MDD group post less tweets than normal group from 6 am in the morning to 6 pm in the evening (Figure 4.6). MDD group post more tweets than normal group in the other time periods, especially between 11 pm in the evening and 3 am in the next morning. One reason may be people suffering from depression often feel lonely and helpless at night. Also, the relationship between sleep and depression illness is complex – depression may cause sleep problems and sleep problems may cause or contribute to depressive disorders. Sleep problems are also associated with more severe depressive illness (Schwartz, Kohler, & Karatinos, 2005).



Note: 0-1 Rate = sum of tweets posted between 0 am –1 am / total tweets
Figure 4.6 Twitter activity by hour of the day

4.4 Conclusion and Limitation

In our research, an advanced text mining method – non-negative matrix factorization, is couple with other statistical methods to detect the temporal pattern of depression rate on Twitter in NY MSA. We find that the depression rate is significantly higher during holiday seasons than normal days. Sundays have the highest depression rate, while Saturdays have the lowest rate, and this difference in depression rate is significant. For depressive users, they post more between 6pm to 6am in any given day, especially from 11pm to 3am, compared to normal users. Our results can help people better recognize the forces behind our mood patterns and make changes to reclaim the day. Being aware of what is likely to affect you so that you do not have to be the passive recipient of life's experiences.

In our research, the Twitter APIs only allows free access to one percent convenience sampling of tweets. Data acquired are restricted to users with public profiles. Also, we only include tweets written in English and users who identified themselves living in the New York MSA, U.S. Furthermore, we only use Twitter data in a half year. Modifying any of those may change the results. In the future, we will run the statistical model for multiple years and consider days having some major event happening to test how the results might change. At this stage, we are not sure whether these statistical results big enough to matter to a patient. In the future, we plan to assess whether statistical significance detected here can lead to clinical significance of depression. We will collaborate with hospitals and clinics and get approval from real patients to participate in the study.

References

American Psychiatric Association. (2014). DSM. http://www.psychiatry.org/practice/dsm

Anscombe, F. (1948). The validity of comparative experiments. Journal of the royal statistical society. series A (General), 111 (3), 181-211

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems* (pp. 3267-3276). ACM.

Domingo-Salvany, A. (2008). The science of real-time data capture: self-reports in health research. *Journal of Epidemiology and Community Health, 62* (5), 471-471.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333* (6051), 1878-1881.

Hillard, J. R., Holland, J. M., & Ramm, D. (1981). Christmas and psychopathology: Data from a psychiatric emergency room population. *Archives of general psychiatry*, *38* (12), 1377-1381.

Kim, S., Li, F., Lebanon, G., & Essa, I. (2012). Beyond sentiment: The manifold of human emotions. *arXiv preprint arXiv:1202.1568*.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401* (6755), 788-791.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics,* 50-60.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)* (pp. 1-8).

Sansone, R. A., & Sansone, L. A. (2011). The christmas effect on psychopathology. *Innovations in clinical neuroscience, 8* (12), 10.

Schwartz, D. J., Kohler, W. C., & Karatinos, G. (2005). Symptoms of depression in individuals with obstructive sleep apnea may be amenable to treatment with continuous positive airway pressure. *CHEST Journal, 128* (3), 1304-1309.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 591-611.

Skinner, B. F., Whiteley, J. M., & Ratner, H. (1990). *BF Skinner*. American Psychological Association.

Stone, A.A., S. Schneider, and J.K. Harter. (2012). Day-of-week mood patterns in the United States: On the existence of "blue Monday," "thank God it's Friday," and weekend effects. *Journal of Positive Psychology* 7(12):306-314.

Velamoor, V. R., Voruganti, L. P., & Nadkarni, N. K. (1999). Feelings about Christmas, as reported by psychiatric emergency patients. *Social Behavior and Personality: an international journal, 27* (3), 303-308.

Welch, B. L. (1947). The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 28-35.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 267-273). ACM.

Yang, W., & Mu, L. (2015). GIS analysis of depression among Twitter users. *Applied Geography*. Volume 60, June 2015, Pages 217-223.

CHAPTER 5

CONCLUSIONS

5.1. Summary and Conclusion

In this dissertation, we have three major research objectives. The first is to identify depressive users portrayed in Twitter and detect their spatial patterns. We then find the association between depression rate and environment socioeconomic status. The second objective is to detect spatiotemporal patterns for depression rate among Twitter users and probe into the effect of climate and seasonality on depression. The third objective is to detect temporal patterns for depression rate among Twitter users, and investigate the effect of U.S. holidays and weekdays on depression rate.

Regarding the first objective, we present a new method for public health research combining GIS with social media. Compared with traditional data collection method, our automated method for detecting MDD users is faster and cheaper for analysis and diagnosis. The system can be applied to some online forum for detecting depression topic and forwarding related questions to psychiatrists. Our GIS results also provide novel knowledge about this disorder by examining the geographic clustering of MDD users and relationship with SES.

Regarding the second objective, intrigued by the facts and discussion from literature that climate and seasonality may play an important but previously overlooked role in the prevalence of depression, we set out to perform the analysis focus on their relationships using Twitter data. We downloaded more than 600 million tweets in a one-year time period in the U.S. In the selected 21 MSAs (top 20 and San Jose MSA) across the U.S., tweets are analyzed at the

aggregation level of MSA, which has never before been conducted in depression related research. Our results show that the relationship between climate, seasonality and the prevalence of depression are localized and different.

Regarding the third objective, we find that the depression rate is significantly higher during holiday seasons than normal days amongst Twitter users in the U.S. Sundays have the highest depression rate, while Saturdays have the lowest rate, and this difference in depression rate is significant. For depressive users, they post higher rate between 6pm to 6am in a day compared to normal users, especially from 11pm to 3am. Our results can help people better recognize the forces behind our mood patterns and make changes to reclaim the day. Being aware of what is likely to affect you so that you do not have to be the passive recipient of life's experiences.

In this dissertation, we also propose and demonstrate the feasibility of a three-stage framework which can semi-automatically detect and analyze geographically distributed health issues using location-based social media data. Social media data can help us gain new, interesting and quite different insights into a specific mental illness, which lie beyond the reach of clinical and social psychology. This implied that social media may have the potential to transform clinical practice because it offers a way to think beyond the box considering some particular disease conditions (Parr, 2004). In addition, this framework can be used to detect other major event outbreaks, such as earthquakes.

In our research, our method can improve diagnosis techniques for depression, however, further detailed clinical contexts are still needed to make a formal diagnosis. Future study should probe into the difference between depression detected online and self-reported depression reported by a professional clinical scale table. The Twitter APIs only allows free access to one percent convenience sampling of tweets. Data acquired are restricted to users with public

profiles. Also, we only include tweets written in English and users who identified themselves living in the U.S. Modifying any of those may change the results. In addition, this study only includes tweets from a few MSAs in each climate zone. Among the seven climate zones in the continental U.S., we could not explore any relationship between tweets and seasonality and climate measures in the zones of humid continental (cool summer) climate and marine west coast climate due to limited data. In the future, we will include all the MSAs in all climate zones to test how the results might change.

5.2. Future Research

For future study, my longstanding goal is to bridge the fields of social media analytics and GIScience, to develop tools that combine the best of both world – a virtual world of online networks and real-space of human activity, and use GIS to analyze spatiotemporal patterns of big data. I have taken the first important steps toward this goal with my thesis research. I want to broaden and deepen this investigation, and extend my work to more domains such as social and behavioral sciences and human geography. Initially, I will focus on the following three research direction.

I plan to use crowdsourcing and VGI for other public health research, not only focus on depression. One example is to study the relationship between human mobility and the spatial spreading of infectious disease. Human interaction and mobility processes are two fundamental aspects to describe a disease spread (Meloni, et al., 2011). With wireless service and location-based techniques, social media can capture the tow aspects perfectly, and provide massive data at different levels of spatiotemporal scales.

I am also interested in geographical topic discovery and comparison. Many interesting concepts, including cultures, scenes, and product sales, corresponding to specialized

geographical distributions. I plan to discover different topics of interests that are coherent in geographical regions. A more interesting thing is to compare several topics across different geographical locations (Yin, Cao, Han, Zhai, & Huang, 2011). People would like to know which medical products are more popular in different regions. Sociologists may want to know the difference opinion towards using certain kinds of drugs across different areas, we can map topics of interests into their geographical distributions and develop effective methods to compute such comparison.

Another interesting research direction is using natural language for spatial query. Many of the web-based mapping tools such as Bing, Google do not identify topological relationships descried by spatial prepositions such as in, on, near. They typically return results based upon bounding boxes or an address geo-coding attribute such as place name (Lampoltshammer, 2012). If we type “Fine the hospital in Athens”, computers have difficulties to understand its meaning within such as spatial search. I plan to use natural language as a search mechanism to identify topological relationships for web-based mapping.

References

Lampoltshammer, T. J. (2012). Natural language processing in geographic information systems-some trends and open issues. *International Journal of Computer Science & Emerging Technologies, 3* (3).

Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports, 1*.

Parr, H. (2004). Medical geography: critical medical and health geography? *Progress in Human Geography*, *28* (2), 246-257.

Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T. (2011). Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web* (pp. 247-256). ACM.

APPENDIX I

LIST OF ACRONYMS

Acronym	Full description
3D	Three Dimensional
AAS	American Association of Suicidology
API	Application Programming Interface
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition
FN	False Negative
FP	False Positive
GIS	Geographic Information System
MDD	Major Depressive Disorder
MSA	Metropolitan Statistical Areas
NMF	Non-negative Matrix Factorization
NOAA	National Oceanic and Atmospheric Administration
OAuth	Open Standard for Authorization
SAD	Seasonal Affective Disorder
SES	Socioeconomic Status
TN	True Negative
TP	True Positive