IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

[1]Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A systematic spatial and temporal sentiment analysis on Geo-tweets

Tao Hu[1,2], Bing She[3], Lian Duan[4,5*], Han Yue[6,*], Julaine Clunis[7]

[1]Center for Geographic Analysis, Harvard University, Cambridge, MA 80305 USA
[2]Geocomputation Center for Social Science, Wuhan University, Hubei, China 430079 China
[3]Institute for social research, University of Michigan, Ann Arbor, MI 48109 USA
[4]Natural Resource and Surveying School, Nanning Normal University, Guangxi, 530001 China
[5]Laboratory of Environment Change and Resources Use in Beibu Gulf, Ministry of Education, Nanning Normal University, Guangxi, 530001 China
[6]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Hubei, China 430079 China
[7]School of Information, Kent State University, OH 44240 USA

Corresponding author: Lian Duan (email: wtusm@163.com) and Han Yue (email: hanyue.geo@gmail.com).

ABSTRACT Sentiment affects every aspect of people's lives and has strong impact on their mental health. This paper explores the sentiments extracted from Geo-tweets data from January to December 2016, analyzed in the spatial and temporal perspective. Considering the noisy data from the Geo-tweets, a workflow is created for extracting the tweets which are posted by local users. The workflow makes it convenient for other researchers to reproduce, replicate or extend the procedures using similar Geo-tweet dataset. The workflow is available at Harvard Dataverse. Using the cleaned data, each tweet's sentiment is classified according to the content. Then, the overall temporal variations of total number of positive, neural, and negative sentiments are analyzed on a monthly, daily and hourly level. From a spatial perspective, the Local Indicators of Spatial Association (LISA) statistical method is employed to discover the spatial clusters. In order to explore the content of positive sentiments, this paper applies the Latent Dirichlet Allocation (LDA) model to classify the Geo-tweets with positive sentiments into different topics. Combining the geospatial information with the topics, some patterns are found which demonstrate the associations between the nature of Twitter content and the characteristics of places and users. For example, weekend events and friend and family gatherings are the time that users prefer to post positive tweets. In the western part of US, users tend to post more photos to share the great moment on Twitter than other parts of the US.

INDEX TERMS Geo-tweet, sentiment, spatial analysis, temporal analysis, health

I. INTRODUCTION

Sentiment analysis is a popular study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [1]. It usually classifies users' attitude into positive, negative or neutral categories based on textual data [2]. Emotions have been shown to play a critical role in health outcomes [3]. For example, negative emotions are tied to self-reports of increased pain, fatigue, and disease [4], whereas positive emotions are tied to decreased pain, fatigue, and disease [5]. Thus, exploring the diversity and depth of people's sentiments is an important research area because of its potential impact on people's health.

Statistics show that one-third of people with a social media profile use Twitter, with 75% accessing it from a handheld device to convey an opinion[6]. The usefulness of Twitter as a tool for sentiment analysis has been verified by comparison to more traditional metrics [7], such as polls [8], and stock market performance [9]. Numerous studies on positive sentiment are published every year, however few studies focus on negative sentiment. The Gallup Report showed that the Negative Experience Index (NEI) score in 2017 had reached new highs. The index includes negative emotions and experiences. In 2018, an investigation about people's feelings on the previous day was reported. Most Americans (55%) experienced stress throughout much of the day, nearly

[1] This work is supported by National Natural Science Foundation of China (No. 41961062) and Natural Science Foundation of Guangxi Province (No. 2018JJA150089) and the each funding program's PI is Lian Duan.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

half (45%) said they felt worried a lot, and more than one in five (22%) said they felt anger a lot [10]. Moreover, in the investigation, younger age and lower income were found to play an important role with worry and stress. Additionally, some researchers found depression and other mental health issues can contribute to digestive disorders, trouble sleeping, lack of energy, heart disease, and other health issues.

Sentiments and space are fundamentally connected: locations have an atmosphere which can evoke strong and diverse emotions in people; places can inspire such sentiments as boring, attractive, calming, scary or dangerous and the loss of a place can be an emotional experience [11]. Geo-tweets are tweets which contain a pair of geographic coordinates from the originating device denoting the location at which the tweet was created.

This research systematically analyzes the spatial and temporal features of sentiments located in the United States. Studying the characteristics of content in the tweets has become important for a number of tasks, such as breaking news detection, personalized message recommendation, friend recommendations, sentiment analysis and others [12]. Assessing the geographical information with the content, will be more helpful for analyzing interesting or trending topics based on region. The paper is organized as follows: section 2 summarizes related work; section 3 describes the data and methodology used in the case study; section 4 presents the results of analysis and includes a comprehensive discussion of the findings. Finally, the conclusion of this study and future work are described.

II. RELATED WORK

A. TWITTER DATA APPLICATIONS

Twitter data has been applied in diversities of research areas, including disaster, crime, health, marketing, and so on. Some researchers focus on identifying the polarity of sentiments expressed by Twitter users during disaster events and explore user's sentiments change according not only to their locations, but also based on their distance from the disaster [13-16]. Other studies have analyzed emotional reactions to terrorist events, including public area attacks [17]. The analysis results can be helpful for developing assistance programs that provide support and help to better cope with terror [18].

Some researchers have used UCR (Uniform Crime Reporting) statistics and twitter data to see how people have reacted towards gun violence and what these words are [19]. Nan and Peter find elevated rates of both pro-gun and anti-gun sentiments on the day of the shooting and different public responses from each state, with the highest pro-gun sentiment not coming from those with highest gun ownership levels but rather from California, Texas and New York [20].

Johan et al. [9] also analyze the text content of daily Twitter posts and measure the mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy), finding an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA (Dow Jones Industrial Average) and a reduction of the MAPR (Mean Average Percentage Error) by more than 6%. Their research group further identify a quantifiable relationship between overall public mood and social, economic and other major events in the media and popular culture, finding that social, political, cultural and economic events are significantly correlated [21].

Geotagged Twitter can be used to understand how the demographic and socioeconomic factors relate to the number of Twitter users at county level [22]. Social sentiments extracted from tweets also play an important role in marketing. Researchers extract positive tweets to improve sales forecasts and to increase sales in marketing campaigns [23, 24]. Similar methods are also applied in movie sales [25], book sales and other products.

B. SENTIMENT ANALYSIS

Sentiment analysis has been very popular in social media applications for both social science research and business marketing [26, 27], especially for analysis of Twitter data [28-31]. In general, sentiment analysis would classify a document as positive, negative, or neutral. Sentiment analysis methods can be generally categorized into rule-based, machine learning-based, and hybrid methods. Rule-based methods combine lexical features and general grammatical and syntactical rules [32], while machine learning methods typically use trained data to learn features and is an active area of research [33]. Hybrid methods will likely achieve even higher accuracy and has become a hot research topic in recent years [34].

Berned et al. [35] propose a human-centered approach for extracting contextual emotional information from human and technical sensors, including wristband sensors and crowdsourced data like Twitter. With the emotion information, it can feed back into urban planning for decision support and for evaluating ongoing planning processes. Bernd and his research team have also worked on a new emotion extraction method from Twitter considering three dimensional properties, space, time and linguistics, based on similarities between each pair of tweets as defined by a specific set of functional relationships in each dimension. The approach bears extensive potential to reveal new insights into citizens' perceptions of the city [36].

C. SPATIAL SENTIMENT ANALYSIS ON TWEETS

Kerla et al. generated a sentiment map of New York City, analyzing the relations between sentiment and points of interests (POIs). It reveals that public mood is generally highest in public parks and lowest at transportation hubs. It identified other areas of strong sentiment such as cemeteries, medical centers, a jail, and a sewage facility. From a temporal perspective, in New York City, more positive tweets are posted on weekends than on weekdays, with a daily peak in sentiment around midnight and a nadir between 9:00 a.m. and noon [7].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

Helen et al. [37] have investigated the spatial and temporal variation of the emotions experienced by individuals whilst using urban green spaces with a case study of 60 urban green spaces in Birmingham being tested. The results show that Twitter data is a viable source of information to researchers investigating human interaction and emotional response to space in cities.

In another study, overall daily variations can be seen, with the morning and late evening having the highest level of happy tweets. Second, geographic variations can be observed, with the west coast showing happier tweets in a pattern that is consistently three hours behind the east coast. Over 300 million tweets (Sep 2006 - Aug 2009) and the Google Maps API were used to infer user locations [38].

Zubair et al [39] measure how common timing and location confounders explain variation in sentiment on Twitter and found it is worthwhile accounting for baseline differences before looking for unexpected changes. However, this research lacks spatial statistics algorithm to explore the patterns. Justin and Henry [40] collected two months of Twitter posts, finding that sentiment in the declining cities does not differ in a statistically significant manner from those in stable and growing cities.

III. DATA COLLECTOIN AND METHODOLOGY

In this section, the methods applied in this research are described, including the workflow for extracting local twitter users, LISA-based spatial autocorrelation and clusters detection method, and LDA-based topic modelling method.

A. STUDY AREA

This study uses Geo-tweet data, which were harvested from the Harvard Center for Geographic Analysis (CGA). The data ranges from January 1st to December 31st, 2016 and the locations encompass the contiguous United States, totaling 69,454,777 records. The internal structure of Geo-tweet data is as follows: tweet_id, time, lat, lon, goog_x, goog_y, sender_id, sender_name, source, reply_to_user_id, reply_to_tweet_id, place_id, and tweet_text [41].

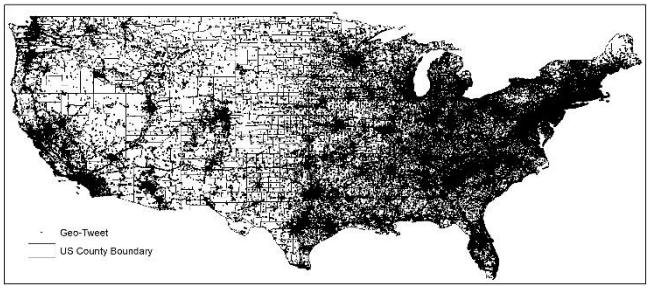

FIGURE 1. Geo-tweets distributions in January 2016.

B. EXTRACTING LOCAL TWITTER USERS

Extracting the local Twitter users is fundamental for analyzing the population bias with census data. Following the method proposed previous researcher [22], we firstly use n-day intervals to eliminate the tweets which are sent from non-local users. Then, we compare the number of tweets posted at night. The county which has the most tweets indicates the hometown of the user. Finally, non-human tweets users are identified based on the average posted tweets number each day. The process is extremely data and computationally intensive due to the massive volume of tweets and users. The details for extracting local twitter users is shown as follows:

1) COMBINE GEO-TWEETS DATA

The raw 2016 Geo-tweet data are saved daily. Thus, it is necessary to combine all the data files into one file. Since each tweet has latitude and longitude, we use spatial operations to assign each Geo-tweet to county.

2) DELETE NON-LOCAL USERS

Obtain the days that each user posts the first and last tweet. If the day interval between the first and last day is smaller than or equals 10, then the user is considered as a non-local user.

3) SELECT USERS' HOME COUNTY

After the previous two steps, a user may still post tweets in a different county. Thus, we calculate the home county based on the maximum total tweets number which are posted from 10 pm to 5 am.

4) DELETE NON-HUMAN USERS

We firstly delete the users who post more than 50 tweets per day. Secondly, there are many bot posts existing in the dataset. The tweeter bot is a type of bot software that controls a Twitter account via the Twitter API. The bot software may autonomously perform actions such as tweeting, re-tweeting, liking, following, unfollowing, or direct messaging other accounts. Thus, we create a blacklist based on the content which are associated with job, advertisement, traffic, weather, and so on. We found that the bots' names include 'TweetMyJOBS', 'SafeTweet by TweetMyJOBS', 'dlvr.it', 'BubbleLife', 'Service Updater', '511NY-Tweets', and 'Cities'. Moreover, if the users' name contains 'Sandaysoft', 'traffic', 'trends', 'weather', and 'sp_', the tweets will be deleted from our dataset.

TABLE 1
NUMBERS OF TWEETS REMAINING AT EACH STEP OF DATA PROCESSING

No	Step	Count
1	Combine Geo-tweets Data	69,454,777
2	Delete non-local users	58,924,735
3	Select users' home county	41,037,822
4	Delete non-human users	24,786,640

Alteryx is a data processing and analysis platform based on workflow, making it is easy for other researchers to reproduce, replicate, or extend the data processing procedures. Thus, we use Alteryx software to create the workflow to extract local twitter users from over 60 million Geo-tweets data. The workflow screenshot is presented below and the generated workflow file is published at Harvard Dataverse. (See link https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6N9VUF). Table 1 shows the number of remaining Geo-Tweets at each step of data processing. In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

testing we found that there are over 16 million machine-generated Geo-tweets, also known as bots. Thus, removing bots from original dataset is an important step, especially for research which focuses on the local users' activities. The existence of non-human users may severely impact the study results.

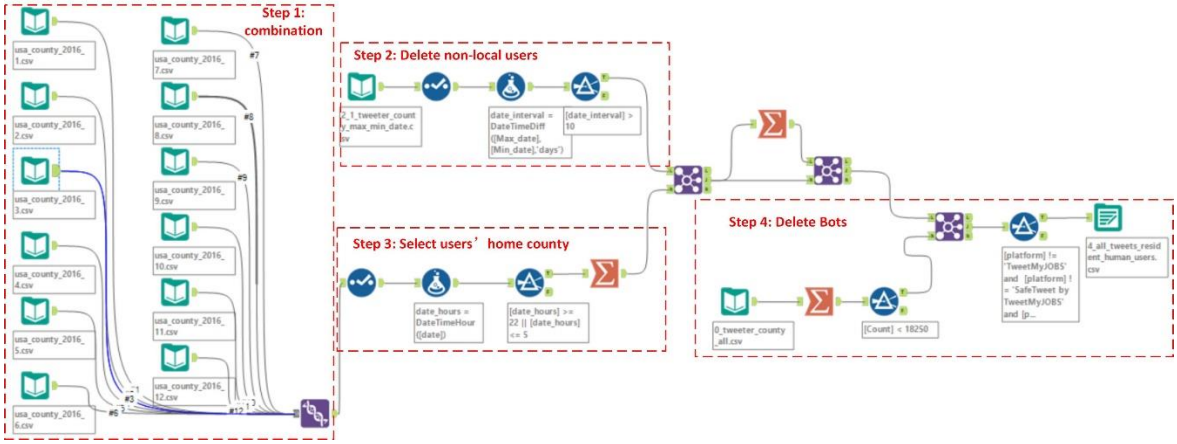

FIGURE 2. The workflow of extracting local Twitter users

C. SENTIMENT ANALYSIS

The sentiment analysis is implemented using the Valence Aware Dictionary and sEntiment Reasoner (VADER), a widely-used rule-based tool specially tuned for sentiment analysis of weblogs and social media text [32]. The VADER leverages the advantages of rule-based modeling methods to construct an advanced sentiment analysis engine, requiring no training data and incorporating a crowd-sourcing approach to improve the lexical feature candidates. In testing, the VADER surpasses individual human raters and performs more favorably across contexts than eleven state-of-practice benchmarks.

The VADER provides a set of scores including the positive score, neutral score, negative score, and the compound score. The compound score is a useful single-value metric that calculates the sum of all the normalized lexicon ratings. In this research, we use the threshold 0.05 to classify a tweet message as positive (score >= 0.05), neutral (-0.05 < score < 0.05), and negative (score <= -0.05).

D. LOCAL INDICATORS OF SPATIAL ASSOCIATION (LISA)

The Local Indicators of Spatial Association (LISA) statistics [42] is a measure for assessing the spatial autocorrelation and clusters. To deal with a variable that is a rate or proportion, the local Moran's I based on Empirical Bayes (EB) standardization is normally used [43]. The LISA statistics and its extensions have been widely used for diverse spatial applications including public health, transportation, ecology, population, and crime [44-48].

The LISA statistic for a variable n of area i is defined as [42]:

$$I_i = m_i \sum_j w_{ij} m_j \quad (1)$$

where m_i and m_j are the normalized value of the variable n. In this research, m represents the count of tweet messages of one specific type of sentiment (positive, negative, and neutral) in area i; m is standardized by the total count of tweet messages in area i using Empirical Bayes (EB) standardization [43]; w_{ij} comes from a spatial weight matrix W that denotes the spatial relationships between all the areas. In this research, we used the widely adopted queen-based contiguity spatial weight, which is a binary symmetric matrix that denotes areas i and j as adjacent when the two area polygons share common sides or vertices. The LISA statistics indicates a spatial clustering tendency when it's positive and spatial dispersion when it's negative. To evaluate the spatial change of Twitter sentiments from weekday to weekend, we simply replace m with the difference in the count of tweet messages of a particular sentiment between weekday and weekend. In this case, the count is first normalized by the length of the time periods (5 for weekdays and 2 for weekends) and then standardized by the total count of tweet messages in area i.

E. TOPIC MODELLING

Topic models are powerful tools to identify latent text patterns in the content. LDA is one of the most popular unsupervised algorithms that models each document as a mixture of topics. The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. Most importantly, LDA makes the explicit assumption that each word is generated from one underlying topic [49].

The number of topics must be specified by the user. In order to reduce the overlap of the extracted topics, we specify 8 groups for the experiment After deciding topics, each Geo-Tweet is assigned to the topic with the highest correlation score. Since Geo-Tweets have geocoordinates and its associated county is easily obtained using spatial operation, each topic was subjectively labelled to ease understanding and interpretation in subsequent geographical analysis.

IV. RESULTS AND DISUCSSION

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

A. OVERALL ANALYSIS

A monthly statistic on Geo-tweets sentiments is calculated as shown in Fig. 3. It is found that more tweets are posted by local users between January and April, than the months of April, October, November, and December. The weather may play an important role in this phenomenon. From February to April, most places in the United States have warm weather, while it gets cold from October to December and people prefer to stay indoors.

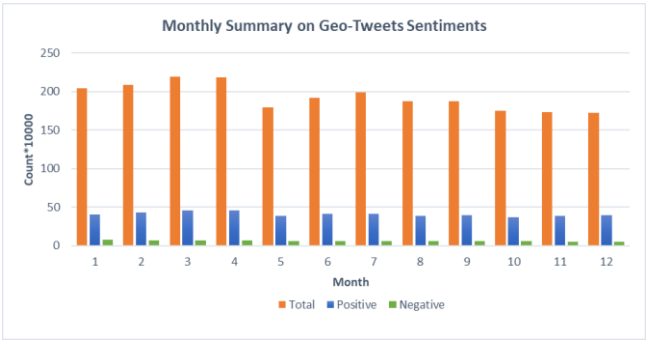

FIGURE 3. **Monthly summary on Geo-tweets sentiments in 2016.**

In order to explore sentiments pattern daily, a daily Geo-tweets statistic was taken as illustrated in Fig. 4. A significant pattern can be found which shows that the total number of posted Geo-tweets increases from Monday to Sunday, and peaks on Saturday. For the positive sentiment, the total number also climbs from the weekdays to the weekend. To clearly presents the differences of negative sentiments on each day, the total number in bold red is added for each day. As was shown with the positive sentiments analysis result, more negative tweets are posted on the weekend than the tweets posted on the weekdays. However, compared to the positive sentiment, the number goes down from Saturday to Sunday. Meanwhile, there's an obvious increase from Thursday to Friday.

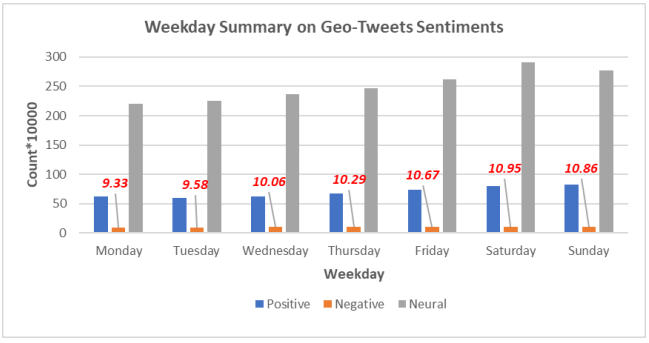

FIGURE 4. **Weekday summary on Geo-tweets sentiments in 2016.**

Figure 5 shows the hourly summary of Geo-tweets sentiments calculated every four hours. The number of tweets steadily increases from time 4 to 24. From 8 am to 12, the positive, negative, and neural sentiment number is much less than other time period. Specifically, three types of sentiment have the same number variations, increasing and decreasing at the same time.

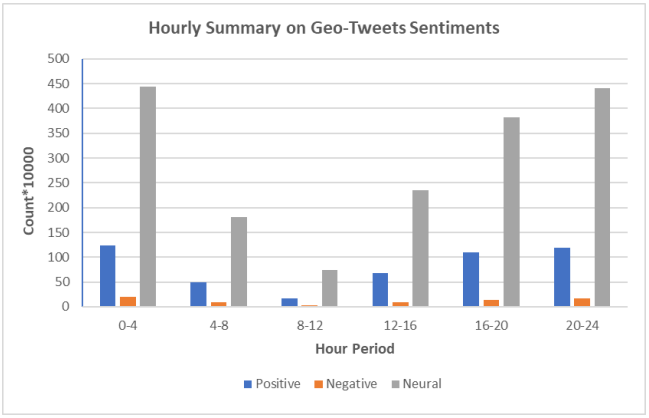

FIGURE 5. **Hourly summary on Geo-tweets sentiments in 2016.**

The spatial pattern is clearly understood when projecting the Geo-tweets on the map. We calculate the total number of positive and negative tweets in each county of the United States. Counties located at the east, west and north coast area with larger populations have more positive and negative tweets. However, it is difficult to discern the relationship between these two sentiments. To address this, the division between positive tweets number and negative tweets number in each county is computed. Fig. 6 illustrates a thematic map with different color representing different value categories. The higher values indicate more positive Geo-tweets are posted. Conversely, lower values indicate more negative Geo-tweets are posted. The result shows there are many counties where the number of negative sentiments is larger than positive sentiments. The counties are located at middle west part of country, such as Montana, North Dakota, South Dakota, and so on. It is also discovered that in most counties, the difference between negative and positive Geo-tweets number is not significant. However, the counties near San Diego in California and Miami in Florida have significant positive sentiment number. These places are famous for its miles of white-sand beaches and amazing weather, offering an abundance of fun attractions for visitors of all ages.

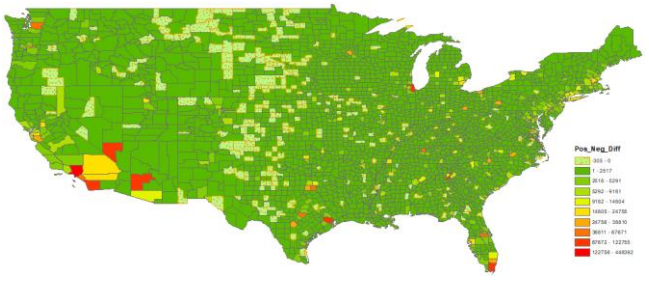

FIGURE 6. **Thematic map of the comparison between positive and negative sentiments.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

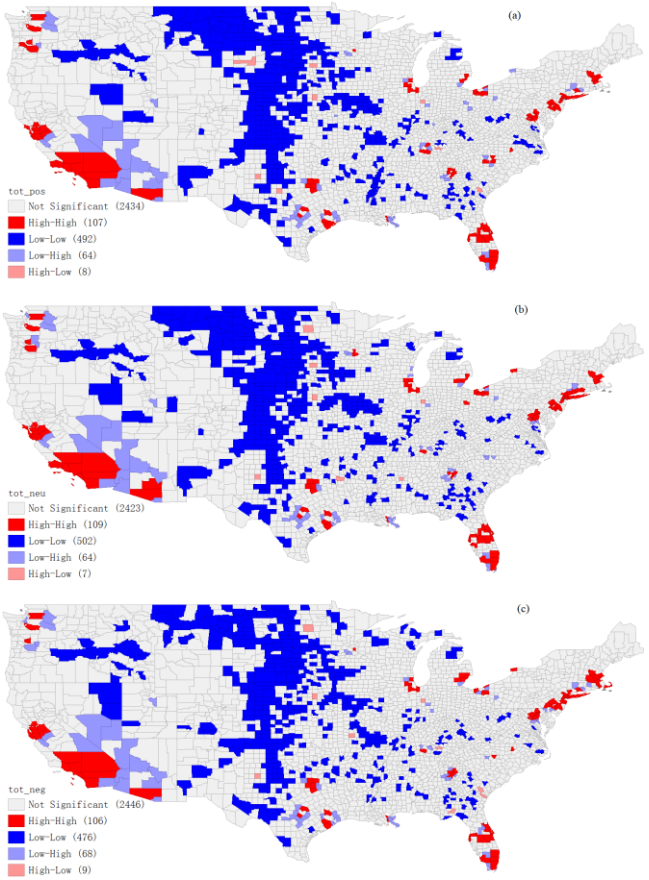

FIGURE 7. **Spatial clustering patterns of three sentiments.**

B. Spatial and Temporal Analysis

Figure 7 shows the spatial clustering patterns of the three types of sentiments (positive, neutral, and negative). In general, the three types of sentiments exhibit similar patterns. Strong low-low clusters exist around Montana, North Dakota, and other states in the central regions. Also, high-high clusters exist around Florida, California, and big cities such as Chicago and New York. Fig. 8 shows the Moran scatter plots for the three sentiments. It shows that all three types of sentiments exhibit mild global spatial autocorrelations.

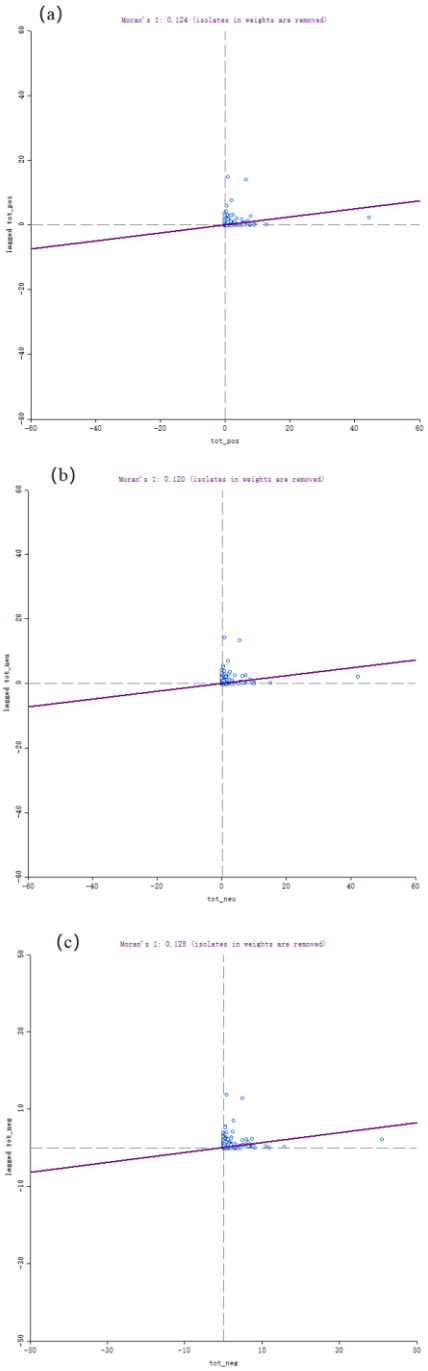

FIGURE 8. **Moran scatter plots of three sentiments.**

C. Semantic Analysis on Positive Sentiment Users' Tweets

This section investigates the semantic analysis result from the positive Tweets based on LDA topic model. Figure 9 presents the top-30 most salient terms in the Geo-tweets with positive sentiments. It is shown that 'love', 'happi', 'thank', 'birthday', 'friend', 'good', 'best', and 'beauti' are the most frequently used words.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

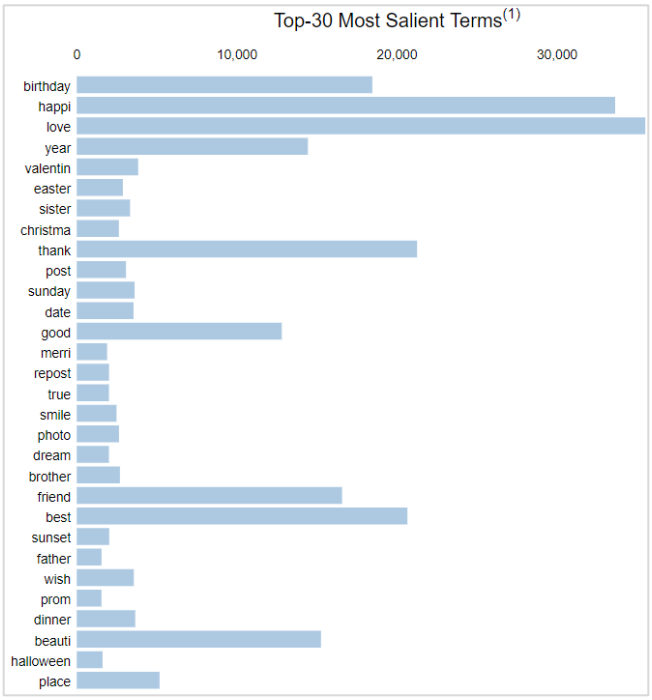

FIGURE 9. **Top-30 most salient terms based on the frequency statistics.**

Based on the LDA model, 8 different topics are extracted as shown in Table 2. The sequence of the words indicates the weight of each word in descending order. Analyzing the correlation of words in each topic, some of the words can be easily associated with specific theme. For example, topic 2 contains the keyword: 'thank', 'best', 'friend', 'today', 'excit', 'come' and 'family'. This topic is talking about 'friend and family gathering'. In topic 5, its top words include 'happi', 'birthday', 'year', 'celebr', 'wish, 'cheer', and 'hour', indicating that the tweets are related to birthday party. In summary, the 8 different topics are related to

- Topic 1: Enjoy the weekend
- Topic 2: Friend and family gathering
- Topic 3: Posting photo
- Topic 4: Birthday
- Topic 5: Great time on Sunday
- Topic 6: Romantics
- Topic 7: Parties at night
- Topic 8: Life and living

Although several topics overlapped to some extent, they still have relatively clear differences between each other.

TABLE 2
THE 7 MOST PORTABLE WORDS SHOWN FOR 8 DIFFERENT TOPICS

Topic	Top 7 Keywords
1	amaz, favorit, christma, enjoy, final, weekend, beauti
2	thank, best, friend, today, excit, come, family
3	beauti, night, super, perfect, post, photo, summer
4	happi, birthday, year, celebr, wish, cheer, hour
5	great, time, morn, good, sunday, drink, nice
6	love, girl, list, know, congratul, sweet, littl
7	tonight, parti, readi, night, game, play, join
8	good, like, life, best, look, feel, live

We map each Geo-tweet into the topics with the highest correlation scores. Exploring the temporal patterns of the topics, the total number of each topic are calculated from Monday to Sunday. Results show that the overall posted positive tweet numbers are increasing from Monday to Sunday and peaks on Sunday. Each topics' number changes in daily distribution has the same pattern. Specifically, topic 1 and 2 have more tweets than other topics, because they are associated with a great time on weekends, as well as friend and family gatherings.

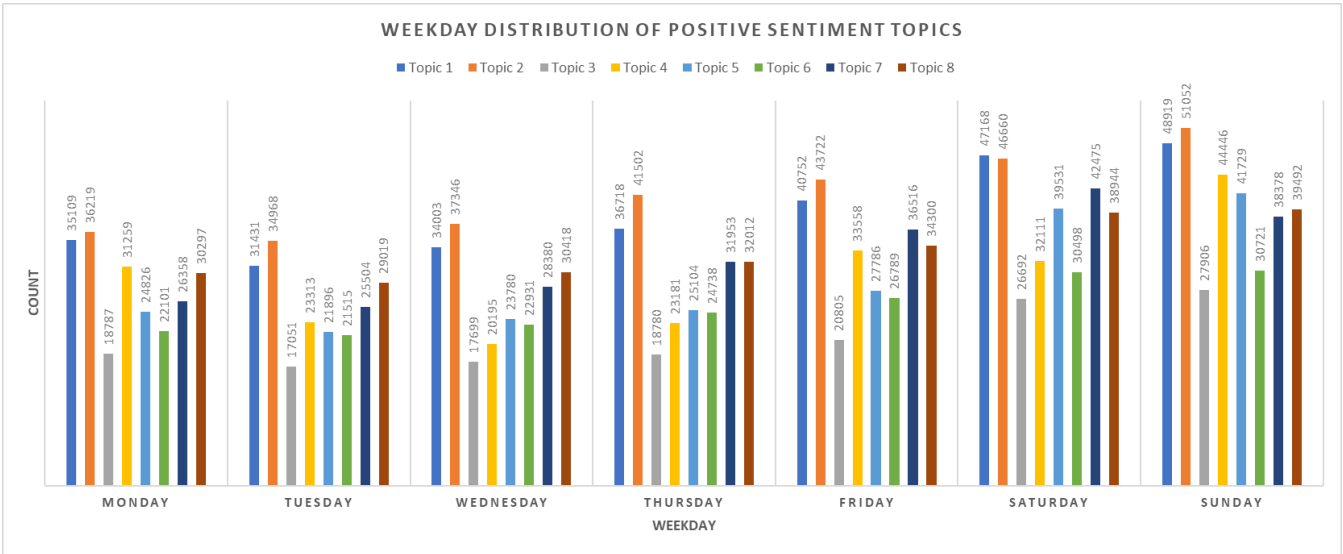

FIGURE 10. **Weekday distribution of positive sentiment topics**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

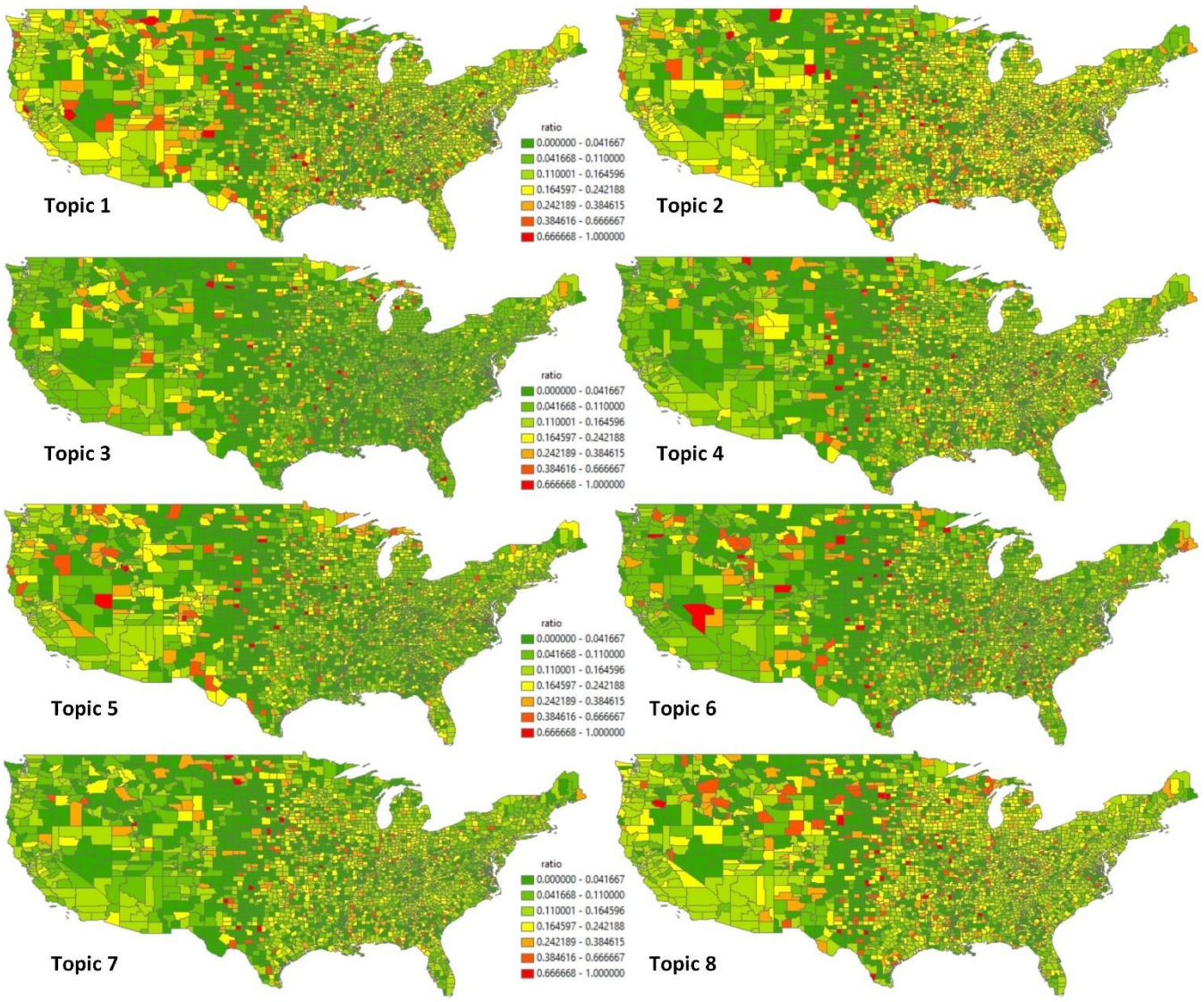

FIGURE 11. Top-30 most salient terms based on the frequency statistics.

From a spatial perspective, the ratio value for each county is calculated, comparing each topic number to total topic number in the county. In order to keep the consistency of spatial visualization, the ratio classifications follow the same rule and it is shown as a legend in Fig. 11, which illustrates the ratio values of each county in US from 8 different topic perspective. More red indicates a higher ratio value, while more green indicates a lower ratio value. Since there are smaller populations in North Dakota, South Dakota, Nebraska, Kansas, and Oklahoma, the posted Tweets are much less than in other places. Table 3 shows the statistical result of each topic from the overall view. Topic 1, topic 2 and topic 8 have relatively higher total ratio values than other topics, implying that Twitter users prefer to post positive tweets when they have a great time on the weekend, with friends and family. The sum of ratio values for topic 3 is much smaller than other topics. Topic 3 is associated with posting photos on Twitter. Since the West Coast has been populated by immigrants and their descendants more recently than the East Coast, its culture is considerably younger. Thus, from Fig. 11 (topic 2), it is shown that the western part of the US has higher values than the other parts.

TABLE 3
THE STATISTICS FOR EACH TOPIC

Topic	Min	Max	Sum	Mean	Median	SD[a]
1	0	1	410.0273	0.1318	0.1224	0.1367
2	0	1	443.3707	0.1426	0.1458	0.1261
3	0	1	235.5290	0.0758	0.0632	0.1010
4	0	1	339.0067	0.1090	0.1020	0.1168
5	0	1	311.6918	0.1003	0.08971	0.1129
6	0	1	322.9230	0.1039	0.0862	0.1256
7	0	1	331.2833	0.1066	0.1042	0.1088
8	0	1	410.1230	0.1319	0.1285	0.1259

[a]SD is the abbreviation of Standard Deviation.

In order to discover which topic play a domain role in the positive sentiment all over the country, we calculate the total number of counties whose ratio value is larger than 0.384816.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2961100, IEEE Access

The result indicates that there are 136 counties where a large amount of the positive sentiments is related to weekend happy time. It can be seen from Fig. 11 (topic 1) that the counties which have higher ratio value have an even distribution cross the country. In contrast, we calculate the total number of counties whose ratio value is smaller than 0.11. The results indicate that in topic 3, there are over 2500 counties whose ratio value is smaller than 0.11, which is much larger than other topics. Thus, users post fewer positive sentiment tweets associated with the topic 'posting photos' compared to other topics. Another reason for this is that topic 3 contains summer related tweets, which emphasizes the seasonal content.

From a regional difference perspective, the standard deviation in table 3 demonstrates how each county is different from each other. The smaller the value is, it indicates less regional difference, and the larger the value is, it indicates more regional differences. There are fewer regional differences for topic 3 and topic 7, which means the ratio of topics related to nighttime parties and posting photos in each county has similar proportions.

V. CONCLUSION AND FUTURE WORK

This research applied spatial, temporal, and semantic analysis methods to the United States Geo-tweets data in 2016. Considering the population bias and noisy data in the original dataset, we built a workflow using Alteryx to clean the data and extract the tweets which are posted by local residential users. The workflow is shared at Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6N9VUF) and any users are able to reproduce, replicate and extend the whole processing procedures based on their own Geo-tweets dataset.

The research reveals that spatial patterns for sentiments do exist, particularly among positive tweet messages. There is also mild spatial autocorrelation from weekdays to weekends. Involving the geospatial information with the topics, some patterns are found to demonstrate an association between the nature of Twitter content and the characteristics of places and users. For example, weekend events and friend and family gatherings are the time that users prefer to post positive tweets. In the western part of the US, users love to post photos on Twitter more than in other parts of the US.

However, there are several limitations in this research. (1) Data Noise. Although we have extracted non-human generated tweets according to basic rules, the cleaned dataset still includes noisy data. More advanced bots detection methods should be employed to refine our dataset in the future. (2) Negative sentiment analysis. This research mainly focuses on the positive sentiment analysis, however discovering the patterns of negative sentiment is also important for further study.

ACKNOWLEDGMENT

(Tao Hu and Bing She contributed equally to this research.) Authors thank to the BOP (Billion Object Project) group at Center for Geographic Analysis, providing the Geo-tweets dataset in this research.

REFERENCES

[1] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data: Springer, 2012, pp. 415-463.
[2] Y. Wang, "Sensing Human Sentiment via Social Media Images: Methodologies and Applications," Arizona State University, 2018.
[3]S. D. Pressman, M. W. Gallagher, and S. J. Lopez, "Is the emotion-health connection a "first-world problem"?," Psychological science, vol. 24, no. 4, pp. 544-549, 2013.
[4] M. E. Geisser, R. S. Roth, M. E. Theisen, M. E. Robinson, and J. L. Riley III, "Negative affect, self-report of depressive symptoms, and clinical depression: relation to the experience of chronic pain," The Clinical journal of pain, vol. 16, no. 2, pp. 110-120, 2000.
[5] S. D. Pressman and S. Cohen, "Does positive affect influence health?," Psychological bulletin, vol. 131, no. 6, p. 925, 2005.
[6] S. Gohil, S. Vuik, and A. Darzi, "Sentiment analysis of health care tweets: review of the methods used," JMIR public health and surveillance, vol. 4, no. 2, p. e43, 2018.
[7] K. Z. Bertrand, M. Bialik, K. Virdee, A. Gros, and Y. Bar-Yam, "Sentiment in new york city: A high resolution spatial and temporal view," arXiv preprint arXiv:1308.5010, 2013.
[8] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Fourth International AAAI Conference on Weblogs and Social Media, 2010.
[9] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of computational science, vol. 2, no. 1, pp. 1-8, 2011.
[10] N. Chokshi, "Americans Are Among the Most Stressed People in the World, Poll Finds," The New York Times, April 25, 2019 2019, [Online]. Available: https://www.nytimes.com/2019/04/25/us/americans-stressful.html, Accessed on: 11/10/2019.
[11] E. Hauthal and D. Burghardt, "Mapping space-related emotions out of user-generated photo metadata considering grammatical issues," The Cartographic Journal, vol. 53, no. 1, pp. 78-90, 2016.
[12] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in Proceedings of the first workshop on social media analytics, 2010: acm, pp. 80-88.
[13] C. Caragea, A. C. Squicciarini, S. Stehle, K. Neppalli, and A. H. Tapia, "Mapping moods: Geo-mapped sentiment analysis during hurricane sandy," in ISCRAM, 2014.
[14] Y. Sano, H. Takayasu, and M. Takayasu, "Emotional changes in Japanese blog space resulting from the 3.11 earthquake," in Proceedings of the International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014, 2015: Springer, Cham, pp. 289-299.
[15] V. K. Neppalli, C. Caragea, A. Squicciarini, A. Tapia, and S. Stehle, "Sentiment analysis during Hurricane Sandy in emergency response," International journal of disaster risk reduction, vol. 21, pp. 213-222, 2017.
[16] Y. Wang and J. E. Taylor, "Coupling sentiment and human mobility in natural disasters: a Twitter-based study of the 2014 South Napa Earthquake," Natural hazards, vol. 92, no. 2, pp. 907-925, 2018.
[17] J. G. Harb and K. Becker, "Comparing Emotional Reactions to Terrorism Events on Twitter," in Workshop on Big Social Data and Urban Computing, 2018: Springer, pp. 107-122.
[18] P. Garg, H. Garg, and V. Ranga, "Sentiment analysis of the Uri terror attack using Twitter," in 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017: IEEE, pp. 17-20.
[19] J. Adil, "American Sentiments Towards Gun Violence," April 30, 2018, [Online]. Available: https://rpubs.com/jayadil47/390889, Accessed on: Nov 10, 2018.
[20] N. Wang, B. Varghese, and P. D. Donnelly, "A machine learning analysis of Twitter sentiment to the Sandy Hook shootings," in 2016 IEEE 12th International Conference on e-Science (e-Science), 2016: IEEE, pp. 303-312.

[21] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[22] Y. Jiang, Z. Li, and X. Ye, "Understanding demographic and socioeconomic biases of geotagged twitter users at the county level," Cartography and Geographic Information Science, vol. 46, no. 3, pp. 228-242, 2019.
[23] R. Dijkman, P. Ipeirotis, F. Aertsen, and R. van Helden, "Using twitter to predict sales: A case study," arXiv preprint arXiv:1503.04599, 2015.
[24] D. Gaikar and B. Marakarkandy, "Product sales prediction based on sentiment analysis using twitter data," International Journal of Computer Science and Information Technologies, vol. 6, no. 3, pp. 2303-2313, 2015.
[25] V. Jain, "Prediction of movie success using sentiment analysis of tweets," The International Journal of Soft Computing and Software Engineering, vol. 3, no. 3, pp. 308-313, 2013.
[26] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in LREc, 2010, vol. 10, no. 2010, pp. 1320-1326.
[27] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1-135, 2008.
[28] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in Proceedings of the Workshop on Language in Social Media (LSM 2011), 2011, pp. 30-38.
[29] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in Fifth International AAAI conference on weblogs and social media, 2011.
[30] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011: ACM, pp. 1031-1040.
[31] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," HP Laboratories, Technical Report HPL-2011, vol. 89, 2011.
[32] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Eighth international AAAI conference on weblogs and social media, 2014.
[33] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, 2011: Association for Computational Linguistics, pp. 142-150.
[34] P. Chikersal, S. Poria, and E. Cambria, "SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 647-651.
[35] B. Resch, A. Summa, G. Sagl, P. Zeile, and J.-P. Exner, "Urban emotions—Geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data," in Progress in location-based services 2014: Springer, 2015, pp. 199-212.
[36] B. Resch, A. Summa, P. Zeile, and M. Strube, "Citizen-centric urban planning through extracting emotion information from twitter in an interdisciplinary space-time-linguistics algorithm," Urban Planning, vol. 1, no. 2, pp. 114-127, 2016.
[37] H. Roberts, J. Sadler, and L. Chapman, "The value of Twitter data for determining the emotional responses of people to urban green spaces: A case study and critical evaluation," Urban Studies, vol. 56, no. 4, pp. 818-835, 2019.
[38] A. Mislove, "Pulse of the nation: US mood throughout the day inferred from twitter," http://www. ccs. neu. edu/home/amislove/twittermood/, 2010.
[39] Z. Shah, P. Martin, E. Coiera, K. D. Mandl, and A. G. Dunn, "Modeling Spatiotemporal Factors Associated With Sentiment on Twitter: Synthesis and Suggestions for Improving the Identification of Localized Deviations," Journal of Medical Internet Research, vol. 21, no. 5, p. e12881, 2019.
[40] J. B. Hollander and H. Renski, Measuring urban attitudes using Twitter: An exploratory study. Lincoln Institute of Land Policy., 2015.
[41] B. Lewis, "Harvard CGA Geo-tweet Archive," V1 ed: Harvard Dataverse, 2016.
[42] L. Anselin, "Local indicators of spatial association—LISA," Geographical analysis, vol. 27, no. 2, pp. 93-115, 1995.
[43] R. M. Assuncao and E. A. Reis, "A new proposal to adjust Moran's I for population density," Statistics in medicine, vol. 18, no. 16, pp. 2147-2162, 1999.
[44] Y. Fan, X. Zhu, B. She, W. Guo, and T. Guo, "Network-constrained spatio-temporal clustering analysis of traffic collisions in Jianghan District of Wuhan, China," PLoS one, vol. 13, no. 4, p. e0195093, 2018.
[45] M. j. Fortin, M. R. Dale, and J. M. Ver Hoef, "Spatial analysis in ecology," Encyclopedia of environmetrics, vol. 5, 2006.
[46] J. D. Morenoff, R. J. Sampson, and S. W. Raudenbush, "Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence," Criminology, vol. 39, no. 3, pp. 517-558, 2001.
[47] L. A. Waller and C. A. Gotway, Applied spatial statistics for public health data. John Wiley & Sons, 2004.
[48] S. J. Rey and B. D. Montouri, "US regional income convergence: a spatial econometric perspective," Regional studies, vol. 33, no. 2, pp. 143-156, 1999.
[49] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, 2009: Association for Computational Linguistics, pp. 248-256.