A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments

Yingjie Hu [1], Chengbin Deng [2], and Zhou Zhou [2]

[1] Department of Geography, University at Buffalo, The State University of New York
[2] Department of Geography, Binghamton University, The State University of New York

Abstract: The perceptions of people toward neighborhoods reveal their satisfactions with their living environments and their perceived quality of life. Recently, there is an emergence of websites designed for helping people to find suitable places to live. On these websites, current and previous residents can review their neighborhoods by providing numeric ratings and textual comments. Such online neighborhood review data provide novel opportunities for studying the perceptions of people toward their neighborhoods. In this paper, we analyze such online neighborhood review data. Specifically, we extract two types of knowledge from the data: 1) semantics, i.e., the semantic topics (or aspects) that people talk about their neighborhoods; and 2) sentiments, i.e., the emotions that people express toward the different aspects of their neighborhoods. We experiment with a number of different computational models in extracting these two types of knowledge and compare their performances. The experiments are based on a dataset of online reviews about the neighborhoods in New York City (NYC), which were contributed by 7,673 distinct Web users. We also conduct correlation analyses between the subjective perceptions extracted from this dataset and the objective socioeconomic attributes of NYC neighborhoods, and find similarities and differences. The effective models identified in this research can be applied to neighborhood reviews in other cities for supporting urban planning and quality of life studies.

Keywords: neighborhood, online review, quality of life, topic modeling, sentiment analysis, geospatial semantics.

1. Introduction

In his landmark paper, Goodchild (2007) proposed the idea of "Citizens as Sensors". He suggested that general individuals can be compared to environmental sensors, who can observe and collect a variety of geographic information. Indeed, the following years witnessed an unprecedented increase in the volume and variety of volunteered geographic information (VGI). There are geotagged photos contributed by people through websites, such as Flickr, Geograph, and Instagram, on which people share what they have seen (Crandall et al. 2009, Purves, Edwardes and Wood 2011, Hochman and Manovich 2013, Hollenstein and Purves 2010). There are geotagged texts, such as Tweets, which show the comments of people toward important events or their everyday experiences (Tsou et al. 2013, Cervone et al. 2016, Huang 2017). There are also volunteer-contributed GPS locations, which record the trajectories of vehicles or movements of animals (Sullivan et al. 2009, Haklay 2013).

Among the many different types of VGI, online neighborhood reviews are a special type of data that emerged in recent years. A typical form of online neighborhood reviews is a combination of numeric ratings and textual comments. For example, a Web user may first assign a *4 star* to a neighborhood and then write a comment to explain this rating. Neighborhood review data are special, because they record the intellectual synthesis performed by people consciously or subconsciously based on the raw information collected via some of the five human senses. Thus, the view of well-groomed lawns in front of houses or scrambled graffiti on walls, the sound of children playing in backyards or the traffic noise from nearby highways, the fragrance of flowers or the odor of trashes, the feel of breeze or humidity, and sometimes the taste of delicious cookies from a kind neighbor all contribute to one's perception of the neighborhood. Neighborhood reviews are not literal recordings of objective environment properties but have added a subjective layer of human cognition. As a result, they offer a unique resource for studying the perceptions of people toward their living environments.

The perceptions of people toward neighborhoods were also investigated in previous research. Questionnaire-based surveys and face-to-face interviews were frequently used to collect the opinions of people (Ceccato and Snickars 1998, Das 2008, Eby, Kitchen and Williams 2012, Sharma 2014, Khaef and Zebardast 2016). While discovering valuable insights, these surveys and interviews were often labor-intensive and limited to small sample sizes. By contrast, online neighborhood reviews allow studies to be scaled up to thousands or even tens of thousands of people in large geographic areas relatively easily. However, challenges exist in effectively and efficiently analyzing large numbers of online reviews and extracting meaningful knowledge.

This paper conducts an analysis on online neighborhood review data. Specifically, we aim to extract two types of knowledge: 1) semantics, i.e., the main semantic topics (or aspects) that people talk about their neighborhoods; and 2) sentiments, i.e., the emotions that people express toward the different neighborhood aspects. We experiment with multiple computational models for extracting these two types of knowledge and compare their performances. A dataset of online reviews focusing on the neighborhoods in New York City (NYC) is used in this study, and these reviews were contributed by 7,673 distinct Web users. The main contributions of this work are as follows:

- We propose to analyze online neighborhood review data for understanding the perceptions of people toward their living environments. To the best of our knowledge, this work is among the first efforts in analyzing such online neighborhood reviews.

- We experiment with multiple models for extracting semantic topics and sentiments from neighborhood reviews. We systematically compare the performances of the models, and identify the most effective models based on the experiment results.
- We conduct correlation analyses between the subjective neighborhood perceptions extracted by our models and the objective socioeconomic attributes of the neighborhoods, and find similarities and differences between the two.

The remainder of this paper is organized as follows. Section 2 provides a literature review on related work. Section 3 presents the core ideas of the models tested in this work for analyzing the semantics and sentiments of online neighborhood reviews. Section 4 presents a case study and related experiments based on a NYC neighborhood review dataset and compares the performances of multiple models. Section 5 conducts correlation analyses between subjective perceptions and objective socioeconomic attributes of neighborhoods. Section 6 summarizes this work and discusses its limitations and future directions.

2. Literature review

One research area that frequently examines the perceptions of people toward neighborhoods is quality of life (QOL). Studies in this area often seek to understand people's satisfactions toward their living environments, as well as the affecting physical, social, and economic factors (Helburn 1982, Sirgy and Cornwell 2002). Many QOL studies were conducted in different cities throughout the world. Ceccato and Snickars (1998) and Ceccato and Snickars (2000) designed questionnaire surveys to investigate the perceptions of people toward QOL in a number of neighborhoods in Sweden. Das (2008) interviewed residents in the city of Guwahati, India to investigate their satisfactions toward different factors of their living environment. Eby et al. (2012) examined people's perceptions of neighborhoods in Hamilton, Ontario, Canada, and identified six themes, such as crime and transportation, with significant impacts on the perceived quality of life. Lee, Gu and An (2016) performed questionnaire-based surveys and interviews to understand the perceptions of people on green space in Jeonju City, South Korea.

Research in urban planning and public participation GIS (PPGIS) also examined the perceptions and opinions of residents (Sieber 2006). Rinner (2001), Keßler, Rinner and Raubal (2005), and Rinner and Bird (2009) designed Argumentation Maps which enables public users to add textual comments and link these comments to locations on a map. Bugs et al. (2010) developed a Web 2.0 PPGIS and applied this system to an urban planning case study in Canela, Brazil. In national park planning, Brown and Weber (2011) used PPGIS to collect and analyze the perceptions of visitors to support decision making. In these studies, the comments of people were usually considered as *qualitative data* and were examined manually.

Natural language processing (NLP) provides useful techniques for analyzing large volumes of text data (Kao and Poteet 2007). By extracting quantitative values such as term frequencies, NLP transforms texts from *qualitative data* to *quantitative data*. Two areas in NLP are closely related to this work: topic modeling and sentiment analysis. Topic modeling aims to discover the main topics discussed in a textual document. For example, a neighborhood review may be discussing two topics related to safety and local transportation, and topic modeling can quickly identify these main topics without requiring one to read the review. A number of topic modeling methods, such as latent Dirichlet allocation (LDA) (Blei, Ng and Jordan 2003) and labeled LDA (LLDA) (Ramage et al. 2009), have been developed and used in various applications (Kling and Pozdnoukhov 2012, Quercia, Askham and Crowcroft 2012, Adams and McKenzie 2013). Sentiment analysis is another sub area in NLP, which aims to extract the opinions and emotions of people (Pang and Lee 2008, Liu 2012). Early research in this area focused on identifying

sentiment polarities (i.e., positive or negative) of whole documents (Pang and Lee 2004, Beineke et al. 2004), while later studies also explored the opinions of people toward particular aspects (Hu and Liu 2004, Wang, Lu and Zhai 2010, Jo and Oh 2011, Cataldi et al. 2013). The target entities whose reviews are frequently examined in sentiment analysis include movies, restaurants, hotels, and products (e.g., MP3 players) on online shopping websites (Kasper and Vela 2011, Feldman 2013, Zhang et al. 2014).

The GIScience community also paid considerable attentions to NLP techniques. This can be partially attributed to the booming of location-based social media, such as Twitter, Flickr, and Foursquare, which generate large volumes of data linking locations and texts (*geo-text data* for short). Many studies have explored these geo-text data. Hu et al. (2015) examined geotagged Flickr photos to extract urban areas of interest (AOI), in which a technique, term frequency and inverse document frequency, was used to find the words that are most representative for an extracted AOI. Adams, McKenzie and Gahegan (2015) performed topic modeling on geotagged travel blogs and Wikipedia, and enabled users to find similar places based on thematic keywords. Ballatore and Adams (2015) studied the emotions related to place types and constructed a place vocabulary to associate place types with sentiment words. Wang and Stewart (2015) mined hazard information from news articles to examine the spatiotemporal impacts of the Hurricane Sandy. Gao et al. (2017) performed a data-synthesis-driven analysis by combining the geo-text data from Twitter, Flickr, and Instagram, and identified the most prominent topics and words associated with different cognitive regions. Martin and Schuurman (2017) applied topic modeling to geotagged Tweets from multiple geographic areas and embedded the extracted topic words into maps.

Despite these previous studies, online neighborhood review data, to the best of our knowledge, have not been examined before. Studies such as Shelton, Poorthuis and Zook (2015) and Jenkins et al. (2016) also discovered interesting properties of neighborhoods, but they focused on identifying the places that are frequently visited by people using geotagged Twitter data rather than examining the perceptions of people based on neighborhood reviews. In addition, this work compares the effectiveness of multiple computational models in extracting semantic topics and sentiments from neighborhood reviews, and identifies the most effective models. We also analyze the spatial autocorrelations of the extracted sentiment ratings under different semantic topics, and compare the subjective perceptions with objective socioeconomic attributes of neighborhoods. In the following section, we describe the methods and models used in this work.

3. Methods

3.1 Problem formalization

We start by formalizing the problem studied in this work. The neighborhood reviews examined here are contributed by Web users, and two components are assumed to be available: 1) numeric ratings and 2) textual comments. Our objectives are to: 1) identify the main semantic topics (or aspects) that people talk about their neighborhoods, and 2) quantify the sentiments that people express toward the identified aspects. The second objective is necessary because we only know the overall rating of a reviewer rather than his/her ratings on different aspects of the neighborhoods.

To give a concrete example, consider the following review: *4 star* (numeric rating) and "*This neighborhood is close to a lot of restaurants and stores. However, I don't feel very safe as there are sometimes suspicious persons walking around.*" (textual comment). We aim to identify the major topics discussed by the reviewer (e.g., topics related to safety and life convenience) and the reviewer's

sentiments toward these aspects (e.g., the reviewer perhaps have a *3.5 star* for safety and a *4.5 star* for life convenience). We formalize this problem as below:

> *Given a set of neighborhoods* $N = \{n_i\}$*, a set of numeric ratings on the neighborhoods* $R = \{r_{ij}\}$*, and a set of review texts* $D = \{d_{ij}\}$*, what are the main semantic topics* $S = \{s_{ijk}\}$ *and the topic-specific ratings* $A = \{a_{ijk}\}$ *of the neighborhood reviewers?*

where n_i represents a neighborhood, r_{ij} represents the jth rating on neighborhood n_i, d_{ij} is the review text associated with rating r_{ij}, s_{ijk} represents the kth semantic topic of the review, and a_{ijk} is the topic-specific rating. We can further aggregate the topic-specific ratings from different reviewers for each neighborhood, and obtain $B = \{b_{ik}\}$, where b_{ik} is the averaged rating on the kth semantic topic (or aspect) of neighborhood n_i. The two terms, *topic* and *aspect*, are used interchangeably in the literature (Wang et al. 2010, Jo and Oh 2011, Liu 2012), and are both used in this paper.

3.2 Semantic topic identification

The first objective of this work is to identify the main topics discussed in the neighborhood reviews. Probabilistic topic models fit this objective with their capability of discovering latent semantic topics from large amounts of unstructured texts (Steyvers and Griffiths 2007, Blei 2012). Specifically, we experiment with latent Dirichlet allocation (LDA) model (Blei et al. 2003), which is a standard topic model used in many studies. We also experiment with multi-grain LDA model, or MG-LDA (Titov and McDonald 2008), which is a variation of LDA tailored for online reviews.

LDA is a generative model which considers one textual document as generated from a probabilistic distribution of topics, and each topic is modeled as a probabilistic distribution of words. In this work, LDA considers each neighborhood review as a document, which is generated from a number of semantic topics, such as safety and life convenience, and each semantic topic is modeled as a distribution over words, such as "safe", "crime", and "police". The per-document topic distributions are drawn from two Dirichlet distributions, and LDA functions by finding a set of parameters that maximize the probability of producing the observed neighborhood reviews. Expectation–maximization (Dempster, Laird and Rubin 1977) and Gibbs sampling (Geman and Geman 1984) are often used for finding the best parameters.

MG-LDA is a variation of LDA with the goal of discovering ratable aspects of objects from online reviews (Titov and McDonald 2008). MG-LDA was initially proposed based on the observations of online MP3 player reviews: the authors of MG-LDA found that LDA often discovers *global topics* that are not directly ratable, such as the unique features related to a brand of MP3 players (e.g., *iPod*), rather than more ratable aspects, such as *sound quality* and *battery life*. MG-LDA addresses this problem by modeling each review as generated from both *global topics* and *local topics*. The global topics provide top-level information, such as brand-specific features, while local topics capture the ratable aspects.

Both LDA and MG-LDA are unsupervised models which can discover semantic topics without requiring labeled data. However, they need input parameters for the estimated number of topics in the texts: LDA requires one parameter K for the total number of topics, while MG-LDA requires two parameters K^{gl} and K^{loc} for the numbers of global and local topics respectively. Finding suitable values for these parameters is often a challenging task (McKenzie et al. 2015, Gao et al. 2017). For LDA, we adopt four methods proposed in the literature to select K. The first is from Griffiths and Steyvers (2004) who used Gibbs sampling to find the best K. Their method is summarized in Equation 1, where $\boldsymbol{w}$

represents the observed words, and the best K is the one that achieves the highest log-likelihood of obtaining the observed words.

$$K^* = \underset{K}{\text{argmax}} \, (\log \, p(\boldsymbol{w}|K)) \tag{1}$$

The second method is from Cao et al. (2009) who selected the suitable K based on the distances among the topics and their densities. A key component of their method is calculating the average topic distance, as shown in Equation 2:

$$avg\_dis(T) = \frac{\sum_{i=1}^{K-1} \sum_{j=(i+1)}^{K} sim(t_i, t_j)}{K \, (K-1)/2} \tag{2}$$

where $t_i, \ t_j$ represent two topics in the topic set T, and $sim(t_i, \ t_j)$ is the cosine similarity between them. A better K discovers topics with smaller values of $avg\_dis(T)$. The third method is from Arun et al. (2010) who used a matrix factorization to find the suitable K. They used Kullback-Leibler (KL) divergence to measure the quality of factorizing a document-word matrix $C_{D\times W}$ into two matrix factors, a document-topic matrix $M1$ and a topic-word matrix $M2$ at different Ks. Their metric is summarized in Equation 3:

$$Divergence(M1, M2) \ = \ KL(C_{M1}||C_{M2}) \ + \ KL(C_{M2}||C_{M1}) \tag{3}$$

where C_{M1} and C_{M2} are two distributions obtained from the matrix factors $M1$ and $M2$. The fourth method is from Deveaud, SanJuan and Bellot (2014) who employed Jensen-Shannon (JS) divergence to identify the suitable value of K. Equation 4 shows the calculation of their metric:

$$Divergence(T) = \frac{\sum_{t_i, t_j \in T} JSD(t_i, t_j)}{K(K-1)/2} \tag{4}$$

where $t_i, \ t_j$ are two topics, and $JSD(t_i, \ t_j)$ is the JS divergence between the topic pair. The four methods are combined to identify a suitable K for LDA. For MG-LDA, we use the parameter recommendation from the authors (Titov and McDonald 2008), in which K^{gl} should exceed K^{loc} by a factor of 2 (e.g., they used $K^{gl} = 30$ and $K^{loc} = 10$ in their paper), while K^{loc} should be the number of ratable aspects. In the Experiments section, we will test and compare the effectiveness of LDA and MG-LDA in identifying topics from online neighborhood reviews.

3.3 Aspect sentiment analysis

The second objective of our work is to quantify the sentiments of a reviewer to the identified aspects of a neighborhood. One naïve approach is to simply assume that the reviewer has the same rating toward each aspect as their overall rating. Such an approach does not capture the cases when a reviewer praises one aspect of the neighborhood while criticizing another. A second approach is to identify the descriptive words associated with each aspect, and then use a sentiment lexicon to quantify the attitudes of the reviewer. However, a sentiment lexicon is typically derived from corpora in certain domains (e.g., movie reviews) which may not well represent the word sentiments in another domain. In addition, people may use the same words to express different review ratings, e.g., one may use the word "good" to express a *3-star* rating whereas another person may use "good" to express a *4-* or *5-star* rating.

In this work, we propose to leverage the latent aspect rating analysis (LARA) model (Wang et al. 2010) to derive aspect-specific ratings from online neighborhood reviews. LARA functions without using an external sentiment lexicon. It assumes that a reviewer has latent ratings on different aspects (e.g.,

safety and life convenience of a neighborhood) in mind and has different weights on these aspects (e.g., one reviewer may care more about safety than other aspects). Such aspect-specific ratings and their weights are hidden (latent), but are implied through the words of the reviewer and the overall rating. The core idea of LARA is summarized in Equation 5:

$$r_d \sim N\left(\sum_{i=1}^{K} \alpha_i \sum_{j=1}^{N} \beta_{ij} w_{ij}, \delta^2\right) \quad (5)$$

where $\beta_{ij} \in \mathbb{R}$ is the sentiment coefficient associated with word w_{ij} for one aspect s_i, and α_i represents the weight of the reviewer on aspect s_i, and K is the total number of aspects in the review dataset. r_d is the overall rating of the review, which is modeled as drawn from a Gaussian distribution with a mean of $\sum_{i=1}^{K} \alpha_i \sum_{j=1}^{N} \beta_{ij} w_{ij}$. LARA discovers the latent aspect ratings by fitting a regression model based on the aspects discussed in the reviews, their latent weights, and the observed overall ratings. With LARA, we can decompose an overall rating into aspect-specific ratings, such as *4.5 star* for life convenience and *3.5 star* for safety.

4. Experiments

In this section, we experiment with the multiple computational models presented above for analyzing the semantics and sentiments of neighborhood reviews. The experimental dataset was collected from Niche (https://www.niche.com), a website that allows users to review neighborhoods. The dataset contains the numeric rating (from *1 star* to *5 star*) and the textual comment of each review. Figure 1(a) shows a screenshot of some neighborhood reviews.

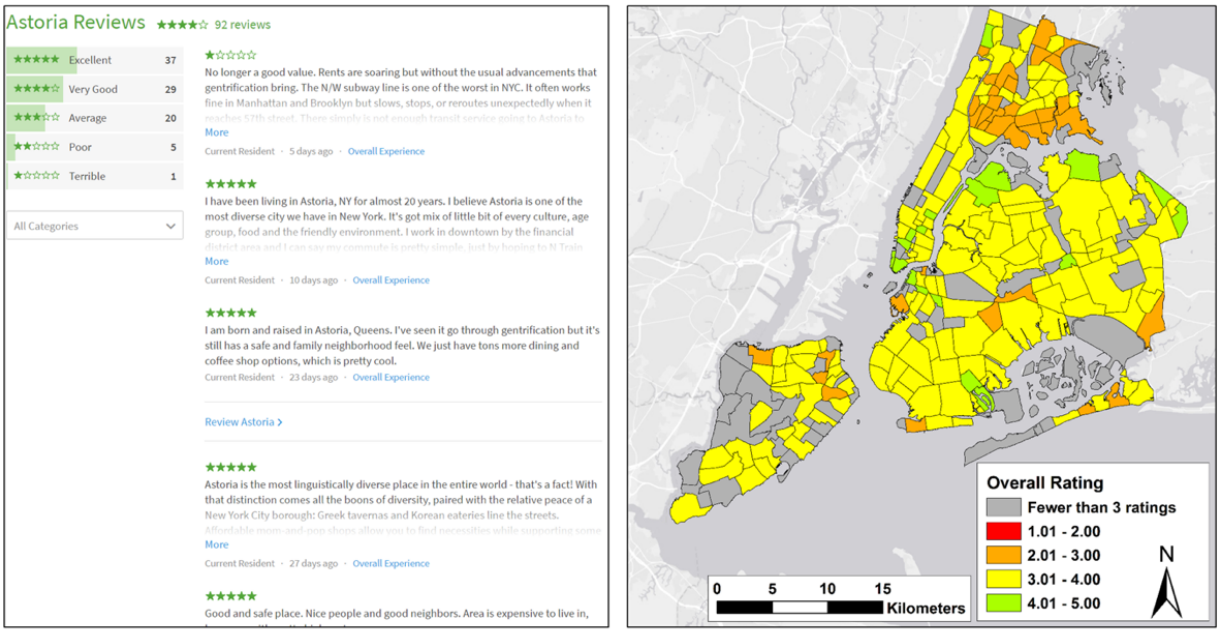

Figure 1. (a) Some neighborhood reviews on Niche; (b) average ratings of NYC neighborhoods based on Niche review data.

NYC was selected for the experiments because its neighborhoods have received many reviews on Niche. In addition, there are rich socioeconomic public data about the neighborhoods in NYC for our comparative study later. We collected review data from Niche on May 2, 2017, and all neighborhood

reviews about NYC published on and before that date were retrieved. In total, we have collected 7,673 review data records covering 233 NYC neighborhoods. These reviews are contributed by 7,673 distinct users (Niche prevents the same user from reviewing the same neighborhood multiple times). By performing an exploratory data analysis, we found that the total lengths of the reviews range from 3 words to 339 words with a median of 21 words. A histogram of the review lengths is plotted in Figure 2. Table 1 shows five example review records. Figure 1(b) shows NYC neighborhoods with their average ratings based on this review dataset.

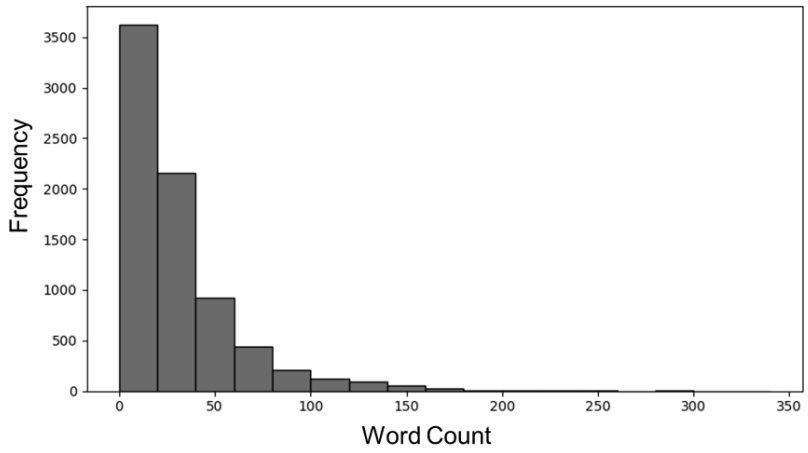

Figure 2. A histogram of the lengths of the neighborhood reviews (the bin width is 20).

Table 1. Five neighborhood review data records.

User ID	Numeric Rating	Textual Comment
60c2cbdd…	4	Crime is very rare here. If there's any crime, it'll have to do with the local homeless guy …
64b5119d…	1	You barely see police around the area solving real crimes and catching criminals.
7322070e…	3	There are a lot of great restaurants cropping up around and there is a corridor of great retail stores, but in general the most accessible salons and grocery stores are below average.
d6d82187…	4	It is clean and a nice community.
62b57671…	5	It's very vibrant if you've never been to new york before. The people here are very helpful when it comes to looking for directions or good places…

4.1 Identifying semantic topics

We use LDA and MG-LDA to identify the semantic topics talked by people about NYC neighborhoods. Before applying topic modeling, we first perform data preprocessing by removing the punctuations and

stopwords in the neighborhood reviews, and all words are converted to lowercase. In addition to typical English stopwords, such as "is", "are", and "of", we also remove the words and phrases, such as "New York City", "NYC", "Manhattan", "Queens", and "Brooklyn", that people frequently use to refer to the city and its sub areas. The implementation of LDA from the R package "topicmodels" is used. To find a suitable K for LDA, we iterate K from 2 to 10 and compute the four metrics discussed in Section 3.2. We use the implementations of the four metrics from the R package "ldatuning". We limit the iteration range of K within 10, since we aim to discover the major topics rather than too detailed ones. At each K, we run 2000 iterations of the Gibbs sampling to identify the semantic topics and then calculate the values of the four metrics, as shown in Figure 3.

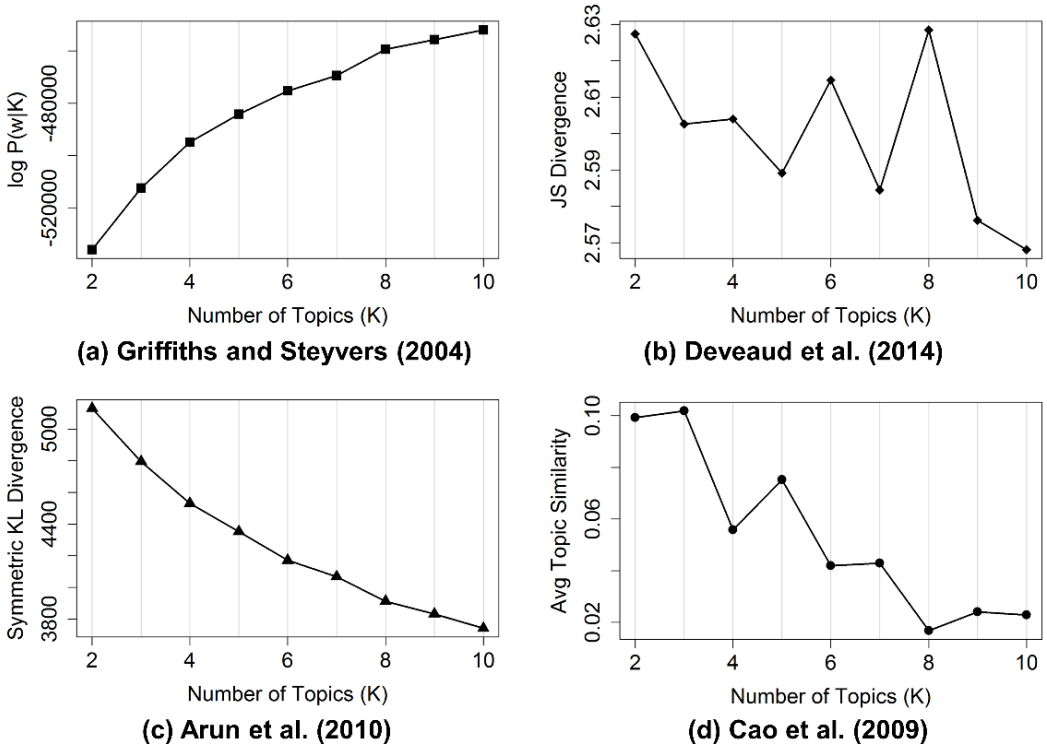

Figure 3. The values of the four metrics at different Ks.

In Figure 3(a) and Figure 3(b), better Ks are at the locations where the metrics achieve higher values. In Figure 3(c) and Figure 3(d), better Ks are at the locations where the metrics achieve lower values. Ideally, we would expect these four metrics to suggest similar results; in reality, however, they do not completely agree with each other: Figure 3(a) and Figure 3(c) show steadily increasing or decreasing values, while Figure 3(b) and 3(d) show more fluctuated values at different Ks. Based on the experiment results, we choose $K = 8$, where (Deveaud et al. 2014) and (Cao et al. 2009) achieve their maximum and minimum values, while (Griffiths and Steyvers 2004) and (Arun et al. 2010) show fairly high and low values. Figure 4 plots the word clouds of the discovered eight topics based on the top 20 terms with the highest probabilities in each topic. These topics are manually labeled as: *Crime and Safety*, *Community Friendliness*, *Cultural Diversity*, *Local Weather*, *Life Convenience*, *Employment Opportunity*, *Transportation Convenience*, and *Housing Conditions*. We also examine and evaluate the discovered topics when K equals to other values, and find that the topics are often intermixed together. We present the discovered topics when K = 3, 7, and 9 in Appendix A.

Figure 4. Eight topics discovered by LDA.

We test the effectiveness of MG-LDA on the same dataset of neighborhood reviews. The MG-LDA Python implementation (https://github.com/m-ochi/mglda) is utilized. To make a fair comparison, we set K^{loc} of MG-LDA to 8 since K^{loc} defines the number of ratable aspects. Based on the parameter recommendation of MG-LDA (Titov and McDonald 2008), K^{gl} is set to 20 which exceeds 2 factors of K^{loc}. We then run 2000 iterations of Gibbs sampling with MG-LDA, and Figure 5 shows the identified eight local topics.

To our surprise, many of the eight local topics discovered by MG-LDA have mixed themes. For example, the topic related to safety is merged into three topics. Meanwhile, some topics discovered by LDA, such as *Local Weather*, are not discovered by MG-LDA. With curiosity, we look into the 20 global topics discovered by MG-LDA, and found several topics related to *Local Weather*, *Cultural Diversity*, and *Job Opportunities* respectively. Meanwhile, there are also abstract global topics which are hard to interpret. This result suggests that LDA may in fact be a more effective approach than MG-LDA in identifying semantic topics from neighborhood review data. One possible explanation is that people tend to care about the same core aspects of neighborhoods, such as location, safety, and life convenience. This can be differentiated from the MP3 player review data examined by Titov and McDonald (2008) when proposing MG-LDA, in which people also care about the special features (e.g., radio recording) provided by different MP3 brands. On the other hand, MG-LDA might show better performances when the review data are about neighborhoods in different cities. In such a situation, MG-LDA may better separate the global topics on city-specific features from those more general and ratable topics.

Figure 5. Eight local topics discovered by MG-LDA.

4.2 Quantifying aspect sentiments

With the identified eight semantic topics, we employ the three methods described in Section 3.3 to quantify the aspect-specific ratings: the naïve approach, the sentiment lexicon-based approach, and LARA. The naïve approach assumes that every discussed aspect has the same rating as the overall rating of a review. For the lexicon-based approach, we employ AFINN-111 (Nielsen 2011) which provides a list of fine-scale sentiment words (the sentiment of each word is quantified from -5 to 5) suitable for Web texts with informal words (Hansen et al. 2011). We re-scale the sentiment of each word in AFINN-111 to *1* to *5 star* to fit our neighborhood review data. LARA derives aspect-specific ratings using a latent rating regression model and does not require an external sentiment lexicon. We implement the naïve approach and the sentiment lexicon-based approach using Python, and use the implementation of LARA from its authors (http://www.cs.virginia.edu/~hw5x/Codes/LARA.zip).

A neighborhood review typically discusses only a subset of the eight semantic topics, and different reviewers often focus on different subsets of the topics. As a result, we need to first identify the particular aspects discussed by one particular review. LARA has its own approach for detecting the aspects of a review based on bootstrapping. We feed LARA with three frequent words from each of the eight semantic topics derived by the LDA model. The three frequent words for each aspect are shown in Table 2. This experiment design is based on the recommendation of the LARA authors (Wang et al. 2010) in providing a few keywords for each topic to inform the model.

Table 2. Three frequent words selected from each LDA topic to inform LARA about the aspects.

Topic	Frequent Words
Crime and Safety	crime, safe, police
Community Friendliness	family, friendly, neighbor
Cultural Diversity	diverse, culture, diversity
Local Weather	weather, winter, summer
Life Convenience	store, restaurant, shop
Employment Opportunity	job, work, employment
Transportation Convenience	transportation, train, bus
Housing Conditions	apartment, housing, rent

For the naïve and sentiment lexicon-based approaches, we identify the aspects of each review using a word probability based method. First, a review is divided into sentences using any punctuations including comma. This is because one reviewer may talk about two aspects in one long sentence, such as "The neighborhood is very quiet and clean, although it does not have many restaurants." Meanwhile, we also concatenate the nearby sentences which have fewer than three words to avoid the issue of misclassifying short phrases as sentences. Second, we use the top 20 words of each topic from the LDA model (shown in Figure 4) and their probabilities to calculate the scores of a sentence belonging to different aspects using Equation 6:

$$s_k = \sum_{i=1}^{N} p_{ik} I_{ik} \quad (6)$$

where s_k is the score of a sentence belonging to the aspect k, p_{ik} is the probability of word i in aspect k from the LDA model, and I_{ik} is an indicator variable indicating whether word i exists in the top 20 words of aspect k. Based on the calculated aspect scores, a sentence will be assigned to the aspect that has the highest score. No tie is found in the scores since the probabilities of the words have sufficient digits to differentiate the aspect scores. In case no aspect receives a score, this sentence is considered as not related to any aspect. Finally, we combine the sentences and their identified aspects together as the aspects of the entire review. Aspect-specific ratings are then derived based on the identified aspects, the aspect related words, and the overall rating of a review.

4.3 Evaluations

In this subsection, we evaluate the quality of the derived aspect-specific ratings. In order to have a set of ground truth data for quantitative evaluations, we randomly select 1,000 neighborhood reviews from our dataset and make use of the crowdsourcing platform Amazon's Mechanical Turk (AMT) to label the aspect-specific ratings of the reviews. Figure 6 shows the Web interface designed to collect human labels.

For each review, an AMT user needs to answer two questions. For the first question, the user can check one or multiple of the eight aspects, and once an aspect is checked, a rating bar will show up asking the user to rate it from *1 star* to *5 star* with an interval of *0.5 star*. To increase the quality of the collected data, we added an attention test which asks the user to always check this option and select a particular rating. The annotations of the users who do not pass the attention test are not included in the final dataset.

Each review is annotated by five different AMT users who passed the attention test. For the second question, the user has the opportunity to add an aspect which is not provided in Question 1, or to choose "None" if no new aspect is identified. It took in total about eight days to complete the labeling of the 1,000 neighborhood reviews, and the labels are provided by 419 different AMT users. With the obtained data, we use the strategy of majority voting by adopting the aspects checked by at least three users and averaging their scores for the adopted aspects. The averaged aspect-specific ratings of the reviews are then used as the ground truth for evaluation.

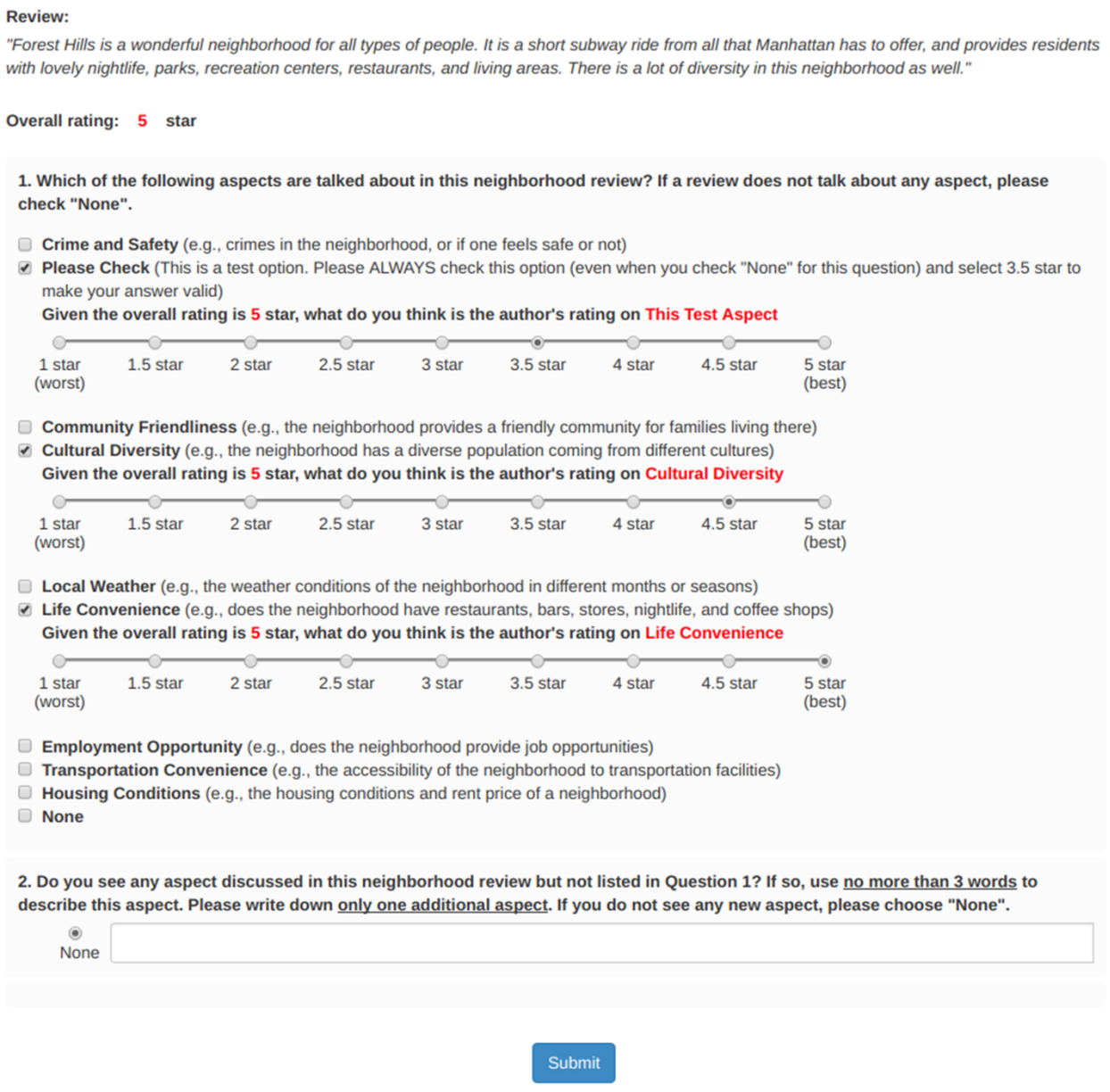

Figure 6. The Web interface designed to collect human annotations for neighborhood reviews using AMT.

To compare the results output by our models with the ground truth data, we use the metric of average rating loss (ARL) defined in Equation 7.

$$ARL = \frac{1}{N}\sum_{i=1}^{N} | r_{label} - r_{predict} | \quad (7)$$

where r_{label} is the average aspect-specific rating provided by human users, $r_{predict}$ is the predicted aspect-specific rating, and N is the total number of aspect-specific ratings in the data. A good model should predict ratings close to the ground truth and therefore should have a low ARL. There are also false positives and false negatives in the aspect detection process. In the former case, a model falsely detects an aspect which was not labeled by humans; in the latter case, a model misses an aspect that was labeled by humans. To account for these false positives and false negatives when calculating the ARLs, we assign them with a penalty of 2.0 which is the average of the maximum rating loss 4.0 and the minimum rating loss 0.0. We compute the ARLs of the three models based on the 1,000 ground-truth data records, and plot the distributions of the ratings in all aspects as well as the ratings in each individual aspect in Figure 7. Their corresponding ARLs are also provided after the histograms.

As can be seen, LARA has lower ARLs for its predicted ratings in both all aspects and each individual aspect, compared with the other two models. While the naïve approach simply assumes that the reviewer has the same rating as the overall rating toward every aspect, this approach is not entirely unreasonable since in some cases a reviewer indeed has similar or even the same aspect-specific ratings as his/her overall rating. In fact, the naïve approach performs better than the sentiment lexicon-based approach which relies on an external sentiment lexicon. As discussed previously, such external sentiment lexicons may not fit the word sentiments in a target domain and cannot reflect the different sentiments expressed by different people using the same words (e.g., "good"). In addition, we can see that the three models perform relatively well and with higher agreements for some aspects, such as *Crime and Safety*, but not so well for some other aspects, such as *Community Friendliness*. This performance difference suggests that there exist varied difficulties in deriving correct ratings for different aspects.

Besides evaluating the predicted ratings, we also look into the additional aspects suggested by the AMT users (via Question 2), and we do find some valid aspects such as *School*, *Park*, and *Outdoor Activity*. In the output of our LDA model, these aspects are merged with others and are not explicitly labeled (see Figure 4). For example, *Park* is merged with the topic of *Transportation Convenience,* while *School* is merged with *Community Friendliness*. Future work could combine computational models and crowdsourcing approaches to obtain more comprehensive topics about neighborhoods.

4.4 Results and discussion

In this subsection, we discuss the results obtained by applying LARA to the neighborhood review dataset. Specifically, we will discuss the decomposed ratings for individual reviews and the average ratings aggregated to the neighborhoods.

4.4.1 The decomposed review ratings

Here, we use a number of examples organized into three groups to demonstrate the aspect-specific ratings decomposed by LARA.

Group 1: Zero-aspect reviews

Some reviews provide only brief comments without addressing a specific aspect of a neighborhood. When no aspect is detected in the review content, the decomposed output is empty (or *Null*). An example is: "*Have been here for 10 years. I like it.*" (User Rating: *4 star*; Decomposed Rating: *Null*).

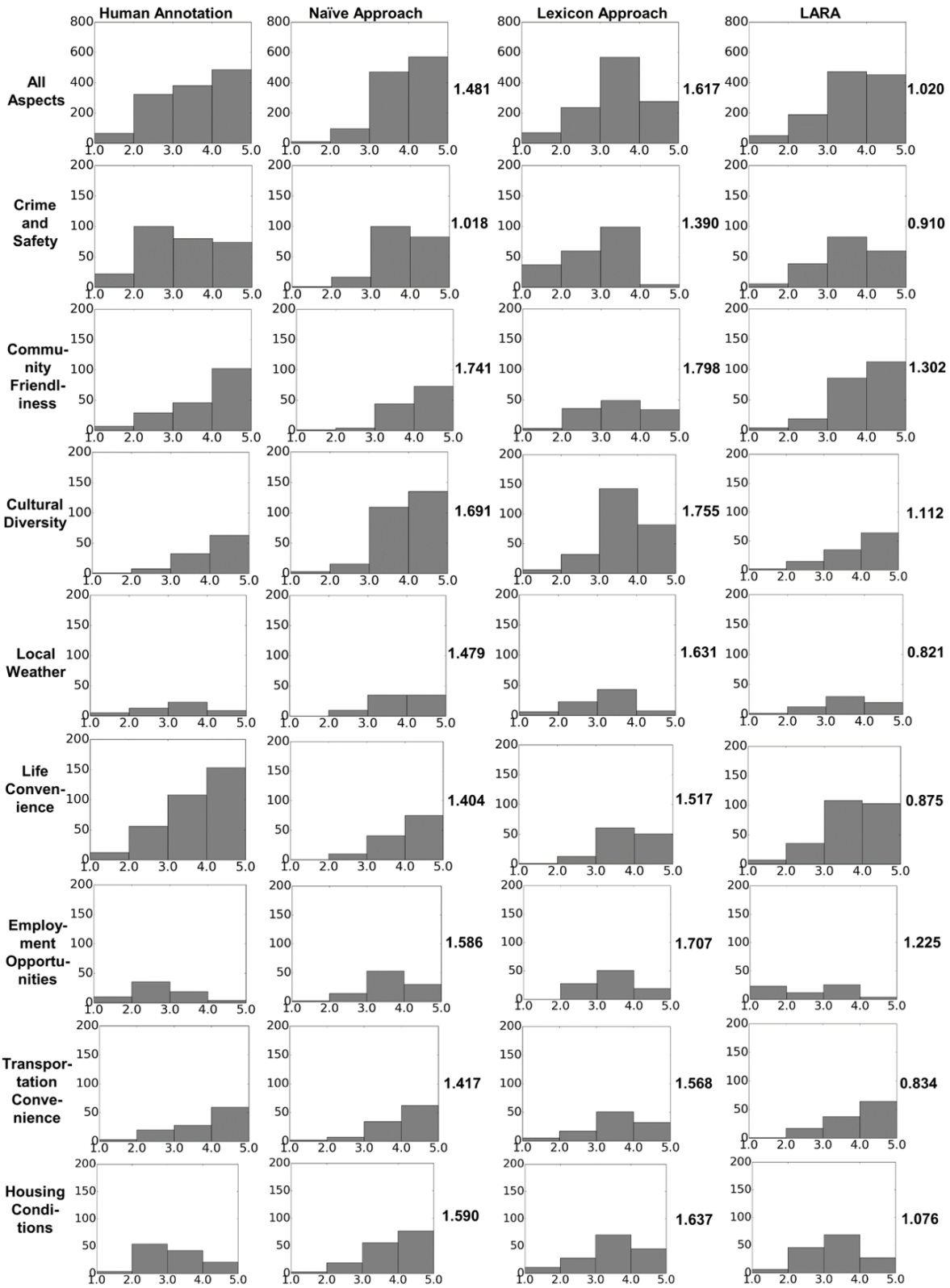

Figure 7. The distributions of the review ratings from the human annotators and the review ratings generated by the three models and their ARLs.

Group 2: Single-aspect reviews

Some reviews focus on one single aspect of a neighborhood. In those cases, the decomposed rating is the same as their overall rating. An example is "*There is a good amount of crime where I live and the cops are unresponsive.*" (User Rating: *1 star*; Decomposed Rating: *Crime and Safety*: *1 star*).

Group 3: Multiple-aspect reviews

Many reviewers comment on more than one aspect of a neighborhood within their reviews. When multiple aspects are detected, LARA decomposes the overall ratings into aspect-specific ratings based on the discussed topics and the sentiments implied in the comment words. An example is shown as follows. "*The accessibility with everything is perfect, everything is around restaurants, notaries, clothing stores, pharmacies. However it is close to neighborhoods with high crime rates but not in this particular area.*" (User Rating: *3 star*; Decomposed Ratings: *Crime and Safety*: *2.705 star*, *Life Convenience*: *3.197 star*).

4.4.2 The aggregated neighborhood perceptions

With the decomposed ratings from different reviewers, we aggregate these ratings to generate average neighborhood perception maps under the identified semantic topics. Specifically, we produce eight aspect-specific neighborhood perception maps based on our experiment results, as shown in Figure 8. To reduce bias, neighborhoods reviewed by fewer than three people in one aspect are not taken into account and are displayed in gray in the maps.

Two interesting observations are obtained from Figure 8. First, the average aspect-specific ratings are different from each other, and they show distinctions from the average overall ratings in Figure 1(b). This result suggests that we can indeed obtain additional information and finer knowledge on aspect-specific neighborhood perceptions by performing semantic and sentiment analysis on the review data. Second, some aspect-specific ratings seem to show spatial autocorrelations with high ratings clustered with other high ratings. To further examine the spatial autocorrelation effect, we compute Global Moran's I for each aspect-specific rating map, and queen's case is used to specify the nearby features of a target feature. Table 3 shows the analysis result.

Table 3. Spatial autocorrelations for the eight aspect-specific ratings.

Neighborhood Aspect	Moran's I
Crime and Safety	0.400 ($p < 0.01$)
Community Friendliness	0.232 ($p < 0.01$)
Cultural Diversity	0.238 ($p < 0.05$)
Local Weather	-0.004 ($p = 0.93$)
Life Convenience	0.360 ($p < 0.01$)
Employment Opportunity	-0.011 ($p = 0.98$)
Transportation Convenience	0.309 ($p < 0.01$)
Housing Conditions	0.259 ($p < 0.01$)

It can be seen that the ratings of six aspects show statistically significant and positive spatial autocorrelations, while two aspects do not show clear spatial autocorrelations. This result suggests that nearby neighborhoods are likely to be perceived similarly by people for some aspects, such as *Crime and Safety* and *Life Convenience*, but not for some other aspects, such as *Local Weather* and *Employment Opportunity*.

5. Comparative analyses between subjective perceptions and objective neighborhood attributes

The derived perceptions on different aspects of neighborhoods reflect the subjective feelings of people toward their living environments. How do these subjective feelings relate to the more objective socioeconomic attributes of neighborhoods, such as the numbers of restaurants, bus stops, crimes, or unemployment rates? In this section, we compare the subjective neighborhood perceptions derived from the review data with the objective neighborhood attributes. Specifically, we make comparisons for the aspects of *Crime and Safety*, *Cultural Diversity*, *Life Convenience*, *Employment Opportunity*, *Transportation Convenience*, and *Housing Conditions*. We do not perform comparisons for the aspects of *Community Friendliness* and *Local Weather,* since there is no objective dataset capturing the friendliness of a community and the weather condition within the same city is relatively homogenous.

5.1 Comparative analyses

Crime and Safety: We compare the subjective perceptions of people toward neighborhood safety with the crime data records from the New York City Police Department (NYPD). This dataset contains the locations of the reported crime incidents and crime categories. We aggregate individual crime incidents to neighborhoods by summing up the crimes falling into each neighborhood and calculating the counts of both total crimes and the crimes in different categories. Different neighborhoods have different areas and populations, and by chance the neighborhoods with larger areas and higher populations are more likely to have higher crime counts even when the likelihood of observing or experiencing a crime at a location is similar. Therefore, we normalize the crime counts by the areas and populations of the neighborhoods. Pearson's correlations are performed, and the results are summarized in Table 4. The result suggests that the total crime and most crime types have a statistically significant and negative correlation with the perceived neighborhood safety. Particularly, the correlation under the crime type *Alcohol and Drug* is the highest with a value of -0.542. Interestingly, the term "drug" is one of the top terms under the topic of *Crime and Safety*, which suggests that many reviewers indeed discuss drug related activities when reviewing neighborhood safety issues.

Cultural Diversity: We compare the perceptions of people toward cultural diversity with 2016 demographics data in NYC from American Community Survey (ACS). This dataset contains information about the populations of white, black or African American, Asian, and other races. Based on this dataset, we compute the Shannon Diversity Index (Jost 2006) to quantify the demographic diversity of each neighborhood (Equation 8).

$$H = -\sum_{i=1}^{R} p_i \ln p_i \tag{8}$$

where p_i is the percentage of a racial group population with regard to the entire population of the target neighborhood, and R is the total number of racial groups in the neighborhood. A higher H indicates that a neighborhood is ethnically more diverse. We then correlate the Shannon Diversity Index with the perceived cultural diversity, and obtain a correlation coefficient 0.235 ($p = 0.09$). This result suggests that

there is no statistically significant relation between the subjective diversity perceptions and the objective diversity index.

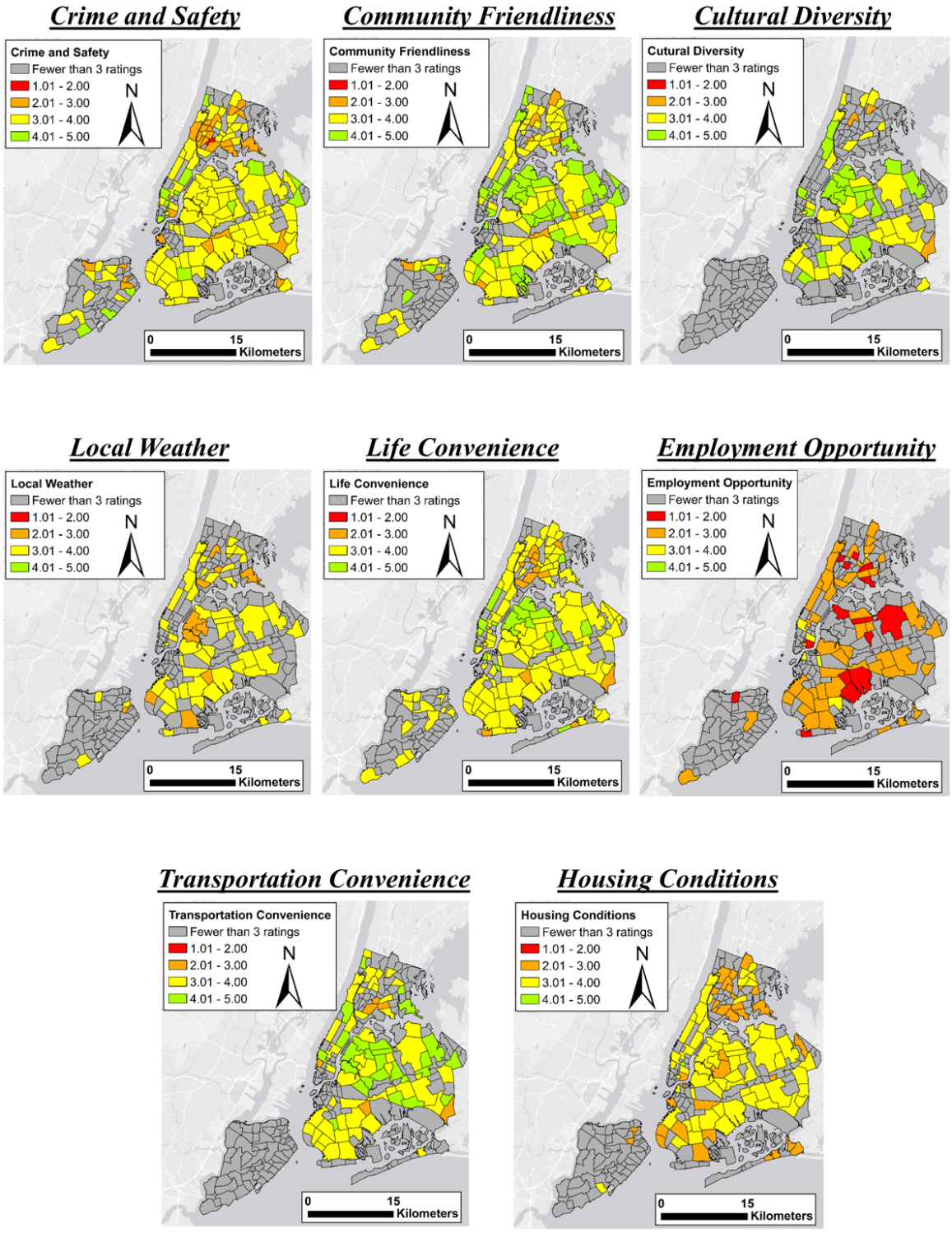

Figure 8. Average neighborhood perception maps for the eight semantic topics.

Table 4. Correlation coefficients between safety perceptions and crime data.

	Normalized by Area	Normalized by Population
Total Crime	-0.275 ($p < 0.01$)	-0.379 ($p < 0.01$)
Violent Crime	-0.369 ($p < 0.01$)	-0.530 ($p < 0.01$)
Property Crime	-0.040 ($p = 0.60$)	-0.131 ($p = 0.21$)
Drug and Alcohol	-0.407 ($p < 0.01$)	-0.542 ($p < 0.01$)
Racket and Gamble	-0.194 ($p < 0.05$)	-0.287 ($p < 0.05$)
Traffic Crime	-0.266 ($p < 0.01$)	-0.402 ($p < 0.01$)
Other Crime	-0.432 ($p < 0.01$)	-0.519 ($p < 0.01$)

Life Convenience: We compare the subjective perceptions of people toward the *Life Convenience* of neighborhoods with the points of interest (POIs) data from Foursquare (Yang et al. 2015). Foursquare is a location-based social media which allows users to *check-in* at POIs such as restaurants, cinemas, and stores. Here, we use only the locations of the POIs from Foursquare, and the numbers of user check-ins are not used. We aggregate the raw POI locations to neighborhoods by summing up their total counts inside each neighborhood. These POI counts are then normalized by the total areas and populations of the neighborhoods respectively. Pearson's correlations are performed and correlation coefficients of 0.267 ($p < 0.01$) and 0.231 ($p < 0.01$) are observed. This result suggests a statistically significant and positive correlation between the POIs in neighborhoods and the perceived life convenience.

Employment Opportunities: We compare people's subjective perceptions of employment opportunities with 2016 ACS block group level unemployment rate from the U.S. Census Bureau. The rationale of using this dataset is that more employment opportunities can lead to low employment rate in a neighborhood. Since the original Census data are at block group level, we aggregate them into neighborhood level using weighted average before the comparison. By performing Pearson's correlation, we obtain a coefficient value -0.116 ($p = 0.30$) which is not statistically significant. Thus, no clear relation is observed between the perceived employment opportunity and the unemployment rates in the neighborhoods.

Transportation Convenience: We compare the perceived transportation convenience with the 2017 NYC bus and subway stops and routes from the Metropolitan Transportation Authority (MTA). We start by counting the number of bus stops and subway stations within a neighborhood, and normalizing the stop counts by the total areas and populations of the neighborhoods. By performing correlation between the normalized stop counts and the perceived transportation convenience, we obtain coefficients of 0.021 ($p = 0.43$) and 0.160 ($p = 0.51$), which are not statistically significant. Considering the situations in which a neighborhood may still be perceived as convenient if it has only one but central station reaching to many other locations, we compute the betweenness of the bus stops and subway stations using Equation 9:

$$g(v) = \frac{C_{st}(v)}{(N-1)(N-2)/2} \tag{9}$$

where v is a stop (a node) in a transportation network, $C_{st}(v)$ represents the count of the shortest paths between any two stops s and t that pass the stop v, and $(N-1)(N-2)/2$ is the total number of possible node pairs excluding node v in the network. We then perform correlation analysis based on the betweenness and the perceived transportation convenience. The resulted correlation coefficients are 0.032 (p = 0.84) and 0.150 (p = 0.42) respectively. This result indicates no significant correlations between the perceived and the objective transportation convenience.

Housing Conditions: We use the Rolling Sales Data from The Department of Finance of NYC as the objective dataset with which the perceived housing conditions are compared. This dataset was selected because housing affordability and housing conditions are mentioned in the reviews. While it is difficult to objectively quantify the conditions of houses, the sales data of residential houses/apartments provide more objective price information. We use the median of the residential sales prices of a neighborhood to represent the general price level in it. Neighborhoods that have fewer than three sales records for dwelling houses are removed. We then correlate the median housing prices with the perceived housing conditions, and obtain a correlation coefficient 0.300 (p = 0.08), which suggests no significant correlation.

5.2 Discussion

We have performed a series of comparative analyses between the subjective perceptions and the objective socioeconomic attributes of the neighborhoods. For the objective data, we normalized the extensive data, such as crime counts, bus stop counts, and POI counts, based on neighborhood areas or populations, while not normalizing the intensive data, such as the diversity indices, unemployment rates, and median housing prices. The topics, *Crime and Safety* and *Life Convenience*, show statistically significant correlations between the subjective perceptions and objective data. The other four topics, *Cultural Diversity*, *Housing Conditions, Employment Opportunities,* and *Transportation Convenience*, do not show significant correlations between the subjective and objective data.

Why do these insignificant or weak correlations happen? To answer this question, we identify three possible reasons. First, the people who wrote online neighborhood reviews were self-selected and may not represent the entire population living in the corresponding neighborhoods. Accordingly, sentiments extracted from their reviews may not represent the sentiments of all of the people in these neighborhoods. Unfortunately, for websites such as Niche, we cannot obtain the demographic information of the users who wrote these reviews. This is a major limitation of online neighborhood review data compared with the comments collected through controlled surveys or interviews. When demographic information of online reviewers is available, we could adjust the derived sentiments accordingly, e.g., based on the ages and genders of the review writers. Second, as pointed out by Rogerson (1995), the perception of people is an internal and complex psychological process which does not accurately reflect the environment. People can have different expectations toward their living environment, depending on their backgrounds, cultures, socioeconomic statuses, and life trajectories. Accordingly, the same neighborhood can be perceived differently by different people. For example, a housing price that is unacceptably high for a low-income individual can be affordable for a middle-class. Similarly, people may have different expectations on what a good *Cultural Diversity* is, and accordingly increases in demographic diversity may not lead to corresponding increases in review ratings on diversity. Third, online neighborhood reviewers may look into neighborhoods from other perspectives which are not measured by objective data. For example, in the aspect of *Transportation Convenience*, our objective dataset on bus and subway stops and routes quantifies the convenience for a person to travel from one neighborhood to other locations, whereas the

reviews show that people also care about the reliability of bus services (e.g., whether the buses arrive on time or are often late), the politeness of bus drivers, and the frequency of traffic congestions. For *Employment Opportunities*, our objective dataset captures the unemployment rate, whereas the reviews show that people also care about the quality of jobs, e.g., a person may have a job (and thus employed) but the job is working at a fast food restaurant. While these other perspectives can lead to weak or insignificant correlations, they can also help identify neighborhood problems that are not captured by existing data. In sum, multiple reasons may have contributed to the weak or insignificant correlations between the extracted subjective perceptions and the objective data. These reasons suggest that online neighborhood review data, while having their values, should be used more critically or as a complementary data source combined with other types of data.

6. Conclusions and future work

In this paper, we performed a semantic and sentiment analysis on online neighborhood review data that emerged in recent years. Such an analysis can help us understand the different aspects of neighborhoods perceived by people and can reveal potential neighborhood problems. We experimented with multiple computational models, including LDA, MG-LDA, naïve sentiment analysis, lexicon-based sentiment analysis, and LARA, for identifying semantic topics and deriving aspect-specific perceptions. An online neighborhood review dataset contributed by 7,673 distinct Web users and covering 233 NYC neighborhoods was collected for the experiments. We conducted both qualitative and quantitative evaluations by comparing the results of different models and benchmarking their performances based on a sample of human-labeled neighborhood reviews obtained via a crowdsourcing platform, Amazon's Mechanical Turk. We also obtained average aspect-specific rating maps for the eight identified neighborhood aspects, performed spatial autocorrelation analyses, and identified the aspects that show significant spatial autocorrelations such as *Crime and Safety* and *Life Convenience*. Finally, we compared the derived subjective perceptions with the more objective socioeconomic datasets.

This work has its limitations which can be improved in future research. First, while we have examined a few thousand reviews, next-step studies can leverage a larger dataset with more neighborhood reviews. This is possible as people are continuously contributing new reviews, and websites similar to Niche may emerge in the coming years. A larger dataset can also enable studies on the temporal changes of the perceptions of people toward neighborhoods. Such studies can be useful for evaluating the effectiveness of urban planning policies by understanding whether a policy indeed improves the satisfactions of people after its implementation. Second, the models used in this work for analyzing neighborhood reviews have their own assumptions and limitations. LDA is a bag-of-words model which ignores the order of words in texts, while LARA assumes that the overall rating of a reviewer follows a Gaussian distribution with its mean as the weighted sum of the aspect-specific ratings. Due to these assumptions and limitations, the used models cannot interpret the reviews in a completely accurately manner. Other methods can be employed to improve the accuracy of parsing and analyzing textual review data. Meanwhile, and as pointed out by Kwan (2016), different algorithms can affect the geographic knowledge generated from the same datasets. Therefore, it is important to experiment with multiple models and compare the obtained results. Third, as revealed in the comparative analyses between subjective perceptions and objective socioeconomic data, online neighborhood reviews are susceptible to data bias issues. Future research may combine online neighborhood reviews with other datasets, such as the data collected via controlled surveys or interviews, to further study the perceptions of people toward their living environments.

References

Adams, B. & G. McKenzie. 2013. Inferring thematic places from spatially referenced natural language descriptions. In *Crowdsourcing geographic knowledge*, 201-221. Springer.

Adams, B., G. McKenzie & M. Gahegan. 2015. Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th International Conference on World Wide Web*, 12-22. ACM.

Arun, R., V. Suresh, C. Veni Madhavan & M. Narasimha Murthy (2010) On finding the natural number of topics with latent dirichlet allocation: Some observations. *Advances in Knowledge Discovery and Data Mining***,** 391-402.

Ballatore, A. & B. Adams. 2015. Extracting Place Emotions from Travel Blogs. In *Proceedings of AGILE*, 1-5.

Beineke, P., T. Hastie, C. Manning & S. Vaithyanathan. 2004. Exploring sentiment summarization. In *Proceedings of the AAAI spring symposium on exploring attitude and affect in text: theories and applications*.

Blei, D. M. (2012) Probabilistic topic models. *Communications of the ACM,* 55**,** 77-84.

Blei, D. M., A. Y. Ng & M. I. Jordan (2003) Latent dirichlet allocation. *Journal of machine Learning research,* 3**,** 993-1022.

Brown, G. & D. Weber (2011) Public Participation GIS: A new method for national park planning. *Landscape and urban planning,* 102**,** 1-15.

Bugs, G., C. Granell, O. Fonts, J. Huerta & M. Painho (2010) An assessment of Public Participation GIS and Web 2.0 technologies in urban planning practice in Canela, Brazil. *Cities,* 27**,** 172-181.

Cao, J., T. Xia, J. Li, Y. Zhang & S. Tang (2009) A density-based method for adaptive LDA model selection. *Neurocomputing,* 72**,** 1775-1781.

Cataldi, M., A. Ballatore, I. Tiddi & M.-A. Aufaure (2013) Good location, terrible food: detecting feature sentiment in user-generated reviews. *Social Network Analysis and Mining,* 3**,** 1149-1163.

Ceccato, V. & F. Snickars. 1998. Objective and subjective indicators to evaluate quality of life (QOL) in two districts in the Stockholm region. In *Urban Ecology*, 273-277. Springer.

Ceccato, V. A. & F. Snickars (2000) Adapting GIS technology to the needs of local planning. *Environment and planning B: Planning and design,* 27**,** 923-937.

Cervone, G., E. Sava, Q. Huang, E. Schnebele, J. Harrison & N. Waters (2016) Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. *International Journal of Remote Sensing,* 37**,** 100-124.

Crandall, D. J., L. Backstrom, D. Huttenlocher & J. Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, 761-770. ACM.

Das, D. (2008) Urban quality of life: A case study of Guwahati. *Social Indicators Research,* 88**,** 297-310.

Dempster, A. P., N. M. Laird & D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)***,** 1-38.

Deveaud, R., E. SanJuan & P. Bellot (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique,* 17**,** 61-84.

Eby, J., P. Kitchen & A. Williams (2012) Perceptions of quality life in Hamilton's neighbourhood hubs: A qualitative analysis. *Social Indicators Research,* 108**,** 299-315.

Feldman, R. (2013) Techniques and applications for sentiment analysis. *Communications of the ACM,* 56**,** 82-89.

Gao, S., K. Janowicz, D. R. Montello, Y. Hu, J.-A. Yang, G. McKenzie, Y. Ju, L. Gong, B. Adams & B. Yan (2017) A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science***,** 1-27.

Geman, S. & D. Geman (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence***,** 721-741.

Goodchild, M. F. (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal,* 69**,** 211-221.

Griffiths, T. L. & M. Steyvers (2004) Finding scientific topics. *Proceedings of the National academy of Sciences,* 101**,** 5228-5235.

Haklay, M. 2013. Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing geographic knowledge*, 105-122. Springer.

Hansen, L. K., A. Arvidsson, F. Å. Nielsen, E. Colleoni & M. Etter. 2011. Good friends, bad news-affect and virality in twitter. In *Future information technology*, 34-43. Springer.

Helburn, N. (1982) Geography and the Quality of Life. *Annals of the Association of American Geographers,* 72**,** 445-456.

Hochman, N. & L. Manovich (2013) Zooming into an Instagram City: Reading the local through social media. *First Monday,* 18.

Hollenstein, L. & R. Purves (2010) Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science,* 2010**,** 21-48.

Hu, M. & B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168-177. ACM.

Hu, Y., S. Gao, K. Janowicz, B. Yu, W. Li & S. Prasad (2015) Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems,* 54**,** 240-254.

Huang, Q. (2017) Mining online footprints to predict user's next location. *International Journal of Geographical Information Science,* 31**,** 523-541.

Jenkins, A., A. Croitoru, A. T. Crooks & A. Stefanidis (2016) Crowdsourcing a collective sense of place. *PloS one,* 11**,** 1-20.

Jo, Y. & A. H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 815-824. ACM.

Jost, L. (2006) Entropy and diversity. *Oikos,* 113**,** 363-375.

Kao, A. & S. R. Poteet. 2007. *Natural language processing and text mining*. Springer Science & Business Media.

Kasper, W. & M. Vela. 2011. Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference*, 45-52.

Keßler, C., C. Rinner & M. Raubal. 2005. An argumentation map prototype to support decision-making in spatial planning. In *Eighth Conference on Geographic Information Science,* ed. P. M. Toppen F., 135-142. Portugal.

Khaef, S. & E. Zebardast (2016) Assessing Quality of Life Dimensions in Deteriorated Inner Areas: A case from Javadieh Neighborhood in Tehran Metropolis. *Social indicators research,* 127**,** 761-775.

Kling, F. & A. Pozdnoukhov. 2012. When a city tells a story: urban topic analysis. In *Proceedings of the 20th international conference on advances in geographic information systems*, 482-485. ACM.

Kwan, M.-P. (2016) Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers,* 106**,** 274-282.

Lee, Y., N. Gu & S. An (2016) Residents' perception and use of green space: Results from a mixed method study in a deprived neighbourhood in Korea. *Indoor and Built Environment,* 26**,** 855 - 871.

Liu, B. (2012) Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies,* 5**,** 1-167.

Martin, M. E. & N. Schuurman (2017) Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers***,** 1-12.

McKenzie, G., K. Janowicz, S. Gao, J.-A. Yang & Y. Hu (2015) POI Pulse: A Multi-granular, Semantic Signature-Based Information Observatory for the Interactive Visualization of Big Geosocial Data. *Cartographica: The International Journal for Geographic Information and Geovisualization,* 50**,** 71-85.

Nielsen, F. Å. (2011) A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pang, B. & L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.

--- (2008) Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval,* 2**,** 1-135.

Purves, R., A. Edwardes & J. Wood (2011) Describing place through user generated content. *First Monday,* 16.

Quercia, D., H. Askham & J. Crowcroft. 2012. TweetLDA: supervised topic classification and link prediction in Twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, 247-250. ACM.

Ramage, D., D. Hall, R. Nallapati & C. D. Manning. 2009. 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*, 248-256. Singapore: Association for Computational Linguistics.

Rinner, C. (2001) Argumentation maps: GIS-based discussion support for on-line planning. *Environment and Planning B: Planning and Design,* 28**,** 847-863.

Rinner, C. & M. Bird (2009) Evaluating community engagement through argumentation maps-a public participation GIS case study. *Environment and Planning B-Planning & Design,* 36**,** 588-601.

Rogerson, R. J. (1995) Environmental and health-related quality of life: conceptual and methodological similarities. *Social science & medicine,* 41**,** 1373-1382.

Sharma, M. (2014) Peoples' perceptions of housing market elements in Knoxville, Tennessee. *southeastern geographer,* 54**,** 137-160.

Shelton, T., A. Poorthuis & M. Zook (2015) Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning,* 142**,** 198-211.

Sieber, R. (2006) Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers,* 96**,** 491-507.

Sirgy, M. J. & T. Cornwell (2002) How neighborhood features affect quality of life. *Social indicators research,* 59**,** 79-114.

Steyvers, M. & T. Griffiths (2007) Probabilistic topic models. *Handbook of latent semantic analysis,* 427**,** 424-440.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink & S. Kelling (2009) eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation,* 142**,** 2282-2292.

Titov, I. & R. McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, 111-120. ACM.

Tsou, M.-H., J.-A. Yang, D. Lusher, S. Han, B. Spitzberg, J. M. Gawron, D. Gupta & L. An (2013) Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science,* 40**,** 337-348.

Wang, H., Y. Lu & C. Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 783-792. ACM.

Wang, W. & K. Stewart (2015) Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Computers, Environment and Urban Systems,* 50**,** 30-40.

Yang, D., D. Zhang, V. W. Zheng & Z. Yu (2015) Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 45**,** 129-142.

Zhang, Y., G. Lai, M. Zhang, Y. Zhang, Y. Liu & S. Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 83-92. ACM.

Appendix A: The discovered topics when K = 3, 7, and 9

In this Appendix, we show the discovered LDA topics when K takes the values of 3, 7, and 9. The results are shown as below:

Appendix Figure 1: The discovered topics when $K = 3$.

Appendix Figure 2: The discovered topics when $K = 7$.

Appendix Figure 3: The discovered topics when $K = 9$.

To facilitate our following discussion, we use (x, y) to refer to a topic in row x and column y. As can be seen, some of the discovered topics in the three examples above are either intermixed, hard to interpret, or duplicated. For example, when $K = 3$, the topics related to restaurant and transportation are intermixed together into one topic $(1, 1)$; when $K = 7$, the topic $(2, 2)$ is hard to interpret; when $K = 9$, we see two topics, $(1, 1)$ and $(3, 1)$, that are both related to community friendliness. Compared with the three results, $K = 8$ produces a set of relatively clear topics.