Occupants' Satisfaction with LEED- and Non-LEED-Certified Apartments Using Social Media Data

Xingtong Guo[a], Kyumin Lee[b], Zhe Wang[c], Shichao Liu[a]

a. Civil and Environmental Engineering, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, 01609, USA

b. Computer Science, Worcester Polytechnic, Institute 100 Institute Road, Worcester, MA, 01609, USA

c. Energy Technology Area, Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA,94720, USA

**Corresponding email: sliu8@wpi.edu*

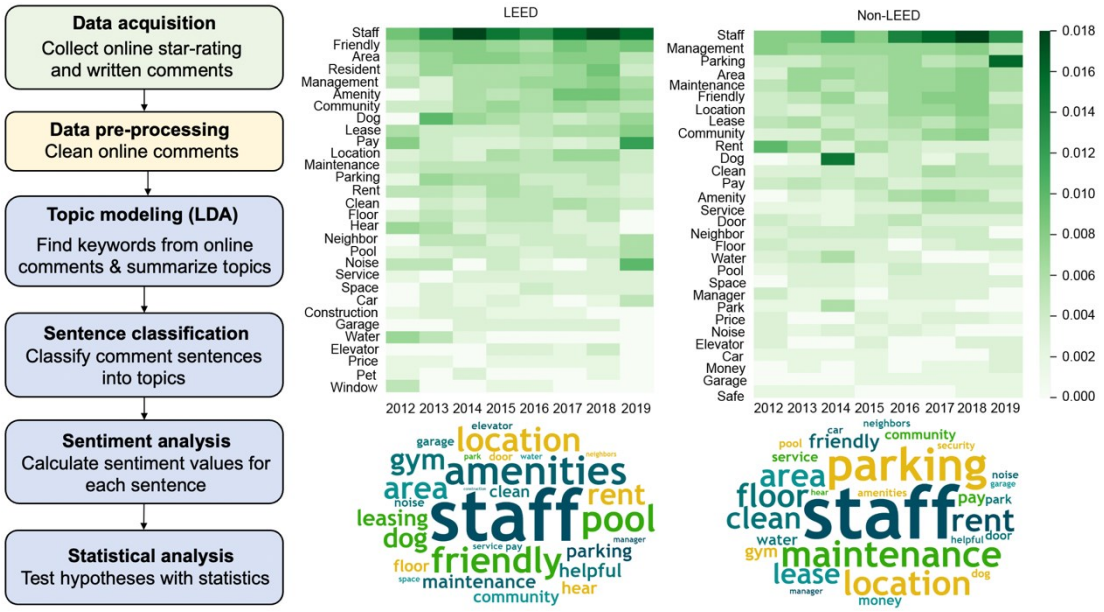
HIGHLIGHTS

1) Proposed a natural language processing method to assess occupants' satisfaction.
2) Identified 3 topics (*transportation & location*, *running cost,* and *health & wellbeing*).
3) The median satisfaction with LEED or non-LEED apartments is positive.
4) A significant but small or negligible uptick in satisfaction with LEED apartments.
5) There is a weak relationship between rent price and star rating.

ABSTRACT

Leadership in Energy and Environmental Design (LEED) certified buildings aim to offer a sustainable and healthy building environment. Previous studies have shown mixed and inconsistent results on whether occupants in LEED-certified buildings are more satisfied than in non-LEED-certified counterparts. Those studies are usually based on surveys or questionnaires for commercial buildings and were limited by sample size and pre-defined question structures. Since most people would spend more time at home after experiencing the COVID-19 pandemic due to the flexibility to work remotely, assessing the satisfaction with residential buildings benefits future environmental design and certification system development. In this work, we propose a natural language processing-based approach for such assessment. The study collected 16,761 online reviews of 232 LEED-certified and 129 non-LEED-certified apartment buildings from social media, then applied topic modeling and sentiment analysis to evaluate occupants' satisfaction. Based on topic modeling, we categorized online comments into three topics, 1) *location and transportation*, 2) *running cost*, and 3) *health and wellbeing.* The subsequent sentiment analysis has shown a statistically significant but small or negligible enhancement in the satisfaction occurring in LEED-certified apartments compared to non-LEED-certified ones concerning all three topics. The "*significant but small or negligible uptick"* has also been found in online star rating and indoor environmental satisfaction. The only exception with a large effect size is lighting which is found to be significantly more satisfying in LEED-certified apartments. Nevertheless, the statistical significance in online star rating disappears when normalized by rent price and property house value.

GRAPHIC ABSTRACT



KEYWORDS

Residential buildings, Post occupancy, Topic modeling, Sentiment analysis, IEQ

1 INTRODUCTION

Leadership in Energy and Environmental Design (LEED), developed by the non-profit U.S Green Building Council (USGBC), is one of the most widely used green building rating systems. As of 2019, there are nearly 100,000 projects registered and certified LEED commercial projects [1]. LEED can support all building types, such as offices, schools, hospitals, and homes. It is comprised of 9 credit categories, from regional priority to indoor environmental quality (IEQ). A project pursuing LEED certification can earn one of four LEED rating levels — Platinum (>80 points), Gold (60-79 points), Silver (50-59 points), and Certified (40-49 points)—based on the total points earned across those categories [2].

Occupants' satisfaction with buildings can be attributed to IEQ (e.g., lighting, temperature, air quality), workplace, and building features such as aesthetic appearance, furniture, cleanliness, level of privacy, and amount of personal control [3][4], in addition to running energy cost [5]. The largest database of occupant indoor environmental quality survey by the Center for the Built Environment (CBE) focuses on seven areas of indoor environmental performance and has been implemented in more than 1000 buildings with over 100,000 individual occupant responses as of March 2017 [6]. Based on the analysis of a subset of the dataset, office buildings with LEED certification outperformed non-LEED-certified buildings in occupants' satisfaction regarding building overall, cleanliness, colors and textures, and air quality, even though the effect sizes of the difference was negligible [4]. Using the same dataset, Lee and Kim [7] concluded that LEED-certified buildings received higher satisfaction with office furnishings, thermal comfort, indoor air quality, cleanliness and maintenance but lower satisfaction with office layout, lighting, and acoustics. The CBE database has revealed marginal advantages of LEED certification in promoting occupants' satisfaction in office buildings.

Besides the results from the CBE database, Table 1 summarizes the previous nine studies on occupants' satisfaction with a variety of factors associated with LEED-certified and non-LEED-certified buildings. There are many studies on LEED buildings in general but not specifically focusing on the comparison of satisfaction between the two building types. Therefore, those studies were excluded in Table 1. Overall, mixed reported findings have been observed on whether LEED-certified buildings are more satisfying or not, which could be attributed to the differences in building location, ages, occupancy period, samples size, or building green features [8]. For instance, teachers in LEED-certified school buildings had a higher satisfaction rate in lighting, thermal comfort, indoor air quality but less satisfaction with acoustics than those in school buildings without LEED certification [9]. However, LEED-certified hospitals have produced elevated satisfaction in terms of all IEQ factors based on quite a small sample size [10].

Four of the nine studies in Table 1 focused on office buildings, but only one study has been reported on residential apartments in terms of occupants' satisfaction with indoor air quality [11]. People spend nearly 90% of their time indoors [12] with half of this time being spent at home [13][14]. The role of residential buildings has become even more crucial especially after the COVID-19 pandemic since many people would have more flexibility to work from home [15]. As a result, occupants' satisfaction with residential buildings has become more important than ever before.

All the reviewed studies relied on questionnaires to determine occupants' level of satisfaction. The design of those questionnaires typically adopts a top-down approach that leaned toward the perspectives of designers, researchers, and policymakers as opposed to occupants. In particular, the structured questions usually have challenges/limitations to reveal occupants' attitudes on aspects not included in the questionnaires, not to mention response rate, timeliness, and longitudinal tracking. Second, it is cost-prohibitive to survey a substantial number of buildings, especially residential ones, using questionnaires. Unlike surveying occupants in large commercial buildings where building managers can easily distribute questionnaires to hundreds of occupants, reaching out to the occupants of residential buildings (e.g., multi-family apartments) is considerably difficult.

One way to avoid those challenges is the bottom-up method. On the Internet, there is an abundance of information from occupants regarding their satisfaction with buildings in the format of star ratings and written comments (Figure 1). Although the comments are unstructured, the information can be processed using text-mining techniques such as Latent Dirichlet Allocation (LDA) [16] and can shed light on occupants' satisfaction with residential buildings.

Text mining and sentiment analysis have been widely applied in many areas to analyze users' satisfaction and attitudes. Berezina et al. [17] used a text-mining method to understand what factors may satisfy or dissatisfy hotels customers; Moreover, Villeneuve et al. [18] employed a text-mining method to study the sentiment feelings of IEQ issues of Airbnb guests; Kar [19] investigated factors affecting user satisfaction in mobile payments based on Twitter tweets through sentiment analysis and topic modeling using LDA. The primary advantages of using social media data are two-folds, 1) it provides a substantial amount of public data with significantly more reviews compared to distributed questionnaires, and 2) occupants' open-ended comments have diminished biases compared with predefined structured questions found

in questionnaires.

The objective of this study is to compare occupants' satisfaction between LEED-certified and non-LEED-certified apartments in the United States through topic modeling and sentiment analysis of publicly posted comments on social media. We selected apartments as the target building type because more online comments are available for apartments than other types of residential buildings. Our goal and contributions are to shed light on the following research questions:

1. Do LEED-certified apartments have a higher star rating than non-LEED-certified apartments?
2. Which latent topics are the most popular and of interest to occupants based on their online comments?
3. How does occupants' satisfaction vary for different factors (e.g., IEQ, running cost) of the apartments both with and without LEED certification?
4. Would rent price and land value affect occupants' satisfaction in LEED-certified and non-LEED-certified apartments?

The subsequent sections include Methodology to retrieve apartment characteristics from USGBC and online review from social media, and to conduct statistical analysis, Results and Discussions to present the major findings to address the above-mentioned research questions, and Conclusions.

Table 1. *Prior studies on occupants' satisfaction with LEED-certified and non-LEED-certified buildings.*

Studies	**Sample size**	**Data collection method**	**Building function and Location**	**Findings**	**Significance level of differences ($p < 0.05$)**
[20]	15 LEED-certified buildings, 6 self-nominated green buildings, and 160 non-LEED-certified buildings	Online questionnaire (Respondents: N = 33285 totally)	Offices, USA	LEED-certified office buildings performed better in most aspects of IEQ, but there was no significant difference in lighting and acoustic quality between LEED-certified and non-LEED-certified buildings.	Statistically significant differences in satisfaction with most aspects of IEQ.
[9]	3 LEED-certified schools ,10 conventional schools, and 20 energy-retrofitted schools	Questionnaire (Respondents: N = 103 totally)	Schools, Canada	LEED-certified school buildings performed better in most aspects of IEQ but worse in acoustics.	Statistically significant differences in teachers' satisfaction with IEQ.
[21]	12 LEED-certified offices and 12 conventional offices	Online questionnaire (Respondents: N = 2545 totally for core the survey module)	Offices, Canada and the north USA	LEED-certified office buildings performed better in most aspects of IEQ and wellbeing, but worse in acoustics and lighting quality.	Statistically significant differences in overall environmental satisfaction, satisfaction with noise from HVAC, thermal preference, and visual and physical symptom frequency.
[7]	15 LEED-certified buildings and 200 conventional buildings	Online questionnaire (Respondents: N= 3769 for LEED, N = 36719 for non-	Offices, USA	LEED-certified buildings performed better in most aspects of IEQ but worse in lighting and acoustics quality.	N/A

		LEED)			
[10]	Two LEED-certified hospitals and one conventional hospital	Questionnaire (Respondents: *N*= 54 for LEED, *N* = 25 for non-LEED)	Health care facilities, USA	LEED-certified hospitals performed better in all aspects of IEQ than the conventional hospital without LEED certification.	Statistically significant differences in the satisfaction with IEQ.
[11]	18 LEED-certified apartments and 13 conventional apartments; 61 home visits	Questionnaire (Respondents: *N* = 37 totally)	Apartments, USA	LEED-certified buildings performed better in most aspects of indoor air quality.	A statistically significant difference in the perception of stuffy air, observation of pests and inadequate ventilation.
[4]	65 LEED-certified office buildings and 79 non-LEED-certified office buildings	Online questionnaire (Respondents: *N*= 10129 for LEED, *N* = 11348 for non-LEED)	Mainly for offices, USA	LEED-certified buildings performed better in air quality, building maintenance, colours and textures, and cleanliness but worse in amount of light, ease of interaction, visual privacy, visual comfort, amount of space, noise, and sound privacy. However, the effect sizes are negligible.	Statistically significant differences have been reported in all investigated factors except for temperature, furniture adjustability, and comfort of furnishing.
[22]	One LEED-certified mix-used building, one conventional mix-used building	Questionnaire (Respondents: *N*= 53 for LEED, *N* = 72 for non-LEED)	Mix-used building, China	LEED-certified buildings performed better in summer temperatures and overall IEQ satisfaction but worse in lighting, noise, and temperatures in winter.	Statistically significant differences in the satisfaction with temperature and lighting. No significant difference in the noise satisfaction.
[23]	One LEED-certified factory and one non-	Questionnaire (Respondents:	Factories, Sri Lanka	LEED-certified factory performed better in views to outside, cleanliness,	Statistically significant difference in thermal comfort,

LEED-certified factory	*N*= 35 for LEED, *N* = 35 for non-LEED)	furniture, privacy, and lighting, while worse in thermal comfort, provision of ventilation for work, and having control over indoor environment.	provision of ventilation of work, having control over indoor environment, views to outside, cleanliness, furniture, privacy, and lighting.

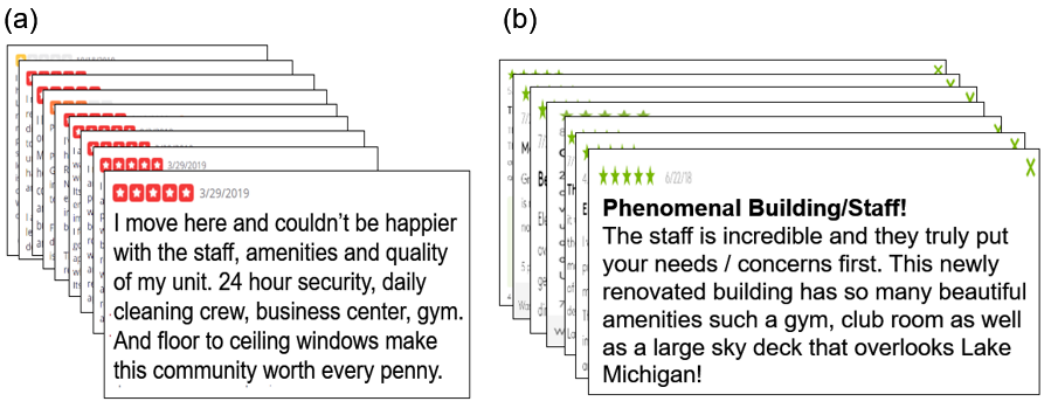


Figure 1. *Online reviews (star rating and written comments) on apartments from social media; (a) Yelp.com; (b) Apartmentratings.com.*

2 METHODOLOGY

Figure 2 depicts the workflow of the methodologies in this work including 1) acquisition of online ratings and comments from social media, 2) data cleaning with regular expression (Regex) matching [24] and stop-words removal [25], 3) topic modeling to extract keywords and latent topics to classify sentences, 4) sentiment analysis at the sentence level, and 5) hypothesis testing on the difference in sentiment values (indicators of satisfaction) between LEED-certified and non-LEED-certified apartments.

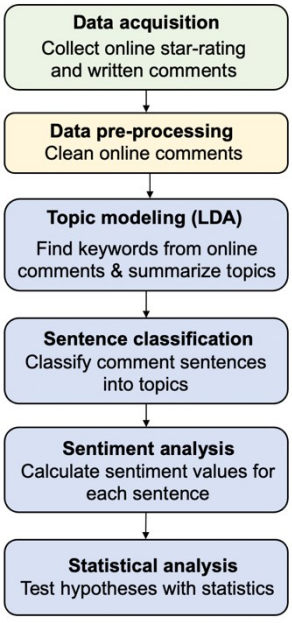


Figure 2. *Methodology workflow*

2.1 Data acquisition

We identified commercial apartments in the United States from the USGBC database by applying a filter of project type "Multi-Family Residential: Apartment" [26]. Under these specifications, there were a total of 490 LEED-certified and 794 non-LEED-certified residential

apartments in the United States until October 2018. The non-LEED-certified apartments had applied for certification but failed. For each building, the database provides the address, project name, LEED system version, rating level (if certified), certification date (if certified), and other information. The database can be publicly accessed on an online repository (http://dx.doi.org/10.17632/hw59ryytdf.1). Next, we searched each building using its address and project name, and from these gathered occupants' posted comments and rating scores (from 1 = worst to 5 = best) using a developed web crawling tool or manually (if necessary). It turned out that most online comments are aggregated on three apartment review websites (*Yelp.com, Apartmentraitngs.com, and Apartment.com*). The search resulted in 8,230 online reviews (1,182,531 words) for 232 LEED apartments and 8,531 online reviews (1,284,763 words) for 129 non-LEED apartments matching the building projects in the LEED database (Table 2). Each review data point included star rating, descriptive written comments, and date of the review.

Table 2. *Statistics of the comments*

	# of apartment	**# of comments**	**Average # of sentences per comment**	**Average # of words per comment**	**Average # of words per sentence**
LEED	232	8230	10	144	15
Non-LEED	129	8531	11	151	14

2.2 Data Pre-processing

Online comments sometimes contain less important or noisy information, so a data pre-processing step is required before conducting an in-depth analysis. Our data pre-processing step removes the following information:

- Information such as symbols and URL links that were removed by Regular Expression (Regex) matching.
- Stop words, like "a", "is", "the", "do" etc. that were taken out by employing the Natural Language Toolkit (NLTK) stop words[27] list. Stop words do not express any emotion [28] or satisfaction but can affect topic modeling results and an optimal topic number [29].

2.3 Topic modeling

Topic modeling is a widely used tool to mine data, analyze latent semantic structures, and find topics in unclassified texts [16]. In this study, we employed an LDA method, a generative probabilistic model of corpus [16]. LDA is the most widely used topic modeling approach [30], which adopts an unsupervised learning technique to uncover hidden topics information from a collection of messages. LDA method applies a hierarchical structure of "document-topic-word." Each document (w) is considered as a probability distribution over topics, and each topic is a probability distribution over words.

Figure 3 shows the graphical model representation of the LDA model. When performing topic modeling, the topic number (k) is an unknown value to be defined by the user. The

selection of a proper *k* is a steppingstone for topic modeling since different *k* values may lead to different topic results.

One metric to optimize *k* is to minimize perplexity that measures the quality of latent topics. A lower perplexity value implies a higher model conditional likelihood, in other words, a better model [31]. For instance, Chen et al. [32] and Ghosh [33] found an ideal *k* in the range of 30 to 50 with the lowest perplexity. However, the selection of *k* cannot solely rely on perplexity since it only crudely indicates an acceptable amount of topic loss. In addition, topic interpretation from a specific discipline (e.g., social science) should be incorporated in the selection process[34]. In this study, we chose *k* according to both perplexity and qualitative exploration on comprehensive information coverage based on our best understanding of LEED and non-LEED buildings.

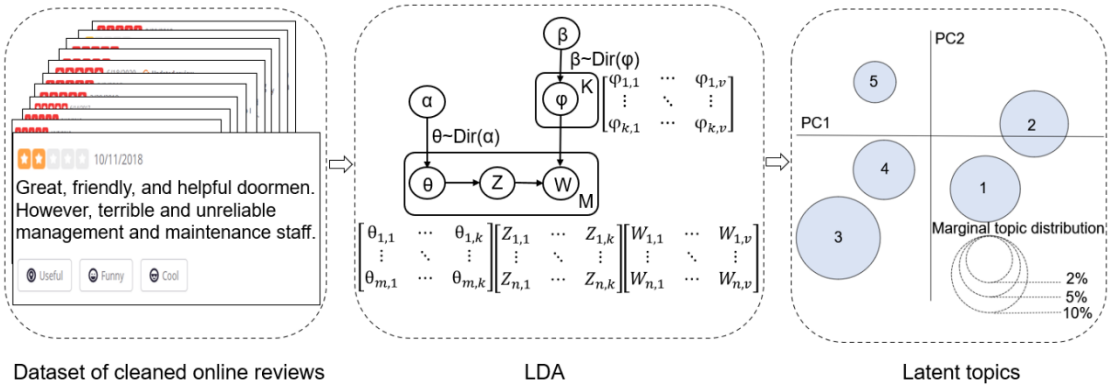


Figure 3. *Graphical model representation of topic modeling of online reviews using LDA. M denotes the number of documents; α is the parameter of the Dirichlet prior on the per-document topic distribution; β is the parameter of the Dirichlet prior on the per-topic word distribution; θ is the topic distribution for a document; φ is the word distribution for the topic; Z is the topic for a word in the document; $W_{n,v}$ is the specific word. The circles in blue shades represent the distribution of each latent topic.*

Preliminary modeling results contained words like "favourite", "name", "anyone" which may be irrelevant to occupants' satisfaction. These suspiciously irrelevant words were removed when the coherence score, which can identify if a topic is semantically interpretable [35], resulted from the topic modeling increased without those words.

2.4 Online Comments Classification Using Supervised Learning

Since a sentence of online comments may contain mixed topics and sentiments, it needs to be classified into each founded latent topic before sentiment analysis. As for sentence classification, we used a seed word dictionary as described below, and matched the dictionary with online comments. If a sentence referred several topics, it will be classified into all relevant topics rather than be counted only once. There are several ways to build a dictionary. The popular method is a heuristic approach of combining human annotation and Wordnet's Synset, which are sets of cognitive synonyms expressing a particular concept [36]. But it doesn't work in our case due to professional specificity. For example, when we try to find Synset of "air",

the Wordnet will find words like “bare”, “beam”, “line” and many other words irrelevant to an apartment. Therefore, we build a relevant expanded word dictionary by selecting topic relative words from online comments and LEED reference guides [37][38][39][40]. The detailed dictionary is described in Table A1 of the appendix.

2.5 Sentiment analysis

Sentiment analysis, also known as opinion mining, is a process of understanding written contexts and is generally used to determine whether the context contains positive, neutral, or negative opinions [41][42][43]. Sentiment analysis in this work is a probabilistic supervised machine learning approach based on the Naïve Bayes algorithm as follows.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d|c)}^{likelihood} \overbrace{P(c)}^{prior} \tag{1}$$

$$C_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{f \in F} P(f|c) \tag{2}$$

Here C denotes a set of all possible classes (negative, neutral, positive) and c is one of the classes; d denotes a document (each sentence of a comment from social media). F means all features value pairs (e.g., location-near, cost-high), and f is one of these feature pairs.

The Naïve Bayes classifier performs well when the output is categorical, while it may do poorly for regression problems by discretizing the target value [44]. To overcome the limitation and guarantee the accuracy of the Naïve Bayes model, we used the following Bayes equation for regression:

$$p(Y|E) = \frac{p(E,Y)}{\int p(E,Y)dY} = \frac{p(E|Y)p(Y)}{\int p(E|Y)p(Y)dY} \tag{3}$$

where the likelihood $p(E|Y)$is the probability density function (pdf) of the evidence E for a given target value Y, and the prior $p(Y)$ is the pdf of the target value before any evidence has been seen.

In this study, the sentiment analysis was conducted with NLTK and the sentiment analysis application programming interface (API) provides polarity values. The final output was a floating-point value from -1 (negative sentiment) to +1 (positive sentiment). In this study, we considered binary classifications, namely, a negative sentiment as dissatisfaction and a positive value to be satisfaction. A higher absolute sentiment value implies that the occupants are more satisfied or dissatisfied with their apartment.

2.6 Statistical analysis

The initial data analysis with the Shapiro-Wilk normality test showed that all datasets were non-normally distributed. Therefore, we assessed the difference of the median with the Wilcoxon signed-rank test. The statistical significance was based on $p < 0.05$(*), $p < 0.01$(**), and $p < 0.001$(***). Cohen’s d was used to calculate the effect size of the difference that $|d| < 0.147$ “negligible”, $|d| < 0.33$ “small”, $|d| < 0.474$ “medium”, otherwise “large”, following our previous method [45].

3 RESULTS and DISCUSSIONS

In this section, we display a geographical map on the distribution of LEED and non-LEED multi-family residential apartments reviewed in social media, in addition to occupants' satisfaction based on the online star rating and sentiment analysis. This section also reports occupants' topics of interest. Besides, the satisfaction with indoor air quality, thermal comfort, acoustics, lighting, and layout are discussed by comparing with prior studies.

3.1 Geographical distribution of LEED and non-LEED multi-family residential apartments

Figure 4 shows the distribution of 232 LEED-certified and 129 non-LEED-certified apartment buildings assessed in this study. The background colors show the median value of housing prices in 2018 [46]. Both LEED and non-LEED certified apartments that applied for but did not obtain certification are more concentrated in coastal states such as California and New York. Those states are generally populous and ambitious to achieve sustainability and environmental goals by motivating developers to voluntarily pursue third-party certifications for their real estate projects [47].

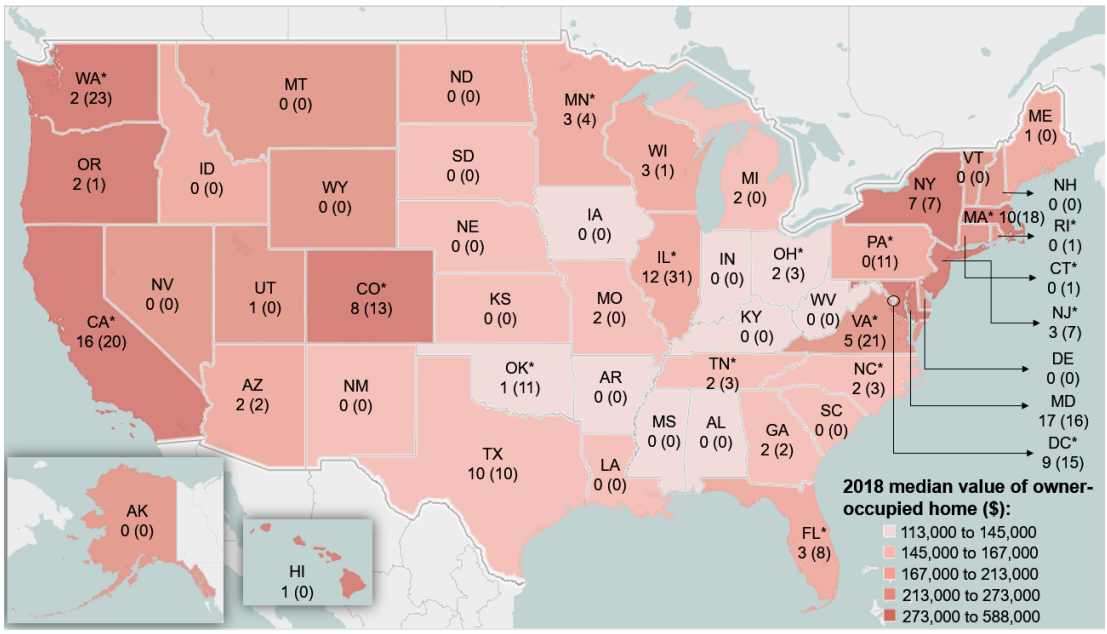


Figure 4. *Distribution of LEED-certified and non-LEED-certified multi-family residential apartments reviewed on social media. The numbers in each state represent # of non-LEED apartments (# of LEED apartments). The asterisk after the name of a certain state means the state has more LEED apartment buildings than non-LEED ones.*

3.2 Occupants' satisfaction with LEED-certified and non-LEED-certified residential apartments

The rating of online reviews in most social media ranged from one (worst) to five (best) stars. In this work, we coded the star rating using a 5-scale Likert scale, very dissatisfied (1 star), dissatisfied (2 stars), neutral (3 stars), satisfied (4 stars), and very satisfied (5 stars) to facilitate the analysis of occupants' overall satisfaction with an apartment building. Figure 5 shows the percentage of the different satisfaction levels for the LEED-certified and non-LEED-

certified apartments.

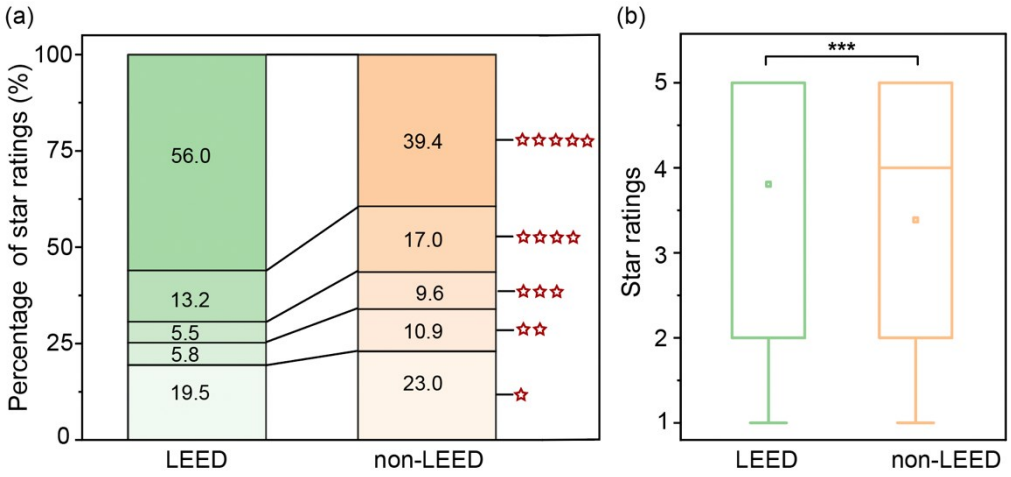


Figure 5. *Star ratings of LEED-certified and non-LEED-certified apartments*

The total percentage of reviews with satisfied (4 stars) and very satisfied (5 stars) is 69.3% for LEED-certified apartments and 56.4% for non-LEED-certified apartments. There are more star ratings in the bins of *Very satisfied (5-star)* and *Very dissatisfied (1-star)* than others, suggesting an under-reporting bias when online reviewers are more motivated to post extreme and negative ratings [48]. It should be noted that such bias could also occur in studies using questionnaires [49]. The median ratings are very satisfied (5 stars) for LEED-certified and satisfied (4 stars) for non-LEED-certified apartments. The Wilcoxon signed-rank test reports a statistically significant difference ($p < 0.001$) between the median rating of the two apartment types, though the effect size of the difference is negligible according to Cohen's d ($|d| < 0.33$). In other words, LEED-certified apartments are perceived slightly more satisfying than non-LEED-certified apartments according to online star ratings. However, the crude rating results could not reveal how occupants feel about specific aspects of the apartments, which necessities a detailed analysis of the contexts and topics of posted comments.

3.3 Topic modeling

The latent semantic structures of occupants' online comments can be revealed through topic modeling. A "topic" in topic modeling is defined as a cluster of keywords that co-occur in the same documents according to certain patterns through unsupervised learning. Perplexity measure is a commonly used computational metric to determine the number of topics summarizing an online review [16]. We identified the number of topics discussed in online comments based on the perplexity calculated with Gensim [50] as well as a manual inspection of keyword semantics since computational algorithms based on perplexity solely can identify nuances that are not semantically meaningful [51].

Generally, the optimal number of topics resulted from an LDA model exists when the perplexity value is the minimum [16], but the method should only serve as the initial selection of models with an acceptable amount of information loss [34]. Figure 6 depicts the lowest perplexity when the topic number is three. However, the manual analysis after reading online

comments found that important information related to “pet-policy” and “amenities” is not included in any identified topic. A similar issue occurs as the topic number is two or four. Therefore, considering both perplexity and interpretability, we found that five topics can generate both comprehensive and semantically meaningful information.

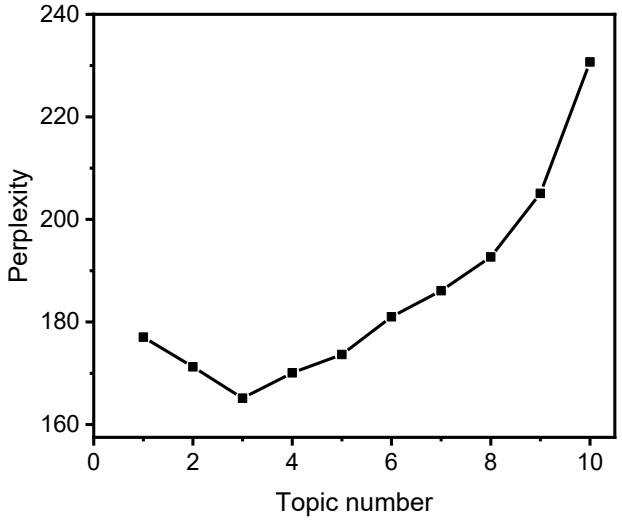


Figure 6. *The perplexity of topic modeling for various topic numbers*

For each of the five topics, Table 3 lists ten keywords and their calculated weight calculated based on frequency and topic relevancy extracted from LDA. We then summarized what themes were covered under each topic by manually examining the semantics of the included keywords. For instance, Topic 1 is comprised of three themes, *Location and transportation* (“car”, “neighbour”), *Pet-policy* (“dog”), and *IEQ* (“floor”, “hear”, “door”, “wall”). Overall, the five topics consist of six themes in total, 1) *Location and transportation*, 2) *IEQ*, 3) *Pet-policy*, 4) *Management service*, 5) *Running cost*, and 6) *Amenities*. A theme can appear in multiple topics (Table 3), such as *IEQ* included in both Topic 1 and Topic 5. Therefore, to facilitate the comparison and analysis of occupants’ satisfaction, we reorganized the five topics based on shared themes and distilled them into three independent topics (*Location and transportation, Running cost,* and *Health and wellbeing)* by consulting with the categories of popular rating systems like LEED, Building Research Establishment Environmental Assessment Method (BREEAM), and Green Building Tool [52] (see Table 4) . The process can be illustrated in Figure A1 of the appendix.

Table 3. *Keywords of initial five topics extracted from comments*

Topic 1 ***(Themes: Location and transportation, Pet-policy, IEQ)***		**Topic 2** ***(Themes: Management service, Running cost,)***		**Topic 3** ***(Themes: Running cost, Management service,)***		**Topic 4** ***(Themes: Location and transportation, Running cost)***		**Topic 5** ***(Themes: Management service, Location and transportation, Amenity, IEQ)***	
Keywords	Weight	Keywords	Weight	Keywords	Weight	Keywords	Weight	Keywords	Weight
people	0.019	apartment	0.039	rent	0.033	parking	0.094	staff	0.037
dog	0.018	maintenance	0.026	pay	0.027	free	0.028	friendly	0.030
apartment	0.018	staff	0.024	management	0.023	park	0.024	location	0.030
floor	0.013	move	0.023	tenant	0.017	store	0.021	apartment	0.030
hear	0.012	resident	0.023	water	0.014	spot	0.023	area	0.030
door	0.012	live	0.023	lease	0.013	food	0.015	building	0.026
building	0.012	property	0.018	money	0.011	shop	0.015	live	0.025
car	0.009	management	0.016	trash	0.009	shopping	0.012	amenity	0.019
neighbour	0.009	service	0.015	unit	0.009	restaurant	0.012	clean	0.018
wall	0.009	office	0.014	office	0.009	close	0.011	build	0.017

Table 4. Comparison of the extracted themes and topics with LEED credit categories

Themes (identified by initial topic modelling)	**Summarized three new topics**	**Coincident LEED credit categories***
Location and transportation	Location and transportation	Location and transportation, Sustainable sites, Regional priority
Running cost	Running cost	Water efficiency, Energy and atmosphere, Material and resources
Management services Pet-policy Amenities IEQ	Health and wellbeing	IEQ

**The LEED credit categories are not completely covered by the identified themes from topic modelling or vice versa, suggesting online comments can shed extra light on occupants' satisfaction that cannot be revealed by a predesigned questionnaire following the LEED categories.*

Figure 7 summarizes the distribution of occupants' topics of interest characterized by the percentage of sentences discussing a topic for both LEED-certified and non-LEED-certified apartments. Occupants in LEED-certified apartments appear to be more (61.0% vs 56.3%) attentive to factors pertaining to *"Health and wellbeing"* than those in non-LEED-certified counterparts. A possible reason is that the topic covers more themes such as *"amenities," "management services," "pet-policy," "appliance,"* and *"indoor environment"* and is related to a larger sample of words like *"noise", "air", "view"* and *"clean"* than the other two topics. Nevertheless, it is surprising that "*running cost"* has been discussed the least of time, only 11.7% for LEED-certified and 12.5% for non-LEED-certified apartments.

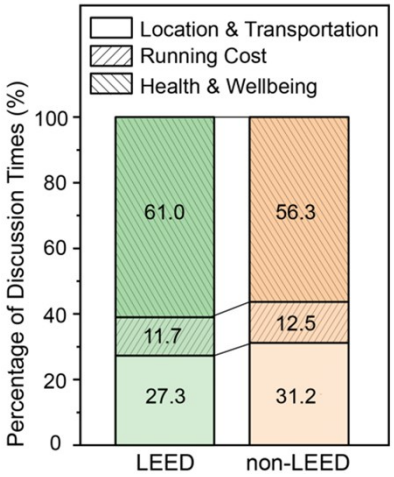


Figure 7. *Occupants' topics of interest based on the percentage of sentences discussed in social media (totally 82,890 sentences for LEED apartments and 92,854 sentences for non-LEED-certified apartments)*

The top three weighted keywords in the online review for LEED-certified apartments are *"staff", "friendly", "area"* as opposed to *"staff", "management",* and *"parking"* for non-LEED-certified apartments (Figure 8). The heat maps indicate that online reviews primarily focus on apartment management service, which is supported by the commonly appeared words (e.g., "*staff*") based on frequency for both apartment types. Nevertheless, leisure facilities are discussed more often for LEED-certified apartments. For example, *"amenities", "pool",* and *"gym"* appear more frequently in those apartments. However, *"maintenance"* is a popular topic in online reviews for only non-LEED-certified apartments. Furthermore, the heat maps show no clear trend of weight change for most keywords.

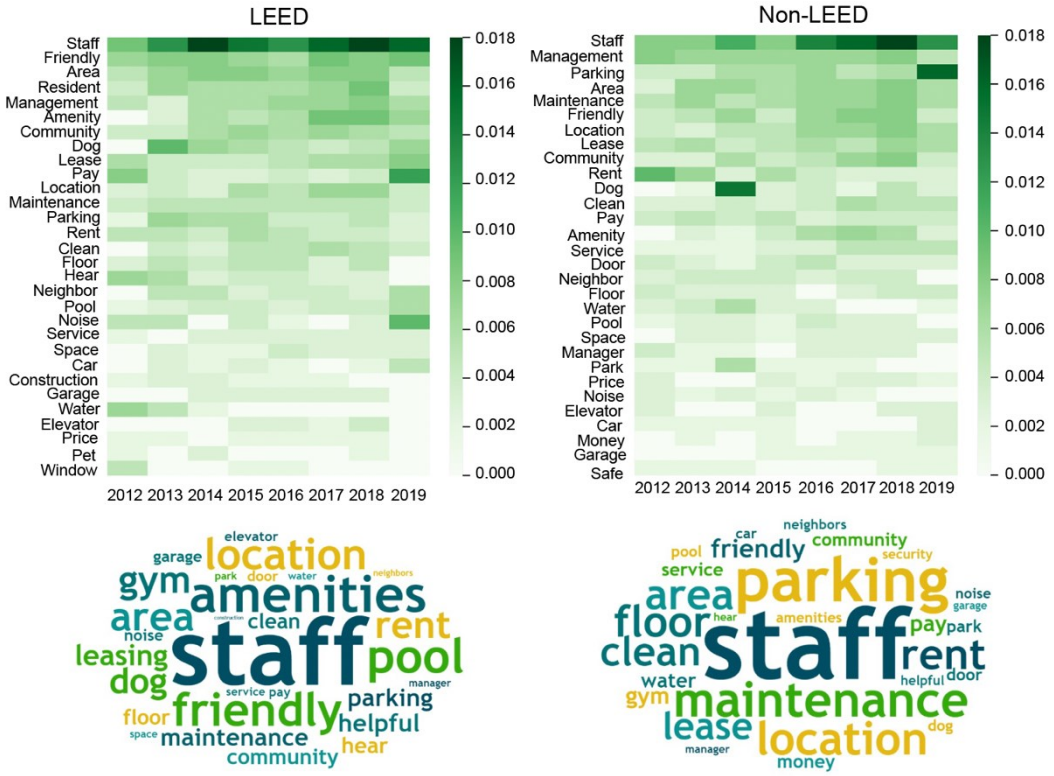


Figure 8. *The evolvement of the 30 most weighted keywords from 2012 to 2019 (heat maps) and the 30 most frequent keywords during the entire duration (word clouds) for LEED-certified and non-LEED-certified apartments. The keywords in the heat maps are ranked from the highest total weight to the lowest.*

3.4 Sentiment analysis of occupants' satisfaction

Occupants' satisfaction with the three topics regarding LEED-certified and non-LEED-certified apartments are evident through sentiment analysis of online review comments. Figure 9 shows the distribution of the sentiment values of all sentences for each topic between LEED-certified and non-LEED-certified apartments. Generally, occupants have been satisfied with the three topics for both apartment types, as indicated by the slightly positive median sentiment values ranging from 0 to 0.25. Occupants are more satisfied with "*location and transportation*," *"running cost"* and *"health and wellbeing"* for LEED-certified apartments than those without

LEED certification. While the differences are statistically significant ($p < 0.001$) in terms of all the three topics, the effect sizes of those differences are negligible according to Cohen's d ($|d|<$ 0.147). It is therefore concluded that LEED-certified apartments generate negligibly higher satisfaction with *"location and transportation," "running cost," and "health and wellbeing."*

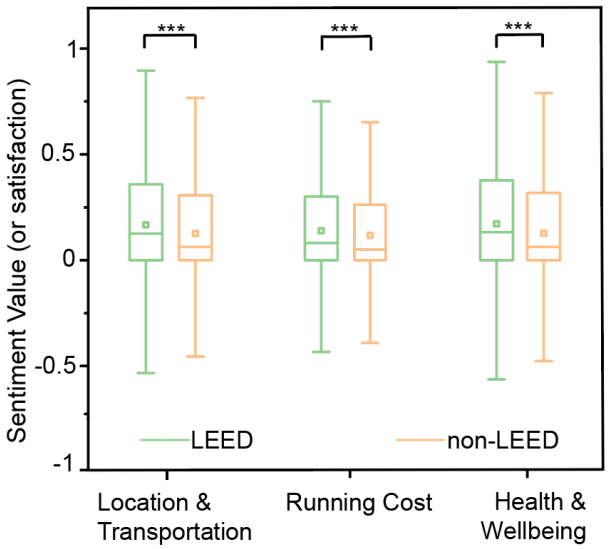


Figure 9. *Sentiment analysis of online comments (statistical significance: *** p<0.001)*

3.5 Satisfaction with various IEQ factors

IEQ is an essential element contributing to people's health and wellbeing. Occupants' sentiment with IEQ factors, such as thermal comfort, acoustics, lighting, indoor air quality, and layout inevitably influences their general satisfaction with the topic *"health and wellbeing."* Figure 10 shows that LEED-certified buildings outperform non-LEED-certified counterparts for all the five IEQ factors generated by topic modeling, with different significant levels and effect sizes. The large distinction in satisfaction occurs for lighting only. It is observed that acoustics is the least satisfying while thermal comfort receives the highest satisfaction rate in apartments both with and without the LEED certification. Similar findings were also reported previously [13].

Furthermore, we compared the results from sentiment analysis with four previous studies conducted in buildings with different functions such as houses and offices (Figure 10). The original scale (-1 to 1) of the sentiment values was adjusted linearly to -3 (very dissatisfied) to 3 (very satisfied) to facilitate the comparison. Albeit the differences in sample size and methodology.

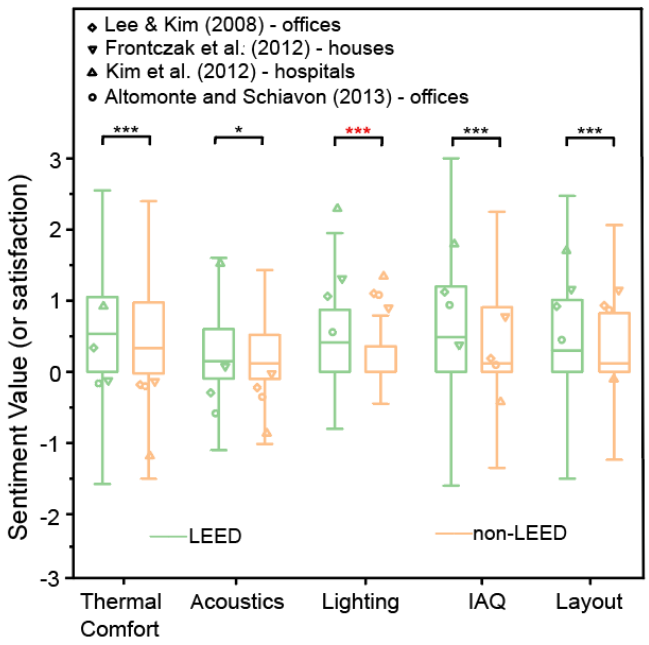


Figure 10. *The comparison of occupants' satisfaction with IEQ in various building types using different methods. The statistical tests were conducted based on the data of this study only. Statistical significance: *** p<0.001, ** p<0.01, * p<0.05, the * in red refers to a large effect size based on Cohen's d (|d|< 0.147 "negligible", |d|< 0.33 "small", |d|< 0.474 "medium", otherwise "large")*

3.6 Rent price, land value, and star rating

Generally, people would expect to get services or products of equal or superior value to what they have paid. Thus, occupants' satisfaction might be related to not only building characteristics (e.g., IEQ) but also rent price, as suggested by studies [53][54] that competitive price can increase customer satisfaction. LEED certification awards buildings in locations that promote less vehicle travel distance and better liveability [55], associated with high property value. This can be observed by higher median house property value where LEED buildings are located in Figure 11a. The difference between the median house property value of the two apartment types is statistically significant with a medium effect size (Cohen's *d*).

Surprisingly, the increment in the property value does not necessarily result in an elevation in the rent price per bedroom number for LEED-certified apartments. Figure 11b depicts that the median rent prices per room are not statistically different between apartments with or without LEED certification.

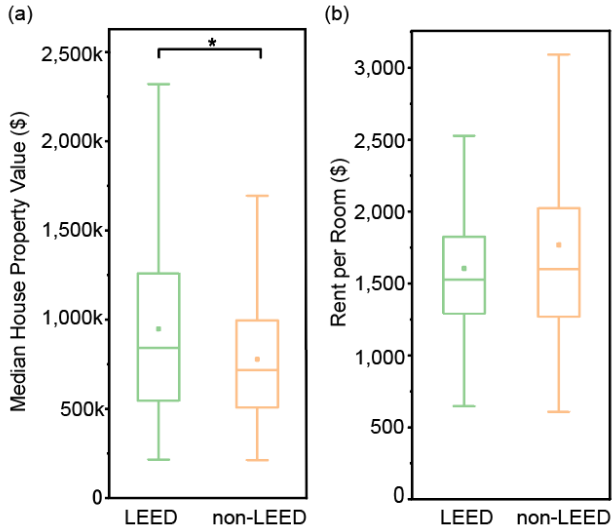


Figure 11. (a) Median house property values and (b) Rent price per room for LEED-certified and non-LEED-certified apartments. The median house property value (in 2018) was retrieved from Federal Housing Finance Agency (FHFA) [56] *based on zip code.*

Impact of rent price and land property value on star ratings

To study how much satisfaction people can get for every dollar they spend, we normalized the star rating (from 1 to 5 stars) of each apartment by the rent price per room to monetize occupants' overall satisfaction. In Figure 12, the normalized star rating of LEED-certified apartments is not statistically (p = 0.073) higher than that of non-LEED-certified apartments, indicating that occupants in the LEED-certified apartment are not statistically more satisfied as for the price they pay for the rent. Besides, the correlations between star rating and normalized price of LEED-certified and non-LEED-certified apartments are -0.04 and 0.16 separately, suggesting a weak relationship between rent price and star rating.

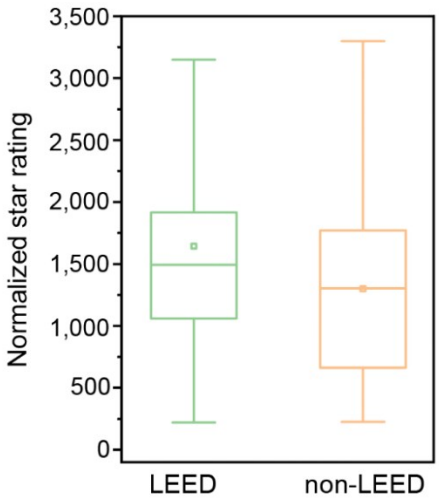


Figure 12. Normalized star ratings by rent price per room for LEED-certified and non-LEED-certified apartments

3.7 Sampling Biases and Limitations

Sampling biases occur when the samples of a stochastic variable cannot represent the true distribution in the population due to non-random reasons. Sampling biases pose a challenge to evaluate occupants' satisfaction with buildings for many research methods, including pre-designed questionnaires and online comments. We discuss common biases during sampling in this section followed by the limitations of this study.

Under-reporting bias is resulted from self-selection or voluntary response. A voluntary questionnaire on occupants' satisfaction might attract more respondents who are more sensitive to incentives or inclined to express negative opinions [49]. In this work with social media data, we found that 1-star ratings and 5-star ratings have a higher weight on the data distribution, partially congruent with a prior finding on under-reporting bias that online reviewers are more motivated to post extreme and negative ratings [48].

Non-response bias can occur when there is a systematic difference between responders and non-responders. As a key indicator of data quality [57], response rate affects survey bias and statistical precision. A questionnaire with a low response rate is more likely to suffer from non-response bias. Some journals [58] request a minimum response rate of up to 60% for a manuscript to be considered for peer review. Despite that the criterion is not necessarily applicable for the research on occupants' satisfaction, obtaining a high response rate is crucial for data validity. The dataset [3] on occupants' satisfaction with IEQ using questionnaires only has a response rate higher than 5%, suggesting a possible skew in the sampled population. Even though a survey using social media data has little relevance to this bias due to the nature of the method, it can suffer from another one: under-coverage bias.

Under-coverage bias in the sampling means that participants cannot adequately represent the population. People who have limited access to the Internet could have difficulties posting their comments online. Online social media data tend to skew towards those created by young, urban, minority individuals [59]. Also, apartment owners while living in the property may not review their building on social media. Similarly, under-coverage bias also exists when questionnaires are administered by email [7][20], since occupants who lack access to email servers are less likely to participate. The under-coverage bias is often coincident with convenience sampling.

The bias resulted from convenience sampling involves samples drawn based on their ready availability in terms of geographical proximity or known contacts. Since distributing questionnaires to occupants could be cost-prohibitive, reaching out to respondents within the same community is a common strategy [11][23]. Compared to questionnaires, online comments may suffer less from this bias but still cannot abstain from it easily. For example, this study does not consider online comments written in languages other than English.

The abovementioned sampling biases are common but may not be exhaustive in the research on occupants' satisfaction. No matter what approach is employed, one should try to take strategies to mitigate potential biases including oversampling [60], post-survey adjustment (imputation and reweighting [61]), encouraging non-responsive participants, making efforts to gain the participation of all intended participants, and so forth.

Besides sampling biases, this study is limited in the categorization of polysemy keywords and instinct drawbacks of the data-driven approach. The applied LDA method may not be robust sufficiently to exploit the contextual semantics of a word. For example, "*area*" could be related

to either apartment location or square footage. In this paper, we had to manually read the comments containing those words to address this limitation. The weaknesses of the data-driven approach are embedded in this work. In particular, the LEED rating system consists of numerous credit categories but many of them are not reflected in online comments.

Despite the methodological shortcomings of this work, using online comments to assess occupants' satisfaction shows multiple advantages in addition to a large sample size. Social media data allow longitudinal analysis over a course of years that is extremely difficult to conduct using questionnaires. When online data carry more refined information on building characteristics, the approach can also be applied to evaluate why occupants' satisfaction is influenced by these factors. This hidden information could be difficult to reveal with a predesigned questionnaire. Nevertheless, online data and questionnaires should supplement each other rather than replacing one with the other because both offer unique advantages from different perspectives. We suggest multimodal approaches, if possible, to better understand occupants' satisfaction in future studies.

4. CONCLUSIONS

The goal of this study is to compare occupants' satisfaction level of LEED-certified and non-LEED-certified apartments by analyzing online reviews posted on social media with natural language processing (NLP). The approach can supplement questionnaires distributed to a selected population for this purpose that are generally limited by sample size and pre-defined question structure.

The online review data regarding apartments can be categorized into three topics, 1) *location and transportation*; 2) *running cost* and 3) *health and wellbeing.* Occupants have discussed *health and wellbeing* more frequently (accounting for 56-61%) on social media. Facilities for leisure (e.g., pool, gym) are discussed more often for LEED-certified apartments compared to non-LEED-certified ones.

The sentiment analysis shows that both apartment types have positive median sentiment values for all three topics, indicating that occupants are satisfied with their apartments in general. Overall, LEED-certified apartments have slightly higher satisfaction than non-LEED-certified counterparts for most investigated perspectives. Additionally, the enhancement is mostly negligible or small according to the calculated effect sizes. In particular, the significant but negligible or small uptick has been found in 1) online holistic star rating, 2) sentiment values of all the three topics, and 3) satisfaction with IEQ factors except for lighting. When the star rating is normalized by the rent price and house property value, no statistical difference (p = 0.073) can be found between the two apartment types.

Both pre-designed questionnaires and online comments in social media could suffer from sampling biases to different extent. Since each method has instinct advantages and shortcomings, if possible, a multimodal approach is suggested to applied to increase research validity.

Acknowledgment

This work was supported by U.S. National Science Foundation (NSF) under Grant No. 2028224.

Declaration of Competing Interest

The authors declare no competing interests.

REFERENCES

[1] No title, (n.d.). https://www.usgbc.org/articles/leed-reaches-new-milestone-surpasses-100000-commercial-green-building-projects.

[2] No title, (n.d.). https://www.usgbc.org/leed.

[3] M. Frontczak, S. Schiavon, J. Goins, E. Arens, H. Zhang, P. Wargocki, Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design, Indoor Air. 22 (2012) 119–131. https://doi.org/10.1111/j.1600-0668.2011.00745.x.

[4] S. Altomonte, S. Schiavon, Occupant satisfaction in LEED and non-LEED certified buildings, Build. Environ. 68 (2013) 66–76. https://doi.org/10.1016/j.buildenv.2013.06.008.

[5] K. Amasyali, N.M. El-Gohary, Energy-related values and satisfaction levels of residential and office building occupants, Build. Environ. 95 (2016) 251–263. https://doi.org/10.1016/j.buildenv.2015.08.005.

[6] No title, (n.d.). https://cbe.berkeley.edu/research/occupant-survey-and-building-benchmarking/.

[7] Y.S. Lee, S. Kim, Indoor environmental quality in LEED-certified buildings in the U.S., J. Asian Archit. Build. Eng. 7 (2008) 293–300. https://doi.org/10.3130/jaabe.7.293.

[8] M. Khoshbakht, Z. Gou, Y. Lu, X. Xie, J. Zhang, Are green buildings more satisfactory? A review of global evidence, Habitat Int. 74 (2018) 57–65. https://doi.org/10.1016/j.habitatint.2018.02.005.

[9] M.H. Issa, J.H. Rankin, M. Attalla, A.J. Christian, Absenteeism, performance and occupant satisfaction with the indoor environment of green Toronto schools, Indoor Built Environ. 20 (2011) 511–523. https://doi.org/10.1177/1420326X11409114.

[10] S.K. Kim, Y. Hwang, Y.S. Lee, W. Corser, Occupant comfort and satisfaction in green healthcare environments: A survey study focusing on healthcare staff, J. Sustain. Dev. 8 (2015) 156–173. https://doi.org/10.5539/jsd.v8n1p156.

[11] M.D. Colton, P. Macnaughton, J. Vallarino, J. Kane, M. Bennett-Fripp, J.D. Spengler, G. Adamkiewicz, Indoor air quality in green vs conventional multifamily low-income housing, Environ. Sci. Technol. 48 (2014) 7833–7841. https://doi.org/10.1021/es501489u.

[12] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, J. V Behar, S.C. Hern, W.H. Engelmann, The national human activity pattern survey, Lawrence Berkeley Natl. Lab. 11 (2011) 231–252. http://exposurescience.org/the-national-human-activity-pattern-survey-nhaps-a-resource-for-assessing-exposure-to-environmental-pollutants.

[13] M. Frontczak, R.V. Andersen, P. Wargocki, Questionnaire survey on factors influencing comfort with indoor environmental quality in Danish housing, Build. Environ. 50 (2012) 56–64. https://doi.org/10.1016/j.buildenv.2011.10.012.

[14] E. Lee, Indoor environmental quality (IEQ) of LEED-certified home: Importance-performance analysis (IPA), Build. Environ. 149 (2019) 571–581. https://doi.org/10.1016/j.buildenv.2018.12.038.

[15] O. Rubin, A. Nikolaeva, S. Nello-Deakin, M. te Brommelstroet, What can we learn from the COVID-19 pandemic about how people experience working from home and commuting?, Certre Urban Stud. Univ. Amsterdam. (2020).
[16] D.M. Blei, A.Y. Ng, M.T. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[17] K. Berezina, A. Bilgihan, C. Cobanoglu, F. Okumus, Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews, J. Hosp. Mark. Manag. 25 (2016) 1–24. https://doi.org/10.1080/19368623.2015.983631.
[18] H. Villeneuve, W. O'Brien, Listen to the guests: Text-mining Airbnb reviews to explore indoor environmental quality, Build. Environ. 169 (2020) 106555. https://doi.org/10.1016/j.buildenv.2019.106555.
[19] A.K. Kar, What Affects Usage Satisfaction in Mobile Payments? Modelling User Generated Content to Develop the "Digital Service Usage Satisfaction Model," Inf. Syst. Front. (2020). https://doi.org/10.1007/s10796-020-10045-0.
[20] S. Abbaszadeh, L. Zagreus, D. Lehrer, C. Huizenga, Occupant satisfaction with indoor environmental quality in green buildings, HB 2006 - Heal. Build. Creat. a Heal. Indoor Environ. People, Proc. 3 (2006) 365–370.
[21] G.R. Newsham, B.J. Birt, C. Arsenault, A.J.L. Thompson, J.A. Veitch, S. Mancini, A.D. Galasiu, B.N. Gover, I.A. MacDonald, G.J. Burns, Do green buildings have better indoor environments? New evidence, Build. Res. Inf. 41 (2013) 415–434. https://doi.org/10.1080/09613218.2013.789951.
[22] Z. Gou, S.S.Y. Lau, Z. Zhang, A comparison of indoor environmental satisfaction between two green buildings and a conventional building in China, J. Green Build. 7 (2012) 89–104. https://doi.org/10.3992/jgb.7.2.89.
[23] S. Ravindu, R. Rameezdeen, J. Zuo, Z. Zhou, R. Chandratilake, Indoor environment quality of green buildings: Case study of an LEED platinum certified factory in a warm humid tropical climate, Build. Environ. 84 (2015) 105–113. https://doi.org/10.1016/j.buildenv.2014.11.001.
[24] L.K. Tee, C. Chee, H.H. Mohamed, O.S. Lee, E-clean: A data cleaning framework for patient data, Proc. - 1st Int. Conf. Informatics Comput. Intell. ICI 2011. (2011) 63–68. https://doi.org/10.1109/ICI.2011.21.
[25] P.K. Prerna Mishra, Ranjana Rajnish, Sentiment analysis of twitter data: Case study on digital India, (2016) 170–184. https://doi.org/10.4018/978-1-5225-3787-8.ch011.
[26] No title, (n.d.). https://www.usgbc.org/projects.
[27] W.J. Wilbur, K. Sirotkin, The automatic identification of stop words, J. Inf. Sci. 18 (1992) 45–55. https://doi.org/10.1177/016555159201800106.
[28] V.S. Pagolu, K.N.R.C. Challa, G. Panda, B. Majhi, Sentiment analysis of Twitter data for predicting stock market movements, (2016) 1345–1350.
[29] Y.S. Chen, L.H. Chen, Y. Takama, Proposal of LDA-Based Sentiment Visualization of Hotel Reviews, Proc. - 15th IEEE Int. Conf. Data Min. Work. ICDMW 2015. (2016) 687–693. https://doi.org/10.1109/ICDMW.2015.72.
[30] W. Zhao, J.J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, A heuristic approach to determine an appropriate number of topics in topic modeling, BMC Bioinformatics. 16 (2015) S8. https://doi.org/10.1186/1471-2105-16-S13-S8.

[31] D. Putthividhya, H.T. Attias, S.S. Nagarajan, Topic regression multi-modal latent dirichlet allocation for image annotation, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2010) 3408–3415. https://doi.org/10.1109/CVPR.2010.5540000.
[32] Q. Chen, L. Yao, J. Yang, Short text classification based on LDA topic model, ICALIP 2016 - 2016 Int. Conf. Audio, Lang. Image Process. - Proc. (2017) 749–753. https://doi.org/10.1109/ICALIP.2016.7846525.
[33] D. Ghosh, R. Guha, What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System, Cartogr. Geogr. Inf. Sci. 40 (2013) 90–102. https://doi.org/10.1080/15230406.2013.776210.
[34] C. Jacobi, W. Van Atteveldt, K. Welbers, Quantitative analysis of large amounts of journalistic texts using topic modelling, Digit. Journal. 4 (2016) 89–106. https://doi.org/10.1080/21670811.2015.1093271.
[35] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring topic coherence over many models and many topics, EMNLP-CoNLL 2012 - 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. Proc. Conf. (2012) 952–961.
[36] C. Fellbaum, WordNet, Wiley Online Library, 1998.
[37] USGBC, LEED v4 for neighborhood development, 2009 (2014) 71.
[38] USGBC, LEED v4 for interior design and construction, (2015).
[39] USGBC, LEED v4 for building operations and maintenance, (2018) 116.
[40] USGBC, LEED v4 credits for building design and construction, (2019) 147.
[41] A. Alsaeedi, M.Z. Khan, A study on sentiment analysis techniques of Twitter data, Int. J. Adv. Comput. Sci. Appl. 10 (2019) 361–374. https://doi.org/10.14569/ijacsa.2019.0100248.
[42] J.A. Morente-Molinera, G. Kou, Y. Peng, C. Torres-Albero, E. Herrera-Viedma, Analysing discussions in social networks using group decision making methods and sentiment analysis, Inf. Sci. (Ny). 447 (2018) 157–168. https://doi.org/10.1016/j.ins.2018.03.020.
[43] A. Bermingham, M. Conway, L. Mclnerney, N. O'Hare, A.F. Smeaton, Combining social network analysis and sentiment analysis to explore the potential for online radicalisation, Proc. 2009 Int. Conf. Adv. Soc. Netw. Anal. Mining, ASONAM 2009. (2009) 231–236. https://doi.org/10.1109/ASONAM.2009.31.
[44] E. Frank, L. Trigg, G. Holmes, I.H. Witten, Technical note: Naive Bayes for regression, Mach. Learn. 41 (2000) 5–25. https://doi.org/10.1023/A:1007670802811.
[45] S. Liu, Z. Wang, S. Schiavon, Y. He, M. Luo, H. Zhang, E. Arens, Predicted percentage dissatisfied with vertical temperature gradient, Energy Build. 220 (2020). https://doi.org/10.1016/j.enbuild.2020.110085.
[46] Tableau, Tableau Prep Help, Tableau. (2020). https://onlinehelp.tableau.com/current/prep/en-us/prep_welcome.htm.
[47] D.A. Prum, T. Kobayashi, Green building geography across the United States: Does governmental incentives or economic growth stimulate construction?, SSRN Electron. J. 43 (2013). https://doi.org/10.2139/ssrn.2276185.
[48] S. Han, C.K. Anderson, Customer Motivation and Response Bias in Online Reviews, Cornell Hosp. Q. 61 (2020) 142–153. https://doi.org/10.1177/1938965520902012.

[49] N. Lassen, F. Goia, S. Schiavon, J. Pantelic, Field investigations of a smiley-face polling station for recording occupant satisfaction with indoor climate, Build. Environ. 185 (2020) 107266. https://doi.org/10.1016/j.buildenv.2020.107266.
[50] L.Y. Heng, R. Logeswaran, B.P. Marikannan, Performance evaluation of analytics models for trends analysis of news, J. Phys. Conf. Ser. 1712 (2020). https://doi.org/10.1088/1742-6596/1712/1/012021.
[51] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, D.M. Blei, Reading tea leaves: How humans interpret topic models, (2009).
[52] Z.M. Darus, N.A. Hashim, E. Salleh, L.C. Haw, A.K.A. Rashid, S.N.A. Manan, Development of rating system for Sustainable building in Malaysia, WSEAS Trans. Environ. Dev. 5 (2009) 260–272.
[53] I. Razak, N. Nirwanto, B. Triatmanto, The impact of product quality and price on customer, J. Mark. Consum. Res. 30 (2016) 59–68.
[54] T. Radojevic, N. Stanisic, N. Stanic, Ensuring positive feedback: Factors that influence customer satisfaction in the contemporary hospitality industry, Tour. Manag. 51 (2015) 13–21. https://doi.org/10.1016/j.tourman.2015.04.002.
[55] USGBC, LEED v4.1 residential BD+C, multifamily homes and multifamily homes core and shell (aparted for India), (2020).
[56] M.A. Davis, W.D. Larson, S.D. Oliner, J. Shui, The price of residential Land for counties, ZIP codes, and census tracts in the United States, J. Monet. Econ. (2020). https://doi.org/10.1016/j.jmoneco.2020.12.005.
[57] L.E.L. Paul P. Biemer, Introduction to Survey Quality, New York: John Wiley, 2003.
[58] JAMA, Instructions for Authors, Am. Med. Assoc. (n.d.). http://jama.jamanetwork.com/Public/Instructionsforauthors.Aspx.
[59] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, J.N. Rosenquist, Understanding the Demographics of Twitter Users, Int'l AAAI Conf. Weblogs Soc. Media. (2011) 554–557.
[60] R. Vaughan, Oversampling in health surveys: Why, when, and how?, Am. J. Public Health. 107 (2017) 1214–1215. https://doi.org/10.2105/AJPH.2017.303895.
[61] A. Culotta, Reducing Sampling Bias in Social Media Data for County Health Inference, Jt. Stat. Meet. Proc. (2014). http://cs.iit.edu/~culotta/pubs/culotta14reducing.pdf%5Cnhttp://tapilab.github.io/public health/2014/08/02/bias/.

APPENDIX

Table A1. Word dictionary for all the topics

Toipcs	Themes	Seed words
Location & Transportation	Location	school, warehouse, center, retail, healthcare, community, public, mail, surrounding, farmland, industry, residential, site, visitability, alleys, safe, security stores, hospital, landscape, landmark, courtyard, plaza, grocery, market, supermarket, bank, mall, theater, pharmacy, gym, laundry, library, clinic, university, restaurant, warehouse, hotel, vendor, Church, club, studio, café, college, healthcare, telecommunication, landfill, waterway, housing, daycare, education, postsecondary, nursery, sun, winds, weather, rain, mapping, slope, stability, flood, wetlands, lakes, streams, shorelines, rainwater;
	Transportation	walking, vehicle, bus, stop, station, bicycle, Uber, car, ferry, sidewalk, pedestrian, transit, corridor, distance, commute, carpool, connectivity, shuttles, pavement, walkways, roadways, bikeway, path, convenience, motor, automobile, streetcar, rail, carshare, rideshare, passenger, route, tour, trips, travel, freight, lane, on-street, off-street, fleet, truck, dock, conveyance;
Running cost	Cost,	Bill, money, capital, rent, payment, charge, maintenance, purchase, price, economic, sustainable, consumption, efficiency, saving, budget, power, depletion, lifecycle, waste, value, burden, discount, finance, benefits, water, gas, fuel, oil, steam, electricity, propane, load, metering, grid, solar, PV;
Health & Wellbeing	Amenities	pool, park, garden, seating, parking, utility, recreation, entertainment, sports, infrastructure, sanitation, flora, fauna, planting, vegetation, trees, greenfield, view
	Management services	Respond, assist, service, lease, office, management, staff, arrangement, agency, housekeeping, regulation, code, policy, trash, stewardship, organize, contractor, hospitality, cleaning, cleanliness, pest, leakage, hygiene, recycling, repair;
	Pet-policy	Dog, cat, pet;
	Appliance	equipment, machine, elevator, fryer, griddle, drawer, cooker, cooking, toaster, refrigerator, freezer, device, dishwasher, steamer, stove, oven, range, drain, tank, disposer, plug, fireplace, woodstove, showerhead, plumbing, network, broiler, faucet, lamp, alarm;

	Indoor environment	Thermal comfort: hot, warm, heat, cold, cool, chill, temperature, radiation; Acoustic: noise, voice, sound, loud; Lighting: daylight, luminaire, luminance, illuminance, sunlight, dim; IAQ: CO_2, smoke, smoking, fume, odour, contaminant, airflow, emission, VOC, moisture, humidity, ozone, particle, smell, pollution, ventilation, air-conditioning, fan, exhaust, combustion, chemical, biological, particulate, toxicity, vent, formaldehyde, tobacco, exposure; Layout: restroom, ceiling, roof, floor, window, bathroom, design, space, cabinet, locking, hallway, stairwell, closet, basement, porch, lavatory, urinals, balcony, rooftop, toilets, shower, furniture, kitchen, storage, doors, material, garage, wallwash, paints, coating, carpet, area, vestibule, refurbish, gate, decorate, cabinetry, lobby, construction;

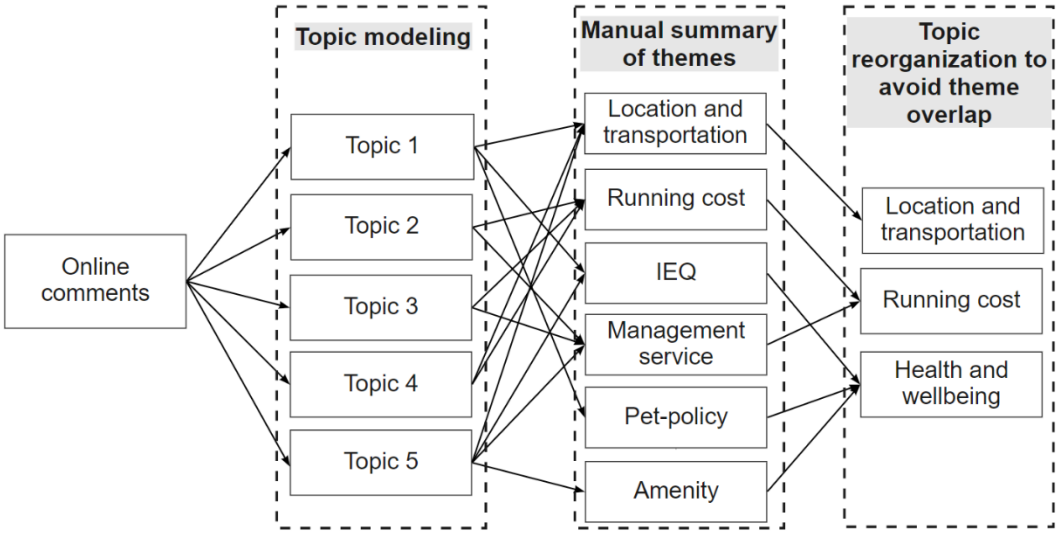


Figure A1. Diagram of topic modelling and topic reorganization