Routledge
Taylor & Francis Group

Building Research & Information

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rbri20

Investigating the indoor environmental quality of different workplaces through web-scraping and text-mining of Glassdoor reviews

Giorgia Chinazzo

To cite this article: Giorgia Chinazzo (2021) Investigating the indoor environmental quality of different workplaces through web-scraping and text-mining of Glassdoor reviews, Building Research & Information, 49:6, 695-713, DOI: 10.1080/09613218.2021.1908879

To link to this article: https://doi.org/10.1080/09613218.2021.1908879

BUILDING RESEARCH & INFORMATION
2021, VOL. 49, NO. 6, 695–713
https://doi.org/10.1080/09613218.2021.1908879

Investigating the indoor environmental quality of different workplaces through web-scraping and text-mining of Glassdoor reviews

Giorgia Chinazzo

Civil and Environmental Engineering Department, Northwestern University, Evanston, IL, USA

ABSTRACT
The analysis of occupants' perception can improve building indoor environmental quality (IEQ). Going beyond conventional surveys, this study presents an innovative analysis of occupants' feedback about the IEQ of different workplaces based on web-scraping and text-mining of online job reviews. A total of 1,158,706 job reviews posted on Glassdoor about 257 large organizations (with more than 10,000 employees) are scraped and analyzed. Within these reviews, 10,593 include complaints about at least one IEQ aspect. The analysis of this large number of feedbacks referring to several workplaces is the first of its kind and leads to two main results: (1) IEQ complaints mostly arise in workplaces that are not office buildings, especially regarding poor thermal and indoor air quality conditions in warehouses, stores, kitchens, and trucks; (2) reviews containing IEQ complaints are more negative than reviews without IEQ complaints. The first result highlights the need for IEQ investigations beyond office buildings. The second result strengthens the potential detrimental effect that uncomfortable IEQ conditions can have on job satisfaction. This study demonstrates the potential of User-Generated Content and text-mining techniques to analyze the IEQ of workplaces as an alternative to conventional surveys, for scientific and practical purposes.

ARTICLE HISTORY
Received 11 November 2020
Accepted 22 March 2021

KEYWORDS
Built environment; data mining; indoor environmental quality; occupant perceptions

Introduction

Investigations on the indoor environmental quality (IEQ) of buildings, comprising thermal, visual, indoor air quality (IAQ), and acoustic aspects, are essential to design and operate buildings that both conserve energy and satisfy occupant needs. Specifically in workplaces, researchers have shown that IEQ significantly affects workers' well-being, health, and productivity (Al horr et al., 2016; Alker et al., 2014; Clements-Croome, 2006; Humphreys & Nicol, 2007). Considering that people's salaries and benefits represent the largest part of the expenses linked to the life cycle of a commercial building (with the other expenses associated with utilities and rent), an improvement of the IEQ of the workplace could result in significant economic benefits for an organization due to the resulting increased productivity and health (Allen & Macomber, 2020). Additionally, such improvements have been shown to have a positive influence on workers even after they leave the office, including improved sleep quality (Aries et al., 2010). Therefore, it is clear that improving the IEQ of the workplace can yield not only environmental benefits but also health and economic ones.

A pivotal method to improve the IEQ of buildings consists of collecting occupants' feedback about the indoor environment during building operations. Conventionally, subjective impressions are gathered through surveys, interviews, and walkthroughs (Li et al., 2018). Surveys represent the most used method of subjective data collection (Li et al., 2018), and they primarily comprise closed-ended questions usually linked to domain-specific response scales (e.g. thermal comfort, thermal sensation). Ongoing work seeks to improve how subjective data are collected through surveys. The principal goal consists of expanding the spatial and temporal granularity of surveys with the use, for example, of micro ecological momentary assessments on a smartwatch platform (Jayathissa et al., 2020) and continuous occupant voting systems (OVS) (Sheikh Khan et al., 2020). Nevertheless, three main limitations of surveys can still be highlighted: (1) the great majority of surveys target office buildings, discarding other types of workplaces; (2) occupants' feedback is collected only if the survey is run in a specific building (a decision made either by researchers or by building managers, but not by building occupants); (3) due to the closed-ended nature of most of the questions, the feedback collected

CONTACT Giorgia Chinazzo giorgia.chinazzo@northwestern.edu

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09613218.2021.1908879.

could be biased towards the specific questions and scales used.

A way to overcome such limitations would be to analyze voluntary and unbiased feedback about the IEQ directly coming from building occupants and not solicited by specific surveys. Online reviews referring to the workplace could contain such type of information due to their open-ended nature. In recent years, online reviews have been recognized as the most common source of information among various User-Generated Content (UGC), which has been indicated as one of the most rapidly growing sources of information (Jung & Suh, 2019). At the same time, advances in natural language processes and software capabilities now allow researchers to analyze unstructured and text-heavy data through text-mining techniques. Thanks to the analysis of a large number of online reviews through text mining, it is possible to work backwards and gain insights about previously unknown information. This method has mainly been adopted in the marketing and hospitality sectors to analyze online reviews from various websites and social media platforms, including Twitter and Facebook (He et al., 2013; Mostafa, 2013), Booking.com (Sutherland et al., 2020; Xu & Li, 2016), TripAdvisor (Alrawadieh & Law, 2019; Berezina et al., 2016; Guo et al., 2017; Kim et al., 2016; Lee et al., 2017; Wong & Qi, 2017; Zhao et al., 2019) and Airbnb (Cheng & Jin, 2019; Sutherland & Kiatkawsin, 2020; Zhang, 2019). In the context of IEQ investigations, text-mining techniques have been used to analyze open-ended responses in conventional surveys and interviews of post-occupancy evaluation studies (Day & O'Brien, 2017; Moezzi & Goins, 2011; Ortiz & Bluyssen, 2019). Only two studies have used text-mining techniques to analyze IEQ information extracted from online reviews (Qi et al., 2017; Villeneuve & O'Brien, 2020). In addition, these studies analyzed online reviews referring to the hospitality sector (Airbnb and hotels) (Qi et al., 2017; Villeneuve & O'Brien, 2020), and not to the workplace. The Glassdoor website has been highlighted as a potential source of information about the IEQ of different workplaces (Allen & Macomber, 2020). However, to date, Glassdoor reviews have only been analyzed from a management perspective (Dabirian et al., 2017; Moro et al., 2020). Therefore, no study has used text-mining techniques to evaluate occupants' feedback about the IEQ of the workplace extracted from online reviews.

This study aims to fill this research gap by analyzing 1,158,706 job reviews posted on Glassdoor from 2008 to 2020 to investigate the IEQ of different workplaces. The online job reviews of 257 firms and institutions worldwide are scraped from the web and analyzed through text-mining techniques. This innovative data collection method proposes to analyze occupants' feedback about the IEQ extracted from online job reviews from a variety of workplaces, rather than from conventional surveys in office buildings. In this way, the limitations of surveys previously highlighted are overcome as (1) online job reviews refer to a variety of workplaces and not only to office buildings, (2) all occupants in any buildings can express their opinions rather than only those in which a survey is conducted, and (3) feedback is unsolicited and not prompted by specific questions about the indoor environment. Consequently, the web-scraping and text-mining of online job reviews allow obtaining worldwide, voluntary and unbiased information about the IEQ of different workplaces that would not be possible to obtain (or too costly and time-consuming) with conventional surveys.

Among the information extracted from Glassdoor, this study focuses on the negative comments of each review (i.e. cons), which, in the following, will be referred to as 'IEQ complaints'. It is important to note that such complaints are open-ended text and are not limited to a set of options or responses as in conventional closed-ended questions. Besides focusing on the results of the analysis of such complaints, this paper describes the method used to extract the IEQ complaints from the online job reviews and to analyze them through text-mining techniques. This approach expands and integrates methods previously used to analyze datasets not related to the workplace environment. Through the use of the described method, this study aims at addressing the following research questions:

- RQ1: Which is the percentage of IEQ complaints in online job reviews compared to other types of complaints referring to the working environment?
- RQ2: Which is the prevailing source of complaint among the IEQ aspects (thermal, visual, IAQ, acoustic) and their combination?
- RQ3: Do the percentage of IEQ complaints and the prevailing source of complaint vary according to the type of workplace?
- RQ4: Which are the reasons for discomfort for each IEQ aspect (e.g. too cold or too hot for thermal discomfort, and too dim or bright for visual discomfort)?
- RQ5: Do online reviews with IEQ complaints differ from those without IEQ complaints, highlighting a potential influence of IEQ complaints on the overall job review (hence, potentially, on overall job satisfaction)?

Methods

A three-phase methodology is used in this study (Figure 1): (1) data selection, (2) data preparation, and (3) data analysis. While phase 1 describes the web-scraping step, phases 2 and 3 represent the two processes that characterize text mining, namely text pre-processing and knowledge extraction (Kumar & Ravi, 2016).

Data selection

This phase includes the selection of the website, organizations, and items to scrape, as detailed in the following (Figure 1).

Website selection

This study analyses data collected from Glassdoor, a job and recruiting website that currently hosts 60 million reviews of more than 1 million companies from more than 190 countries (Glassdoor, n.d.). The users of the Glassdoor website are both employees posting reviews and information about job positions, firms and institutions, as well as people interested in knowing such information. Glassdoor offers a variety of user-generated information, ranging from company reviews to salary and benefits information. Within the company reviews, users can provide quantitative and qualitative information through numerical scores of specific categories and free input text. The textual information shared by users is divided into positive and negative feedback (i.e. pros and cons), and sometimes into an additional category indicating advice to management.

Organizations selection

From the extensive database of companies and institutions reviewed on Glassdoor, the job reviews of 257 organizations posted from April 2008 to July 2020 are scraped (i.e. extracted from the web). The companies and institutions with more than 10,000 employees are selected to focus the investigation on large organizations. Preference is given to organizations with a larger number of reviews (greater than 1000) to reduce review bias. Only industry sectors with a sufficient number of organizations satisfying the selection criteria (i.e. size and number of reviews) are included in the analysis. The majority of the companies and institutions analyzed are public (67.3%), followed by college and universities (12.8%), private (8.2%), non-profit organization (4.7%), subsidiary or business segment (4.3%), hospital (2.3%), and government (0.4%). More than half of the organizations have a revenue of $10+ billion (USD) per year (60.3%), followed by $2 to $5 billion (USD) per year (13.2%), $5 to $10 billion (USD) per year (11.7%), and $1 to $2 billion (USD) per year (5.4%).

Items to scrape selection

The web-links to all the companies and institutions webpages are manually inserted in a customized R script used for scraping specific information from English reviews: date, title of the review, employment status, pros, cons, and author information (Figure 2(c)). This semi-automated process is adopted as Glassdoor does not allow running scripts to automatically extract all employees' reviews (Dabirian et al., 2017). However, the designed script allows to automatically open the webpages with the reviews of each organization, extract the selected information based on its html pattern, and save it in a CSV file.

Scraped reviews

A total of 1,158,706 English job reviews results from the data selection phase. This number includes reviews that report the date as those without it (i.e. 111 reviews) are excluded from the dataset before the analysis. Figure 2 (a) and (b) reports the distribution of the scraped job reviews according to the industry sector and publication year, respectively. More than half of the reviews refer to companies in IT (23.0%), retail (17.3%), and finance (15.4%) sectors. An increasing number of reviews are posted over the years, as illustrated in Figure 2(b), with a peak number of reviews in 2017. The smaller number of reviews in 2020 is linked to the fact that the web scraping was performed in July 2020.

Figure 2(d) illustrates the geolocation of the scraped reviews indicating the job location extracted from the author's information. Not all reviews indicate the location in the author's information (not reported in 39.5%). Reviews with a job location are collected from 173 countries, with the majority of the reviews located in the U.S. (68.8%), India (12.7%), U.K. (4.2%), and Canada (4.0%). The uneven distribution of reviews in the world is most likely because only English reviews are analyzed. A discussion on this subject is reported in Section '*Limitations*'.

Data preparation

The data preparation consists of three steps: (1) pre-processing, (2) iterative cleaning, and (3) extraction, categorization, and tokenization (Figure 1).

Pre-processing

The data preparation starts from the pre-processing of the text reported in the cons section and consists of the removal of stop words such as 'a', 'the', 'of', 'and',

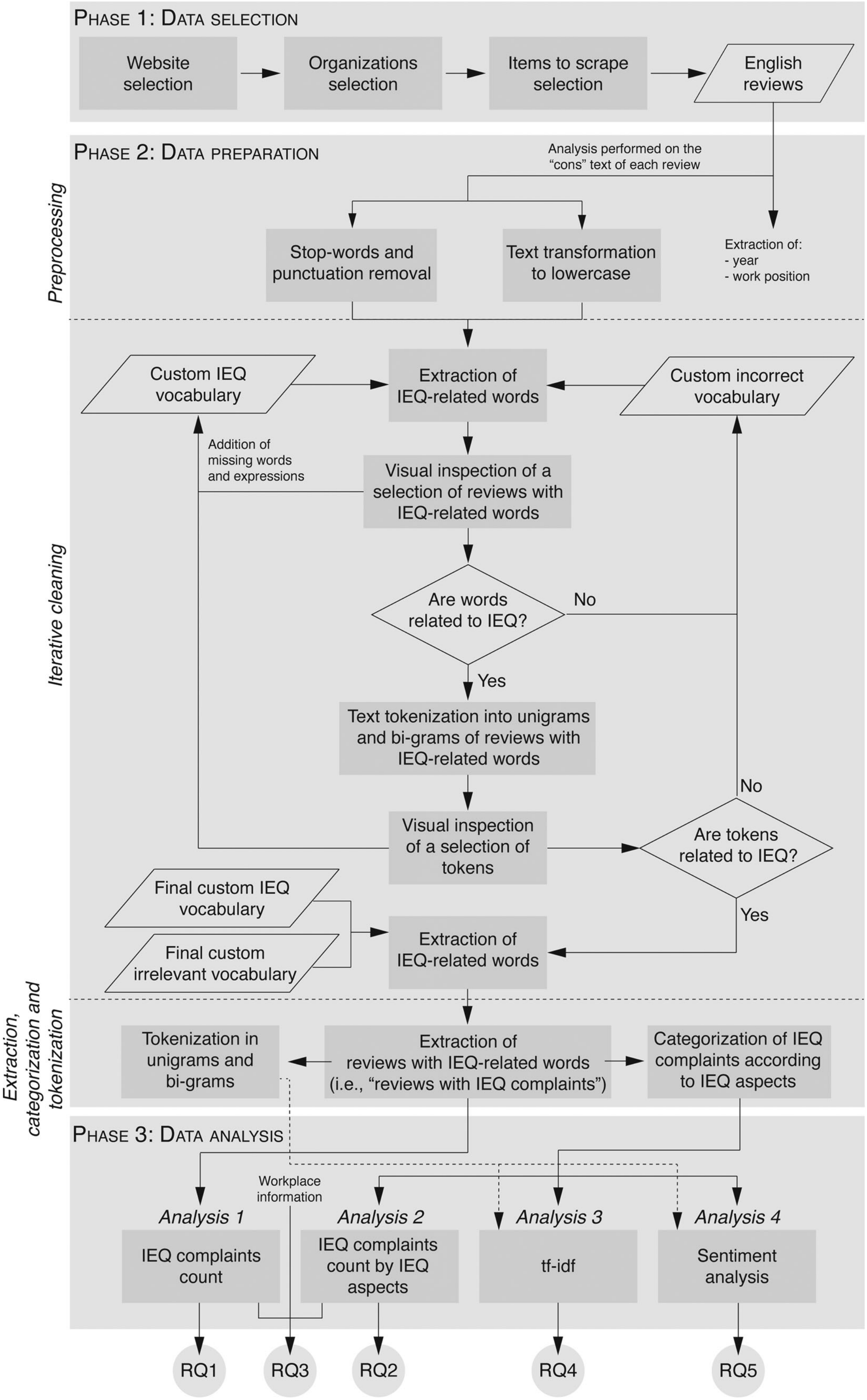
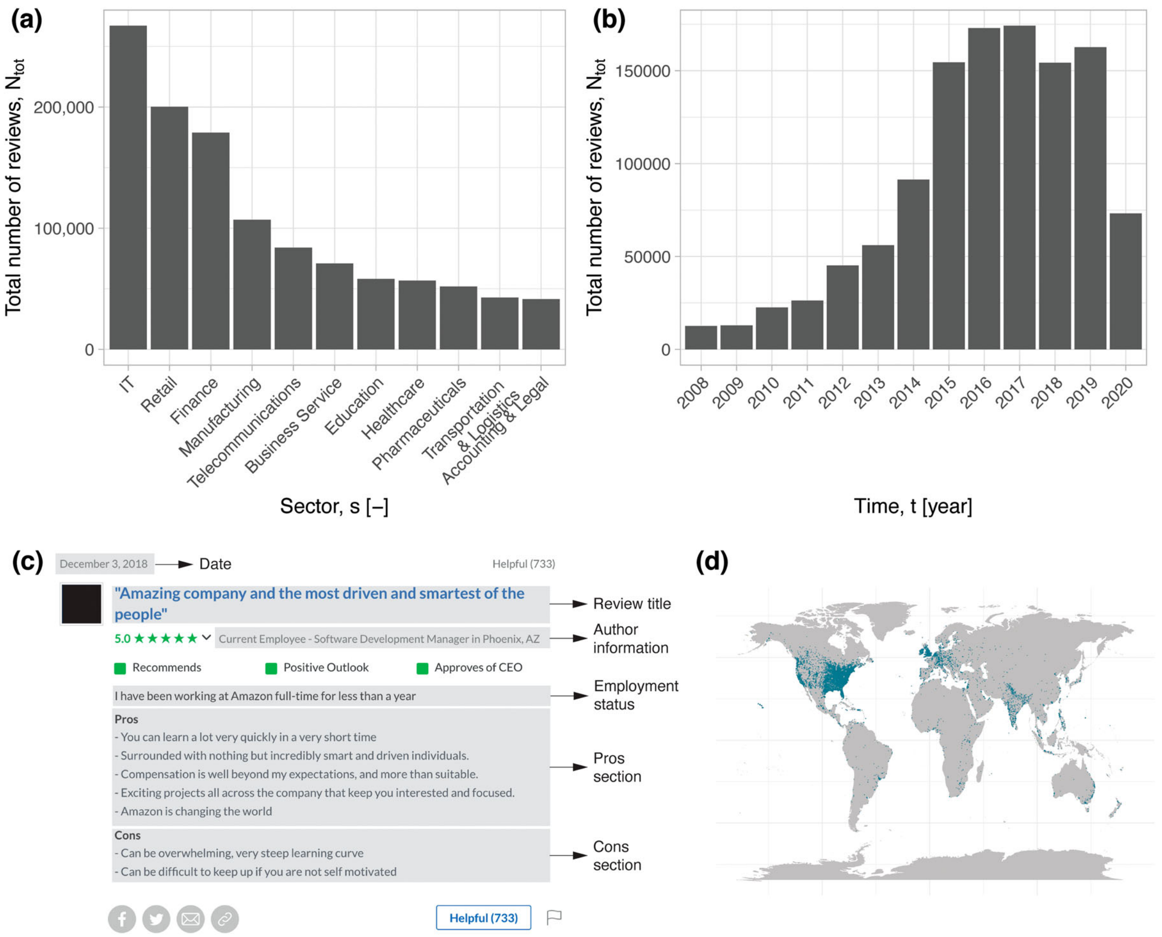

Phase 1: Data selection
Website selection
Organizations selection
Items to scrape selection
English reviews
Phase 2: Data preparation
Preprocessing
Analysis performed on the "cons" text of each review
Stop-words and punctuation removal
Text transformation to lowercase
Extraction of:
- year
- work position
Iterative cleaning
Custom IEQ vocabulary
Extraction of IEQ-related words
Custom incorrect vocabulary
Addition of missing words and expressions
Visual inspection of a selection of reviews with IEQ-related words
Are words related to IEQ?
No
Yes
Text tokenization into unigrams and bi-grams of reviews with IEQ-related words
Visual inspection of a selection of tokens
Are tokens related to IEQ?
No
Yes
Final custom IEQ vocabulary
Final custom irrelevant vocabulary
Extraction of IEQ-related words
Extraction, categorization and tokenization
Tokenization in unigrams and bi-grams
Extraction of reviews with IEQ-related words (i.e., "reviews with IEQ complaints")
Categorization of IEQ complaints according to IEQ aspects
Phase 3: Data analysis
Workplace information
Analysis 1
IEQ complaints count
Analysis 2
IEQ complaints count by IEQ aspects
Analysis 3
tf-idf
Analysis 4
Sentiment analysis
RQ1
RQ3
RQ2
RQ4
RQ5

Figure 1. Methodology for web-scraping and text mining of online reviews.



December 3, 2018 Helpful (733)

"Amazing company and the most driven and smartest of the people"

5.0 ★★★★★ Current Employee - Software Development Manager in Phoenix, AZ

🟩 Recommends 🟩 Positive Outlook 🟩 Approves of CEO

I have been working at Amazon full-time for less than a year

Pros
- You can learn a lot very quickly in a very short time
- Surrounded with nothing but incredibly smart and driven individuals.
- Compensation is well beyond my expectations, and more than suitable.
- Exciting projects all across the company that keep you interested and focused.
- Amazon is changing the world

Cons
- Can be overwhelming, very steep learning curve
- Can be difficult to keep up if you are not self motivated

Helpful (733)

Figure 2. Summary of scraped job reviews from Glassdoor. (a) Number of reviews scraped from Glassdoor according to the industry sector. (b) Number of reviews scraped from Glassdoor according to the year of publication. (c) Items scraped from each online job review. (d) Geolocation of the scraped job reviews (each point refers to one review).

etc., and the transformation of all words into lowercase. In this phase, other information is extracted from the job reviews: the year from the date and the work position from the author's information (Figure 2(c)).

Iterative cleaning

This step aims to extract for each job review the words associated with IEQ (i.e. IEQ-related words) if any. The word extraction is performed with the use of a custom vocabulary (i.e. 'IEQ vocabulary') as performed in other studies extracting IEQ information from online reviews in the hospitality sector (Qi et al., 2017; Villeneuve & O'Brien, 2020). The vocabulary wordings are chosen to reduce the probability of extracting terms that could refer to both IEQ and work-related issues. One of the most outstanding examples is the word 'stink.' This word is excluded from the IEQ vocabulary as its search yields to 487 words, most of which refer to the description of a management/job situation (e.g. 'the leadership stinks'). Similar cases are 'bright', 'light', 'dry', and 'sound' (e.g. 'bright ideas', 'the light at the end of the tunnel', 'work can be dry', 'technically sound'). These types of words are inserted in the IEQ vocabulary as expressions (e.g. 'bright lights' and 'sound pollution') and not as single words to avoid extracting work-related words. However, an automated exclusion method is used for other words, such as for 'hot' and 'noise'. Non-IEQ expressions such as 'hot water' and 'make noise for yourself' are included in an additional custom vocabulary (i.e. 'incorrect vocabulary'), used together with the IEQ vocabulary to extract IEQ-related words from each review automatically. The purpose of the iterative cleaning is to include as many words and expressions as possible in these two vocabularies to allow a more precise IEQ-related word extraction. The iterative cleaning process follows these steps:

(a) IEQ-related words are extracted with the use of an initial IEQ vocabulary, created based on the author's knowledge and previous literature;
(b) 1200 reviews (300 per IEQ aspect) are visually inspected to detect the incorrect classification of

words into IEQ-related words and expressions and potential new IEQ-related words or expressions. The new words and expressions are included in the two vocabularies;

(c) The job reviews are analyzed again with the extended IEQ and incorrect vocabularies, extracting IEQ-related words again for each cons review;
(d) Step (b) is repeated with 400 reviews (100 per IEQ aspect);
(e) Step (c) is repeated;
(f) The reviews containing IEQ-related words are divided into a list of unigrams and bi-grams through tokenization (see explanation about tokens in section 'Extraction, Categorization, and Tokenization');
(g) Approximately 2500 unigrams and bi-grams are visually inspected to further enrich the IEQ vocabulary and irrelevant vocabularies;
(h) The expanded vocabularies are finally used to analyze the initial dataset of cons reviews and extrapolate the final reviews with IEQ-related words.

Table 1 summarizes the final IEQ vocabulary used in the analysis. The incorrect vocabulary is reported in Appendix A.

Extraction, categorization, and tokenization

The goal of the last phase of the data preparation is threefold: extraction of reviews containing IEQ complaints, categorization of such reviews according to the four IEQ aspects and their combination, and creation of word tokens of such reviews.

A review is considered to report an IEQ complaint if it contains at least one IEQ-related word. Once these types of reviews are extracted, they are categorized according to the four IEQ aspects and their combination, concerning the IEQ-related word(s) present in the review. For example, one review containing two words referring to a thermal complaint (e.g. 'warm temperature') is classified as 'Thermal', and another review presenting one word for a visual complaint and one for an IAQ complaint (e.g. 'dark and dusty environment') is classified as 'Combined'. At the end of the categorization process, each review is classified into one of the following six categories: 'Thermal', 'Visual', 'IAQ', 'Acoustic', 'Combined', and 'No IEQ complaint' (referring to reviews not reporting any IEQ-related word). Also, reviews classified as 'Combined' are further divided according to the combination of IEQ aspects: 'Thermal & Visual', 'Thermal & IAQ', 'Thermal & Acoustics', 'Visual & Acoustic', 'Visual & IAQ', 'IAQ & Acoustic', 'Three IEQ' and 'four IEQ' (if a review contains three or four IEQ-related words referring to three or four IEQ aspects). Finally, from reviews containing IEQ complaints, tokens are created. A token is defined as an 'instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing' (Manning et al., 2010). Through tokenization, the process of dividing a text into tokens, it is possible to extract both unigrams (i.e. single words) and bi-grams (i.e. two subsequent words) composing a text. From the reviews containing IEQ complaints, two lists are extracted, one for unigrams and the other for bi-grams.

Data analysis

The data analysis phase includes four main analyses used to address the research questions previously introduced (Figure 1).

Analysis 1: IEQ complaints count

Analysis 1 intends to count the reviews containing IEQ complaints, first considering the entire dataset and then within each industry sector and work position. This analysis is used to investigate the percentage of IEQ complaints reported in online job reviews posted on Glassdoor (RQ1) and to study if the rate of IEQ complaints varies according to the type of workplace (RQ3). To address RQ1, the total number of reviews containing IEQ complaints, N_{IEQ}, is counted as well as the total number of scraped reviews, N_{tot}. Then, these

Table 1. IEQ vocabulary indicating the words and expressions for each IEQ aspect.

IEQ aspect	Words and expressions
Thermal	thermal comfort, hot, heat, warm, humid, cold, draughty, chilly, freezing, HVAC, AC, A/C, draft, drafty, boiling, sticky, frostbite(s), cool off, warm up, temperature(s), climate control, air conditioner(s)
Visual	visual comfort, dark, dim, no window(s), few windows, small window(s), windowless, without window(s), glare, glary, bright light(s), blinding, flicker(s), flickering, lighting, room light(s), fluorescent light(s), indoor light(s), without light(s), no daylight, lack of daylight, little daylight, natural light, no sunlight, no sun light, screen reflections, lamp(s), lamp, artificial light, artificial lights, artificially lit, office(s) without window(s), poorly lit, store lights, office lights, warehouse lights, overhead lights, room lights, lights are (very) bright, flashlights, space(s) has(have) no light, office(s) has(have) no light, no sun/day light, too much light, too little light
IAQ	ventilation, odour, odor, smell(s), smelly, musty, dust, dusty, stench, rick, stale office, stale air, dry air, office is a little dry, damp, dingy, hazardous chemicals, feces, fresh air, air quality, unhealthy space, unhealthy indoor environment, healthy indoor environment, stuffy
Acoustic	acoustic comfort, noise, noisy, loud, blaring, sound pollution, loud sound, soundproof, racket, ear-splitting, squeaky, creak, vibration, beeping, quiet

two variables are used to determine the absolute percentage of reviews containing IEQ complaints, quantified as follows:

$$AP = \frac{N_{IEQ}}{N_{tot}} \cdot 100 \qquad (1)$$

The total number of reviews and the absolute percentage are calculated over the years to study their temporal variation due to the uneven number of reviews posted each year (Figure 2(b)).

To address RQ3, a series of relative percentages of reviews containing IEQ complaints are determined according to the following formula

$$RP_i = \frac{N_{IEQ,i}}{N_{tot,i}} \cdot 100 \qquad (2)$$

where i represents the different industry sectors and work positions. The normalization of the IEQ complaints reported for a specific industry sector or work position, $N_{IEQ,i}$ (e.g. the reviews containing IEQ complaints of the IT sector) with the total number of scraped reviews within the same industry sector or work position, $N_{tot,i}$ (e.g. the total scraped reviews of the IT sector) is necessary considering the uneven distribution of reviews across industry sectors (Figure 2(a)) and work positions.

Analysis 2: IEQ complaints count by IEQ aspects

Analysis 2 has the objective to count each review containing IEQ complaints according to the IEQ aspects and their combination, first considering the entire dataset and then within each industry sector and work position. This analysis is used to determine the source of complaints. More specifically, it is used to investigate the relative distribution of complaints concerning the IEQ aspects (RQ2) and to study if the distribution of IEQ complaints across IEQ aspects varies according to the type of workplace (RQ3). To address RQ2, the relative percentage of reviews containing IEQ complaints within IEQ aspects is determined as follows:

$$RP\_IEQ_k = \frac{N_{IEQ,k}}{N_{IEQ}} \cdot 100 \qquad (3)$$

where $N_{IEQ,k}$ indicates the number of reviews within a specific IEQ aspect (k). Besides for the five IEQ aspects, the relative percentage is calculated within the 'Combined' aspect to determine the distribution of sources of complaints when more than one aspect is reported.

To address RQ3, a series of relative percentages of reviews containing IEQ complaints concerning the IEQ aspects are determined according to the following formula

$$RP\_IEQ_{k,i} = \frac{N_{IEQ,k,i}}{N_{IEQ,i}} \cdot 100 \qquad (4)$$

where $N_{IEQ,k,i}$ is the number of reviews containing IEQ complaints referring to a specific IEQ aspect, k, for a given industry sector or work position, i, and $N_{IEQ,i}$ is the number of reviews containing IEQ complaints in the same industry sector or work position.

Analysis 3: term frequency-inverse document frequency (tf-idf)

Analysis 3 is performed to understand the reasons mentioned as source of discomfort for each IEQ aspect and reply to RQ4. Topic modelling such as the Latent Dirichlet Allocation (LDA) method (Blei et al., 2003) has been used in other text-mining studies of online reviews to understand patterns of topics in a text, such as topics of interests in reviews (Jia, 2020; Jung & Suh, 2019; Sutherland & Kiatkawsin, 2020; Sutherland et al., 2020; Zhang, 2019). This unsupervised classification is used in topic modelling (Silge & Robinson, 2017) and this method can potentially be used to extract the reasons for discomfort for each IEQ aspect. However, this method does not perform well with the present dataset, probably because IEQ complaints are only a small part of all the complaints, even within the reviews containing IEQ complaints (e.g. in the same review, a person might complain about the indoor temperature but also about the inadequate salary and a disrespectful manager). In this study, comments classified as IEQ complaints are not reduced to contain only the sentences referring to the IEQ as done in other studies such as in Villeneuve and O'Brien (2020). This choice is made to retain the entire text of each review, without risking excluding important parts of the text, as well as to analyze the entire comment in the sentiment score analysis (see Section 'Analysis 4: Sentiment Analysis').

The term frequency-inverse document frequency (*tf-idf*) is used for the analysis. The numerical statistic *tf-idf* is calculated by multiplying the term frequency $tf_{t,d}$, defined as the frequency of the term t in document d, with the inverse document frequency $idf_{t,D}$ of the same term t (Silge & Robinson, 2017)

$$tfidf_{t,d,D} = tf_{t,d} \cdot idf_{t,D} \qquad (5)$$

where $idf_{t,D}$ indicates the logarithmically scaled fraction of the total number of documents in a collection D and the number of documents d of the collection D in which the term t appears, according to the following formula

$$idf_{t,D} = \log \frac{N}{|\{d \in D{:}t \in d\}|} \tag{6}$$

If a specific word appears in all documents of a collection, $idf_{t,D} = 0$ as it results from the natural log of 1. The more unique a term is in a collection of documents ($idf_{t,D} > 0$) and the more often it appears in a specific document ($tf_{t,d} > 0$) the higher the value of $tfidf_{t,d,D}$ is.

Through the calculation of $tfidf_{t,d,D}$, it is possible to identify terms that characterize a document within a collection of documents (e.g. words specific to a single book within a collection of books). In this study, reviews containing IEQ complaints are considered as the collection of documents, with IEQ aspects determining the separate documents of the collection. As a consequence, the calculation of $tfidf_{t,d,D}$ is used to determine the distinctive words for each IEQ aspect (i.e. those with the highest $tfidf_{t,d,D}$). The method is applied to both unigrams and bi-grams and allows defining the most common words and combinations of two words for each IEQ aspect. However, due to its mathematical formulation, this method misses potential words that could represent reasons for discomfort, but that occur for all IEQ aspects (hence, that are not 'unique' for each IEQ aspect). Examples are the words 'people' and 'office'. For this reason, a term frequency for each unigram and bi-gram, m, ($tf_{m,k}$) is further calculated for each IEQ aspect as the number of times that the token appears within that IEQ aspect, $N_{m,k}$, concerning the total number of tokens referring to the same aspect, N_k

$$tf_{m,k} = \frac{N_{m,k}}{N_k} \tag{7}$$

$tf_{m,k}$ is then divided by the total number of reviews within each IEQ aspect to calculate the relative term frequency of each token ($Rtf_{m,k}$), according to the following formula:

$$Rtf_{m,k} = \frac{tf_{m,k}}{N_{IEQ,k}} \tag{8}$$

Tokens are then ordered from the most common to the least common according to $Rtf_{m,k}$. All tokens appearing in at least 1.5% of each IEQ aspect are visually inspected to detect potential terms missed with the $tfidf_{t,d,D}$ statistic, but that could refer to complaints about IEQ. The results of this semi-automated method are also used to verify the results of the $tfidf_{t,d,D}$ statistic (i.e. are the same words extracted? Which ones are missed?).

Analysis 4: sentiment analysis

Analysis 4 is performed to investigate whether reviews with IEQ complaints differ from those without IEQ complaints, highlighting a potential influence of IEQ complaints on the overall job review and potentially on overall job satisfaction (RQ5). The analysis consists in the comparison of the reviews containing IEQ complaints with reviews with no IEQ complaints, considering the categorization according to the IEQ aspects. For this analysis, a text-mining technique is used to estimate the 'sentiment score' of each review. The sentiment score is calculated with the sentiment analysis, a natural language processing (NLP) application (Zagal et al., 2012). This method has been used in text mining of general online reviews (Cheng & Jin, 2019; Jung & Suh, 2019; Lee et al., 2017; Mostafa, 2013) as well as in text mining of online reviews referring to IEQ in the hospitality sector (Villeneuve & O'Brien, 2020). The technique allows estimating the positive and negative attitude of a text by comparing the words of the text with those of a specific lexicon consisting of positive and negative words in English (Silge & Robinson, 2017).

The AFINN lexicon (Nielsen, 2011) is used in this analysis to estimate the sentiment score of each review, divided into unigram tokens. Compared to other lexicons that only categorize words in a binary fashion (i.e. negative and positive), the AFINN lexicon assigns a sentiment score between −5 and +5 to each token corresponding to a word that is present in the vocabulary, SS_m. The sentiment score for each review, n, characterized by J tokens, is therefore calculated as the sum of the sentiment score of its tokens following the formula:

$$SS_n = \sum_{m=1}^{J} SS_m \tag{9}$$

A score close to zero indicates a relatively neutral review, whereas positive and negative scores indicate positive and negative reviews, respectively. As it has been remarked in the literature (Silge & Robinson, 2017), the lexicons are based on unigrams and cannot, therefore, detect more complicated sentence structures, such as those using qualifiers before a word to indicate a negative sentiment (e.g. 'not good', 'not happy'), nor can detect sarcasm.

Statistical analyses are performed to compare the calculated sentiment scores across reviews through an analysis of variance with one independent variable (i.e. review type). First, data is tested for normality and homogeneity of variances with Anderson–Darling normality test (Anderson & Darling, 1954) and Levene's test

(Levene, 1949), respectively. Considering that these assumptions are violated, the use of the non-parametric Kruskal–Wallis test (Kruskal & Wallis, 1952) is adopted. To compare the results of the different review types in case of significant results, the post-hoc Dunn test for multiple comparisons is used, applying the Benjamini-Hochberg adjustment to the *p-values* to control the false discovery rate (Zar, 2013). Effect sizes are calculated with ε^2 and Vargha and Delaney's A (VDA) (Vargha & Delaney, 2000) to estimate the degree to which the sentiment scores of the review types are different from each other (overall and pairwise, respectively). The interpretation values for the effect size ε^2 are 0.01–0.06 (small effect), 0.06–0.14 (moderate effect) and $\geq$0.14 (large effect) (Khalilzadeh & Tasci, 2017). The interpretation values for Vargha and Delaney's A are 0.56–0.64 & 0.34–0.44 (small effect), 0.64–0.71 & 0.29–0.34 (moderate effect) and $\geq$0.71 & $\leq$0.29 (large effect) (Vargha & Delaney, 2000). Besides being beneficial in all statistical investigations, the calculation of the effect size is particularly suggested for large sample sizes as in the present study, as they can result in deflated *p-values* (Lin et al., 2013).

R version 3.6 (R Core Team, 2017) with the RStudio integrated development environment (RStudio Team, 2020) is used for data scraping, handling, screening, pre-processing, analysis, and visualization. The R packages used are *rvest*, *purr*, *curl* for the web scraping, *tidytext*, *stringr*, *dplyr,* and *textdata* for data handling, pre-processing and analysis, *car*, *nortest*, *rcompanion, FSA* and *coin* for the statistical analysis, and *ggplot2* (Wickham, 2009) for data visualization.

Results

IEQ complaints analysis

IEQ complaints number and trend over time

A total number of 10,593 reviews are categorized as IEQ complaints. This number corresponds to 0.91% of the total reviews scraped from Glassdoor. Figure 3(a) shows the trend of the IEQ complaints in the analyzed period (2008–2020). The trend indicates an increasing number of reviews over the years. However, the same trend was observed for the entire sample of reviews (Figure 2(b)), and the increasing number of IEQ complaints could simply be due to the increasing number of reviews posted online. Figure 3(b) displays the absolute percentage of IEQ complaints calculated considering the total number of reviews posted each year. The graph displays an overall decrease in the absolute percentage of IEQ complaints over the years, passing from a higher value than 1.1% to a lower value than 0.8%, corresponding to a decrease of about 30% in 12 years.

IEQ complaints by workplace

Figure 4(a) indicates the relative percentages of IEQ complaints by industry sector. The highest value occurs in the transportation & logistics sector, with about 3% of total reviews indicating a complaint about the IEQ.

Figure 4(b) indicates the relative percentage of IEQ complaints by work position. It must be remarked that the work position is not given in 14% of the total reviews (and in 12% of the reviews containing IEQ complaints). The majority of the IEQ complaints refer to an 'Anonymous Employee' (about 20%), which is not included in the analysis. The majority of IEQ complaints occur for the following positions: full-time and part-time package handler, followed by cook, material handler, warehouse associate, and fulfilment associate. These work positions do not correspond to traditional office jobs, as they are primarily performed in warehouses, kitchens, trucks, and stores. Only three office jobs are displayed in the figure, namely graduate research assistant, software developer, and customer service representative.

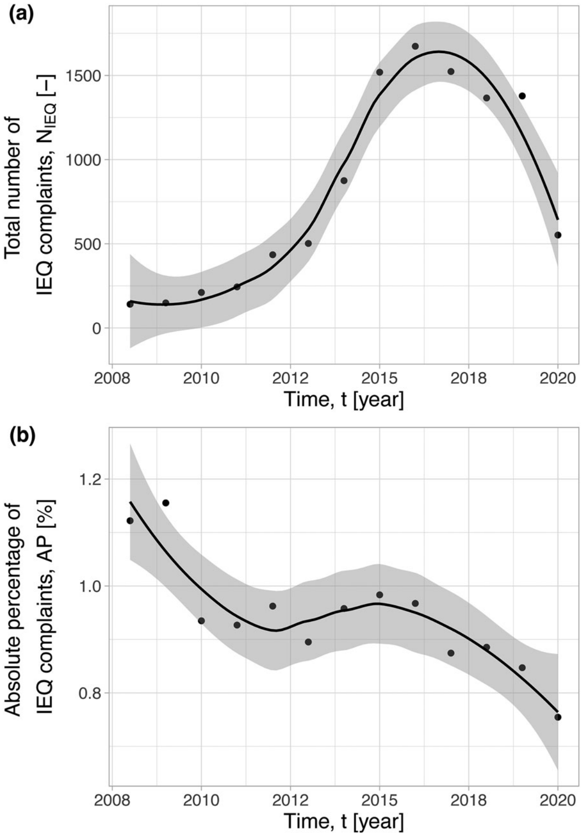


Figure 3. (a) Total number of IEQ complaints over time and (b) absolute percentage of IEQ complaints over time.

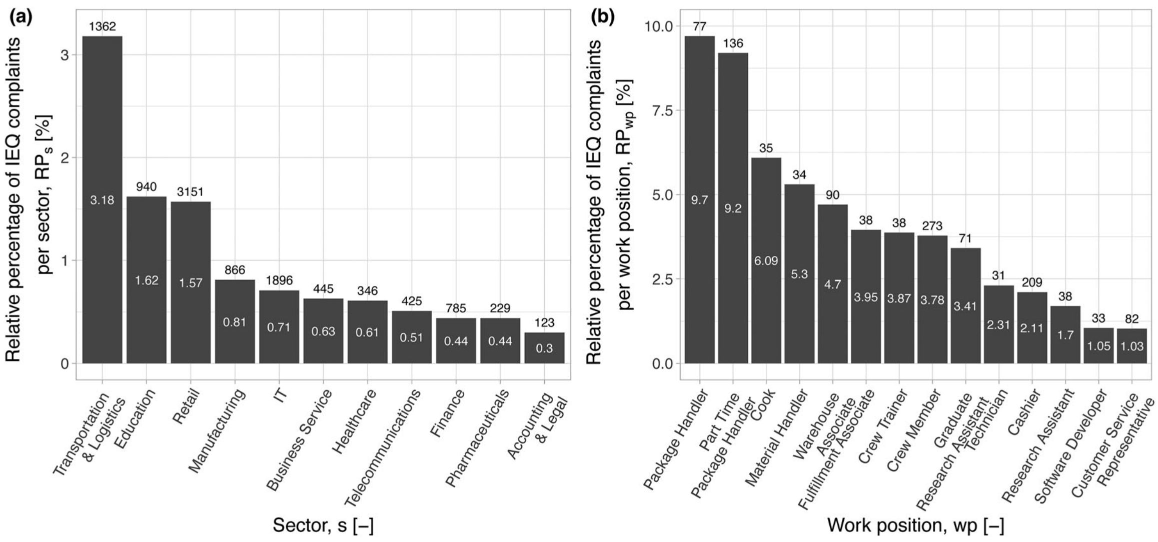


Figure 4. Relative percentage of IEQ complaints by (a) sector and (b) work position. The dark grey numbers on top of each bar indicate the total number of IEQ complaints in the considered group. Not all reported work positions are indicated in figure (b) but only those indicating at least 30 IEQ complaints and corresponding to at least 1% of the total reviews.

Source of IEQ complaints

IEQ complaints source relative distribution

Figure 5 summarizes the results of the sources of IEQ complaints with respect to the entire dataset, showing the average distribution of sources of complaints, independently of the workplace. The great majority of IEQ complaints refer to thermal aspects (54%), followed by acoustic (19%), IAQ (16%), visual (7%), and combined aspects (4%). Thermal complaints also result in the most reported ones in combination with other factors (thermal and IAQ and thermal and acoustic).

IEQ complaints source relative distribution by workplace

Figure 6(a) shows the relative percentage of sources of complaints among the IEQ aspects and their combination for each industry sector. The bars are ordered according to the number of IEQ complaints in each sector, from the most numerous (left) to the least numerous (right). The percentage of thermal complaints, the principal source of discomfort when the entire dataset is considered, results even higher for two sectors, namely transportation & logistics and education. A reduced complaint rate about thermal aspects can be observed for the finance (41%) and the pharmaceutical (39%) sectors. Together with the IT and telecommunication sectors, these sectors report almost a double rate of acoustic complaints compared to the overall trend (from 19% of the entire dataset to about 31–34%). IAQ complaints are larger than the average trend primarily for one sector, namely for retail. The large number of IEQ complaints in this sector and the large percentage referring to IAQ determine the large amount of IAQ complaints overall. Visual complaint rates are double for the pharmaceutical (13%), healthcare (13%) and the accounting and legal sectors (11%).

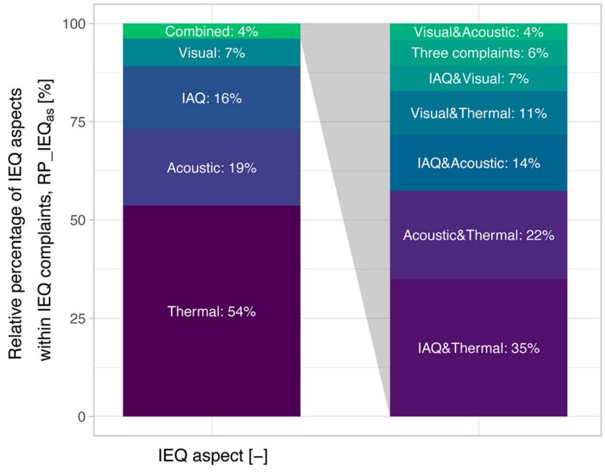


Figure 5. Relative percentage of sources of IEQ complaints according to IEQ aspects and their combination.

Figure 6(b) reports the relative percentage of sources of complaints among the IEQ aspects and their combination for each work position. The distribution of sources of IEQ complaints across the selected work positions reflects the one discussed for the different industry sectors. Work positions referring to the retail sector report a higher percentage of complaints related to IAQ. Work positions referring to the transportation and logistics sectors (i.e. warehouse associate, fulfilment associate, technician, and package handler) report a higher percentage of complaints related to the thermal

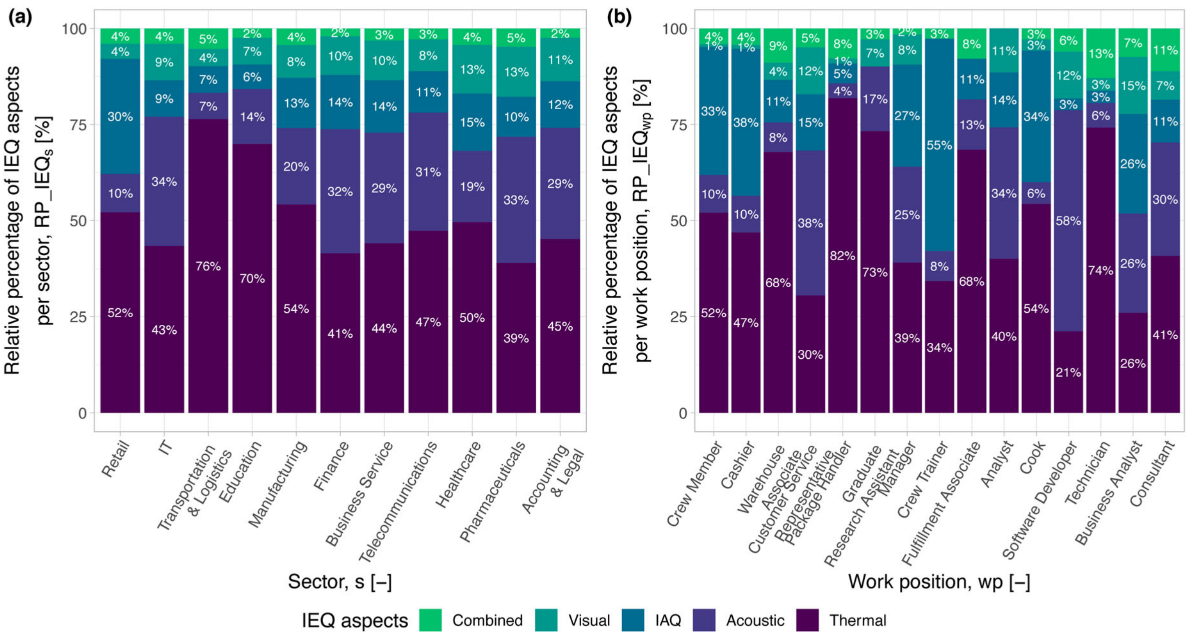


Figure 6. Relative percentage of IEQ aspects by (a) sector and (b) work position. The x-axis is ordered according to the number of IEQ complaints in each group, from the largest to the smallest. Not all work positions are displayed in the graph, but only those reporting at least 25 IEQ complaints.

environment. Another interesting result is the increasing rate of acoustic complaints for customer service representatives, analysts and software developers, and jobs conventionally conducted in offices.

The reviews referring to three sectors, namely education, retail, and transportation & logistics, are the ones that could skew the results about the source of complaints of the entire dataset due to their large number and their emphasis on a specific IEQ aspect. For this reason, the calculation of the source of complaints among the IEQ aspects is performed again by excluding one or more of these three sectors. The exclusion of the reviews from the education sector from the analysis does not compromise the overall distribution of the IEQ aspects, with the thermal complaints resulting in the most reported ones (52%), followed by acoustic (20%), IAQ (17%), visual (7%) and combined (4%). The same can be concluded with the complete exclusion of the transportation & logistics sector. By excluding both the retail and the transportation & logistics sectors from the analysis (the sectors that are not associated with a conventional office job), the resulting relative percentages vary, presenting a slightly reduced rate of thermal complaints (49%), an increased rate of acoustic complaints (27%), and an almost equal rate of complaints about IAQ and visual aspects (11% and 9%, respectively). This last distribution of IEQ sources of discomfort can, therefore, be associated with the office workplace.

Reasons for discomfort

Figure 7 illustrates the results of the term frequency-inverse document frequency statistic (*tf-idf*) for the analysis of the unigrams. The most common reasons for discomfort (i.e. the most characteristic words for each IEQ aspect) are ‘noise’, ‘quiet’ and ‘loud’ for acoustic, ‘smelly’, ‘dusty’ and ‘stuffy’ for IAQ, ‘hot’, ‘heat’ and ‘cold’ for thermal and ‘dark’ for visual (complaints about a windowless space are the second most common ones, together with those related to poor lighting). For the combined category, the most used words refer to the thermal (‘hot’ and ‘cold’) and IAQ aspects (‘dusty’ and ‘smelly’), a result that is consistent with the largest type of combination among IEQ aspects (Section ‘IEQ Complaints Source Relative Distribution’).

The calculation of the relative term frequency ($Rtf_{m,k}$) for the unigrams of each IEQ aspect (reported in Appendix B) confirms the results of the *tf-idf* statistic, validating the use of this method for this type of analysis. The relative term frequency also expands beyond the *tf-idf* statistic, indicating words that were not found with the *tf-idf* statistic as they are not unique to a specific category. For thermal discomfort, additional reasons include words related to the season and climate (‘summer’, ‘winter’, and ‘weather’), and the workplace (‘store’, ‘warehouse’, ‘office’, ‘truck’, and ‘kitchen’). For visual discomfort, it appears that the more considerable discomfort occurs in offices (and somewhat in cubicles) and such discomfort is linked to a lack of windows and natural light and bad lighting, resulting in

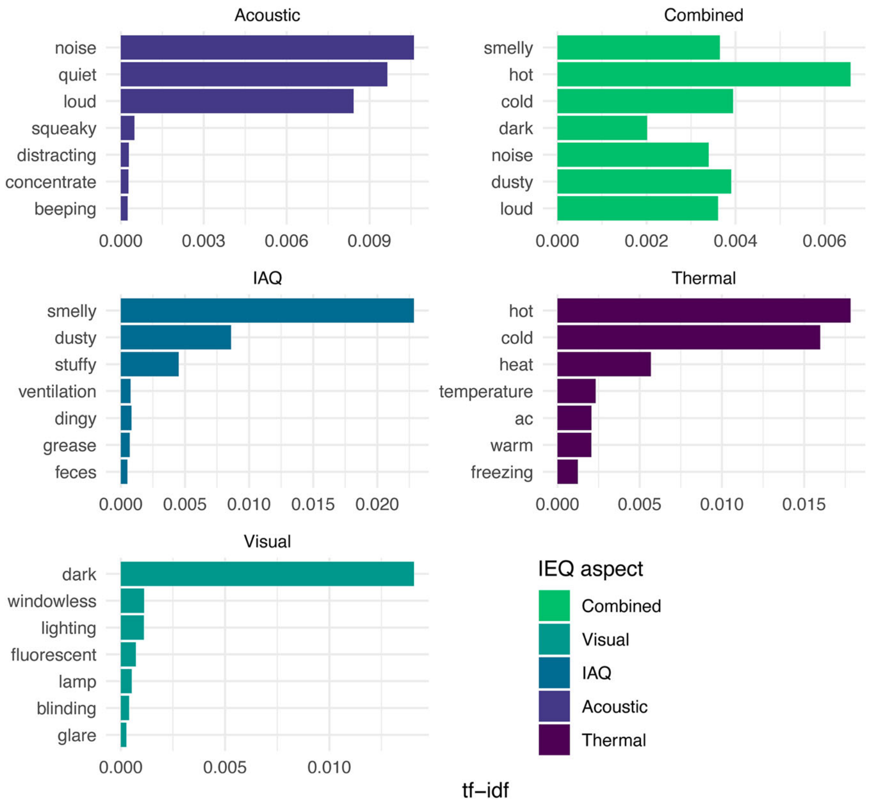


Figure 7. Term frequency-inverse document frequency for the unigrams derived from the tokenization of IEQ complaints, divided by IEQ aspects. For the creation of the graph, some of the tokens and their occurrences are merged (i.e. smells and smell merged into smelly, noisy merged into noise, temperatures into temperature, lights into lighting, and dust into dusty).

depressing environments. The additional words for IAQ discomfort are primarily linked to the workplace ('office', 'warehouse', 'store' and 'kitchen') and a lack of cleanliness ('dirty', 'bathrooms' and 'cleaning'). For acoustic discomfort, it results that the most significant discomfort arises in offices (mostly in cubicles) and is linked to people talking, having meetings and phone calls, and listening to music. The additional words for the combined category refer to the season and climate, similarly to thermal discomfort, and highlight how such combinations usually occur in offices, stores, and warehouses.

Figure 8 illustrates the results of the analysis for the bi-grams. In general, the results confirm those reported for the analysis of the unigrams. Only the bi-grams referring to visual complaints provide some additional insights, reporting the 'natural light' as one of the most present combinations of words in the visual category. It also confirms the complaints about the inadequate and unsatisfactory light conditions (especially the fluorescent lights).

Impact of IEQ complaints on overall job reviews

To understand the sentiment scores of the analyzed reviews, the sentiment scores of the cons reviews are compared with those of the pros reviews (which were not included in the analysis until now). Figure 9(a) shows the mean (grey diamond), median (thick black horizontal line), and the interquartile of the distribution of the sentiment scores of the cons reviews (mean = -0.63, sd = 3.37) and of the pros reviews (mean = 3.23, sd = 3.15). The sentiment scores of the two types of review are shown to be significantly different following the analysis performed with the Wilcoxon rank sum test ($W = 1.0273e + 11$, $p < 0.001$). This difference is also shown to be large, with an effect size of Cohen's d of 1.19.

Figure 9(b) illustrates the boxplots for the reviews without IEQ complaints and the reviews with IEQ complaints categorized according to the IEQ aspects. We reiterate the fact that the comments containing IEQ complaints are not reduced to retain the IEQ complaints only, but they also encompass other types of job complaints (e.g. salary, management, etc.). It is possible to see that most of the sentiment scores have negative values and that most of the reviews with IEQ complaints have a lower sentiment score than the reviews without IEQ complaints. The results of the Kruskal–Wallis test indicate a statistically significant difference between the review types

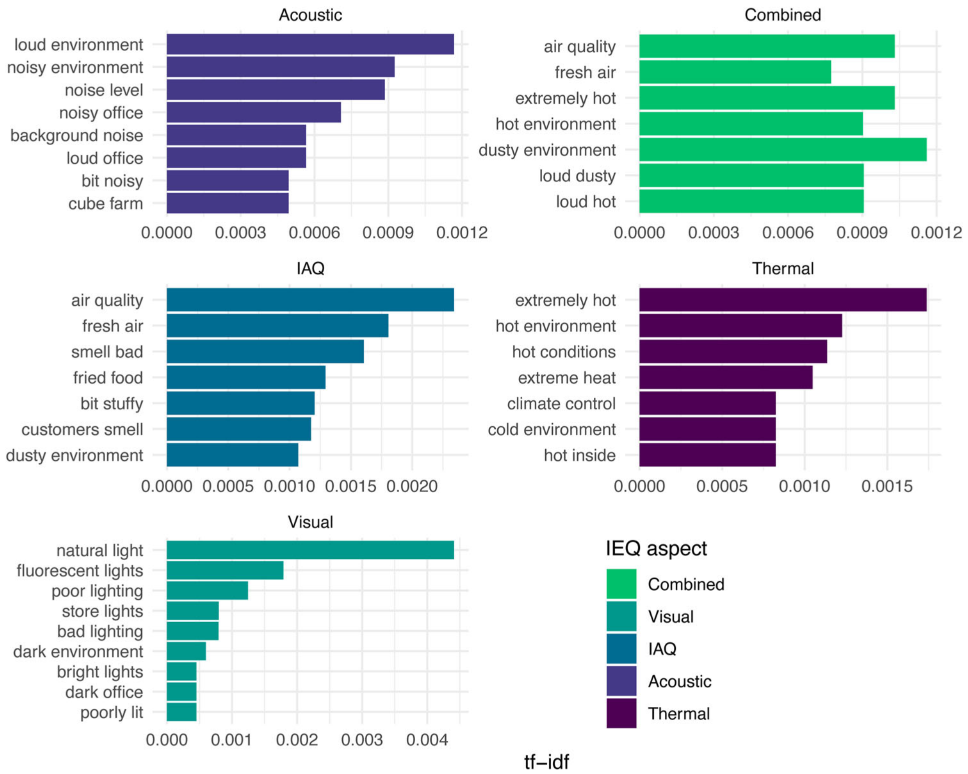


Figure 8. Term frequency-inverse document frequency for the bi-grams derived from the tokenization of IEQ complaints, divided by the IEQ aspect.

(chi-squared = 743.17, df = 5, $p < 0.001$). However, the small p-values are likely linked to the large sample size, considering the non-significant effect size ($\varepsilon^2 = 0.001$). Such a result implies that differences across all groups have no practical significance despite the large statistical significance of the performed test. However, the results of the Vargha and Delaney's A (VDA) indicate significant, albeit small, effect sizes for pairwise comparisons between reviews categorized according to the IEQ aspects. Table 2 illustrates the results of the VDA analysis and the results of the post-hoc Dunn test for multiple pairwise comparisons. The pairwise comparisons of both types of tests indicate statistical and practically significant differences between the sentiment scores of all the reviews containing IEQ complaints (for each IEQ aspects) and those of reviews without IEQ complaints. The most negative sentiment scores result for reviews reporting complaints for more than one IEQ aspects (the 'combined' category). It can be concluded that reviews with IEQ complaints result in lower sentiment scores than reviews without IEQ complaints, especially if more than one IEQ complaint is reported. The lower sentiment score of reviews reporting IEQ complaints could be correlated to a higher overall job dissatisfaction.

Discussion

This study presents an innovative analysis of occupants' feedback about the IEQ of their workplace. No comparable studies exist on the IEQ of the workplace; hence, results are discussed with reference to studies adopting a similar approach to investigate the IEQ in the hospitality sector and more traditional IEQ investigations performed with the use of conventional surveys in the workplace.

Indoor environmental quality complaints in online job reviews

This study indicates that less than 1% of the total number of scraped job reviews report IEQ complaints. The percentage is smaller than the average 4%–4.84% reported in similar studies in the hospitality sector (Qi et al., 2017; Villeneuve & O'Brien, 2020). The smaller percentage of IEQ complaints could be explained by the nature of the two types of reviews: hospitality reviews are primarily focused on the place of the stay and the service received, whereas job reviews are primarily focused on job-related aspects (e.g. salary, management, etc.). In an analysis of Glassdoor reviews, it has been shown that interest, management, and economic aspects are the most common topics in negative reviews

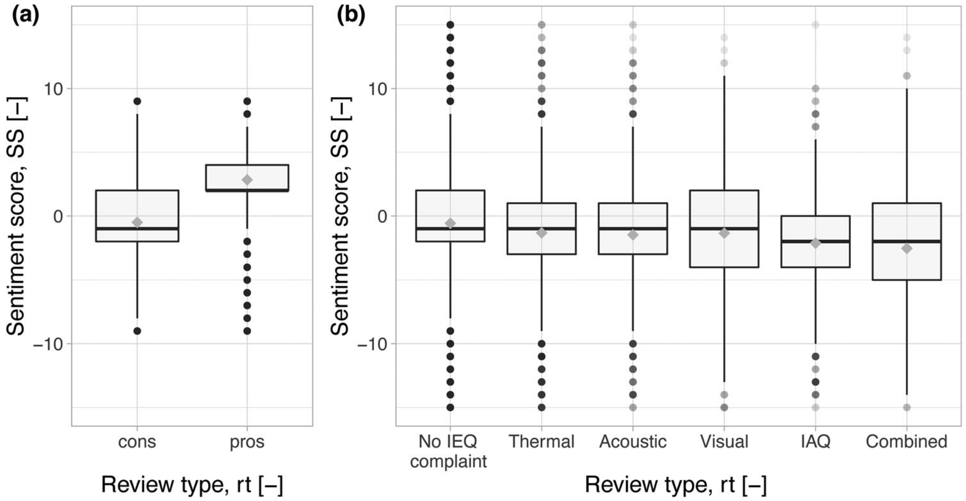


Figure 9. Sentiment score comparisons (a) for the cons and pros reviews, and (b) for the IEQ complaints according to the IEQ aspects and the reviews with no IEQ complaints.

(Dabirian et al., 2017). Consequently, IEQ aspects are not the principal focus of many job reviews. A complaint about the IEQ in a job review could be caused by prolonged exposure to uncomfortable indoor conditions, well beyond acceptable variations to which people can usually adapt (e.g. by wearing noise-cancelling headphones in a noisy environment or turning off disturbing lights). The dissatisfaction linked to uncomfortable indoor conditions exceeding the occupants' adaptive ability has been reported in the thermal comfort field (Li et al., 2020) but could also apply to the other comfort spheres.

This study also reports a decreasing absolute percentage of IEQ complaints over the years, which could have two interpretations: significant improvements in workplace conditions or an increasing number of concerns about other job aspects, such as job stability, compensation, and perspective. The improvement of the workplace conditions could be interpreted as increasing awareness of the importance and implications of good IEQ in the workplace or a direct response to the IEQ complaints by the organizations' facility management.

The results of the present study show that the number of IEQ complaints vary in the workplaces considered. A larger number of complaints are observed for jobs not conventionally performed in offices. This result highlights the importance of not limiting IEQ investigations to office spaces, currently the most commonly studied spaces within non-residential buildings (Li et al., 2018).

Sources of IEQ complaints: distribution, reason, and impact on overall job reviews

Thermal aspects are the most reported source of complaints, independently of the workplace. The number

Table 2. Vargha and Delaney's A (VDA) and Dunn test results of the pairwise comparisons across sentiment scores of different types of job reviews.

Pairwise comparison	Dunn test	VDA	VDA effect size interpretation
No IEQ complaint-Combined*	$p < 0.001$	0.343	small
No IEQ complaint-IAQ*	$p < 0.001$	0.365	small
Combined-Thermal*	$p < 0.001$	0.408	small
Combined-Visual*	$p < 0.001$	0.416	small
No IEQ complaint-Acoustic*	$p < 0.001$	0.417	small
IAQ-Thermal*	$p < 0.001$	0.438	small
IAQ-Visual	$p < 0.001$	0.445	n.s.
Combined-IAQ	n.s.	0.461	n.s.
Acoustic-Thermal	n.s.	0.486	n.s.
Acoustic-Visual	n.s.	0.488	n.s.
Thermal-Visual	n.s.	0.5	n.s.
IAQ-Acoustic	$p < 0.001$	0.549	n.s.
No IEQ complaint-Visual*	$p < 0.001$	0.561	small
No IEQ complaint-Thermal*	$p < 0.001$	0.568	small
Acoustic-Combined*	$p < 0.001$	0.579	Small

Note: The symbol * indicates that results are both statistically and practically significant. n.s. = non significant result.

of thermal complaints is even higher for people who do not have a traditional office job, such as those working in stores, warehouses, kitchen, or trucks. The predominance of complaints about the thermal environment confirms the findings of previous investigations based on traditional survey (Frontczak & Wargocki, 2011; Lai et al., 2009), and justifies the vast number of studies on thermal comfort (Mishra & Ramgopal, 2013; Mishra et al., 2016; Rupp et al., 2015). The methodology presented here confirms those findings across thousands of comments spanning several countries, years and workplaces, a result that would have been much more difficult to achieve with conventional surveys techniques.

Acoustic aspects represent the second most reported source of complaints. The acoustic complaint rate increases when reviews primarily associated with office jobs are retained. The results highlight that the acoustic disturbances (mainly from people talking to co-workers, having phone calls, or listening to music) are distracting and lead to a lower concentration level, confirming results found in conventional investigations performed in open-plan offices (Sundstrom et al., 1987; Sundstrom et al., 1994). Currently, acoustic conditions have been reported to be the least recorded environmental parameter in post-occupancy evaluation studies (Li et al., 2018), and consequently, the least investigated factor.

IAQ aspects are the third source of workplace complaints when all workplaces are considered. The percentage of IAQ complaints, however, double in the retail sector. The main reasons for IAQ discomfort are linked to the bad smell and unclean conditions of the workplace, more specifically of warehouses, and stores. This result highlights the importance of IAQ investigations in workplaces other than offices, such as in the retail sector. To date, the great majority of investigations on IAQ and workers are performed in office buildings (Al horr et al., 2016), primarily focusing on the sick building syndrome (Burge et al., 1987) and their associated effects and symptoms (Fisk & Rosenfeld, 1997; Wargocki et al., 2000).

Visual aspects represent the least reported source of complaints when all workplaces are considered. However, the percentage doubles for specific sectors that are generally associated with traditional office jobs. The lack of windows is reported as the principal reason for visual discomfort, followed by the poor electric lighting installations. Numerous investigations have highlighted the beneficial presence of windows and daylight in the built environment, contributing to occupants' health and well-being (Andersen, 2015). In workplaces, daylight has been reported as the preferred source of light (Galasiu & Veitch, 2006), and having access to a window or daylight has been shown to strongly improve satisfaction with lighting (Veitch et al., 2003). Also, daylight in schools has been correlated with increased student performance, a finding which may also hold true for adults in office buildings (Heschong et al., 2002). In addition to the daylight it provides, a window's view has also been shown to be beneficial for people (Aries et al., 2010; Heschong and Mahone, 2003). The lower percentage of visual complaints compared to the other IEQ aspects might be due to a lack of awareness of the impact of the visual conditions on people's health and productivity, or to a higher tolerance of people to poor visual conditions compared to thermal, acoustic and air quality ones.

Reviews with complaints about more than one IEQ aspect cover only 4% of the total IEQ complaints, with most of them referring to thermal aspects in combination with IAQ or acoustic aspects. The presence of IEQ complaints (independently of the source of complaint) results in reviews with lower sentiment scores than reviews without IEQ complaints. In addition, reviews reporting complaints about multiple IEQ aspects result in the most negative sentiment scores. This outcome highlights how multiple sources of discomfort can lead to a higher discomfort that, in turn, can potentially have a negative influence on the perception of the job itself (reflected in the low sentiment score of the job review). This result is in line with the literature indicating how satisfaction with the indoor environment correlates with job satisfaction (Veitch et al., 2007).

Practical implications

For the scientific community, the results of this study highlight important research directions and areas of focus for future IEQ investigations:

- Investigations beyond office environments are encouraged to improve the working conditions of people that suffer the most from inadequate IEQ. This specifically applies for investigations about thermal and IAQ factors;
- Given the importance of the acoustic environment in office spaces, acoustic measurements and investigations are strongly encouraged in future studies;
- Further studies should be conducted to evaluate the influence of uncomfortable indoor conditions on overall job satisfaction;
- Investigations on the simultaneous presence of multiple indoor environmental factors (Bluyssen, 2019; Schweiker et al., 2020) are strongly advised due to

the potential detrimental effect of multiple discomfort sources on job satisfaction;

The results also highlight the most critical IEQ aspects in several industry sectors and job positions, an outcome that can be of considerable value for companies, especially in sectors where IEQ complaints appeared to be most frequent.

The described method combining User-Generated Content from online job reviews and text-mining techniques can be exploited in several ways by many stakeholders. According to the four types of occupants' feedback applications illustrated for OVS (Sheikh Khan et al., 2020) (i.e. control, facility management, research, and social), the unconventional feedback from online reviews can serve all the applications except 'control' due to the nature of the feedback (i.e. non-continuous and asynchronous). For facility management applications, this type of occupants' feedback can be used to improve the operation of specific buildings. For research applications, it can be used as an alternative or an addition to feedback collected through conventional surveys and can be extremely valuable if studied in combination with environmental measurements collected from the investigated buildings. For social applications, it allows building occupants from all over the world to express their feedback about the IEQ of their workplace (unlike conventional surveys that are performed in selected buildings). To facilitate this latter application, the evaluation of the workplace could become one of the evaluations in company review websites such as Glassdoor.

Limitations

Despite the great potential for understanding the indoor environmental quality of different workplaces innovatively, the described analysis has some limitations, discussed hereafter.

First, the automated information extraction through the iterative cleaning process cannot provide all the correct information. Despite the use of the IEQ and incorrect vocabularies, it is believed that some unwanted reviews were included, and other relevant ones were missed in the final sample analyzed. This is because misspelled words cannot be captured as well as other more complex sentence structures and idioms. Only a visual screening and analysis of all the reviews would allow performing the correct selections, but this process cannot be automated. A manual assessment would not be efficient and discard the benefits of text-mining techniques. Second, the great majority of the reviews are located in English-speaking countries. A similar analysis of reviews posted in different languages and on different online platforms would be beneficial to extend the results of this paper. Third, this study only focuses on large firms and institutions. The same analysis could be performed for smaller organizations to investigate whether they offer a better or worse IEQ than bigger organizations. Fourth, this study does not consider the confounding effects that climate and location can have on the perception of the IEQ. Despite the fact that not all reviews include the location from which it is possible to deduce the climate and the country, additional analyses should be performed on this matter considering the reviews that indicate their location. Next, this study only considers comments posted online by employees and, due to its nature, cannot include physical environmental measurements of the considered workspace. As a consequence, this analysis only encompasses the employees' perception of their workplace. However, the results of occupant perception studies could initiate more detailed investigations involving physical measurements. Finally, the demographic of people posting online reviews could be restricted by age, and access and acquaintance to technology, potentially limiting the random distribution of samples. However, compared to conventional surveys usually limited in number, performed in specific buildings chosen by researchers or employers and mostly tied to specific closed-ended questions, the proposed methodology allows to gather a larger number of feedbacks coming voluntarily from all the people that wish to express their opinion online. Therefore, the chances of obtaining a random distribution of samples and the objectivity of the responses are higher than in conventional surveys. The increasing use of the internet and of online platforms will only increase these chances.

Conclusions

This study utilizes an innovative way to collect occupants' feedback about the indoor environmental quality (IEQ) of different workplaces by analyzing a large amount of voluntary and open-ended online job reviews. The value and novelty of this study stem from the analysis of a large number of comments and their reference to several workplaces, an effort that would have been extremely costly and time-consuming with conventional survey techniques.

This work presents two fundamental results:

- IEQ complaints mostly arise in workplaces that are not office buildings, especially regarding poor thermal and indoor air quality conditions in warehouses, stores, kitchens, and trucks;

- The reviews reporting IEQ complaints are more negative than reviews without IEQ complaints (despite they still report other negative aspects of the work environment, such as salary or management). Moreover, the most negative reviews are those reporting complaints about more than one IEQ aspect.

The first result challenges whether offices should be the primary focus of most IEQ investigations and highlights that the workplaces that need the most consideration to improve their IEQ have mostly been overlooked. Such workplaces, encompassing warehouses, stores, kitchens, and trucks are those in which most of the low-income jobs are conducted. This result opens up an equity problem as comfort currently appears to be a privilege of a restricted number of people rather than a right that should be guaranteed to all employees, independently of their salary and work position. The second result highlights the potential detrimental impact of uncomfortable indoor conditions on overall job satisfaction, especially when more than one uncomfortable condition from different IEQ aspects is reported. This outcome strengthens the importance of indoor environmental conditions in the workplace and how they might affect not only the well-being and productivity of people but also their overall evaluation of their job. In turn, this fact may foster improvements of indoor conditions by employers that wish to retain talents in their companies and create better working environments.

The method and the associated results presented in this study show that valuable information about the IEQ in the workplace can be extracted from User-Generated Content from online job reviews. This type of feedback can be considered cheaper, globally available and less biased than occupants' feedback conventionally collected with surveys. The use of the described method and the inclusion of additional data scraped from the web (e.g. positive comments, location and photos) could allow researchers, practitioners, and facility managers to investigate the indoor conditions in an innovative way and improve the indoor environmental conditions of millions of workplaces worldwide.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Giorgia Chinazzo http://orcid.org/0000-0003-0842-8493

References

Al Horr, Y., Arif, M., Kaushik, A., Mazroei, A., Katafygiotou, M., & Elsarrag, E. (2016). Occupant productivity and office indoor environment quality: A review of the literature. *Building and environment*, *105*, 369–389.

Alker, J., Malanca, M., Pottage, C., & O'Brien, R. (2014). Health, wellbeing & productivity in offices: The next chapter for green building. *World Green Building Council*.

Allen, J. G., & Macomber, J. D. (2020). *Healthy buildings: How indoor spaces drive performance and productivity*. Harvard University Press.

Alrawadieh, Z., & Law, R. (2019). Determinants of hotel guests' satisfaction from the perspective of online hotel reviewers. *International Journal of Culture, Tourism and Hospitality Research*, *13*(1), 84–97. https://doi.org/10.1108/IJCTHR-08-2018-0104

Andersen, M. (2015). Unweaving the human response in daylighting design. *Building and Environment*, *91*(September), 101–117. https://doi.org/10.1016/j.buildenv.2015.03.014

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, *49* (268), 765–769. https://doi.org/10.1080/01621459.1954.10501232

Aries, M. B. C., Veitch, J. A., & Newsham, G. R. (2010). Windows, view, and office characteristics predict physical and psychological discomfort. *Journal of Environmental Psychology*, *30*(4), 533–541. https://doi.org/10.1016/j.jenvp.2009.12.004

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, *25*(1), 1–24. https://doi.org/10.1080/19368623.2015.983631

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bluyssen, P. M. (2019). Towards an Integrated analysis of the indoor environmental factors and its effects on occupants. *Intelligent Buildings International*, *0*(0), 1–9. https://doi.org/10.1080/17508975.2019.1599318

Burge, S., Hedge, A., Wilson, S., Bass, J. H., & Robertson, A. (1987). Sick building syndrome: a study of 4373 office workers. *The Annals of Occupational Hygiene*, *31*(4A), 493–504. https://doi.org/10.1093/annhyg/31.4A.493

Heschong, L., & Mahone, D. (2003). Windows and offices: A study of office worker performance and the indoor environment. *California Energy Commission*, 1–5.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*(January), 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004

Clements-Croome, D. (2006). *Creating the productive workplace*. Taylor & Francis.

Dabirian, A., Kietzmann, J., & Diba, H. (2017). A great place to work!? Understanding crowdsourced employer branding. *Business Horizons*, Crowdsourcing, *60*(2), 197–205. https://doi.org/10.1016/j.bushor.2016.11.005

Day, J. K., & O'Brien, W. (2017). Oh behave! Survey stories and lessons learned from building occupants in high-performance buildings. *Energy Research & Social Science*, Narratives and Storytelling in Energy and Climate

Change Research, *31*(September), 11–20. https://doi.org/10.1016/j.erss.2017.05.037

Fisk, W. J., & Rosenfeld, A. H. (1997). Estimates of improved productivity and health from better indoor environments. *Indoor Air*, *7*(3), 158–172. https://doi.org/10.1111/j.1600-0668.1997.t01-1-00002.x

Frontczak, M., & Wargocki, P. (2011). Literature survey on how different factors influence human comfort in indoor environments. *Building and Environment*, *46*(4), 922–937. https://doi.org/10.1016/j.buildenv.2010.10.021

Galasiu, A. D., & Veitch, J. A. (2006). Occupant preferences and satisfaction with the luminous environment and control systems in daylit offices: A literature review. *Energy and Buildings*, Special Issue on Daylighting Buildings, *38* (7), 728–742. https://doi.org/10.1016/j.enbuild.2006.03.001

Glassdoor. n.d. About us. Glassdoor about us. Retrieved August 14, 2020, from https://www.glassdoor.com/about-us/

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. *Tourism Management*, *59*(April), 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Heschong, L., Wright, R. L., & Okura, S. (2002). Daylighting impacts on human performance in school. *Journal of the Illuminating Engineering Society*, *31*(2), 101–114. https://doi.org/10.1080/00994480.2002.10748396

Humphreys, M. A., & Nicol, J. F. (2007). Self-assessed productivity and the office environment: Monthly surveys in five European countries. *ASHRAE Transactions*, *113*, 606. https://search.proquest.com/scholarly-journals/self-assessed-productivity-office-environment/docview/192515114/se-2?accountid=147023

Jayathissa, P., Quintana, M., Abdelrahman, M., & Miller, C. (2020). Humans-as-a-sensor for buildings—intensive longitudinal indoor comfort models. *Buildings*, *10*(10), 174. https://doi.org/10.3390/buildings10100174

Jia, S. S. (2020). Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management*, *78*(June), 104071. https://doi.org/10.1016/j.tourman.2019.104071

Jung, Y., & Suh, Y. (2019). Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. *Decision Support Systems*, *123*(August), 113074. https://doi.org/10.1016/j.dss.2019.113074

Khalilzadeh, J., & Tasci, A. D. A. (2017). Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management*, *62*(October), 89–96. https://doi.org/10.1016/j.tourman.2017.03.026

Kim, B., Kim, S., & Heo, C. Y. (2016). Analysis of satisfiers and dissatisfiers in online hotel reviews on social media. *International Journal of Contemporary Hospitality Management*, *28*(9), 1915–1936. https://doi.org/10.1108/IJCHM-04-2015-0177

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*(December), 128–147. https://doi.org/10.1016/j.knosys.2016.10.003

Lai, A. C. K., Mui, K. W., Wong, L. T., & Law, L. Y. (2009). An evaluation model for indoor environmental quality (IEQ) acceptance in residential buildings. *Energy and Buildings*, *41*(9), 930–936. https://doi.org/10.1016/j.enbuild.2009.03.016

Lee, M., Jeong, M., & Lee, J. (2017). Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. *International Journal of Contemporary Hospitality Management*, *29*(2), 762–783. https://doi.org/10.1108/IJCHM-10-2015-0626

Levene, H. (1949). On a matching problem arising in genetics. *Annals of Mathematical Statistics*, *20*(1), 91–94. https://doi.org/10.1214/aoms/1177730093

Li, P., Froese, T. M., & Brager, G. (2018). Post-occupancy evaluation: State-of-the-Art analysis and state-of-the-practice review. *Building and Environment*, *133*(April), 187–202. https://doi.org/10.1016/j.buildenv.2018.02.024

Li, P., Parkinson, T., Schiavon, S., Froese, T. M., de Dear, R., Rysanek, A., & Staub-French, S. (2020). Improved long-term thermal comfort indices for continuous monitoring. *Energy and Buildings*, *224*(October), 110270. https://doi.org/10.1016/j.enbuild.2020.110270

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: Large samples and the *p*-value problem. *Information Systems Research*, *24*(4), 906–917. https://doi.org/10.1287/isre.2013.0480

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16*(1), 100–103. https://doi.org/10.1017/S1351324909005129

Mishra, A. K., Loomans, M. G. L. C., & Hensen, J. L. M. (2016). Thermal comfort of heterogeneous and dynamic indoor conditions—an overview. *Building and Environment*, *109*, 82–100. https://doi.org/10.1016/j.buildenv.2016.09.016

Mishra, A. K., & Ramgopal, M. (2013). Field studies on human thermal comfort—an overview. *Building and Environment*, *64*(June), 94–106. https://doi.org/10.1016/j.buildenv.2013.02.015

Moezzi, M., & Goins, J. (2011). Text mining for occupant perspectives on the physical workplace. *Building Research & Information*, *39*(2), 169–182. https://doi.org/10.1080/09613218.2011.556008

Moro, S., Ramos, R. F., & Rita, P. (2020). What drives job satisfaction in IT companies? *International Journal of Productivity and Performance Management*, *70*(2), 391–407. https://doi.org/10.1108/IJPPM-03-2019-0124

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251. https://doi.org/10.1016/j.eswa.2013.01.019

Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *ArXiv Preprint ArXiv:1103*.2903.

Ortiz, M. A., & Bluyssen, P. M. (2019). Developing home occupant archetypes: First results of mixed-methods study to understand occupant comfort behaviours and energy use in homes. *Building and Environment*, *163* (October), 106331. https://doi.org/10.1016/j.buildenv.2019.106331

Qi, M., Li, X., Zhu, E., & Shi, Y. (2017). Evaluation of perceived indoor environmental quality of five-star Hotels in China: An application of online review analysis. *Building and Environment*, *111*(January), 1–9. https://doi.org/10.1016/j.buildenv.2016.09.027

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

RStudio Team. (2020). *RStudio: Integrated development environment for R*. RStudio, PBC. http://www.rstudio.com/

Rupp, R. F., Vásquez, N. G., & Lamberts, R. (2015). A review of human thermal comfort in the built environment. *Energy and Buildings*, *105*(October), 178–205. https://doi.org/10.1016/j.enbuild.2015.07.047

Schweiker, M., Ampatzi, E., Andargie, M. S., Andersen, R. K., Azar, E., Barthelmes, V. M., Berger, C., Bourikas, L., Carlucci, S., Chinazzo, G., Edappilly, L. P., Favero, M., Gauthier, S., Jamrozik, A., Kane, M., Mahdavi, A., Piselli, C., Pisello, A. L., Roetzel, A., … Zhang, S. (2020). Review of multi-domain approaches to indoor environmental perception and behaviour. *Building and Environment*, *176*(June), 106804. https://doi.org/10.1016/j.buildenv.2020.106804

Sheikh Khan, D., Kolarik, J., & Weitzmann, P. (2020). Design and application of occupant voting systems for collecting occupant feedback on indoor environmental quality of buildings – a review. *Building and Environment*, *183*(October), 107192. https://doi.org/10.1016/j.buildenv.2020.107192

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.

Sundstrom, E., Stokols, D., & Altman, I. (1987). *Handbook of environmental psychology*. Work Environments: Offices and Factories, USA: Wiley.

Sundstrom, E., Town, J. P., Rice, R. W., Osborn, D. P., & Brill, M. (1994). Office noise, satisfaction, and performance. *Environment and Behavior*, *26*(2), 195–222. https://doi.org/10.1177/001391659402600204

Sutherland, I., & Kiatkawsin, K. (2020). Determinants of Guest experience in Airbnb: A topic modeling approach using LDA. *Sustainability*, *12*(8), 3402. https://doi.org/10.3390/su12083402

Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020). Topic modeling of online accommodation reviews via latent Dirichlet allocation. *Sustainability*, *12*(5), 1821. https://doi.org/10.3390/su12051821

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size Statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. https://doi.org/10.3102/10769986025002101

Veitch, J. A., Charles, K. E., Farley, K. M. J., & Newsham, G. R. (2007). A model of satisfaction with open-plan office conditions: COPE field findings. *Journal of Environmental Psychology*, *27*(3), 177–189. https://doi.org/10.1016/j.jenvp.2007.04.002

Veitch, J. A., Charles, K. E., Newsham, G. R., Marquardt, C. J., & Geerts, J. (2003). Environmental satisfaction in open-plan environments: 5. Workstation and physical condition effects. *NRC Institute for Research in Construction*. Retrieved May 13, 2014, from https://doi.org/10.4224/20378817

Villeneuve, H., & O'Brien, W. (2020). Listen to the guests: Text-mining Airbnb reviews to explore indoor environmental quality. *Building and Environment*, *169*(February), 106555. https://doi.org/10.1016/j.buildenv.2019.106555

Wargocki, P., Wyon, D. P., Sundell, J., Clausen, G., & Ole Fanger, P. (2000). The effects of outdoor air supply rate in an office on perceived air quality, sick building syndrome (SBS) symptoms and productivity. *Indoor Air*, *10*(4), 222–236. https://doi.org/10.1034/j.1600-0668.2000.010004222.x

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag. www.springer.com/us/book/9780387981413

Wong, C. U. I., & Qi, S. (2017). Tracking the evolution of a destination's image by text-mining online reviews – the case of Macau. *Tourism Management Perspectives*, *23* (July), 19–29. https://doi.org/10.1016/j.tmp.2017.03.009

Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management*, *55*(May), 57–69. https://doi.org/10.1016/j.ijhm.2016.03.003

Zagal, J. P., Tomuro, N., & Shepitsen, A. (2012). Natural language processing in game studies research: An overview. *Simulation & Gaming*, *43*(3), 356–373. https://doi.org/10.1177/1046878111422560

Zar, J. H. (2013). *Biostatistical analysis: Pearson new international edition*. Pearson Higher Ed.

Zhang, J. (2019). What's yours Is mine: Exploring customer voice on Airbnb using text-mining approaches. *Journal of Consumer Marketing*, *36*(5), 655–665. https://doi.org/10.1108/JCM-02-2018-2581

Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big Data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, *76*(January), 111–121. https://doi.org/10.1016/j.ijhm.2018.03.017