City and Environment Interactions 10 (2021) 100058

Contents lists available at ScienceDirect

City and Environment Interactions

journal homepage: www.elsevier.com/locate/carcint

# Leveraging artificial intelligence to analyze citizens' opinions on urban green space ☆

Mohammadhossein Ghahramani [a], Nadina J. Galle [a,b,*], Fábio Duarte [b,c], Carlo Ratti [b], Francesco Pilla [a]

[a] Spatial Dynamics Lab, College of Engineering and Architecture, University College Dublin, Ireland
[b] Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA.
[c] Pontifícia Universidade Católica do Paraná, Curitiba, Brazil

ARTICLE INFO

*Keywords:*
Sentiment analysis
Supervised learning
Natural language processing
Imbalanced classification

ABSTRACT

Continued population growth and urbanization is shifting research to consider the quality of urban green space over the quantity of these parks, woods, and wetlands. The quality of urban green space has been hitherto measured by expert assessments, including in-situ observations, surveys, and remote sensing analyses. Location data platforms, such as TripAdvisor, can provide people's opinion on many destinations and experiences, including UGS. This paper leverages Artificial Intelligence techniques for opinion mining and text classification using such platform's reviews as a novel approach to urban green space quality assessments. Natural Language Processing is used to analyze contextual information given supervised scores of words by implementing computational analysis. Such an application can support local authorities and stakeholders in their understanding of–and justification for–future investments in urban green space.

## 1. Introduction

Urban Green Space (UGS) such as parks, woods, and wetlands represent a fundamental component of any urban ecosystem. In addition to the many ecological, economic, and psychological benefits, since the 1800s, UGS have been recognized for their ability to offer refuge from pervasive air pollution, and congestion [2,10,43]. Today, the ecological benefits of green in the city are well-documented, but there is also a growing body of evidence of its positive impact on human health and well-being [43]. Green space offers citizens more opportunities for social contact and stress relief - whether impromptu or planned [48]. Studies show UGS should be of critical importance to public mental health, especially from an urban planning perspective [21,40].

For many citizens, UGS has become an extension of, or in many cases the replacement of, the traditional backyard, meaning more people are sharing less green space. Despite appeals for green space's place in the city's master plans and worldwide urban population growth, UGS has decreased in several cities [13,19,23,33]. Lucrative urban development and construction are often to blame for its demise. To meet demand, studies have suggested the quality of green space significantly contributes to neighborhood satisfaction and well-being, independent of the quantity of green space [48]. How to measure the quality of UGS has been hotly debated in urban forestry and planning fields, with several attempts made to streamline and standardize quality assessments of UGS [6,8,18,20]. However, with current methods relying on expert assessments, some warn it discredits the experience of local users; who are likely more qualified to assess their own UGS than outside experts [25].

The definition of quality UGS is still contested, and their role remains undervalued. Measuring UGS quality is also a tedious process; observational techniques are often criticized as they require extensive repeat measurements at the same location, incurring large time and cost expenditures [39]. Even when data is collected, it is quickly outdated, leaving progress out of reach [36].

How can a city ensure it provides safe, inclusive, high-quality UGS for all? Emerging technologies are gaining traction as a way to gain up-to-date information on—and engage local users in—the planning and improvement of UGS [14,32,17]. There are several quantitative approaches to analyzing UGS such as drones [31], satellite imagery [11], and Google Street View images [42], but there is a need for reproducible qualitative analysis. One such technology, Natural Language Processing (NLP), combines computer science and linguistics

☆ This paper is the result of a research project funded by Connecting Nature (Grant Agreement No. 730222) under the European Community's Framework Program Horizon 2020 and the Netherlands Fulbright Association through the Fulbright U.S. Doctoral Student Program. This research was made possible by the support of the Senseable City Laboratory at the Massachusetts Institute of Technology, which hosted Nadina Galle during her fieldwork.

* Corresponding author.
*E-mail addresses:* sepehr.ghahramani@ucd.ie (M. Ghahramani), nadina.galle@ucd.ie (N.J. Galle), fduarte@mit.edu (F. Duarte), ratti@mit.edu (C. Ratti), francesco.pilla@ucd.ie (F. Pilla).

https://doi.org/10.1016/j.cacint.2021.100058
Available online 4 February 20212590-2520/

to understand the language in a piece of text. One of its applications, Sentiment Analysis (SA), can extract and categorize positive, negative, or neutral sentiment from a chunk of text. While NLP can work on any written text, performing SA on georeferenced crowd-sourced data sources such as TripAdvisor, Twitter, Yelp, Booking.com, and Airbnb have shown particular promise [5,36,26]. Applications range from understanding consumers' attitudes toward their products to the socioeconomic status of communities to hospitality organizations' performance [29,30,37]. It has been suggested that in most of these applications, sentiment analysis should become a complementary tool for quality assessment and evaluation [15].

TripAdvisor is a particularly popular platform with a rich and publicly accessible database on attractions, destinations, and landmarks, including UGS [29]. While relevant to this research, studies of the demographic makeup of TripAdvisor are limited. Some groups are likely over- and/or underrepresented on TripAdvisor, but it is still advantageous over population-based surveys, a costly and tedious method to acquire representative population samples. TripAdvisor offers a viable, complementary method to harvest local opinion and feedback on UGS.

This paper presents a novel NLP application using TripAdvisor to assess the quality of UGS. The corpus collected were TripAdvisor reviews of St. Stephen's Green, the most popular public park in Dublin (Ireland). St. Stephen's Green, in the middle of Dublin's city center, is a 10-hectare park with over eight million visitors on an annual basis. The park has well-maintained facilities on the grounds, including over 3 km of walking paths and public restrooms inside the park.

Experimental, computational analyses were implemented via two scenarios, and different phases have been included to address identifying the sentiment expressed in reviews. The proposed method allows the extraction and interpretation of sentiment with minimal human effort by applying Artificial Intelligence (AI) and Machine Learning (ML) algorithms. The contributions of this work are as follows:

- We present a novel application of NLP and text mining using TripAdvisor to assess the quality of UGS.
- We present how a self-contained sentiment analysis model can be implemented to evaluate people's attitudes toward various entities given a class imbalanced issue.

The paper is organised as follows: some related work in the field of sentiment analysis and opinion mining is presented in Section 2; the proposed approach with its associated discussions is presented in Section 3; Section 4 shows the experimental results; Section 5 details the discussion; and Section 6 concludes the paper.

## 2. Related work

The value of UGS remains underestimated due to a lack of information about what quality green space entails and how existing spaces within the city score on important social-quality parameters. Measuring the quality of UGS is a tedious process. Observational approaches are often criticized as they require extensive repeat measurements at the same location, incurring large time and cost expenditures [6,39]. Even when data is collected, it is quickly outdated, leaving progress out of reach.

### 2.1. Web-based civic participation platforms

A recent improvement is web-based civic participation platforms. In an effort to gain insights into how people perceive a park's quality, several cities have released apps. Amsterdam recently launched "MyPark", an app that asks the user questions about specific areas of a specific park. Once the results are analyzed, the feedback is incorporated into a park redesign to better meet local user needs.

FixMyStreet is another example. The map-based app acts as a liaison between residents and their local authority on problems such as potholes and broken street lights needing their attention. The app was launched in the United Kingdom (UK) and has proliferated across the country. FixMyStreet also has an open-source platform that helps people run similar websites all over the world. Although mainly used for reporting common street problems, the app could also be used to highlight issues facing urban parks and woodlands.

### 2.2. Microblogging platforms

Microblogging platforms challenge users to summarise their thoughts in a limited amount of words. Twitter, arguably the world's microblogging pioneer, allows 280 characters per post (or "tweet"), a recent upgrade from the iconic 140 characters they used to enforce. In her inspiring paper [39], Roberts proposes the use of crowdsourced, geotagged social media data, such as tweets, to inform how, when, and why people use UGS. This method overcomes some issues with previous approaches, such as report based methods, which are difficult to validate, and observational methods, which require multiple observations over different days and seasons to ensure reliability [39]. It can even be used to derive seasonal variation in physical activity in UGS [39]. Crowdsourcing data from Twitter offers an alternative as it is publicly available and instantly accessible, incurring no additional time or costs.

Yet, both web-based civic participation platforms and social media data face limitations. FixMyStreet, with over 12,000 reports sent to UK councils every month, has much less usability than Twitter, which recorded 17 million monthly active British users in the first quarter of 2018. It is unlikely any significant amount of these tweets are actually about issues on the street, but it is a much broader data source.

There are also socio-demographic concerns regarding the user base of both FixMyStreet and Twitter. In 2017, [34] analyzed over 30,000 FixMyStreet reports, compared them to a range of socio-demographic indicators, and revealed crowdsourced civic participation platforms tend to marginalize low-income and ethnically diverse communities.

In the same way, the elderly population, who show lower levels of engagement with these forms of technology, are disregarded explicitly in such research [4]. This is especially concerning as urban parks are supposed to be a shared public space for all ages. Roberts also reports Twitter data lacks demographic information about Twitter users such as their age, occupation, or ethnicity [39]. Although not crucial to determine opinions, these parameters are useful for further examination of where particular attitudes may originate. Evidently, there is a need for inclusive, unrestricted, unbiased, and freely-solicited opinions about UGS.

### 2.3. TripAdvisor

NLP is used to understand the language in a piece of text and reveal the sentiment behind it. The method combines computer science and linguistics. In recent years, the popularity of virtual assistants like Siri, Alexa, and Google Home has accelerated the demand for voice user interfaces. And, as such, increased research on how computers understand speech and speak themselves.

NLP can also work on written text, like user-generated reviews on the world's largest travel website, TripAdvisor. The open, online community reaches 390 million unique visitors each month and lists 465 million reviews and opinions about more than seven million accommodations, restaurants, and attractions in 49 markets worldwide. TripAdvisor is a treasure trove of sentiment. When writing a review, a reviewer is prompted to describe their first-hand experience causing both tourists and locals to flock to TripAdvisor to express their opinions. Whereas Twitter offers a platform for sharing an occasional opinion, TripAdvisor explicitly asks for the sentiment.

The overall user base of both TripAdvisor and Twitter is still poorly understood. In 2007, Gretzel [22] found frequent travel review readers tend to be younger, have slightly higher incomes, and are more likely to contribute to online content. They are also more likely to post reviews themselves. Thus, one could assume the reviewers share a similar demographic to the reader, at least in 2007.

Gender differences can also play a role. Blumenthal [3] found little to no gender differences amongst reviewers on TripAdvisor in 2014. Twitter, on the other hand, did exhibit gender differences. According to Statista, an online statistics, market research, and business intelligence portal, during a 2018 study period, 42.8 percent of global Twitter users were female, and 57.2 percent were male (Global Twitter User Distribution by Gender 2018 — Statistic, n.d.).

The use of Twitter tends to drop as age increases. In the United States, those under 50, especially those 18–29, are most likely to use Twitter. And only 6 percent of Twitter users constitute the 65+ age group. TripAdvisor's TripBarometer report showed a slightly more even distribution of the travel site's user base.

Research to validate these demographic claims is limited, and studies comparing TripAdvisor with Twitter's user base are non-existent. Although some groups remain over- and/or underrepresented on TripAdvisor, it is advantageous over Twitter as it generally covers a broader demographic spectrum. In fact, the only known method to encompass a general population is population-based surveys, where an experiment is administered to a representative population sample. However, this process is costly and lengthy, and as such, TripAdvisor offers a viable, complementary method to harvest local opinion and feedback on UGS.

So far, sentiment analysis using TripAdvisor as a data source has only been applied in the hospitality and tourism sectors. Here, shallow NLP techniques are applied to extract sentiment [15] automatically. These simple expressions, which are derived from the reviews, can be used to evaluate the quality of hotels or restaurants. García-Barrio canal's preliminary study [15] was able to identify emotion types with reasonable effectiveness and suggested sentiment analysis using TripAdvisor reviews should become a complementary tool for hospitality evaluation.

### 2.4. Supervised text classification

Sentiment classification is an example of a supervised machine learning task, a process of assigning text documents into two or more predefined classes. In this process, an algorithm takes any observation (text document) as input and assigns a label from the class labels [16,24]. Different data-driven supervised approaches have been used to deal with such a classification problem. Sentiment classification has raised much attention in recent years and has undergone many changes. Generally speaking, three techniques can be used to construct a sentiment lexicon, i.e., dictionary-based, corpus-based, and hybrid methods. Dictionary-based methods use word matching based on the lexicon. However, since sentiment words in the lexicon might be difficult to recognize, many texts cannot be analyzed by utilizing such classifiers. Corpus-based methods use labeled data, and lexicons are not effectively taken into account in such approaches. To alleviate the discussed shortcomings, a hybrid approach (i.e., a combination of machine learning methods and lexicons) can help improve the sentiment classification performance. Since the text classification problem is a supervised learning task in which the class observations is predicted based on some feature values, a wide range of ML algorithms (e.g., Support Vector Machine (SVM) [12,1], Naive Bayes (NB) [12,38], decision tree [12], random forest [12,1], logistic regression, and neural networks [27,47,9,7]) can be incorporated.

As explained, in this work, people reviews as to UGS are taken into account. These texts are unstructured; thus, manually analyzing them can be tedious and time-consuming. In this type of data mining, people's opinions, sentiments, and attitudes are analyzed. The main objective is to computerize the process of reading reviews and evaluate them. It should be mentioned that the most crucial task in sentiment analysis is the pre-processing phase, including different operations. Due to differences in data characteristics, these tasks might differ from one sentiment analysis approach to another. Because of the complexity of feature dependency, ML methods may achieve different results. Given this work's characteristics, we aim to propose an appropriate approach to deal with various issues to be explained next. It is worth mentioning that a self-contained model consisting of multiple phases is implemented in this paper. In the last stage of the model (the classification phase), different ML algorithms are tested, and their corresponding results are compared.

## 3. Method

There have been different approaches to perform sentiment analysis. However, choosing a proper method is highly related to the nature of a given work. This paper analyzes people's opinions and sentiments to identify different positive and negative polarities on urban green space. Different from book articles and news reports, review texts are often short and ambiguous. Various models, i.e., fully supervised and semi-supervised methods, have been considered to analyze review comments. The methods in the former category use manually labeled data. Their approach is very time-consuming to create lexicons manually. Some specific supervised methods have been introduced to train sentiment classifiers on emoticons and hashtags. Because of such shortcomings, a semi-supervised model has been considered in this paper. It should be noted that the classification phase of the model is based on a supervised technique, while an unsupervised method is used in the pre-processing phase of the model. Most of the concerns related to opinion mining and sentiment analysis of reviews can be addressed by implementing effective pre-processing techniques. However, there are no effective pre-processing methods for all datasets and algorithms. For instance, in this work, we deal with an imbalanced classification issue since most comments in the dataset used are positive. A multi-layer approach consisting of different phases (i.e., web scrapping, data cleanings, imbalanced classification, and supervised ML) is implemented to address all concerns. Fig. 1 illustrates different phases of the proposed model. The following sections present all details regarding each stage of the model.

### 3.1. Data extracting

TripAdvisor reviews for St. Stephen's Green (Dublin, Ireland) were scarped using Selenium and Python. The pseudocode is presented in Algorithm 1. The reviews were subsequently processed to focus on English texts, because of the mass availability of English language text analysis tools and dictionaries. The reviews were collected from the period of May, 2006 to November, 2020, for a total of 16,613 reviews; in contrast, Dublin's second most popular park had 4,753 reviews for this time period. Of the St. Stephen's Green reviewers, 5,622 were from the United States, 3491 were from Ireland, 2,835 from the United Kingdom (UK), 796 from Canada, 483 from Australia, 105 from Germany, 92 from the Netherlands, 91 from South Africa, 48 from New Zealand, 43 from Denmark, 39 from Greece, 32 from United Arab Emirates, and 30 from China. The remaining 2,906 reviewers were from other countries, had misspelled their country, or left their location blank.

**Algorithm 1.** Pseudo-code for extracting Tripadvisor reviews

**Input :** $CSV\,file \leftarrow$ [Score, Date, Title, Review]; $driver \leftarrow$ webdriver.Chrome();
$Fn \leftarrow$ Selenium.find_element_by_xpath; $URL \leftarrow$ https://www.tripadvisor.ie/*.st_stephens_green
**Output:** Review Comments

```
# Exception Function
try:
  driver.find_element_by_xpath(xpath)
except NoSuchElementException:
  return False
return True
n = number of web pages;
# HTML elements
e0= 'taLnk ulBlueLinks'; e1= 'ui_bubble_rating bubble_'; e2= '_34Xs-BQm'; e3= 'glasR4aX'; e4= 'IRsGHoPm';
e5= 'Dq9MAugU T870kzTX LnVzGwUB';
for i ← 1, 2, ...., n do
    if ExceptionFunction(''//span[@class = e0]'') then
        driver.Fn(''//span[@class = e0]'').click();
    end
    df ← driver.Fn(''//div[@class = e5]'');
    num ← len(df);
    for j ← 1, 2, ..., num do
        Score = df[j].Fn(''.//span[contains(@class, e1)]'').get_attribute(''class'').split(''_'')[3]);
        Date = df[j].Fn(''.//span[@class = e2]'').get_attribute(''title'');
        Title = df[j].Fn(''.//div[@class = e3]'').text;
        Review = df[j].Fn(''.//q[@class = e4]'').text.replace(''\n'','' '');
    end
end
Return: Score, Date, Title, Review.
```

The following review fields were extracted: review-title (written title of review); review-body (written review about the destination); rate-value (1 is the lowest evaluation, 5 is the best); review-location (where a reviewer is from); and review-date (date review was written). Only review-body and rate-value data fields were used in this experiment. This dataset can be considered as a sequences of text, i.e., $D = \{X1, X2, ..., Xn\}$ where $X_i$ refers to the $i^{th}$ review. Each review is also labeled as positive or negative, depending on its corresponding rate value.

### 3.2. Pre-processing

As the quality of data affects the analysis, it is essential to employ a data pre-processing procedure. To that end, feature extraction was performed, and a structured set from the reviews is created for the model-training purposes. A dimensionality reduction operation is also considered by applying the Term Frequency-Inverse Document Frequency (TF-IDF) technique [45]. These pre-processing steps help us convert unstructured text sequences into a structured feature space. Data cleansing operations were performed, and punctuations and stop words were omitted. To make transformations (removing punctuations, stop words, and other cleansing operations) implemented in this work, libraries from the Natural Language Toolkit (NLTK) were used. This Python library has been written for modeling text and provides various tools for loading and cleaning texts. This library's different functions were used for filtering punctuation, stemming, normalizing, extracting text from HTML, decoding Unicode characters, locating typos, and handling numbers.

Data normalization techniques (e.g., Stemming and Lemmatisation) were applied, each review was converted to a numeric representation (corpus), and the *n*-grams approach (with two different measures like Word Counts and TF-IDF) was implemented. The former is based on mapping more than just one word (unigrams) onto the corpus. We have also included word counts into our model. To that end, the number of times a given word or a sequence of words appear is counted. The latter, term TF-IDF, is a weighting measure to be used instead of word count representations. This measure is considered to lessen the effect of implicitly common words in the corpus. The weight of a term in a review can be defined as:

$$w(r,t) = TF(r,t) * log(\frac{n}{df(t)}) \quad (1)$$

where $n$ is the number of reviews and $df(t)$ is the number of reviews consisting of the term $t$ in the corpus.

Negation handling could be another challenging task for sentiment classification. However, since we deal with a two-class classification task, such concern can be easily addressed. The model negates the predicted class of observations, as there are only two classes to choose from. Such negation recognition can be a complicated process in cases where there are more than two possible classes. In our case, the negation handling procedure is considered as an Exclusive-OR problem. As far as the negation scope detection is concerned, different negation keywords are defined, and the regular expression-based NegEx method is used. Moreover, the negation implicitly is captured via n-grams.

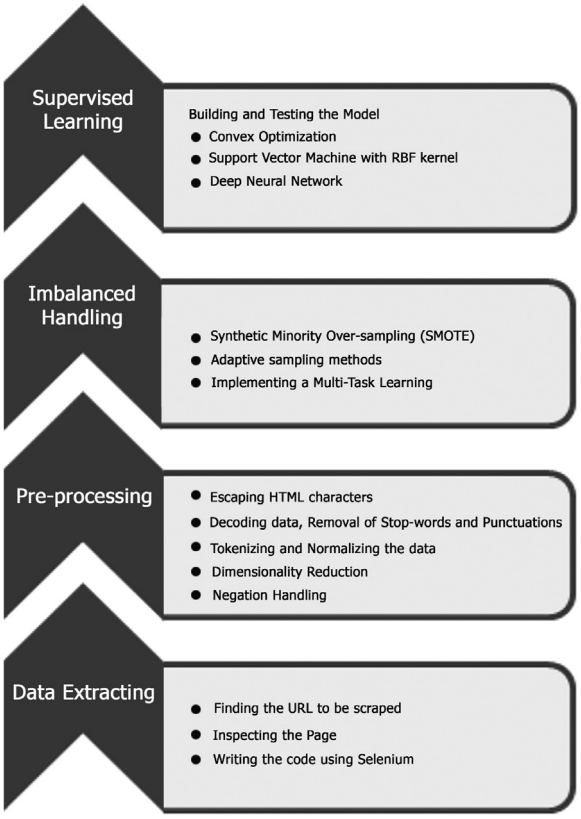


**Fig. 1.** Four phases of the sentiment analysis model used in this work.

Although the dataset has been cleaned after performing the explained operations, there is still a considerable concern. The most challenging data pre-processing task in this work is an imbalanced class issue. This procedure is usually regarded as a pre-processing task; however, we consider it as a separate task to be explained next due to its importance in our work.

### 3.3. Imbalanced handling

The observations that were labeled as negative are relatively rare as compared to the positive class (less than 10%). Positive and negative labels are determined according to the reviewers' ratings. Should a rating be higher than four, the corresponding review is considered as a positive one; otherwise, it is treated as a negative review (Table 1).

Hence, we face an imbalanced classification issue. In other words, positive class (the majority) outnumbered negative class (the minority), and both classes do not make up an equal portion of our dataset. The conventional classifiers such as Decision Tree ([41]) and Logistic Regression ([35]) do not accurately measure model performance when faced with imbalanced datasets. They usually have a bias towards the majority class, and the minority class observations are treated as noise. To handle this issue, an SVM that performs well against highly imbalanced datasets are used to train our model. This classifier is also equipped with a class weight measure to alleviate the situation. Moreover, a separate imbalanced classification phase is embedded into the model. In doing so, different types of algorithms, i.e., multi-task learning [28], adaptive sampling [44], and synthetic oversampling method [46], were integrated and tested.

Generally speaking, there are two distinctive approaches for handling the mentioned issue: (1) skew-insensitive techniques and (2) re-sampling approaches. The former deals with a class imbalanced problem by assigning a cost measure to the training data. The latter adjusts the original dataset such that a more balanced class distribution is achieved. Re-sampling methods ([46]) have become standard approaches and have been dominantly utilized recently. They can be classified into different categories, e.g., sampling strategies, wrapper approaches, and ensemble-based methods. Implementing a proper method is crucial; otherwise, it can be problematic, e.g., data loss and overfitting, and can result in a poor outcome. This phase aims to balance class distribution relatively. As stated, three different techniques have been tested. We have found that a synthetic oversampling algorithm ([46]) performs better than the other two methods (i.e., adaptive sampling and multi-task learning). It is worth mentioning that the two other methods used are also computationally expensive. The synthetic oversampling algorithm creates synthetic samples based on the nearest neighbor approach. By implementing the method, *Failure class* instances are synthetically created, and the distribution is more balanced. The procedures are as follow:

- Let $A$ be the set of all elements of the minority class. The algorithm detects k-nearest neighbors of all observations ($S \in A$) of this class. In doing so, the Euclidean distance between each observation and other elements is measured.
- A sampling rate (e.g., 60%) is defined based on the imbalanced proportion. Given such a pre-defined rate, 60% of k-nearest neighbors of each observation in the minority class are randomly selected. Let $A'$ be the set of k-nearest neighbors.
- For each element in the obtained set ($S' \in A'$) the following formula is used to create new samples.

$$S_{new} = S + \alpha * |S - S'| \tag{2}$$

where $\alpha$ is a random number between 0 and 1.

The pseudo-code of the procedure integrated into the model to handle class imbalanced issue is presented in Algorithm 2.

**Table 1**
Examples of positive and negative TripAdvisor reviews about St. Stephen's Green (Dublin, Ireland).

| Bubble_rating | Review_body | Label |
|---|---|---|
| 5 | Stephens Green is a great place to visit nice walk around the park there also in the summertime music playing you can have a picnic there watch the ducks and the swans in the pond there is also boards giving a bit of history… | positive |
| 5 | Great place to relax in the city. Beautiful gardens and paths to walk around. If you need to just sit a bit this is a great place to do so. | positive |
| 4 | St. Stephen's Green park is perfect to step away from the hustle and bustle of Dublin. The scenery is beautiful and calming. Then when you are refreshed you just step back into the action of the city. | positive |
| 3 | Only downside is anti-social behaviour. Always somebody hassling for money or asking for a smoke. Wouldn't mind it on my own but with kids it's terrible. Should be better patrolled. Europe's… | negative |
| 1 | While on a visit to Dublin we brought our children to the park. The place is really nice but we were really shocked when we went to the playground. Near the playground entrance there were about 200 teens drinking and causing trouble. | negative |
| 2 | While we were walking across the park, a young man tried to take my husband's laptop. It was zipped inside a shoulder bag. I yelled and this person went away. | negative |

**Algorithm 2.** Pseudo-code for handling class imbalanced issue

**Input** : $m \leftarrow$ number of minority class observations;
$r \leftarrow$ amount of over-sampling (%);
$k \leftarrow$ number of neighbors;
**Output:** r*m synthetic observations

```
# Exception Function
f = len(features) # number of features;
S ← [ ] # observations in the minority class;
S' ← [ ] # synthetic observations;
Γ ← [ ];
for i ← 1, 2, ..., m do
    D ← Compute Euclidean distance;
    D.sorted() # sort in an ascending order
    Index ← Find k-nearest neighbors and indices
    Γ.append(Index)
    Populate(i, r, Index)
end
p ← a number between 1 and k
α ← a number between 0 and 1
for j ← 1, 2, ..., f do
    Λ = S[j] − S'[j];
    S_new[j] = S[j] + α * Λ;
end
Return S_new
```

After all operations explained above are done, three more steps are required before the pre-processed data is given to a supervised algorithm.

1. Tokenisation: each review is broken into words (called tokens).
2. Vectorisation: each review is converted into a numeric representation (called corpus).
3. Transformation: each review is transformed into one row (including 0 or 1) where 1 is the word in the corpus corresponding to that column appearing in that review.

### 3.4. Supervised learning

Text normalization was used to convert text into more convenient, standard forms. Tokenisation was used to separate words from running text. Each review has a rate_value between 1 and 5. Any rate_value of 1, 2, and 3 is considered a negative review; 4 and 5 are considered a positive review. Thus, there are two classes in this work, as the methods implemented are binary classification models. Then tokenized words are converted into a numeric representation, a process known as vectorization. After the data is processed, two approaches were applied: unigrams and n-grams. An *n*-gram is a contiguous sequence of *n* words collected from our reviews. When *n* is equal to 1, it refers to as a unigram. Their corresponding models are probabilistic language models for predicting every word's ratio (in a unigram approach) or sequence of words (in an *n*-gram approach). After all the described operations are done, the pre-processed data is trained on a supervised ML method, i.e., SVM.

SVM is incorporated as a discriminative classifier for document categorization in this work. As explained in the prior section, it is less sensitive to the class imbalanced problems. This technique is based on the Structural Risk Minimisation principle. SVM's task is to learn and generalize an input-output mapping by finding separation between hyperplanes defined by classes of data. In our case, the set of reviews is the algorithm input, and their respective labels are the output. SVM searches for a separating hyperplane, which separates positive and negative reviews from each other with maximal margin; in other words, the distance of the decision surface and the closest review is maximal (Fig. 2).

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), y_i \in \{positive, negative\}$ be our training observations. The SVM classifier is implemented by solving the following optimisation problem:

$$maximise \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j \phi(x_i, x_j) \tag{3}$$

$$f(x) = \sum_{i=1}^{n} y_i \mu_i \phi(x_i, x_j) + \xi$$
$$\forall i : 0 \leqslant \mu_i \leqslant C \quad and \quad \sum_{ij=1}^{n} \mu_i y_i = 0 \tag{4}$$

where $\phi$ is a pisa kernel function, $\mu$ is a weight value, $\xi$ is a threshold and $C$ is a misclassification cost. The algorithm offers an optimal hyperplane, which is a decision boundary between the two classes.

## 4. Results

Supervised machine learning approaches are about conducting algorithms that precisely project a given input features to an output space. Each of these methods operates in two stages. First, an algorithm is trained based on a training dataset. Then, the algorithm is evaluated over various metrics based on a test dataset. Splitting the dataset is essential for an unbiased evaluation of prediction performance. Hence, the dataset used in this work was divided into two subsets. The testing dataset includes 3000 reviews, consisting of 2714 positive and 286 negative comments. As explained above, this dataset is used for the evaluation of all models implemented in this work. It should be mentioned that the imbalanced handling phase is implemented when the training data is fitted.

Given the above discussion, the training set was applied to train models. Computational analyses were implemented based on two scenarios, i.e., a traditional approach and the model proposed in this work. Both scenarios include all the data handling steps explained earlier, i.e., data pre-processing and supervised learning. However, our proposed model includes an additional imbalanced handling phase described in the previous section. As far as the first scenario is concerned, various supervised algorithms, including Deep Neural Network (DNN), Recurrent Neural Network (RNN), Quadratic Discriminant Analysis (QDA), and Random Forest (RF), are tested and their results are compared with the proposed model.

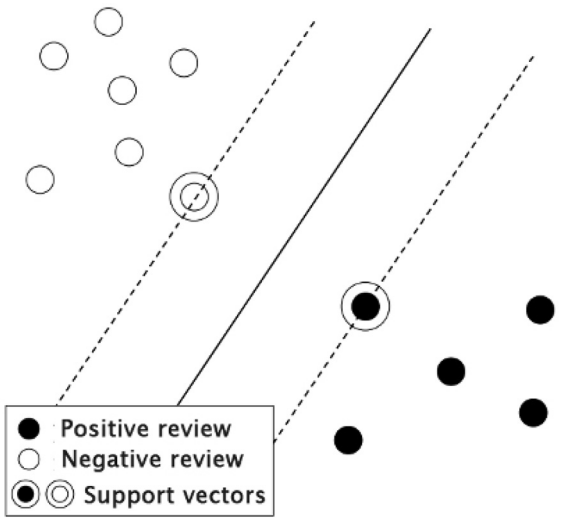


**Fig. 2.** The support vectors given positive and negative classes.

Given the extracted features, all the mentioned algorithms were fitted. After training all models, we have evaluated them to verify their applicability. Understanding how a model performs is essential to the use and development of text classification methods. To do so, the area under the Receiver Operating Characteristics (ROC) curves are used for comparing the accuracy of algorithms. These curves reveal a trade-off between the true positive rate and the false positive rate. The evaluation metric is based on a confusion matrix that comprises true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The significance of these four elements may vary based on the classification application. In this work, the fraction of correct predictions overall predictions is considered.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

Fig. 3 illustrates the ROC curve given the proposed model in this paper with and without the imbalanced handling phase. As shown, the model's predictive accuracy, assessed using the area under the curve (AUC), is over 97%. The confusion matrix is also presented in Fig. 4.

As stated, the proposed model has been experimentally validated and compared with four different approaches. Their corresponding performances have been evaluated according to their classification accuracies. The results are depicted in Fig. 5. The ability of each method to accurately predict the correct class is measured and expressed as a percentage. ROC curves have been used to determine the predictive performance of the examined classification algorithms. The area under a ROC has been considered as an evaluation criterion to select the best classification algorithm. When the area under the curve is approaching 1, it indicates that the classification was carried out correctly. We have also tested three more metrics, i.e., Precision, Recall, and $F1$-Score (Table 2). The Recall metric is the measure of the correctly predicted positive reviews from all the actual positive ones (Recall $= \frac{TP}{TP+FN}$). Hence, it is a good indicator for evaluating models (given the cost of False Negatives) dealing with the imbalanced class issue.

All experimental results show that our proposed model is superior to those tested. The additional imbalanced handling phase incorporated improves the fit of the model.

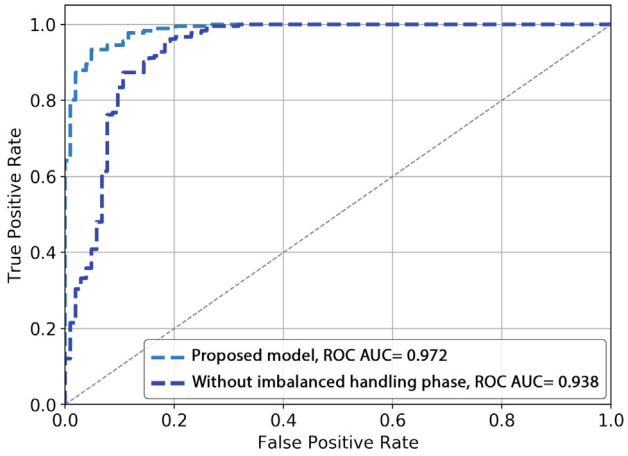


**Fig. 3.** Proposed model ROC curves with and without imbalanced handling phase.

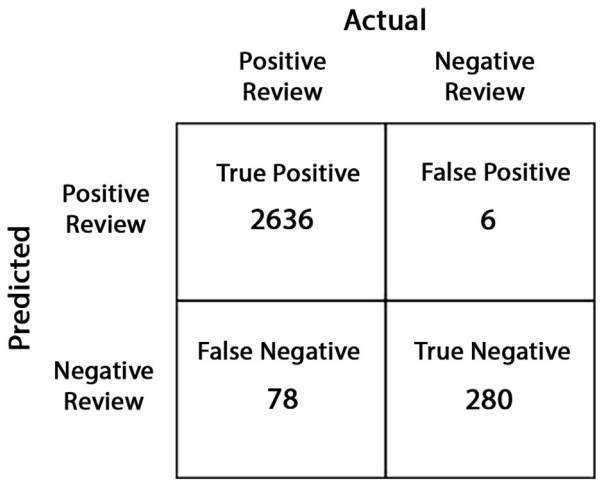


**Fig. 4.** Confusion matrix for the positive and negative classes.

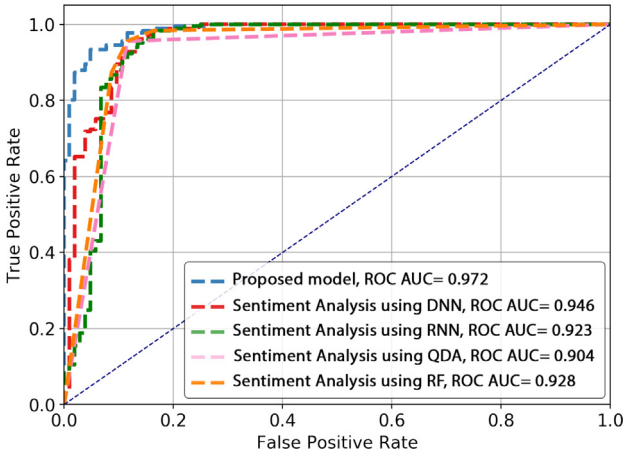


**Fig. 5.** ROC curves for different approaches.

**Table 2**
Comparisons of different classification metrics given 5 tested approaches.

| Models | Precision | Recall | F1 score |
|---|---|---|---|
| Proposed approach | 0.971 | 0.997 | 0.983 |
| DNN-based model | 0.946 | 0.993 | 0.968 |
| RNN-based model | 0.92 | 0.994 | 0.955 |
| QDA-based model | 0.901 | 0.991 | 0.942 |
| RF-based model | 0.927 | 0.992 | 0.957 |

## 5. Discussion

TripAdvisor reviews reveal a treasure trove of global, comparative data. To date, no other crowdsourced data was widespread enough to allow for such comparison between green spaces, both within the city and beyond. The most common criticism of observational approaches was time and cost expenditures spent on repeat measurements at the same locations ([6]). Like TripAdvisor, Twitter-based methods overcame this hurdle; tweets can be captured easily and frequently, offering greater measurement and (even longitudinal) analysis opportunities, saving time and costs ([39]). However, TripAdvisor, for a city's most popular parks, offers more data. The TripAdvisor reviews collected for this study averaged out to be 80 reviews/-month/park for St. Stephen's Green. Roberts ([39]) study collected about 11 tweets/month/park for her study area. Furthermore, tweets and reviews are not the same; while a tweet can be about UGS, a TripAdvisor review explicitly asks for a reviewer's experience (i.e., a person's sentiment).

The elderly population, who show lower levels of engagement with technology in general, are specifically overlooked in crowdsourced

data-based research ([5]). This is especially concerning as UGS are intended to be a shared public space for all ages. Roberts also reports Twitter data lack demographic information about Twitter users, such as their age, occupation, or ethnicity ([39]). These parameters, although not crucial to determine opinions, are useful for further examination of where particular attitudes may originate. Research to validate these demographic claims is limited, and studies comparing TripAdvisor with Twitter's user base are non-existent. Although some groups remain over- and/or underrepresented on TripAdvisor, there is an option to collect some demographic information such as gender, nationality, and age. We acknowledge the bias most crowdsourced data has and understand it is both a contested and fertile research area.

TripAdvisor enabled us to utilize the abundance of open-source reviews. The accessibility of the reviews makes the proposed method highly scalable–especially for popular parks. In Dublin, 33 of its 50 parks are listed on TripAdvisor. However, besides the most popular St. Stephen's Green (16,613 reviews), Phoenix Park (4,753 reviews), and St. Anne's Park (244 reviews), the remaining 30 parks have 1–66 reviews. Worldwide, thousands of UGS, from large to small, are listed on TripAdvisor. However, Dublin follows a similar pattern as other cities, where the most popular parks have significant reviews, and the lesser common parks have significantly fewer reviews. Therefore, we suggest the proposed method to be used only on a city's most popular parks, as a proxy for UGS in cities, and then compare UGS between cities worldwide. By leveraging machine learning techniques for opinion mining and text classification, hundreds of thousands of opinions previously overlooked can now be heard in an effort to improve these vulnerable public spaces.

## 6. Conclusions

Research to inform both policy and design of UGS is critical to protect these vulnerable areas while simultaneously ensuring access to the potential health and well-being benefits these spaces provide. Green spaces play a pivotal role across all aspects of city life, and as cities densify, the importance of accurately and effectively measuring the quality of UGS has never been greater.

This paper presents an experiment's results to use NLP to extract citizen opinion on the quality of UGS, a highly novel application of automatic text classification on TripAdvisor reviews. The results indicate that the proposed method performs better, at 97 accuracy, which is better than other approaches tested in this work.

Citizens, collectively, can enact meaningful change by acting as "ground agents" and providing valuable insights directly from the front lines. In this regard, citizens' insights are a goldmine of data that organizations can use to make their cities smarter. The results presented in this paper hold the potential to harness those opinions and give urban planners and local authorities greater choice to identify, analyze, and improve the sentiment behind specific UGS, and allow UGS comparisons between cities worldwide.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## CRediT authorship contribution statement

**Mohammadhossein Ghahramani:** Methodology, Software, Validation, Formal analysis, Writing - original draft, Visualization. **Nadina J. Galle:** Conceptualization, Methodology, Methodology, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Fábio Duarte:** Resources, Writing - review & editing. **Carlo Ratti:** Supervision. **Francesco Pilla:** Resources, Data curation, Writing - review & editing, Supervision.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Al Amrani Y, Lazaar M, El Kadiri KE. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput Sci 2018;127:511–20. https://doi.org/10.1016/j.procs.2018.01.150. pROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2017.
[2] Barton J, Rogerson M. The importance of greenspace for mental health. BJPsych Int 2017;14:79–81. https://doi.org/10.1192/s2056474000002051.
[3] Blumenthal M. Reviewer demographics-facebook has more women, yelp has more men; 2014. http://blumenthals.com/blog/2014/07/31/reviewer-demographics-facebook-has-more-women-yelp-has-more-men/.
[4] Chen Y, Lee B, Kirk RM. Internet use among older adults. Engaging Older Adults Modern Technol 124–41.
[5] Chen Y, Liu X, Li X, Liu X, Yao Y, Hu G, et al. Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k -medoids method. Landscape Urban Plan 2017;160:48–60.
[6] Cohen DA, Sehgal A, Williamson S, Marsh T, Golinelli D, McKenzie TL. New recreational facilities for the young and the old in Los Angeles: Policy and programming implications. J Public Health Policy 2009;30:S248–63.
[7] Colón-Ruiz C, Segura-Bedmar I. Comparing deep learning architectures for sentiment analysis on drug reviews. J Biomed Inform 2020;110:103539. https://doi.org/10.1016/j.jbi.2020.103539.
[8] Daniels B, Zaunbrecher BS, Paas B, Ottermanns R, Ziefle M, Rob-Nickoll M. Assessment of urban green space structures and their quality from a multidimensional perspective. Sci Total Environ 2018;615:1364–78.
[9] Dong J. Financial investor sentiment analysis based on fpga and convolutional neural network. Microprocess Microsyst 2020:103418. https://doi.org/10.1016/j.micpro.2020.103418.
[10] Douglas O, Lennon M, Scott M. Green space benefits for health and well-being: A life-course approach for urban planning, design and management. Cities 2017;66:53–62.
[11] Fang F, McNeil B, Warner T, Dahle G, Eutsler E. Street tree health from space? an evaluation using worldview-3 data and the washington dc street tree spatial database. Urban Forestry Urban Greening 2020;49.
[12] Fitri VA, Andreswari R, Hasibuan MA. Sentiment analysis of social media twitter with case of anti-lgbt campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. Procedia Comput Sci 2019;161:765–72. https://doi.org/10.1016/j.procs.2019.11.181. the Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.
[13] Fuller RA, Gaston KJ. The scaling of green space coverage in european cities. Biol Lett 2009;5:352–5.
[14] Galle NJ, Nitoslawski SA, Pilla F. The internet of nature: How taking nature online can shape urban ecosystems. Anthropocene Rev 2003;6:279–87. https://doi.org/10.1177/2053019619877103.
[15] Garcia-Barriocanal E, Sicilia M, Korfiatis N. Exploring hotel service quality experience indicators in user-generated content: a case using tripadvisor data. In: Mediterranean Conference on Information Systems. p. 200–5.
[16] Ghahramani M, Qiao Y, Zhou MC, O'Hagan A, Sweeney J. Ai-based modeling and data-driven evaluation for smart manufacturing processes. IEEE/CAA J Automatica Sinica 2020;7:1026–37. https://doi.org/10.1109/JAS.2020.1003114.
[17] Ghahramani M, Zhou M, Wang G. Urban sensing based on mobile phone data: approaches, applications, and challenges. IEEE/CAA J Automatica Sinica 2020;7:627–37.
[18] Gidlow CJ, Ellis NJ, Bostock S. Development of the neighbourhood green space tool (ngst). Landscape Urban Plan 2012;106:347–58.
[19] Giezen M, Balikci S, Arundel R. Using remote sensing to analyse net land-use change from conflicting sustainability policies: The case of amsterdam. ISPRS Int J Geo-Inform 2018;7. https://doi.org/10.3390/ijgi7090381.
[20] Girling C, Kellett R. Skinny streets and green neighborhoods: Design for environment and community. Island Press; 2005.
[21] Grahn P, Stigsdotter UK. The relation between perceived sensory dimensions of urban green space and stress restoration. Landscape Urban Plan 2010;94:264–75.
[22] Gretzel U. Online travel review study: Role and impact of online travel reviews. Laboratory for Intelligent Systems in Tourism; 2007.
[23] Haaland C, van den Bosch CK. Challenges and strategies for urban green-space planning in cities undergoing densification: A review. Urban Forestry Urban Greening 2015;14:760–71. https://doi.org/10.1016/j.ufug.2015.07.009.
[24] Hu S, O'Hagan A, Sweeney J, Ghahramani M. A spatial machine learning model for analysing customers' lapse behaviour in life insurance. Ann Actuarial Sci 2020. https://doi.org/10.1017/S1748499520000329.
[25] Hur M, Nasar JL, Chun B. Neighborhood satisfaction, physical and perceived naturalness and openness. J Environ Psychol 2010;30:52–9. https://doi.org/10.1016/j.jenvp.2009.05.005.

[26] Tieskens KF, Van Zanten BT, Schulp CJ, Verburg PH. Aesthetic appreciation of the cultural landscape through social media: An analysis of revealed preference in the Dutch river landscape. Landscape and Urban Planning 2018;177:128–37.
[27] Jin N, Wu J, Ma X, Yan K, Mo Y. Multi-task learning model based on multi-scale cnn and lstm for sentiment classification. IEEE Access 2020;8:77060–72. https://doi.org/10.1109/ACCESS.2020.2989428.
[28] Jin N, Wu J, Ma X, Yan K, Mo Y. Multi-task learning model based on multi-scale cnn and lstm for sentiment classification. IEEE Access 2020;8:77060–72. https://doi.org/10.1109/ACCESS.2020.2989428.
[29] Kourtit Nijkamp, Romao. Cultural heritage appraisal by visitors to global cities: The use of social media and urban analytics in urban buzz research. Sustainability 2019;11:3470. https://doi.org/10.3390/su11123470.
[30] Ma E, Cheng M, Hsiao A. Sentiment analysis – a review and agenda for future research in hospitality contexts. Int J Contemp Hospitality Manage 2018;30:3287–308. https://doi.org/10.1108/ijchm-10-2017-0704.
[31] Nasi R, Honkavaara E, Blomqvist M, Lyytikäinen-Saarenmaa P, Hakala T, Viljanen N, Holopainen M. Remote sensing of bark beetle damage in urban forests at individual tree level using a novel hyperspectral camera from uav and aircraft. Urban Forestry Urban Greening 2018;30:72–83.
[32] Nitoslawski SA, Galle NJ, Van Den Bosch CK, Steenberg JWN. Smarter ecosystems for smarter cities? a review of trends, technologies, and turning points for smart urban forestry. Sustain Cities Soc 2019;51:101770. https://doi.org/10.1016/j.scs.2019.101770.
[33] Nowak DJ, Greenfield EJ. Declining urban and community tree cover in the united states. Urban Forestry Urban Greening 2018;32:32–55. https://doi.org/10.1016/j.ufug.2018.03.006.
[34] Pak B, Chua A, Moere AV. Fixmystreet brussels: Socio-demographic inequality in crowdsourced civic participation. J Urban Technol 2017;24:65–87.
[35] Pampel FC. Logistic regression: A primer. SAGE; 2000.
[36] Plunz RA, Zhou Y, Vintimilla MIC, Mckeown K, Yu T, Uguccioni L, et al. Twitter sentiment in New York city parks as measure of well-being. Landscape Urban Plan 2019;189:235–46. https://doi.org/10.1016/j.landurbplan.2019.04.024.
[37] Rahimi S, Mottahedi S, Liu X. The geography of taste: Using yelp to study urban culture. ISPRS Int J Geo-Inform 2018;7:376. https://doi.org/10.3390/ijgi7090376.
[38] Rakhmanov O. A comparative study on vectorization and classification techniques in sentiment analysis to classify student-lecturer comments. Procedia Comput Sci 2020;178(194–204):2020. https://doi.org/10.1016/j.procs.2020.11.021. 9th International Young Scientists Conference in Computational Science, YSC2020, 05-12 September 2020.
[39] Roberts H, Sadler J, Chapman L. Using twitter to investigate seasonal variation in physical activity in urban green space. Geo: Geography Environ 2017;4.
[40] Roe JJ, Thompson CW, Aspinall PA, Brewer MJ, Duff EI, Miller D, et al. Green space and stress: evidence from cortisol measures in deprived urban communities. Int J Environ Res Public Health 2018;10:4086–103.
[41] Rokach L, Maimon O. Top-down induction of decision trees classifiers—a survey. IEEE Trans Syst Man Cybernet Part C (Appl Rev) 2005;35:476–87.
[42] Seiferling I, Naik N, Ratti C, Proulx R. Green streets- quantifying and mapping urban trees with street-level imagery and computer vision. Landscape Urban Plan 2017;165:93–101.
[43] Swanwick C, Dunnett N, Woolley H. Nature, role and value of green space in towns and cities: An overview. Built Environ 2003;29:94–106. https://doi.org/10.1016/j.cities.2017.03.011.
[44] Tra V, Duong B, Kim J. Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data. IEEE Trans Dielectr Electr Insul 2019;26:1325–33. https://doi.org/10.1109/TDEI.2019.008034.
[45] Yahav I, Shehory O, Schwartz D. Comments mining with tf-idf: The inherent bias and its removal. IEEE Trans Knowl Data Eng 2019;31:437–50. https://doi.org/10.1109/tkde.2018.2840127.
[46] Yang X, Kuang Q, Zhang W, Zhang G. Amdo: An over-sampling technique for multi-class imbalanced problems. IEEE Trans Knowl Data Eng 2018;30:1672–85. https://doi.org/10.1109/TKDE.2017.2761347.
[47] Yao F, Wang Y. Domain-specific sentiment analysis for tweets during hurricanes (dssa-h): A domain-adversarial neural-network-based approach. Comput Environ Urban Syst 2020;83:101522. https://doi.org/10.1016/j.compenvurbsys.2020.101522.
[48] Zhang Y, Van den Berg AE, Van Dijk T, Weitkamp G. Quality over quantity: Contribution of urban green space to neighborhood satisfaction. Int J Environ Res Public Health 2017;14. https://doi.org/10.3390/ijerph14050535.