COGITATIO

Urban Planning (ISSN: 2183–7635)
2018, Volume 3, Issue 1, Pages 21–33
DOI: 10.17645/up.v3i1.1231

Article

Investigating the Emotional Responses of Individuals to Urban Green Space Using Twitter Data: A Critical Comparison of Three Different Methods of Sentiment Analysis

Helen Roberts [1,*], Bernd Resch [2,3], Jon Sadler [1], Lee Chapman [1], Andreas Petutschnig [2] and Stefan Zimmer [2]

[1] School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK; E-Mails: hxr133@bham.ac.uk (H.R.), j.p.sadler@bham.ac.uk (J.S.), l.chapman@bham.ac.uk (L.C.)
[2] Department of Geoinformatics, University of Salzburg, 5020 Salzburg, Austria; E-Mails: bernd.resch@sbg.ac.at (B.R.), andreas.petutschnig@sbg.ac.at (A.P.), stefan.zimmer@sbg.ac.at (S.Z.)
[3] Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA; E-Mail: bresch@fas.harvard.edu

* Corresponding author

Submitted: 19 October 2017 | Accepted: 22 January 2018 | Published: 29 March 2018

Abstract
In urban research, Twitter data have the potential to provide additional information about urban citizens, their activities, mobility patterns and emotion. Extracting the sentiment present in tweets is increasingly recognised as a valuable approach to gathering information on the mood, opinion and emotional responses of individuals in a variety of contexts. This article evaluates the potential of deriving emotional responses of individuals while they experience and interact with urban green space. A corpus of over 10,000 tweets relating to 60 urban green spaces in Birmingham, United Kingdom was analysed for positivity, negativity and specific emotions, using manual, semi-automated and automated methods of sentiment analysis and the outputs of each method compared. Similar numbers of tweets were annotated as positive/neutral/negative by all three methods; however, inter-method consistency in tweet assignment between the methods was low. A comparison of all three methods on the same corpus of tweets, using character emojis as an additional quality control, identifies a number of limitations associated with each approach. The results presented have implications for urban planners in terms of the choices available to identify and analyse the sentiment present in tweets, and the importance of choosing the most appropriate method. Future attempts to develop more reliable and accurate algorithms of sentiment analysis are needed and should focus on semi-automated methods.

Keywords
emotions; sentiment analysis; Twitter; urban green space; urban planning

Issue
This article is part of the issue “Crowdsourced Data and Social Media for Participatory Urban Planning”, edited by Bernd Resch (University of Salzburg, Austria), Peter Zeile (Karlsruhe Institute of Technology, Germany) and Ourania Kounadi (University of Salzburg, Austria).

1. Introduction

1.1. Twitter, Sentiment Analysis and Urban Green Space

Sentiment analysis describes the field of study concerned with analysing the opinions, attitudes and emotions of individuals towards entities such as products, services, organisations, locations and events (Liu, 2012). Over the last two decades, the field has become increasingly active given the vast real-world applications to a plethora of disciplines, such as politics, economics, business, healthcare and urban planning. Increased engagement with sentiment analysis has also coincided with the rapid growth in social networks, without which a lot of

the recent research would not have been possible. For the first time in human history researchers have access to huge volumes of freely accessible data published by individuals online.

The increase in social media sites such as Twitter has led to the internet becoming a place of increased expression and opinion sharing on a vast range of topics (Pak & Paroubek, 2010). This phenomenon is providing new sources of text which can be used to gauge public opinion through sentiment analysis (Zhang, Riddhiman, Dekhil, Hsu, & Liu, 2011). Recent studies have indicated the potential and versatility of tweets in examining emotions. These include: a variety of economic (Bollen, Mao, & Zeng, 2011; Chung & Liu, 2011; Jansen, Zhang, Sobel, & Chowdury, 2009) and social (Thelwall, 2014) contexts, examining emotional responses to specific events, such as political elections (Bruns & Burgess, 2011; Tumasjan, Sprenger, Sandner, & Welpe, 2010; Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012), natural disasters (Mandel et al., 2012; Shalunts, Backfried, & Prinz, 2014) and terrorism events (Cheong & Lee, 2011); and exploring new ways to measure happiness (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Mitchell, Frank, Harris, Dodds, & Danforth, 2013; Quercia, Ellis, Capra, & Crowcroft, 2012). Recent research by Roberts, Sadler and Chapman (in press) identified how Twitter data can be successfully used to identify both emotions in tweets; and the cause of these emotions, in relation to green space experience. Following the success of this work, this study investigates the use of three different methods of sentiment analysis in this context. In doing so, different methodologies are explored and their limitations discussed.

The information made available by individuals in their tweets has the potential to provide insights into how urban landscapes are perceived by individuals as they navigate them. The urban landscape is being experienced by an increasing number of individuals as global urban populations continue to expand (UN Habitat, 2016), leading some to question the long-term sustainability of cities (Grimm, Grove, Pickett, & Redman, 2000). Understanding how individuals are responding and relating to city landscapes is a key element for facilitating their design, implementation and management. Urban green spaces in cities provide the opportunity for individuals to have contact with the natural environment (Daniel et al., 2012), a fundamental influence on human wellbeing (Fuller & Gaston, 2009; Kellert & Wilson, 1995; Wilson, 1984), while the benefits associated with nature and green spaces are a vital component of the ecosystem services they provide to human populations (Costanza et al., 1997; Daily, 1997; Ehrlich & Ehrlich, 1981; MEA, 2005). Despite broad agreement that these cultural ecosystem services are beneficial to urban dwellers (World Health Organisation, 2017) there remains limited methodological progress in capturing the transfer and receipt of these services to populations, largely due to their intangible nature and difficulty in assigning economic value to the benefits they provide (Daniel et al., 2012; Milcu, Hanspach, Abson, & Fischer, 2013). Studies have only recently emerged that consider the effect of number and duration of encounters on ecosystem service receipt (Shanahan, Fuller, Bush, Lin, & Gaston, 2015; Shanahan, Lin, Gaston, Bush, & Fuller, 2014), and at present they remain small scale and highly contextualised. Twitter data have the potential to offer a wider spatial and temporal lens through which responses of people to urban green spaces can be captured.

While environmental cues have a significant impact on how individuals respond to and experience space (Ulrich, 1983), a wide range of other factors are also influential, including weather conditions, group dynamics, types of activities and what people observe happening around them. These factors are hard to study successfully due to limitations on experiment size and cohort selection, so capturing their high spatial and temporal variability has proved challenging (Cohen et al., 2009). As a result, studies lack explorations of the emotional responses of people to urban green spaces and the range of sentiments they can elicit in individuals. Twitter data offers the potential to overcome these limitations and can provide information about how individuals feel while experiencing urban green spaces. The information provided in tweets also has the potential to contextualise why an individual may be experiencing certain emotions and what activities they are engaging in that result in the given response. Such information has significant utility for urban planning. For example, data which provides evidence for the beneficial effects of urban green spaces for urban dwellers can be used to justify their continued presence in the urban landscape amidst intense development pressures. Furthermore, the successful identification of the causes of positive and negative emotions experienced by users of urban green space using Twitter data (Roberts et al., in press), could be used to develop an evidence base from the which planners can create and manage green spaces to promote positive emotional experiences and minimise and remove features which cause negative responses.

Despite the benefits Twitter data offers to researchers, sentiment analysis studies obtained from tweets are not common, especially in an urban context. Nonetheless, studies have utilised tweet text to investigate how public mood varies both spatially (Bertrand, Bialik, Virdee, Gros, & Bar-Yam, 2013) and temporally (Martinez & González, 2013) in urban areas, and to compare how the positivity of Twitter posts by urban citizens varies between different cities (Hollander et al., 2016). Others have used Twitter data alongside additional sources (such as biosensors) to assess how individuals perceive and emotionally respond to cities (Resch, Summa, Zeile, & Strube, 2016), in order to develop more citizen centric approaches to urban planning. For tweets to be a useful source of emotional data to urban planners, methods of sentiment analysis are required which enable the fast, accurate and replicable annotation of tweets.

1.2. Methods of Sentiment Analysis

The possibility of accurately extracting emotions from tweets has been demonstrated in recent studies (e.g., Roberts, Roach, Johnson, Guthrie, & Harabagiu, 2012), which have classified tweets according to a range of readily identifiable and distinct emotions. However, working with such an informal text genre presents new challenges for language processing as the language used by the twitter community is often informal with creative punctuation and spelling, slang, abbreviations and URLs (Rosenthal, Ritter, Nakov, & Stoyanov, 2014). The use of emoticons/emojis also provides an additional challenge for analysts as the emotions they convey can be highly subjective and often context dependent. Debate on how to develop methods which address these challenges and capture the fullest range of responses possible, and how best to mine people's opinions and sentiments is an increasing body of literature.

To compensate for the range of challenges inherent in using Twitter data, approaches to identifying sentiment and emotion are varied, but can broadly be placed into three commonplace methodologies. Firstly, manual annotation requires human annotators to categorise tweets into emotion categories (Jansen et al., 2009; Roberts et al., in press). Fully automated annotation can also be undertaken, relying on algorithms and rules to annotate the emotion in tweets. Many different approaches to fully automated annotation exist, but methods typically rely on n-gram analysis (Barbosa & Feng, 2010) to annotate the emotion in a tweet. Significant limitations have been identified with using both manual and automated sentiment analysis on tweets (and are discussed in detail in subsequent sections). As a result, novel methodologies are being developed to progress tweet sentiment analysis. This study presents one such method, drawing on semi-supervised or machine learning annotation. There are a number of machine learning techniques which can be employed to annotate tweets including Naïve Bayes classification (Go, Bhayani, & Huang, 2009; Pak & Paroubek, 2010), maximum entropy classification (Go et al., 2009), graph based propagation algorithms (Resch et al., 2016) and semantic orientation (Turney, 2002). The method presented herein relies on a graph based semi-supervised learning algorithm (Resch et al., 2016) and is described in full in Section 2.5. The variety of approaches undertaken within these three methodological approaches reflects the complexity inherent in the task.

This article uses tweets relating to urban green spaces to evaluate three different sentiment analysis methods, focusing on the variation in sentiment they indicate, in order to facilitate discussion around the limitations and benefits of each approach. However, this article does not attempt to identify the most effective method for tweets. Instead, the aims of this article are twofold:

1) To compare the outcomes of manual, fully automated and semi-supervised learning methods of sentiment analysis on the same corpus of tweets;
2) To evaluate each method in the context of urban green space research.

The three methods of sentiment analysis presented and compared herein have been chosen as each one is derived from one of the three broad methodologies of sentiment analysis: manual, automated and semi-automated. In this way, a comparison can be made between these differing methodologies in the context of urban green space research; and their potential contribution in providing ways for urban planners to engage meaningfully with social media derived data.

2. Methodology

2.1. Case Study Location

The tweets collated for analysis relate to 60 urban green spaces located in Birmingham, United Kingdom (Figure 1). With a population of approximately 1.1 million people (Office for National Statistics, 2014) the 600 public parks, open spaces and nature reserves within the Birmingham metropolitan area (Birmingham City Council, 2016) provide an important resource for urban citizens in terms of their contribution to cultural ecosystem service provision.

The 60 green spaces were chosen to reflect the diversity of spaces found across the city in terms of their size, habitat type, available facilities and amenities and locations within different types of neighbourhoods. Alongside 46 parks, 14 linear green features were also included for investigation consisting of the footpaths along 4 rivers and 7 canals and 3 cycle ways.

2.2. Tweet Corpus Creation

The tweets used in this study were obtained via Twitters publically accessible REST API. The REST API provides access to a 1% sample of tweets published by users with public profiles, and allows queries to be used to search for specific tweets. Searches made using the REST API are based on relevance and therefore this source of tweets was most appropriate for use in this article. To create the tweet corpus used in this study, English language tweets were downloaded every 10 days from the REST API. During preparation for the tweet data collection various different time scales were used to collect tweets to ascertain the most effective frequency for harvesting tweets. Tests were carried out over a three month trial period to look at which frequency worked best to harvest tweets in terms of minimising duplications and ensuring sufficient capture of the available tweets. Frequencies of 3, 5, 7 and 10 days were tested. This showed that using frequencies of 3, 5 and 7 days were too frequent and re-

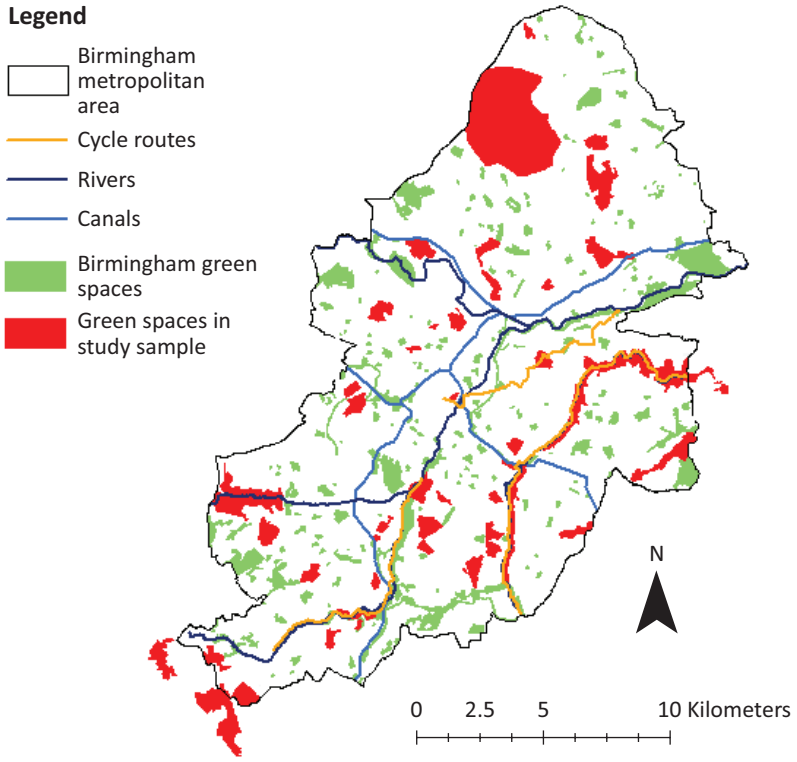

Figure 1. The locations of the green space study sites included in this article.

sulted in large duplications and made unnecessary pre-processing work to remove the duplications. Using the 10 day frequency, there was no lack of tweets compared to searching every 7 days, and given the benefits of this frequency in harvesting the tweets this frequency was used throughout the subsequent data collection period ensuring maximal temporal coverage over a period of 12 months, from June 2015 to May 2016. A search query was used to ensure that the tweets downloaded related to one of the 60 sites included in the study. Therefore, each tweet in the corpus contains specific reference to one of the sixty green spaces included in the sample. Any duplicated tweets were removed during pre-processing. In this way, a corpus of 10268 tweets was generated for use in this study.

2.3. Manual Annotation

During manual annotation, tweets were first assigned into one of three categories: positive, negative or neutral. This annotation was based on the presence of emotive words, emoticons/emojis or meaning. Subsequently, the positive and negative tweets were categorised into distinct emotions. The higher level emotions chosen included five of Ekman's basic emotions (anger, disgust, fear, sadness and happiness (Ekman, 1999; Ekman & Friesen, 1971)), in line with previous research using Twitter data (Roberts et al., 2012; Resch et al., 2016). These emotions are arranged into the ontology shown in Figure 2. In this study, beauty was included an additional sub-category to the positive tweets but outside of the emotions to account for the large amount of tweets referencing the beauty of nature and the landscape (as to be expected for green space). Each tweet could only be assigned into one of these emotion categories based on the strongest present emotion.

Five annotators were used to annotate a random sample of 1,000 tweets, in order to ensure there was sufficient agreement between different annotators in how tweets were categorised. A metric of comparison was

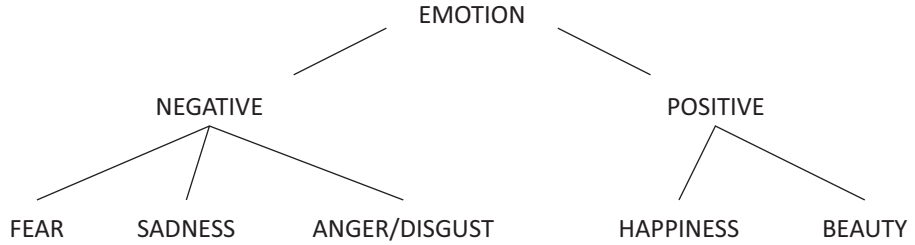

Figure 2. High level emotion ontology for the emotions used in manual and semi-automated tweet annotation.

derived (K = 0.666) suggesting sufficient agreement to assume inter-annotator reliability (Landis & Koch, 1977). Given the identification of sufficient inter-annotator reliability between annotators, and the time required for the task, the remaining tweets were annotated by one annotator. To the authors' knowledge this is largest manually annotated dataset of sentiment present in tweets, providing a robust test set against which other methods can be compared.

2.4. Fully Automated Annotation

For the automated method, an Affective Norms for English Words (ANEW) resource was used as the basis for emotion annotation. The ANEW resource utilised here derives from Warriner, Kuperman and Brysbaert (2013) in which over 13,000 English lemmas were assigned valence scores. Using an automated process these scores were used to annotate the valency of each tweet in the corpus. After assigning each word in each tweet with a valence score, an average valence score was created for each tweet based on the number of words present. Thresholds were then used to place the tweets into positive, neutral and negative categories. Following the thresholds used by Warriner et al. (2013) tweets with scores of $\geq$6.0 were categorised as positive, scores between 5.9 and 4.9 were categorised as neutral and scores of $\leq$4.9 were categorised as being negative. Given there remains no robust way to determine specific emotions from numeric scores, this method only annotated the tweets in terms of their positivity as opposed to annotating each with a discrete emotion. The implications of this are discussed in greater detail further on.

2.5. Graph Based Semi-Supervised Learning Annotation

In this method (Resch et al., 2016), a sample of manually annotated tweets was used to train a graph based semi-supervised learning algorithm which annotated the remaining tweets. A sample of 1,000 tweets from the corpus, known as the gold standard, were annotated manually (as described in Section 2.3) and used to train and evaluate the annotation algorithm. This was done to compromise between manual and automated analysis and capture the benefits of each, namely the accuracy of manual annotation and the quickness of automated annotation.

In order to classify tweets according to the emotions they contain a similarity computation was first undertaken, where similarity is defined as the likelihood that two tweets contain the same emotion. The similarity computation comprises three dimensions; linguistic similarity (defined through proven emotion emotion-related linguistic features such as co-occurring words, part-of-speech tags, punctuation, spelling, emojis and n-grams), spatial similarity and temporal similarity (defined through spatial and temporal decay functions according to recent literature). It should be noted that the results presented in this article only used the linguistic feature groups because not all tweets were geolocated, thus lacked the necessary spatial information.

Once the similarity between tweets has been computed, the graph, which creates the input for the semi-supervised learning approach is constructed and is defined by the tweets (nodes) and pairwise similarity values (weighted edges). Assigning emotions to the tweets was undertaken by applying the graph-based semi-supervised learning algorithm Modified Absorption (MAD) using a subset of the gold standard (training dataset) as this method is found to be most effective for graphs where nodes connect to many other nodes (Talukdar & Pereira, 2010). The assigned emotions were then validated using the rest of the gold standard (test dataset) through computing statistical measures including precision, recall, f-measure and micro average precision. The results prove to be better than random and majority baselines which in the understanding of the field of computational linguistics, demonstrates that the methods works well. The developed algorithm outperforming the majority baseline is considered assuring, whereas the better performance compared to random baseline provides strong evidence that the method works well because it demonstrates that the results are not produced by chance, but that significant similarities have been found between pairs of tweets.

Once each tweet had been assigned a discrete emotion using this method, it was then possible to reverse the process and place the tweets into positive, neutral and negative categories using the same ontology as shown in the manual method.

2.6. Analysis

Following presentation of the relevant descriptive statistics for each method, various statistical tests were undertaken to assess the significance of any differences in the assignment of the number of positive, neutral and negative tweets by each of the three methods. Fleiss and Cohen Kappa Indexes were then generated to assess inter-method reliability of tweet assignment into each category between the three methods alongside percentage agreement assessments of the three methods in their annotation of each individual tweet.

3. Results

3.1. Assignment of the Tweets into Positive, Neutral and Negative Categories

Variation existed in the numbers of tweets assigned to into the 'positive', 'neutral' and 'negative' categories by each of the methods (Figure 3). Although for all three methods, the majority of tweets were placed into the 'neutral' category, categorisation of tweets into 'positive' and 'negative' categories showed to be more variable between the three methods (Table 1).

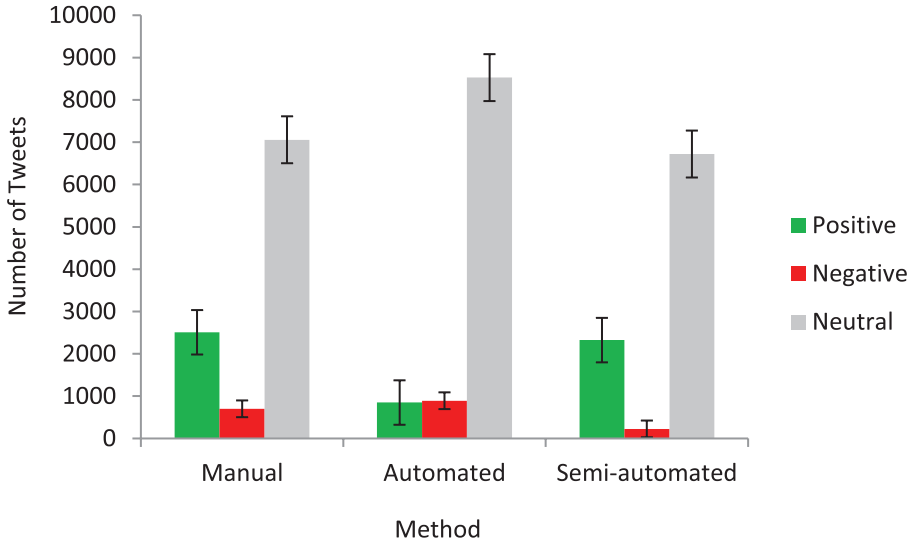

Figure 3. The number of tweets assigned by each method into positive, neutral and negative categories with standard error bars displayed. N (Number of tweets analysed) = 10268, for all methods.

Table 1. The percentage (%) of tweets assigned by each method to positive, neutral and negative categories.

	Manual	Automated	Semi-Automated
Positive	24.4	8.2	25.1
Neutral	68.8	83.0	72.5
Negative	6.8	8.8	2.4

Given that all three methods show some similarity in the numbers of tweets assigned to each category (Figure 3), statistical analysis was undertaken to investigate the significance of the differences identified between the three methods of classification for all three classes: 'positive', 'negative' and 'neutral'. Given that the assumption of homogeneity of variance was not met by the 'positive' datasets, a Welch ANOVA test was used and identified significant difference in the number of tweets annotated as positive by each of the three methods ($F(2,17.867)=39.343$, $p<0.001$). Post hoc Tukey analysis identified specific significant differences between manual and automated analysis ($p < 0.001$) and automated and semi-automated analysis ($p = 0.001$). There was no significant difference in the number of tweets annotated as 'positive' by the manual and semi-automated methods ($p = 0.76$). Using a one-way ANOVA, no significant differences were identified between the number of tweets classified as being 'neutral' by each method ($F(2,33)=3.216$, $p = 0.053$). Finally, a Kruskal-Wallis H test, given the violated assumption of normality, identified significant differences between the number of tweets classified as 'negative' by the three methods ($\chi^2(2)=16.176$, $p < 0.001$) . These were largest between the automated and semi-automated annotations of negativity.

By making adjustment to the thresholds (Table 2) used to assign the automated tweet scores into the 'positive', 'neutral' and 'negative' categories, it was possible to generate very similar outputs for the manual and fully automated methods (Figure 4), and identify no significant differences in the number of tweets each method assigned to each category.

3.2. Inter-Method Reliability

Consideration of inter-method reliability however, shows a more complex picture. A Fleiss Kappa Index identified very little inter-method agreement (k = 0.0445) between the three methods, highlighting that the annotation of each individual tweet into the three different categories by each method differed substantially. Indeed,

Table 2. Original and adjusted thresholds used to assign automated tweet scores into positive, neutral and negative categories.

	Original threshold adapted from Warriner et al. (2013)	Adjusted threshold
Positive assigned tweets	≥6.0	≥5.73
Neutral assigned tweets	≥5.0	≥4.931
Negative assigned tweets	≤4.99	≤4.93

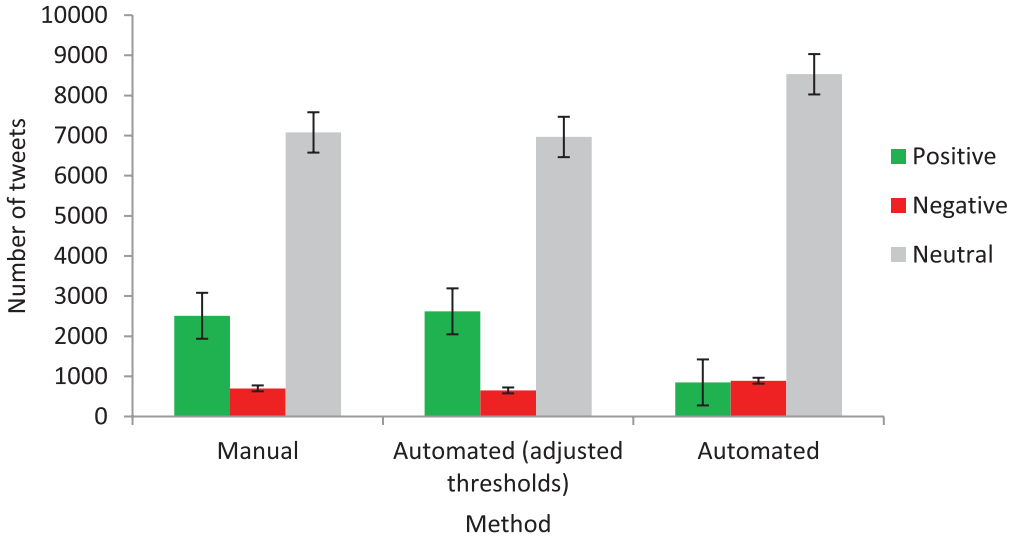

Figure 4. Comparisons of the numbers of tweets assigned to positive, neutral and negative categories by the manual and automated methods using two different thresholds.

only 44.5% of tweets were found to have been assigned the same category by all three methods, with 5.5% of tweets being assigned different categories by all three methods, indicating wide misallocation.

The relatively high percentage agreement compared to the low Fleiss Kappa Index is due to a large number of tweets being annotated as neutral by all three methods. Indeed, further investigation of the 44.5% of tweets which were annotated the same by all three methods revealed the vast majority to have been assigned to the 'neutral' category (98.1%). However, annotations of positive and negative tweets were less similar, suggesting that where emotions were present, the methods showed more variance in identifying them, either annotating them as neutral or with the incorrect polarity of positivity. Positive and negative annotation agreement between all three methods was extremely low at 1.9% and 0% respectively.

Interestingly, the low percentage in the agreement of tweets remained following the adjustment of the automated thresholds. The adjusted threshold annotations showed most similarity with the manual annotations. Again, however, only 56.8% of tweets were placed in the same category by both methods; showing that despite increasing similarity in number of tweets assigned to each category by each method, altering the thresholds used to assign tweets into 'positive', 'neutral' and 'negative' categories had no effect on increasing the percentage agreement of tweet assignment between the manual and fully automated methods.

Cohen Kappa tests were undertaken to see if the inter-method reliability was higher between any two specified annotation methods. The highest inter-method reliability was found to be between the manual and semi-automated methods (K = 0.136), compared to similarity between manual and automated (K = 0.0814), and semi-automated and automated methods (K = –0.00784). However, all these Kappa Indices are low (McHugh, 2012) and there remains large variation in the way each method assigns individual tweets into 'positive', 'neutral' or 'negative' categories, despite the appearance of similarity in Figure 3.

3.3. Quality Control Using Character Emojis

By way of a quality control measure, assessment was undertaken on just the tweets containing objective character emojis for the manual and semi-automated methods (automated annotation did not include character emojis in the lexicon). This was done as tweets containing such characters clearly belonged to either the positive or negative categories. All tweets containing positive or negative character emojis were assigned as 'positive' or 'negative' respectively by the manual method indicating a complete success rate of allocating these tweets into the correct emotion category. Compared to this, the ability of the semi-automated method was less successful. 54.4% of tweets containing positive character emojis were misallocated by the semi-automated method as either 'neutral' or 'negative'; while 75% of the tweets containing negative character emojis were misallocated as 'neutral' or 'positive'.

3.4. Assignment of Tweets into Discrete Emotion Categories

Using the manual and semi-automated methods of annotation it was possible to assign tweets into a number of emotion categories. A comparison of the number of tweets assigned into each of these categories again highlights substantial variation between the methods (Figure 5). Both methods showed variation in the number of tweets they identified as belonging to each emotion category. Substantially higher numbers of tweets were annotated as anger/disgust, fear and beauty by the manual method compared to the semi-automated method.

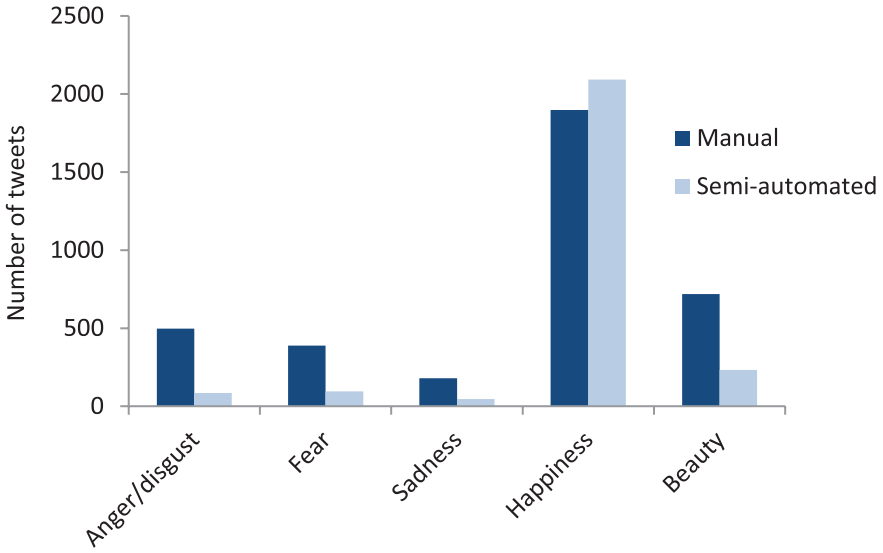

Figure 5. The number of tweets assigned by the manual and semi-automated methods into discrete emotion categories.

Percentage agreement between the two methods was found to be 44.5% when undertaken on all tweets. However, when tweets which were allocated as 'neutral' by both methods were removed, this figure falls substantially to 3.91%. This indicates that the methods show higher levels of variance when allocating an emotion to a tweet as opposed to just identifying the presence of an emotion, and that the presence of neutrality in a dataset can affect how the results of agreement between the assignment of tweets can first appear. A Cohen's Kappa Index of 0.0157 further emphasises the low level of agreement in allocation of tweets to discrete emotions.

4. Discussion

4.1. Comparison of the Outputs of Manual, Automated and Semi-Automated Analysis

The results presented show that detecting sentiments from tweets is a highly complex task, and importantly, that the method of analysis employed determines the categorisation of positivity, neutrality or negativity, despite using the same corpus of tweets. Moreover, the comparison of the manual and semi-automatic methods illustrated considerable variability in Ekman's specific emotion classes.

All three methods were found to assign variable yet similar numbers of tweets into the positive, neutral and negative categories, with the majority of tweets being annotated as neutral, followed by smaller numbers of positive and negative tweets respectively. Despite this analysis suggesting similarities between the three methods, assessment of inter-method reliability found percentile agreement between the assignment of tweets into the three categories by the methods to be only 444.5%.

The adjustment of thresholds used to assign automated tweet scores into positive, neutral and negative categories improved the similarity in the number of tweets assigned to each category between the manual and fully automated methods; however, it did not improve the percentage agreement between the two methods.

Manual annotation has previously been cited as providing the most reliable method of sentiment analysis, given that human annotators have the best chance of identifying the emotion present in a tweet (Saif, Fernez, He, & Alani, 2013). However, a dataset resulting from manual annotation is not unambiguous given that labelling tweets with an emotion remains a subjective task (Resch et al., 2016). Different human annotators may interpret the same text differently for many reasons—for example, sarcasm, slang or ambiguous use of emojis. This issue is also relevant for the semi-supervised learning method used here, given that the 'gold standard' tweet dataset used to train the algorithm relied on initial manual annotation of 1,000 tweets. To ensure that annotation was reliable between human annotators, a metric of comparison was derived suggesting agreement between them to be sufficient to assume inter-annotator reliability (Landis & Koch, 1977). Kappa Indexes enable the assessment of inter-annotator reliability between manual annotators and allow the variation in annotation by different annotators to be quantified.

Setting aside inherent subjectivity, the most significant limitation of manual sentiment analysis of tweets is the researcher time needed to examine each tweet. Given that Twitter generates large volumes of tweets in very short time periods, manual annotation is simply not viable. For this reason, automated and semi-automated methods are often employed.

Automated methods of sentiment analysis offer a quick and easy means of annotating large tweet datasets. Methodologically, however, there remains no robust way to derive discrete emotions from numeric scores, thus the granularity of the automated method demonstrated herein is limited to assessment of positivity

rather than identifying specific emotions from tweet text. In this study, a large lexicon of words was used to enhance the reliability in the scores generated for each tweet. Despite this, the limitations seem to outweigh the benefits. Low inter-method reliability was prevalent and there was a particularly low percentage agreement between annotations of positive and negative suggesting that this method is unlikely to reliably identify the correct polarity of sentiment in tweet text. Additionally, while the large lexicon used provides robustness for scoring words, it does not include emojis which are increasingly common ways to express sentiment in short social media posts (Pavalanathan & Eisenstein, 2015). Previous research has shown that emojis can be successfully used to inform automated analysis of tweets (Go et al., 2009). Indeed, the creation of an emoji lexicon in which each is given a score would be of significant use to future research and enable the combined use of words and emojis in the annotation of sentiment from tweet text. Such an undertaken would need to overcome the challenge of interpreting emojis in their different representational forms:

Unicode (e.g. "U+1F642"), Kaomojis (e.g. "(◕‿◕)"), a sequence of ASCII characters (e.g. ":-)") or a specific code used by Twitter (e.g., "<ed><a0><bd><ed><b2><af>" or "<ed><U+00A0><U+00BC><ed><U+00BC><U+009E>").

An issue of spatial variation in language use was also identified associated with the automated method of annotation. Despite the large lexicon used, it cannot account for regional/local dialect. Given the location for this study was Birmingham, where some language used by local populations is not used elsewhere, these words will not have been included and scored and a proportion of sentiment in the tweets, albeit small, will not have been captured by this method. Provided that manual annotators are native to the language and region from which the tweets have been captured, this should not be an insurmountable issue.

The semi-automated method generated similar numbers of neutral, positive and negative tweets as the other two methods. However, Kappa Indices indicate that the placement of individual tweets into each of these categories showed low levels of agreement. Differences were also identified in how semi-automated annotation assigned tweets to discrete emotion categories, when compared to manual annotation. The notion of beauty is not a basic emotion as defined in emotion psychology; indeed, it is usually subsumed under happiness. This makes it difficult for the algorithm to identify beauty in written text because it is often expressed in comparatively subtle terms.

For the experiment presented in this article, it was possible to identify a limitation in the semi-supervised method, in that the full range of emojis in the dataset could not be captured by the algorithm. The method is designed for character-wise emojis (e.g. ":-)"), however unicode emojis are widely used alongside character-wise emojis in tweet texts. In fact, the semi-supervised learning method was not able to interpret unicode emojis, increasing the likelihood that essential elements of tweets were missed by this method, diluting the precision of assigning emotions and polarities.

The quality control measure, which used character emojis to assess the allocation of tweets into the correct category, highlighted that the semi-automated method was often unable to recognise emotion, despite these being included in the assessment of linguistic similarity undertaken during analysis.

The parameter choices of semi-automated approaches make such methods highly sensitive; the number of seeds used, the seed distribution, details of similarity computation, edge weight threshold and the emotion categories used strongly influence the results. A significant issue is that no formalised method exists to perform an *a priori* estimation for these parameters. In most cases, 'optimal' parameter settings can only be found through empirical experiments, which in turn means it cannot be stated with certainty how good any results are in relation to the best achievable results. Thus, the parameter choices require a substantial amount of expert knowledge and experience, particularly because random permutations cannot be performed due to the computational complexity of the algorithms. This opens up debate as to how a training dataset should be generated. In this article, 1,000 tweets were randomly chosen. It may be more appropriate to actively identify tweets which cover all the discrete emotion categories so the algorithm can learn most effectively.

Finally, in this article, for all the methods of emotion annotation used, it was assumed that one tweet contains a maximum of one emotion. However, in reality tweets can be inherently more complex and contain a variety of emotions over a short space of characters. This is a finding that future methods looking to classify the emotion in tweet text will need to consider and overcome to provide the most accurate interpretation of the emotional information that tweets contain.

4.2. Implications of These Findings for Urban Planners

The availability of emotional data to urban planners has significant utility in the creation, management and justification of urban green spaces which promote positive emotional experiences and minimise features which may elicit negative emotional responses (Roberts et al., in press). The provision of such emotional data through social networks, such as Twitter, provides the opportunity for planners to gain access to this information in inexpensive, time efficient and replicable ways. However, in order to be used meaningful, methodologies are required which can accurately annotate any emotion present in a tweet relating to an urban green space.

This article has identified that challenges remain to this end. Indeed, none of the three methods presented herein are appropriate in their current form to provide sentiment analysis of tweet text for urban planners.

Whilst manual analysis can be used to accurately identify any emotion present, the amount of time taken to undertake this method on a large corpus of tweets makes it unsuitable in the context of urban planning where resources and individuals are often limited.

Similarly, the current inability of automated and semi-automated methods to accurately identify emotion, make them dubious approaches to employ where the identification of such emotion and their causes could have significant implications for the management and creation of green spaces.

However, the authors tentatively suggest that pursuing a semi-automated method, like the one presented herein is the most appropriate way forward. The development of a method through which the accuracy of manual annotation can be achieved, in much shorter time is doubtless of interest to urban planners. This is of particular relevance because manual annotation of tweets is a time-consuming and expensive method. This article suggests that the development of a gold standard training data set should be a priority, enabling algorithms to learn the variety and complexity with which emotions can be conveyed in tweets.

Without a doubt, Twitter data presents a useful and abundant source of easily accessible emotion information which is generated by users as they experience specific urban green spaces. Such a source of data presents vast opportunities for urban planners; however there remains a need for increased innovation and development in the methodologies which would enable this data source to be engaged with most effectively.

5. Conclusion

This paper has presented a comparison of three approaches to sentiment analysis undertaken to collate the sentiment and emotion present in tweet text. Despite their utility, significant differences exist in the outcomes of three methods of sentiment analysis on the same corpus of tweets. The discrepancies in how tweet text is analysed by different methods is thus a critical consideration for future research.

It was possible to identify differences in positivity annotation between all three methods in terms of the numbers of tweets assigned to each category as well as inter-method reliability in assignment. Using the manual and semi-automated methods, discrete emotions can be annotated, but again significant differences were identified in this process, particularly for beauty and anger/disgust tweets.

Overall, whilst this article is positive about the role of Twitter in providing a useful and substantial data source for urban planners on which to undertake sentiment analysis, it suggests caution is needed in interpreting the outputs of sentiment analysis and an understanding of the process can help place the results in an appropriate context. A critical discussion of the limitations identified through the undertaking of all three methods in this research has been presented. In doing so, it adds to the debate surrounding annotation of sentiment and emotion from tweets and identifies methodological constraints which should be taken into account in future work. Given the utility of the sentiment information captured by tweets relating to urban green space for planners and decision makers, it is of important that an efficient and reliable method is established through which these can be identified and annotated. Despite its reliability, manual annotation is unfeasible for large volumes of data. However, automated and semi-automated methods are hampered by a number of limitations associated with each, and this work shows that methodological progression is necessary before either can be used robustly to annotate sentiments from large tweet datasets.

The findings presented here suggest that automated methods of sentiment analysis are not able to accurately identify the emotion present in tweet text and that manual analysis, whilst accurate, is impractical for use on large tweet corpi given the time taken to undertake such analysis. As a result, this research suggests that future attempts to develop methods of sentiment analysis should focus on semi-automated methods, with particular focus given to how the gold standard dataset is selected. Successful algorithms should aim to include Unicode as well as character emojis in order to best capture the emotion represented by these in tweets.

Acknowledgments

We would like to express our gratitude to the Austrian Science Fund (FWF) for supporting the project "Urban Emotions", reference number I-3022. We would also like to thank Dr. Wendy Guan from Harvard University's Center for Geographic Analysis for her support through providing us with the Twitter data for our study. We would also like to thank the Engineering and Physical Sciences Research Council (EPSRC), reference number EP/L504907/1.

Conflict of Interests

The authors declare no conflict of interests.

References

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 36–44). Stroudsburg, PA: Association for Computational Linguistics.

Bertrand, K., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in New York City: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*

Birmingham City Council. (2016). Parks and nature conservation. Retrieved from http://www.birmingham.gov.uk/parks

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bruns, A., & Burgess, J. (2011). The use of Twitter hashtags in the formation of ad hoc publics. In *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*.

Cheong, M., & Lee, V. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, *13*(1), 45–59.

Chung, S., & Liu, S. (2011). *Predicting stock market fluctuations from twitter*. Retrieved from https://www.stat.berkeley.edu/ aldous/157/Old_Projects/Sang_Chung_Sandy_Liu.pdf

Cohen, D., Sehgal, A., Williamson, S., Marsh, T., Golinelli, D., & McKenzie, T. (2009). New recreational facilities for the young and old: Policy and programming implications. *Journal of Public Health Policy*, *30*(1), S248–S263.

Costanza, R., d'Arge, R., Groot, R., Farber, S., Grasso, M., Hannon, B., . . . Belt, M. (1997). The value of the world's ecosystem services and natural capital. *Nature*, *387*, 253–260.

Daily, G. (Ed.). (1997). *Nature's services: Societal dependence on natural ecosystems*. Washington, DC: Island Press.

Daniel, T., Muhar, A., Arnberger, A., Aznar, O., Boyd, J., Chan, K., . . . von der Dunk, A. (2012). Contributions of cultural services to the services agenda. *PNAS*, *109*(23), 8812–8819.

Dodds, P., Harris, K., Kloumann, I., Bliss, C., & Danforth, C. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLOS ONE*, *6*(12), e26752.

Ehrlich, P., & Ehrlich, A. (1981). *Extinction: The causes and consequences of the of the disappearance of species*. New York: Random House.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). New Jersey: Wiley and Sons Ltd.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Fuller, R., & Gaston, K. (2009). The scaling of green space coverage in European cities. *Biology Letters*, *5*(3), 352–355.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

Grimm, N., Grove, J., Pickett, S., & Redman, C. (2000). Integrated approaches to longterm studies of urban ecological systems. *Bioscience*, *50*, 571–584.

Hollander, J., Graves, E., Renski, H., Foster-Karim, C., Wiley, A., & Das, D. (2016). *Urban social listening: Potential and pitfalls for using microblogging data in studying cities*. Basingstoke: Palgrave Macmillan.

Jansen, B., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of The American Society for Information Science and Technology*, *60*(11), 2169–2188.

Kellert, S., & Wilson, E. (1995). *The biophilia hypothesis*. Washington, DC: Island Press.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *1977*, 159–174.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., & Rodrigue, J. (2012). A demographic analysis of online sentiment during hurricane Irene. In *Proceedings of the Second Workshop on Language in Social Media* (pp. 27–36). Stroudsburg, PA: Association for Computational Linguistics.

Martinez, V., & González, V. (2013). Sentiment characterization of an urban environment via Twitter. In *Ubiquitous computing and ambient intelligence. Context-awareness and context-driven interaction* (pp. 394–397). Berlin: Springer Verlag.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochem Med*, *22*(3), 276–282.

MEA. (2005). *Ecosystems and human well-being* (Vol. 5). Washington, DC: Island Press.

Milcu, A., Hanspach, J., Abson, D., & Fischer, J. (2013). Cultural ecosystem services: A literature review and prospects for future research. *Ecology and Society*, *18*(3), 44.

Mitchell, L., Frank, M., Harris, K., Dodds, P., & Danforth, C. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, *8*(5), e64417.

Office for National Statistics. (2014). *Mid-2014 mid-year population estimates, BDB2015/04 Birmingham demographic briefing*. London: ONS.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREC*, *10*.

Pavalanathan, U., & Eisenstein, J. (2015). Emoticons vs. emojis on Twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking gross community happiness from tweets, In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 965–968). New York: ACM.

Resch, B., Summa, A., Zeile, P., & Strube, M. (2016). Citizen-centric urban planning through extracting emotion information from Twitter in an interdisciplinary space-time linguistic algorithm. *Urban Planning*, *1*(2), 114–127.

Roberts, K., Roach, M., Johnson, J., Guthrie, J., & Harabagiu, S. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. *LREC*, 3806–3813.

Roberts, H., Sadler, J., & Chapman, L. (in press). The

value of Twitter data for determining the emotional responses of people to urban green spaces: A case study and critical evaluation. *Urban Studies*.

Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 73–80).

Saif, H., Fernez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold. In *1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*, Turin, Italy.

Shalunts, G., Backfried, G., & Prinz, P. (2014). Sentiment analysis of German social media data for natural disasters. In *Proceedings of the 11th International Conference on Information Systems for Crisi Response and Management (ISCRAM)*.

Shanahan, D., Fuller, R., Bush, R., Lin, B., & Gaston, K. (2015). The health benefits of urban nature: How much do we need? *Bioscience*, biv032.

Shanahan, D., Lin, B., Gaston, K., Bush, R., & Fuller, R. (2014). What is the role of trees and remnant vegetation in attracting people to urban parks? *Landscape Ecology*, *30*(1), 153–165.

Talukdar, P., & Pereira, F. (2010). Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1473-1481). Uppsala, Sweden, Association for Computational Linguistics.

Thelwall, M. (2014). Sentiment analysis and time series with Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 83–95). New York: Peter Lang.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment, *ICWSM*, *10*(1), 178–185.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 417–424). Stroudsburg, PA: Association for Computational Linguistics.

Ulrich, R. (1983). Aesthetic and affective response to natural environment. In *Behavior and the natural environment*. Boston, MA: Springer.

UN Habitat. (2016). *World city report 2016. Urbanisation and development: Emerging futures*. Nairobi: UN Habitat.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 US presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 115–120). Stroudsburg, PA: Association for Computational Linguistics.

Warriner, A., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal and dominance for 13,915 English lemmas. *Behavioural Research*, *45*, 1191–1207.

Wilson, E. (1984). *Biophilia*. Boston: MA: Harvard University Press.

World Health Organisation. (2017). *Urban green spaces: A brief for action*. Denmark: WHO Regional Office for Europe.

Zhang, L., Riddhiman, G., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining lexicon-based and learning-based methods for Twitter sentiment analysis* (HPL-2011-89). Palo Alto, CA: HP Laboratories.

About the Authors

Helen Roberts is a Doctoral Researcher at the University of Birmingham using social media data to investigate a variety of urban interactions between people and green spaces, particularly parks and corridors. A socio-ecological perspective frames this work and examines key how certain aspects of ecosystem service provision are dependent on human use of green spaces.

Bernd Resch is an Assistant Professor at University of Salzburg's Department of Geoinformatics—Z_GIS and a Visiting Fellow at Harvard University (USA). His research interests revolve around fusing data from human and technical sensors, including the analysis of social media. Amongst a variety of other functions, Bernd Resch is Editorial Board Member of the *International Journal of Health Geographics* and the *International Journal of Geo-Information*, Associated Faculty Member of the doctoral college "GIScience", and Executive Board member of Spatial Services GmbH.

Jon Sadler is a Biogeographer and Ecologist whose research focuses on species population and assemblage dynamics in animals (sometimes plants). His work is highly interdisciplinary, bisecting biogeography, ecology, urban design, riparian management and island Biogeography. It uses approaches that combine detailed field studies, field and laboratory experimentation, sometimes with social science to examine the links between environmental variability and species (including humans) responses. His research has significant blue skies and applied implications for understanding and responding to the impacts of climate and environmental change variability on urban and island ecosystems, hydrological systems, riparian/riverine ecology, the management/conservation of freshwaters. He is a fan of numbers and coding (especially using open source software such as R).

Lee Chapman is a Professor of Climate Resilience at the University of Birmingham focussed on researching the impact of weather and climate on the built environment. His research has a particular emphasis on critical infrastructure systems in urban areas and includes significant knowledge transfer to the meteorological market place.

Andreas Petutschnig received his BEng in Cartography and Geomedia-technology at the University of Applied Sciences in Munich, Germany and his MSc in Applied Geoinformatics at the University of Salzburg, Austria. His research interests include the analysis and visualisation of spatiotemporal data flows, spatial statistics, scaling issues of spatial data clusters, and reproducible research. His PhD work has a focus on spatiotemporal point pattern analysis. Specifically, this includes the detection of refugee camps in crisis regions based on traces of social media data.

Stefan Zimmer studied Geoinformatics (BSc) in Germany and currently finalises the Masters progamme in Geoinformatics. In the research group, he is involved in building and improving the social media crawler infrastructure. Furthermore, Stefan optimizes the performance of the emotion classification algorithm for social media posts by exploiting massive parallelism on graphic processing units with Set Similarity Joins.