

Integrating Geospatial Data and Social Media in Bidirectional Long-Short Term Memory Models to Capture Human Nature Interactions

ANDREW LARKIN* AND PERRY HYSTAD

College of Public Health and Human Sciences, Oregon State University, Corvallis, OR 97331

**Corresponding author: larkin@oregonstate.edu*

Contact with nature has been linked to human health, but little information is available for how individuals utilize urban nature. We developed a bidirectional long short-term memory model for classifying whether tweets describe the proposed pathways through which nature influences health: exercise, aesthetic stimulation, stress reduction, safety, air pollution mediation, and/or social interaction. To adjust for regional variations in urban nature context, we integrated OpenStreetMap data on nature and non-nature features for each long-short term memory cell. Training ($n = 63073$), development ($n = 5000$), and test ($n = 5000$) sets consisted of labeled tweets from Portland, Oregon. Tweets from New York City (NYC) ($n = 5000$) were also labeled to test generalizability. The model was applied retrospectively to 20 million tweets from 2017 and continuously to Meetup posts for 7,708 cities in North America. F1Scores ranged from 0.54 to 0.82 in the NYC dataset, a 24% to 92% improvement over current methods. Precision ranged from 0.58 to 0.83, while recall ranged from 0.39 to 0.81. Adding OpenStreetMap features led to greater percent and absolute F1Scores in NYC compared to Portland. Average F1Scores were greater in models with a nature label in addition to human behavior labels (0.59 vs. 0.65), suggesting health behaviors are influenced by urban nature.

Keywords: Social Media; Long-Short Term Memory; Nature; Environmental Health; Georeferenced

Received 3 December 2019; Revised 13 April 2020; Accepted 19 June 2020

Handling editor: Fionn Murtagh

1. INTRODUCTION

Natural components of the built environment (e.g. parks, tree-lined streets, backyards), are associated with multiple health benefits, including reduced blood pressure [1], reduced rates of childhood obesity [2], depression and anxiety [3] and type-2 diabetes [4], and increased rates of attention restoration [5], positive attitude [6], ability to cope with stress [7], and social cohesion [1, 8]. Previous studies suggest these health benefits are attributable to multiple human-nature interactions, including promoting physical exercise [9], aesthetic stimulation [10, 11], promoting social activities [8, 10, 12], air pollution mitigation [13, 14], and providing safe environments [2, 15].

While most studies support these hypothesized pathways of action, identifying and creating ideal urban nature environments remains challenging for public health researchers and urban planners, respectively. Associations between nature

and health might differ with respect to land use classification (e.g. parks, backyard, neighborhood) [11] [16], nature composition (e.g. grass, trees, lake) [17], socioeconomic status (e.g. income, education level) [16], surrounding nature density (small forested city vs. sprawling urban metropolis) [18], and personal affinity for nature [19]. Further, human-nature interactions are dynamic, influenced by both repeatable time-series [20] (e.g. seasonal park access) and unanticipated events (e.g. wildfires, public protests) [21]. Cumulatively, the diversity of urban nature composition and intra-population behaviors makes it difficult to standardize, quantitatively measure, and predict ideal urban nature environments, and modify environments as population needs change both temporarily and permanently over time.

Social media has significant potential to capture behaviors and perceptions that drive human interactions with nature.

Eighty-one percent of American teenagers [22] and 69% of all American adults [23] use one or more social media channels. Social media has become an essential component of young adult communications, and provides rich textual, visual, and meta data for capturing self-reported human-nature interactions. For example, Twitter is a publicly available social media data stream which includes up to 240 characters of text, hashtags for highlighting keywords and topics of interest, emoticons for emphasizing emotional states, and hyperlinks. With more than 500 million tweets per day, Twitter has significant potential to provide insight on human-nature interactions.

Unfortunately, methods for analyzing nature-related tweets (and other forms of social media) are limited. There are several candidate methods for nature-related social-media analytics. First, you could collect a large continuous sample of social media posts and label records by hand or via crowdsourcing methods such as Mechanical Turk. Mechanical Turk involves paying workers a small sum of money for performing simple and subjective tasks, such as filling out a short survey or labeling images (<https://www.mturk.com/>). This method has significant limitations for large datasets, restricted budgets, and ongoing analyses, all of which are common in public health research and policy. Second, sample records can be collected that include the name of a nature body (e.g. Grand Canyon, a frequently used abbreviation for Grand Canyon National Park, a popular tourist attraction and UNESCO World Heritage Site) [24]. Restricting tweets to known entities has high precision, but potentially low recall. Furthermore, behaviors at named entity locations such as parks are likely to differ from behaviors in informal nature locations (e.g. backyards, gardens, unnamed trails and forests). Third, you could collect tweets with hashtags that indicate nature-related themes (e.g. #nature, #lastchildinthewoods, #lake, etc.) [25]. However, hashtag sampling has limited recall, as a large percentage of nature-related tweets don't include popular hashtags. Hashtag sampling requires frequent updates to the hashtag list, and misses hashtag posts that occur before new relevant hashtags are added to the list. Finally, semantic analyses can be used that applied part of speech (POS) tagging, bag of words, and/or topic modeling (see the following github repository for an example application of POS tagging and topic modeling with a set of social media posts related to urban nature: https://github.com/larkinandy/Portland_UrbanNature_Twitter). These semantic grouping-based methods have a delicate balance between precision and recall, as inclusion of words or semantic structures with multiple meanings may be essential for high recall, but have reduced precision. For example, running is an essential keyword for capturing exercise behaviors, but in addition to humans, animals, cars, and noses also run.

Bidirectional long-short term memory (BiLSTM) models have significant potential to capture complex grammatical context and generate inference in language-related tasks. BiLSTMs include one or more layers of long-short term memory (LSTM) cells with connections between previous and following cells to

integrate past and future states into cell outputs (e.g. previous and following words in a sentence). LSTM models have been successfully developed for multiple domains, including speech recognition [26], text translation [27], sentiment analysis [28], and POS tagging [29]. To date, LSTM models have not been implemented in large scale, automated processing of nature-related tweets.

We set out to develop a BiLSTM model which can infer human-nature interactions from social media text. We chose to focus on identifying human-nature interactions as a test case because it represents an understudied area that is difficult to measure with traditional data. This approach was designed to achieve the following objectives:

- 1) Classify social media text records with multiple labels, including whether the record describes urban nature and/or identified pathways through which nature either directly (e.g. reducing stress) or indirectly (e.g. improving air quality) influences health.
- 2) Find an optimal balance between precision and recall, measured by the F1Score (harmonic mean of precision and recall) for the different human-nature interaction pathways.
- 3) Generalize across diverse populations and geographical extents with minimal variation in precision and recall.
- 4) Develop model architectures that capture interactions between labels.
- 5) Georeference tweets to the finest spatial extent possible. Overall, this approach will be able to leverage Twitter data to accurately identify how individuals interact with nature in cities, which can inform future urban planning, sustainability and environmental health research.

2. METHODS

All scripts for model training and evaluation, along with sample training records are available at http://github.com/larkinandy/GreenTweet_MultivariateBiLSTM. Tables and figures which supplement but are not essential to the manuscript body (Supplemental Tables and Supplemental Figures, respectively) are appended to the end of the manuscript.

2.1. Twitter Data Collection

Twitter data was collected using Python (v. 2.7) [30] and Tweepy (v. 3.5) (<http://docs.tweepy.org/en/latest/>). From January 1st to December 31st 2017, tweets containing one or more of the search keywords in Table 1 were continuously downloaded from the Twitter data stream and stored in MySQL (v. 5.7) [31] (the corresponding SQL data dictionary, which describes the variables and format of the collected Twitter data while stored in SQL format, is available in Supplemental Table 1). Downloaded records include time of initial post. Post times among the downloaded Tweets cover 99.7% of the minutes in 2017.

TABLE 1. Twitter Search Keywords

Twitter Search Keywords				
backyard(s)	forest(s)	lawn(s)	pasture(s)	stream(s)
bush(es)	garden(s)	leaves	plant(s)	trail(s)
crop(s)	grass	mountain(s)	prairie(s)	tree(s)
field(s)	hay	nature	river(s)	wood(s)
flower(s)	lake(s)	park(s)	riverside	yard(s)

Tweet Coding Examples


Original Text	Hiking with the family at Silver Falls today  #yeosinthewild @SilverFallsStatePark
Coded Text	Hiking with the family at Silver Falls today evergreen tree yeos in the wild @ Silver Falls State Park
Hashtag Flag	0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0
Emoticon Flag	0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
OSM Flag	0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1
Original Text	Me talking to my brother. Me: "Have you been to Fire on the Mountain?" Him: "No, but there's one on the mountain right now 🙄"
Coded Text	Me talking to my brother . Me : " Have you been to Fire on the Mountain ? "
Hashtag Flag	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Emoticon Flag	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
OSM Flag	0 0 0 0 0 0 0 0 0 0 0 0 0 -1 -1 -1 -1 0 0
Coded Text	Him : " No , but there is one on the mountain right now eyes
Hashtag Flag	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Emoticon Flag	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
OSM Flag	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

FIGURE 1. Example Tweets From the Training Dataset Before and After Preprocessing. Preprocessing steps include substituting emoticons with their descriptions, removing hashtag symbols, padding punctuations, and adding flags for words and punctuations that originated from hashtags, emoticons, and match a named entity in OpenStreetMap records (e.g. Silver Falls State Park, a restaurant named Fire on the Mountain). The meaning of the word 'yeos' in the original tweet text is unknown. Red, blue, and purple colors correspond to hashtag, emoticon, and OSM values that differ from the default value of 0.

Tweets were screened to remove retweets and tweets which contained common phrases that contain search keywords but don't correspond to nature (e.g. 'snake in the grass', 'data stream', 'George Bush').

2.2. Data Subset and Labeling

Tweets specific to Portland, Oregon were identified from tweet authors self-identified as living in the Portland, Oregon metropolitan region. These tweets were labeled by the manuscript authors for the following labels (tweet examples of positive and negative labels for each outcome are in Supplemental Table 2).

Nature—Does the tweet include a non-anthropogenic object that contributes to the composition of the built environment? Weather, animals, and nature references within similes or metaphors were not included.

Air—does the tweet mention air quality or respiratory function directly influenced by the atmospheric composition?

Aesthetic—does the tweet describe the visual aesthetic quality of an object or overall visual aesthetic quality of the environment? Does not include quality descriptions for other senses (e.g. beautiful music).

Exercise—are there any descriptions of physical activity? Includes short moments of exercise (e.g. running through the yard).

Safety—are there perceptions of overall safety, events that signify a safety failure (e.g. murder) or success (e.g. rescue) or descriptions of acts that increase or decrease safety?

Social—interactions between individuals, activities designed for community engagement, or activities performed by groups.

Stress—descriptions of increased or decreased anxiety, workload, mental pressure, or general unease.

In addition to tweets from Portland, OR, 5000 tweets from New York City (NYC) were randomly sampled from the 2017 dataset and coded by the lead author. The NYC dataset was used to evaluate the potential generalizability of model inferences,

as the NYC population and urban environment significantly differs from Portland, OR.

2.3. Data Preprocessing

Data preprocessing was performed in Python (v. 2.7). During preprocessing, hashtags symbols were removed, and emoticons were replaced with Emoji 5.0 descriptions (a full list of descriptions and corresponding emojis are available at <https://emojipedia.org/emoji-5.0/> [5.0]). Tweets were then parsed using the snowball parser algorithm in the natural language toolkit (NLTK) Python module, which consists of a suite of well-established tools for natural language processing [32].

Tweet texts were transformed into word vector arrays using the Stanford GloVe (Global Vectors for Word Representation) Common Crawl dictionary (<https://nlp.stanford.edu/projects/glove/>). The GloVe dictionary contains 300 dimensional (300d) numerical vector representations of words derived using an unsupervised algorithm applied to data from website crawl of 840 billion tokens (2.2 million vocabulary words, differing between cased and non-cased words) [33]. To account for unknown words and variable sentence lengths, two 300d vectors were initialized with random weights from a uniform distribution (0,0.1) and added to the dictionary with 'UNK' and 'PAD_TOKEN' keys. Words present in tweets but missing from the GloVe dictionary were assigned the UNK vector, and PAD_TOKENS were appended so all tweets contained the same number of word vectors as the longest tweet. Each word vector array was appended with two binary units, indicating whether the corresponding word from the tweet text originated from a hashtag or emoticon (Figure 1).

To increase the likelihood that tweets contained enough information for inferring labels, tweets were restricted to those that met the following inclusion criteria: 1) Word length > 5 ; 2) contained three or more words from the Stanford GloVe dictionary, excluding words originating from hashtags and emoticons; 3) differed from temporally preceding tweet records by at least one trigram, excluding words originating from hashtags and emoticons.

The North America OpenStreetMap (OSM) data was used to link specific names in Tweets to geolocated nature and non-nature objects already classified in OSM data. The OSM release for March 12, 2019 was downloaded from <http://download.geofabrik.de/> [34]. Records were converted into shapefiles using GDAL (Geospatial Data Abstraction Library) (v. 2.4.0) [35]. A categorical variable or 'flag' was added to the shapefiles, in which records with a nature-related json (JavaScript Object Notation) tag (e.g. natural.*, tourism.picnic site) were assigned a value of 1, records with a nature keyword in the name but without a nature-related tag (e.g. Columbia River High School) were assigned a value of -1, and all other records were assigned a value of 0. Portland and NYC tweet text records were then searched for references to named OSM

objects within 100 km of each respective city center. Search parameters included ignoring upper case and allowing for common abbreviations (e.g. rd for road). Abbreviations are listed in Supplemental Table 3. Each word vector array was then appended with a categorical integer or 'flag' (referred to as OSM flag in Figure 1), with the value of the derived OSM categorical value if the word referenced an OSM object, or 0 otherwise (Figure 1).

Finally, Portland tweets were partitioned into three datasets. Ten thousand tweets were randomly selected and evenly partitioned into development (dev) ($n = 5000$) and test ($n = 5000$) datasets. The remaining Portland tweets ($n = 63,073$) make up the training dataset. NYC tweets that met inclusion criteria ($n = 4850$) collectively make up the fourth dataset used to test generalizability of model inferences.

2.4. Neural Network Model Structure

Models were built, trained, and evaluated using TensorFlow (v. 1.15) [36]. Additional system properties of interest include Windows 10 Operating System and NVIDIA Titan Volta GPUs. The BiLSTM model structure is shown below in Figure 2. Input features consist of a 303-element vector for each LSTM cell: 300 elements for the word vector, and three elements for the binary hashtag, binary emoticon, and categorical OSM flags. Output from the forward and reverse passes of the bi-directional LSTM cells are fed into a fully connected neural network layer, which is followed by n (tunable hyperparameter) additional fully connected layers. The final layer consists of seven nodes with sigmoid activation functions to predict binary labels for the seven classes of interest.

2.4.1. Cost Function

The cost function J_{total} is the unweighted arithmetic mean cost J_c for all c outcomes in a given candidate model.

$$J_{\text{total}} = \frac{1}{m} \sum_{c=1}^m J_c(w, b) \quad \text{Eq. 1.}$$

$$J_c = \frac{1}{n} \sum_{i=1}^n L(\hat{y}_{c,i}, y_{c,i}) \quad \text{Eq. 2.}$$

$$L(\hat{y}_{c,i}, y_{c,i}) = (1 - y_{c,i}) \log(1 - y_{c,i}) - y_{c,i} \log(\hat{y}_{c,i}) \quad \text{Eq. 3.}$$

Where.

$\hat{y}_{c,i}$ = predicted label for label c and sample i .

$y_{c,i}$ = actual label for label c and sample i .

$L(\hat{y}_{c,i}, y_{c,i})$ = loss for outcome c of sample i .

J_c = loss for outcome c .

m = number of outcomes labels.

n = number of records.

Green Tweet BiLSTM Model Structure

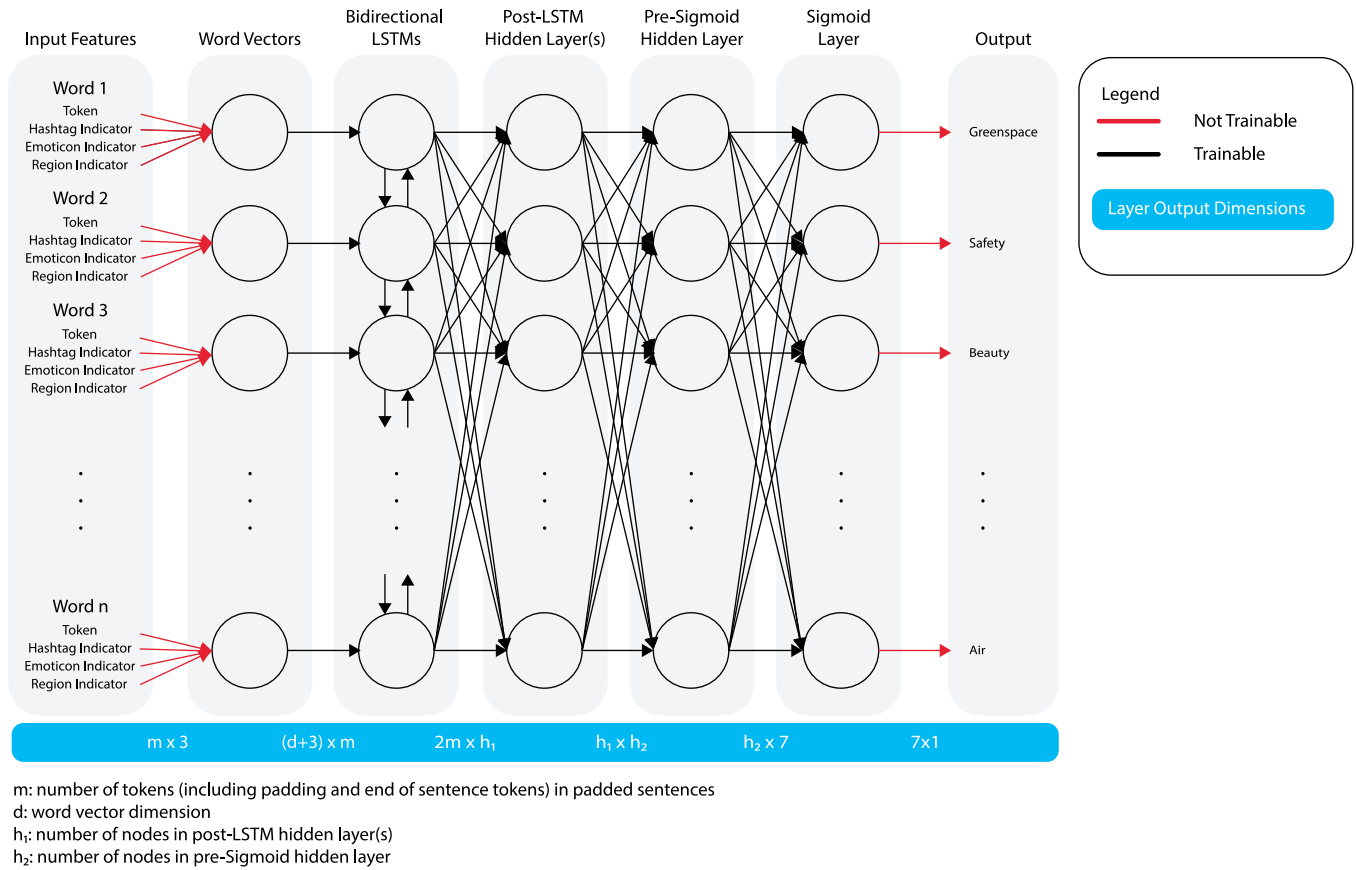


FIGURE 2. Neural Network Model Structure. Input features into BiLSTM cells include word vectors, binary flags for hashtag and emoticon, and a categorical flag for OSM references. BiLSTM cells are followed by several fully connected layers, and finally a layer of sigmoid nodes to predict seven binary class labels. Note that the first layer in the model diagram, transforming words into word vectors, is part of preprocessing and not trainable. The diagram is designed to match the underlying Tensorflow [36] models, which takes an embedding dictionary and words as input and transforms words into vectors as the first step in the Tensorflow session.

2.4.2. Hyperparameter Tuning

Model hyperparameters and candidate hyperparameter values considered during tuning are shown below in Table 2. Model overfitting was addressed using two dropout rates, one for BiLSTM cells and another for the fully connected layers between the BiLSTM cells and the sigmoid layer. Number of epochs was set prior to hyperparameter tuning, and default parameters were used for the Adam Optimizer (an adaptive algorithm used during gradient descent optimization to dynamically adjust the model learning rate). Hyperparameter values were selected which maximized F1Score and minimized cost for the training set without overfitting the dev set.

Once hyperparameter values were selected, the model was trained for 250000 epochs, with dev cost evaluated every 500 epochs. Weights from the epoch with the lowest dev cost were selected as the final model.

2.4.3. Model Evaluation

Model performance metrics include the confusion matrix, precision, recall, F1Score, and MCC for all seven nature-human interactions labels for the train, dev, test, and NYC datasets. The training dataset was used to train model weights, the dev dataset was used for hyperparameter tuning, and the test dataset was used to evaluate model performance. To evaluate the relative contributions of the hashtag, emoticon, and OSM flags, three sensitivity models were created with one of the three feature flags removed from each model. To test for interactions between nature pathway labels, we trained an additional sensitivity model with six output labels instead of seven, excluding the nature output label.

To compare model performance against published methods, F1Score, precision, and recall were calculated for the NYC dataset using the hashtag, and bag of words classification methods described in the introduction section. For hashtag-based

TABLE 2. Candidate Hyperparameter Values Considered During Training and Chosen Hyperparameter Values in the Final Model

Hyperparameter	Selected Value	Tested Values
Learning Rate	9e-05	{1e-5, 5e-5, 7.5e-5, 9e-5, 1e-4, 1.1e-4, 1.25e-4,}
Minibatch Size	64	{32, 64, 128, 256}
Dropout Rate LSTM Layers	0.1	{0, 0.1, 0.25, 0.75, 0.9}
Dropout Rate Post LSTM	0.5	{0.1, 0.5, 0.9}
# Hidden Post-LSTM Layers	2	{1, 2, 3}
# Hidden Nodes in Post-LSTM Layer	256	{32, 64, 128, 256}
# Hidden Nodes in Pre-sigmoid Layer	14	{14, 28, 56}
Activation Function	tangent	{tangent, ReLu**, leaky ReLu**}
# epochs	120000	NA
momentum rate of Adam Optimizer	0.9	NA
RMSProp of Adam Optimizer	0.999	NA
small constant of Adam Optimizer	1.00E-08	NA
# BiLSTM cells	$n + 1^*$	NA

Values with NA in the Tested Values column were set prior to model tuning.

*n consists of the number of words in the longest tweet among the training, dev, test, and NYC datasets.

**ReLU: Rectified Linear Unit, a ramp function where negative input values output either 0 (ReLU) or a small positive gradient (leaky ReLU).

classification, precision was assumed to be 100%, which is an overly optimistic scenario but allows for best case comparisons against the trained BiLSTM.

2.5. Model inference

The trained BiLSTM model was used to generate labels for the collection of 2017 tweets originating from towns and cities in the United States and Canada with populations greater than 10,000 ($n = 7708$). To demonstrate the feasibility of near real time analytics and utilizing other social media data streams, the model was also applied to the social media site ‘Meetup’ events data stream for the same set of US and Canada towns and cities. Meetup records were streamed in real time via Kafka (a low-latency platform for ingesting and transferring data in real time), to a Python docker container for pre-processing, followed by model inference via Tensorflow serving (a platform for deploying Tensorflow models in a production pipeline). Labels for both Twitter and Meetup records were joined with OpenStreetMap records and stored in a backend Geoserver (v. 2.14.4) and accessed via a spatial-temporal website GUI, built using React (a JavaScript library for responding to complex user input and dynamically updating website content, v. 16.7).

3. RESULTS

3.1. Labeled Data Descriptive Statistics

Descriptive statistics for the train, dev, test, and NYC datasets are shown in Table 3, stratified by nature labels. Additional descriptive statistics are available in Supplemental Table 4. Percent tweets with a positive nature label in the train, test, and dev

datasets (tweets originating from Portland) range from 58.04 to 58.74, while percent tweets in the NYC dataset is 49.35. Percent tweets with positive safety labels are greater for tweets with positive nature labels, and between Portland and NYC datasets. Similarly, percent tweets with positive exercise labels is greater for Portland compared to NYC datasets and positive compared to negative nature labels, respectively. Percent tweet with positive social labels are greater for NYC compared to Portland and positive compared to negative nature labels. Collectively, these summary statistics suggest self-reported nature utilization levels differ between Portland and NYC, and self-reported behaviors differ both between nature and non-nature tweets.

3.2. Model Precision and Recall

Precision and recall across all datasets are shown below in Table 4. Precision for nature, safety, aesthetic, exercise, and social labels range from 0.81 to 0.92, while precision for stress and air labels range from 0.77 to 0.89 and 0.58 to 0.95. Differences in precision between train, test, and NYC datasets are within 0.06 for nature, safety, aesthetic, social, and stress labels, while exercise and air precision is 12% and 37% lower, respectively, in the NYC compared to the training dataset.

Recall across all outcomes and datasets range from 0.33 to 0.91. Recall is above 0.70 across all datasets for nature, safety, aesthetic, and exercise. Recall for the social label ranges from 0.50–0.72, while recall for the air label ranges from 0.45 to 0.72. The average absolute difference in recall is 9% and 14% for test and NYC datasets, respectively, compared to the training dataset, with greatest differences in safety and exercise.

TABLE 3. Descriptive Statistics for Train, Test, Dev, and NYC Datasets, Stratified by Nature Label

	Nature Label = Yes				Nature Label = No			
	Train	Test	Dev	NYC	Train	Test	Dev	NYC
Sample Size	36541	2835	2937	2397	26422	2165	2063	2460
Percent	58.04	56.70	58.74	49.35	41.96	43.30	41.26	50.65
Yes Label								
Frequency (%)								
Safety	7.77	9.38	8.07	12.22	2.50	3.19	2.62	8.29
Beauty	6.74	7.97	7.29	6.13	1.08	1.43	1.50	1.50
Exercise	10.07	10.69	10.49	6.76	0.91	0.88	0.63	0.93
Social	13.40	15.66	15.73	22.11	4.52	4.85	5.14	6.46
Stress	1.44	1.52	1.16	1.29	0.46	0.37	0.58	0.41
Air	1.12	1.06	1.40	0.25	0.27	0.37	0.15	0.16
Emoticon	10.08	10.44	12.39	9.85	5.86	5.91	6.20	5.49
Frequency (%)								
Hashtag	23.89	24.02	24.65	29.37	18.18	19.08	17.45	26.83
Frequency (%)								
OSM Record	2.80	2.72	2.52	4.34	7.57	7.81	6.79	21.67
Frequency (%)								
Mean Tweet	18.44	18.65	18.83	18.06	19.62	19.78	19.38	19.02
Length (words)								

TABLE 4. Precision and Recall Scores for the Trained Model

	Precision				Recall			
	Train	Dev	Test	NYC	Train	Dev	Test	NYC
Nature	0.89	0.86	0.85	0.83	0.91	0.88	0.89	0.81
Safety	0.82	0.83	0.88	0.86	0.78	0.72	0.76	0.47
Aesthetic	0.86	0.82	0.87	0.92	0.80	0.73	0.69	0.71
Exercise	0.92	0.81	0.82	0.80	0.82	0.72	0.73	0.62
Social	0.81	0.83	0.83	0.89	0.72	0.60	0.62	0.50
Stress	0.86	0.77	0.87	0.89	0.33	0.37	0.39	0.39
Air	0.95	0.92	0.77	0.58	0.66	0.52	0.45	0.70

3.3. Model Generalizability

Comparing model performance in Portland and New York datasets provides insight into how well the model generalizes beyond the geographical extent of the training dataset (based on tweets originating from Portland only). Precision is similar for the test and NYC datasets, except for the air label. Aesthetic, social, and stress labels have greater precision in the NYC dataset compared to the test dataset, while differences in nature, safety, and exercise label precision between the two datasets are within 2%. However, recall is significantly lower in the NYC compared to the test dataset for four out of seven labels.

3.4. Input Feature Sensitivity Analysis

Figure 3 shows precision vs. recall for all labels, sensitivity models, and datasets. X and y axis limits differ to highlight

intra label differences. Precision vs. Recall graphs with fixed axes are available in [Supplemental Figure 1](#), and performance scores for all models in [Figure 3](#) are available in [Supplemental Tables 5–9](#). Precision is strongly positively correlated with recall for nature ($r = 0.8$) and exercise ($r = -0.9$) and negatively correlated with recall for safety ($r = -0.71$). Removing the hashtag input feature has the greatest impact on exercise, aesthetic, and stress. Removing the emoticon input feature has the greatest impacts on safety and stress labels (-0.09 and -0.12 change in NYC F1Score, respectively). Removing the OSM location input feature has the greatest impact on safety and stress (-0.07 and -0.10 , respectively). Removing all three flag input features has the greatest impact on safety and stress (-0.08 and -0.16 change in NYC F1Score, respectively). Removing the nature outcome label from the model has the greatest impacts on safety and stress (-0.12 and -0.14 change in NYC F1Score, respectively). Adding greenspace as an out-

TABLE 5. Comparison of Word List, Hashtag-based, and LSTM Performance in the NYC Dataset

	F1Score					Precision			Recall		
	Word	Hash	LSTM	% Word*	% Hash**	Word	Hash	LSTM	Word	Hash	BiLSTM
Nature	0.66	0.45	0.82	24	82	0.49	1	0.83	1.00	0.29	0.81
Safety	0.32	0.48	0.60	88	25	0.55	1	0.86	0.23	0.32	0.47
Aesthetic	0.48	0.52	0.80	67	54	0.36	1	0.92	0.71	0.35	0.71
Exercise	0.43	0.50	0.70	63	40	0.31	1	0.8	0.70	0.33	0.62
Social	0.35	0.43	0.64	83	49	0.59	1	0.89	0.25	0.27	0.50
Stress	0.03	0.42	0.54	1700	29	0.02	1	0.89	0.12	0.27	0.39
Air	0.17	0.33	0.64	276	94	0.09	1	0.58	1.00	0.20	0.70

LSTM F1Score is between 0.12 and 0.47 units greater than Word and Hash models.

*Percent Improvement of LSTM model relative to bag of words classification.

**Percent Improvement of LSTM model relative to hashtag-based classification.

TABLE 6. Number of Twitter Labels for Five Select Cities in the US and Canada

Name	Nature	Safety	Exercise	Beauty	Social	Stress	Air
Vancouver, BC	137072	13963	10090	8595	30174	7538	2410
Detroit, MI	108223	11047	3813	4499	30168	7689	2454
Orlando, FL	185017	17888	7288	8288	48582	8396	2896
Dallas, TX	275045	27369	11016	13980	79388	16108	5828
Wichita, KS	34850	4422	1457	1332	8252	2256	597

Tweet numbers correspond to tweets continuously collected from the Twitter data stream throughout 2017. Note that collected tweets are a representative sample rather than a complete set of records.

come label improves NYC F1Score for other labels on average by 0.06.

3.5. Model Comparison to Alternative Methods

Model performance for the BiLSTM, hashtag-based, and bag of word labeling methods are shown in Table 5. For all labels, BiLSTM model F1Scores are between 0.12 to 0.31 units greater than the best score from the comparative labeling methods. BiLSTM model advantage is greatest for the air label and smallest for safety and stress.

3.6. Model Inference

An example screenshot of model application and inferences derived from 2017 tweets and the real time Meetup events datastream is shown below in Figure 4. Labels were successfully derived for more than 21 million nature-related tweets. For example, during 2017 there were 275,045 nature-related tweets from Dallas, TX (Table 6). Note that these are total number of nature labels, unadjusted for population size. From June 10th to November 21st, 2019, the Tensorflow server inferred labels for more than 12,000 Meetup events per day. Events were processed within 5 to 15 minutes of initial post, depending on server load (virtual machine resources allocated to Tensorflow serving consisted of one NVIDIA Titan Volta GPU, 32GB ram, and four virtual CPU threads).

Tweets were successfully joined to georeferenced OSM records using automated entity name recognition (left side of the graphical user interface in Figure 4). Similarly, the right side of the graphical user interface (GUI) demonstrates that an automated pipeline can capture fine temporal variations in self-reported behaviors at specific locations, in this case Forest Park in Portland, OR. Readers can access the pilot project GUI and corresponding records at spatialhealthsocialmedia.com (planned availability through 2022).

4. DISCUSSION

4.1. Rationale for Model Development

Urban nature can promote behaviors which influence health and wellness. Social media is a temporally fine and contextually rich data source with self-reported behaviors that can be used to characterize urban-nature interactions, as well as other complex behavioral and environmental conditions that are not easy to measure using traditional data sources. Current methods for automated classification of social media records have limited precision and/or recall, with little potential for improvement. We developed a BiLSTM model to test the potential of deep learning for improving classifications. This method can be applied to all social media data to better understand a range of human-environment interactions.

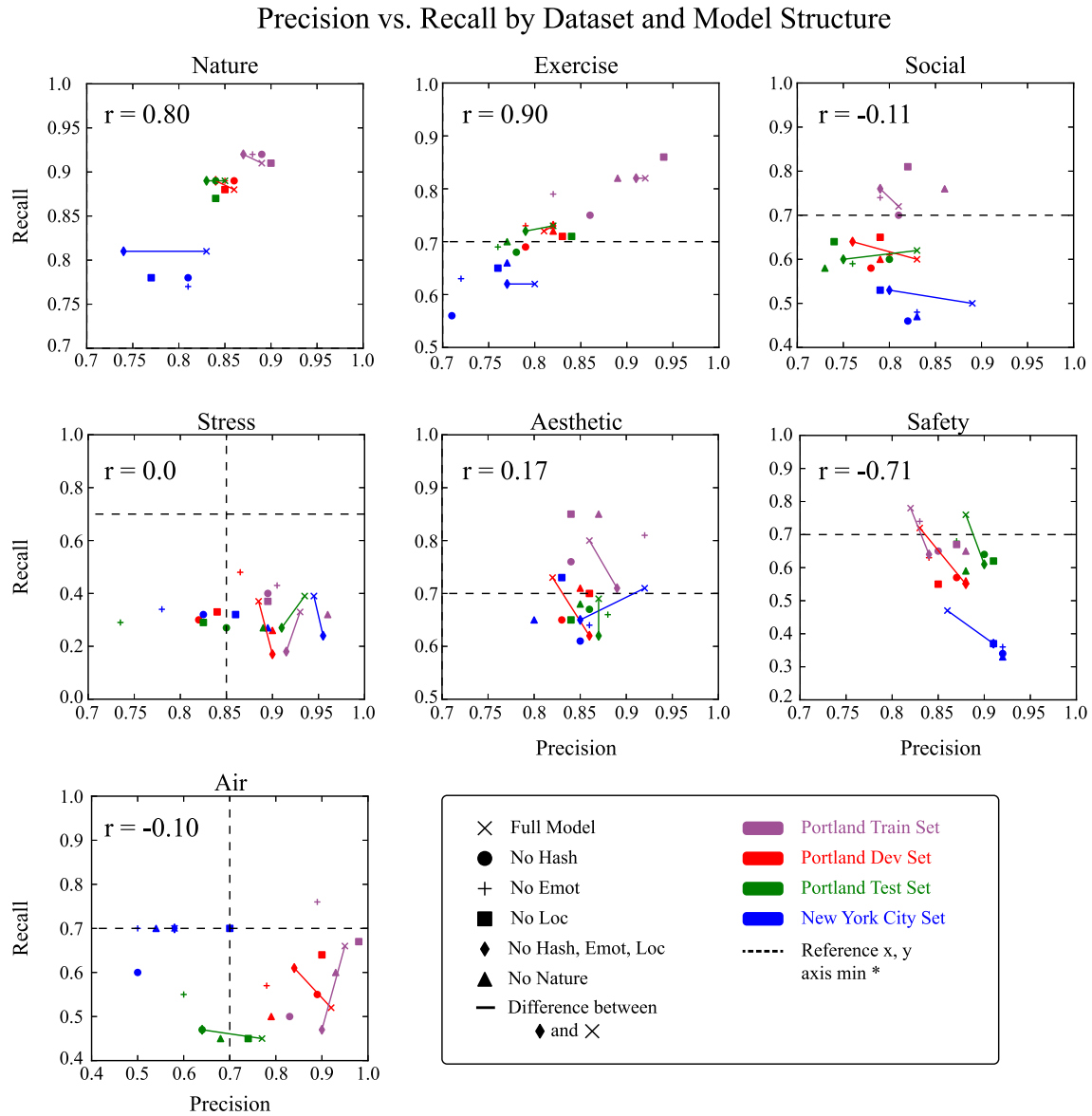


FIGURE 3. Precision vs. Recall for Each Outcome, Dataset, and Sensitivity Model. Marker symbols correspond to LSTM model structure, while colors correspond to dataset. Abbreviations: No Hash—no hashtag indicator in model. No Emot—no emoticon indicator in model. No Loc—no OSM flag in model. No Nature—no nature outcome in model. *Reference lines correspond to the precision and recall performance space for the nature label.

For all labels of interest, F1Scores in the developed BiLSTM model are noticeably greater than alternative methods, even under the assumption that hashtag-based classification has 100% precision. This is the first attempt to develop a deep learning model for nature-related tweet classification. We chose to focus on identifying human-nature interactions as a test case because it represents an understudied area that is difficult to measure with traditional data and there are a diverse range of pathways that test the application of our modelling approach in different domains. Advantages of a deep learning model approach are likely to increase further as larger datasets,

deeper neural network models, and alternative neural network architectures are evaluated.

4.2. Model Performance

Comparison of precision between NYC and Portland test datasets suggest model precision generalizes well across geographical extents: NYC precision is greater than or within 2% of Portland test precision for all outcome labels except air. This is partly attributable to the OSM feature flag, which increased nature precision in the NYC dataset by 9%. The

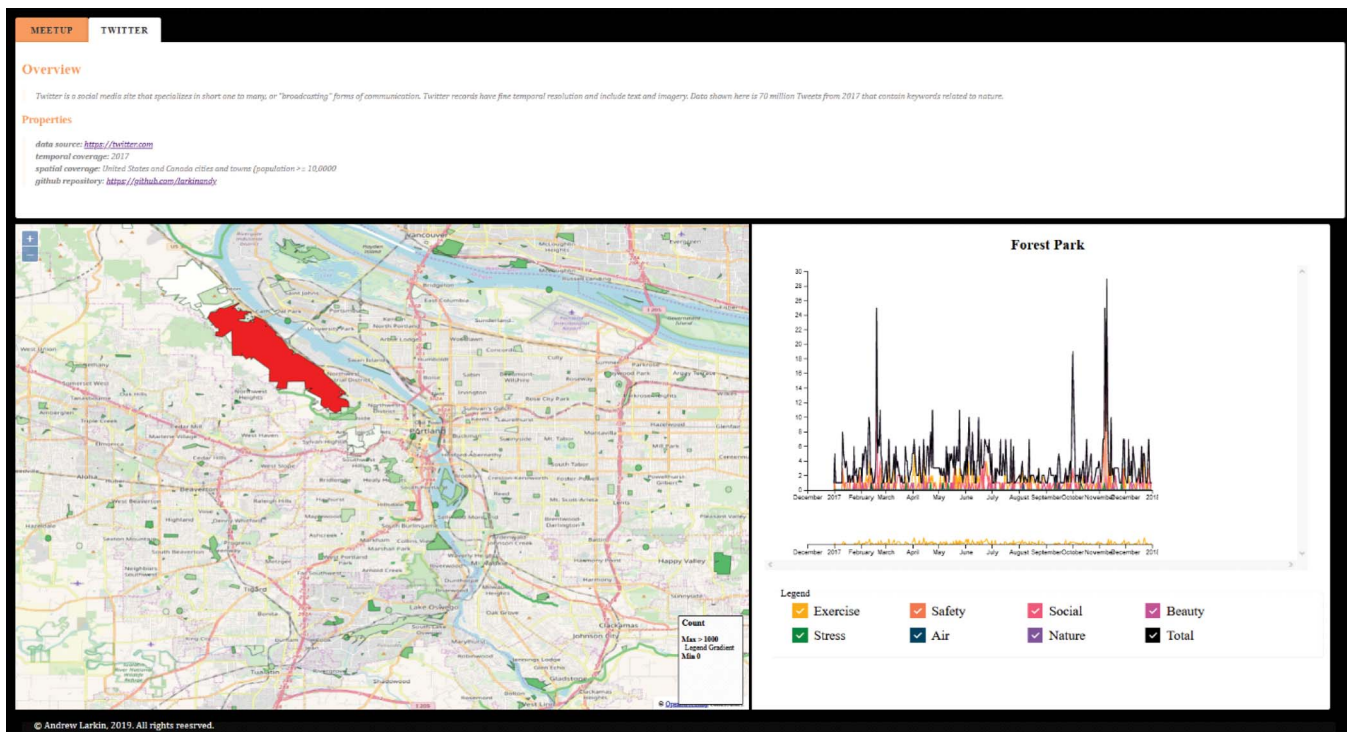


FIGURE 4. Spatial and Temporal Distribution of Automated Large-Scale Model Inferences.

OSM flag increases the model's ability to classify tweets with named entities that weren't in the training dataset, possibly through reducing the weight of the literal inference of a name (e.g. Sunset Park) in favor of greater weight on the surrounding context (e.g. apartments for rent in Sunset Park).

Recall in the NYC dataset was significantly lower than the Portland test dataset, particularly for safety and social outcomes. This is partly attributable to behaviors encountered in the NYC dataset, but not the training dataset. For example, nearly 20% of the false negative safety predictions in our NYC dataset were describing the 2017 shooting of United States Senator Scalise at a baseball field. While our model correctly predicted safety labels for multiple descriptions of violent actions in parks, we hypothesize the word baseball likely led the model to predict the word shoot was being used in relation to sports (e.g. shooting hoops), and the model incorrectly predicted the safety outcome as false. In a follow up sensitivity analysis, when 50% of the Scalise shooting-related tweets were added to the training dataset the model correctly predicted the remaining 50% as positive for safety. Together, these results suggest that the model's ability to accurately predict labels for contexts similar to those in the training dataset generalizes well across geographical extents, but the ability to capture behaviors outside of the contextual coverage space provided by the training dataset is limited.

Increasing model recall in the near term is highly probable. Recall of all model-outcome combinations in Figure 3 are

moderately correlated with the number of positive labels in the training dataset ($r = 0.34$). Most notably, precision and recall for nature, which contains 4x greater number of positive labels in the training dataset than any other outcome, were highly correlated ($r = 0.80$) and greater than 80% across all datasets. It's also noteworthy that less than 1% of the training dataset contains positive labels for stress and air, which provides few positive examples for the model to learn from. Increasing the number of positive labels in the training dataset is very likely to increase recall and, to a lesser extent, precision. Recall can also be improved by integrating data records across geographical regions, providing greater diversity of described behaviors, named entities, and local dialects within the training dataset.

4.3. Limitations and Future Directions

There are several limitations to the current model and modelling approach that should be highlighted. First, OSM records include unverified contributions from the OSM community, and the quality of feature naming and tag-labeling can vary. Second, cultural references to nature objects can often inflate false positive labels. For example, tweets referencing the song 'Supermarket Flowers' by Ed Sheeran led to 41 false positive nature labels in the Portland test dataset within a very short amount of time surrounding the song's release date. Term Frequency-Inverse Document Frequency and other Natural Language Processing methods can potentially identify

temporary spikes in these cultural references. In our comparison between methods, we did not include POS tagging, topic modeling, or other NLP-based methodologies. In a related study, the authors evaluated the Portland dataset using POS and topic modeling. While the study is not published at the time of this writing, scripts and supplemental materials are available at the GitHub repository https://github.com/larkinandy/Portland_UrbanNature_Twitter. Third, model performance for behavior labels is slightly greater when concomitantly predicting nature (average F1Score improves by 0.16). While the improvement is slight at present, it's possible that future models with improved precision and recall will be able to better capture nature-behavior interactions. Integrating convolutional layers (e.g. pooling subsets of trigram word vectors), attention-retention layers, and similar model structures with high interpretability can potentially better capture and infer nature-behavior interactions. Fourth, in this model application we chose to focus on Twitter datasets. However, we believe this model is a prime candidate for transfer learning to generate models and inferences for other text-based social media platforms (e.g. Meetup, Reddit, etc.). Social media streams have distinct geographical and demographic distributions and capturing multiple social media streams is essential for accurate population inferences. Each social media stream has unique dataset properties, but there is significant potential for the development of transfer learning models. For example, here we demonstrated that the trained BiLSTM model can infer labels for Meetup events, albeit without validation.

In this model we chose to focus specifically on urban nature, as this where we have expert domain knowledge for optimizing model labels and identifying potential biases related to the subject matter. This is just a first step, as human-environment interactions are influenced by many additional environmental features (e.g. weather) and non-urban land use properties (e.g. nearby countryside, water bodies, mountains). Future models would ideally integrate domain knowledge from multiple fields to capture a more holistic set of human-environmental interactions.

5. CONCLUSIONS

We developed the first deep learning model for classifying human-nature interactions from Twitter records, integrating OSM data to provide additional contextual information. In our most stringent test dataset, model performance improved F1Scores between 24 to 94% for class labels compared to hashtag-based and bag of word classification methods. The model was able to capture a range of complex pathways (e.g. safety, exercise, beauty) frequently observed in primary data collection studies. We also demonstrated how this model could be applied in near-real time to collect, analyze and display human-nature data from Twitter and Meetup, which can inform future urban planning, sustainability and environmental

health research. This modelling approach has significant potential to be expanded to other social media data streams and questions.

FUNDING

This work was performed without funding.

REFERENCES

- [1] Shanahan, D. *et al.* (2016) Health benefits from nature experiences depend on dose. *Sci. Rep.*, 6, 28551.
- [2] Lovasi, G. *et al.* (2013) Neighborhood safety and green space as predictors of obesity among preschool children from low-income families in New York City. *Prev. Med.*, 57, 189–193.
- [3] Beyer, K., Kaltenbach, A., Szabo, A., Bogar, S., Nieto, F. and Malecki, K. (2014) Exposure to neighborhood green space and mental health: evidence from the survey of the health of Wisconsin. *Int. J. Environ. Res. Public Health*, 11, 3453–3472.
- [4] Astell-Burt, T., Feng, X. and Kolt, G. (2014) Is neighborhood green space associated with a lower risk of type 2 diabetes? Evidence from 267,072 Australians. *Diabetes Care*, 37, 197–201.
- [5] Bratman, G., Hamilton, J. and Daily, G. (2012) The impacts of nature experience on human cognitive function and mental health. *Ann. N. Y. Acad. Sci.*, 1249, 118–136.
- [6] Takano, T., Nakamura, K. and Watanabe, M. (2002) Urban residential environments and senior citizens' longevity in megacity areas: the importance of walkable green spaces. *J. Epidemiol. Community Health*, 56, 913–918.
- [7] Van den Berg, A., Maas, J., Verheij, R. and Groenewegen, P. (2010) Green space as a buffer between stressful life events and health. *Soc. Sci. Med.*, 70, 1203–1210.
- [8] Maas, J., Van Dillen, S., Verheij, R. and Groenewegen, P. (2009) Social contacts as a possible mechanism behind the relation between green space and health. *Health Place*, 15, 586–595.
- [9] Toftager, M. *et al.* (2011) Distance to green space and physical activity: a Danish national representative survey. *J. Phys. Act. Health*, 8, 741–749.
- [10] Sabbion, P. (2018) *Green Streets Social and Aesthetic Aspects. In Nature Based Strategies for Urban and Building Sustainability.* Butterworth-Heinemann, Oxford.
- [11] Szilagyi, B., Zaharia, D. and Dănilă-Guidea, S. (2015) The landscape of parks in the municipality of baia mare from an aesthetic-urban perspective. *Sci. Pap.-Ser. B Hortic.*, 59, 409–417.
- [12] Rasidi, M., Jamirsah, N. and Said, I. (2011) Urban green space design affects urban residents' social interaction. *Soc. Behav. Sci.*, 68, 464–480.
- [13] Blum, J. (2017) Contribution of ecosystem services to air quality and climate change mitigation policies: the case of urban forests in Barcelona, Spain. In Blum, J. (ed) *Urban Forests*. Apple Academic Press, Palm Bay, FL.
- [14] Franchini, M. and Mannucci, P. (2018) Mitigation of air pollution by greenness: A narrative review. *Eur. J. Intern. Med.*, 55, 1–5.
- [15] Maas, J., Spreeuwenberg, P., Van Winsum-Westra, M., Verheij, R., Vries, S. and Groenewegen, P. (2009) Is green space in the

- living environment associated with people's feelings of social safety? *Environ. Plan. A*, 41, 1763–1777.
- [16] Barbosa, O. *et al.* (2007) Who benefits from access to green space? A case study from Sheffield. *UK. Landsc. Urban Plan.*, 83, 187–195.
- [17] Neal, S., Bennett, K., Jones, H., Cochrane, A. and Mohan, G. (2015) Multiculture and public parks: Researching super-diversity and attachment in public green space. *Popul. Space Place*, 21, 463–475.
- [18] Shackleton, C. and Blair, A. (2013) Perceptions and use of public green space is influenced by its relative abundance in two small towns in South Africa. *Landsc. Urban Plan.*, 113, 104–112.
- [19] Hinds, J. and Sparks, P. (2008) Engaging with the natural environment: The role of affective connection and identity. *J. Environ. Psychol.*, 28, 109–120.
- [20] Rossi, S., Byrne, J. and Pickering, C. (2015) The role of distance in peri-urban national park use: Who visits them and how far do they travel? *Appl. Geogr.*, 63, 77–88.
- [21] Lubin, J. (2012) The 'occupy' movement: emerging protest forms and contested urban spaces. *Berkeley Plan. J.*, 25, 184–197.
- [22] Rideout, V. and Robb, M. (2018) *Social media, social life: Teens reveal their experiences*. San Franc, CA Common Sense Media, San Francisco, CA.
- [23] Perrin, A. and Anderson, M. (2018) Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center*, Washington, DC.
- [24] Schwartz, A., Dodds, P., O'Neil-Dunne, J., Danforth, C. and Ricketts, T. (2019). Exposure to urban parks improves affect and reduces negativity on Twitter. *People Nat.*, 0, 1–10.
- [25] Palomino, M., Taylor, T., Göker, A., Isaacs, J. and Warber, S. (2016) The online dissemination of nature–health concepts: Lessons from sentiment analysis of social media relating to 'nature-deficit disorder'. *Int. J. Environ. Res. Public. Health*, 13, 142.
- [26] Graves, A., Fernández, S. and Schmidhuber, J. (2005) Bidirectional LSTM networks for improved phoneme classification and recognition. In Duch, W., Kacprzyk, J., Zadrozny, S. (eds) *Artificial Neural Networks: Formal Models and Their Applications-ICANN 2005*. Springer, Berlin, Heidelberg.
- [27] Bérard, A., Pietquin, O., Servan, C. and Besacier, L. (2016) *Listen and translate: A proof of concept for end-to-end speech-to-text translation*. *arXiv*, 1612, 01744.
- [28] Zhou, X., Wan, X. and Xiao, J. (2016) Attention-based LSTM network for cross-lingual sentiment classification. *Proceedings of the 2016 conference on empirical methods in natural language processing*, Austin, TX, 1–5 November, pp. 247–256. Association for Computational Linguistics, Stroudsburg, PA.
- [29] Plank, B., Søgaard, A. and Goldberg, Y. (2016) Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, August 7–12, pp. 412–418. Association for Computational Linguistics, Stroudsburg, PA.
- [30] Van Rosum, G. (2007) Python Programming Language. *Proceedings of the USENIX annual technical conference*, Santa Clara, CA, 20–22 June, pp. 36. USENIX Association, Berkeley, CA.
- [31] DuBois, P. (2004) *MySQL Language Reference*. MySQL Press, Cupertino, CA.
- [32] Loper, E. and Bird, S. (2004) NLTK: the natural language toolkit. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Barcelona, Spain, 21–26 July, pp. 7. Association for Computational Linguistics, Stroudsburg, PA.
- [33] Pennington, J., Socher, R. and Manning, C. (2014) Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar, 25–29 October, pp. 1532–1543. Association for Computational Linguistics, Stroudsburg, PA.
- [34] Haklay, M. and Weber, P. (2008) Openstreetmap: user-generated street maps. *IEEE Pervasive Comput.*, 7, 12–18.
- [35] Warmerdam, F. (2008) The geospatial data abstraction library. In Michael, G. (ed) *Hall, B. Open source approaches in spatial data handling*, Springer-Verlag, Berlin Heidelberg.
- [36] Abadi, M. *et al.* (2016) Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, GA, 2–4 November, pp. 265–283. USENIX Association, Berkeley, CA.