Urban Planning (ISSN: 2183-7635)
2016, Volume 1, Issue 2, Pages 114-127
Doi: 10.17645/up.v1i2.617

Article

Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm

Bernd Resch [1,2,3,*], Anja Summa [4], Peter Zeile [5] and Michael Strube [6]

[1] Department of Geoinformatics - Z_GIS, University of Salzburg, 5020 Salzburg, Austria; E-Mail: bernd.resch@sbg.ac.at
[2] Center for Geographic Analysis, Harvard University, MA 02138 Cambridge, USA; E-Mail: bresch@fas.harvard.edu
[3] Institute of Geography (GIScience), Heidelberg University, 69120 Heidelberg, Germany; E-Mail: bernd.resch@uni-heidelberg.de
[4] Department of Computational Linguistics, Heidelberg University, 69120 Heidelberg, Germany; E-Mail: summa@cl.uni-heidelberg.de
[5] Computergestützte Planungs und Entwurfsmethoden (CPE), University of Kaiserslautern, 67663 Kaiserslautern, Germany; E-Mail: zeile@rhrk.uni-kl.de
[6] NLP Group, Heidelberg Institute for Theoretical Studies gGmbH, 69118 Heidelberg, Germany; E-Mail: michael.strube@h-its.org

* Corresponding author

Submitted: 8 February 2016 | Accepted: 9 June 2016 | Published: 5 July 2016

Abstract
Traditional urban planning processes typically happen in offices and behind desks. Modern types of civic participation can enhance those processes by acquiring citizens' ideas and feedback in participatory sensing approaches like "People as Sensors". As such, citizen-centric planning can be achieved by analysing Volunteered Geographic Information (VGI) data such as Twitter tweets and posts from other social media channels. These user-generated data comprise several information dimensions, such as spatial and temporal information, and textual content. However, in previous research, these dimensions were generally examined separately in single-disciplinary approaches, which does not allow for holistic conclusions in urban planning. This paper introduces *TwEmLab*, an interdisciplinary approach towards extracting citizens' emotions in different locations within a city. More concretely, we analyse tweets in three dimensions (space, time, and linguistics), based on similarities between each pair of tweets as defined by a specific set of functional relationships in each dimension. We use a graph-based semi-supervised learning algorithm to classify the data into discrete emotions (happiness, sadness, fear, anger/disgust, none). Our proposed solution allows tweets to be classified into emotion classes in a multi-parametric approach. Additionally, we created a manually annotated gold standard that can be used to evaluate *TwEmLab*'s performance. Our experimental results show that we are able to identify tweets carrying emotions and that our approach bears extensive potential to reveal new insights into citizens' perceptions of the city.

Keywords
integrated space-time-linguistics methodology; participatory planning; semi-supervised learning; Twitter emotions

Issue
This article is part of the issue "Volunteered Geographic Information and the City", edited by Andrew Hudson-Smith (University College London, UK), Choon-Piew Pow (National University of Singapore, Singapore), Jin-Kyu Jung (University of Washington, USA) and Wen Lin (Newcastle University, UK).

1. Introduction

Traditional urban planning processes typically take place in offices and behind desks, and thus oftentimes neither fully comply with citizens' needs nor sufficiently account for neogeographic and Web 2.0 phenomena like participatory planning or online participation (Brenner, Marcuse, & Mayer, 2012). This is increasingly problematic as citizen participation initiatives become more demanding and clearly articulate their claim for participation in urban planning and decision-making processes. The recent developments mentioned above are highly suitable for assessing citizens' subjective emotions and observations, which are a key element in participatory planning (Nold, 2009). In this context, participatory sensing approaches like "People as Sensors", Collective Sensing and Volunteered Geographic Information (VGI) (Resch, 2013) can undoubtedly play a key role, but their potential has not been fully exhausted so far.

These citizen-centric approaches are critical for the future of urban planning because the weighting process of all public and private interests is one of the core elements of urban planning (Zeile, Resch, Exner, & Sagl, 2015). It ideally considers all public and private parties and minimises conflicts to achieve optimal planning results, preferably for all citizens. Thus, all available information and knowledge sources should be considered in the planning process (Pahl-Weber, Ohlenburg, Seelig, von Bergmann, & Schäfer, 2013).

The sources of user-generated data introduced above are therefore potentially of significant interest for urban planning processes. In fact, they have been used in a variety of disciplines throughout the last decade, ranging from urban planning and sociology to geoinformatics, computer science and computational linguistics. This is because these data inherently cover a range of information dimensions such as, for instance, spatial and temporal information, as well as the textual content. In previous research these dimensions were generally examined separately in single-disciplinary approaches. As such, text analysis, geospatial interpolation, time series analysis, etc. have not been combined into a single joint method. However, using such separate research approaches severely limits the significance of the results as no holistic conclusions can be drawn for urban planning (see Section 2).

A further issue in capitalising on user-generated data such as social media posts in urban planning is that previous approaches do not work reliably because they have been designed for edited text. Examples of these approaches include Capdevila, Arias and Arratia (2016), Kouloumpis, Wilson and Moore (2011), or Hauthal and Burghardt (2013). These previous approaches do not perform well with social media posts like Twitter tweets as these data are characterised by a higher level of uncertainty and dimensionality (Steiger, Resch, & Zipf, 2015). More concretely, social media posts contain a large portion of slang words, abbreviations, emoticons, irregular punctuation, "yoof speak", or other words that cannot be found in standard dictionaries, which most previous approaches work with (Eisenstein, 2013).

In consequence, new approaches need to be found to analyse user-generated text content. Rather than analysing text in traditional ways, such as like rule-based methods, string comparison, word-matching, or phrase detection, more intelligent ways have to be designed to reliably analyse social media posts. In this context, self-learning systems (neural networks, semi-supervised labelling mechanisms, etc.) seem to be the most promising approaches (Eisenstein, 2013). This shift towards more complex text analysis algorithms also necessitates close collaboration between researchers from urban planning, geoinformatics, and computational linguistics.

This paper introduces a citizen-centric urban planning approach that uses tweets to assess citizens' perceptions of the city and associated emotions in an interdisciplinary manner. More precisely, we extract emotions from tweets in geo-space, time, and linguistic space in a semi-supervised learning algorithm by labelling posts, i.e. by assigning a distinct emotion class (see next paragraph) to each post. Therefore, we leverage the concept of similarity, which exists in all three dimensions. Our proposed solution, *TwEmLab* (Twitter Emotion Labeller), constitutes a full-fledged implementation pipeline that allows for the classification of tweets into emotional classes in a multi-parametric approach. Our experimental results show that emotions can be conditionally detected in an integrated space, time, and linguistics method (validated through a gold standard) and that the approach can potentially significantly enhance urban planning processes. In contrast to numerous previous approaches, our research does not aim to use conventional ways of assessing emotions or purely map them (see Section 2). Thus, our paper does not deal with the general topic of emotion mapping, but rather presents a specific approach for extracting emotions from social media for use in urban planning. Furthermore, our integrated space-time-linguistics approach goes beyond previous research methods, which have oftentimes been presented as "spatio-temporal analysis", while merely being methods for emotion extraction and subsequent spatial or temporal analysis.

As a basis for extracting emotions we use a modified version of the emotion model by Ekman and Friesen (1971), which defines six basic emotions. However, recent research found that two pairs of emotions can be merged into only two emotions due to their high similarity. This results in four basic emotions: happiness, anger (including disgust), sadness, and fear (Jack, Garrod, & Schyns, 2014). In addition, our research defines the class "none" (no emotion is present

or can be unambiguously detected in a tweet). Furthermore, we use the subdivision of these basic emotions by Shaver, Schwartz, Kirson and O'Connor (1987), which assigns more granular emotions to the four emotion classes. The structure of this paper is as follows: This introduction is followed by a section on related work in emotion detection in social media posts and citizen-centric urban planning. Thereafter, we present our approach from a theoretical viewpoint, i.e., the process of generating a set of labelled tweets from unlabelled ones. Section 4 then lays out the case study and our results, before Section 5 presents the evaluation of our results together with a discussion of the approach and the results. Finally, the paper ends with a set of key conclusions.

2. Related Work

Our presented approach addresses the overarching topic of citizen-centric urban planning, for which we concretely developed a method for extracting emotions from user-generated data, leveraging the concept of "similarity" in three dimensions (geo-space, time, linguistics). The following paragraphs describe related work in these areas.

2.1. Citizen-Centric Urban Planning

Jane Jacobs was one of the pioneers of a bottom-up and citizen-centric planning approach (Jacobs, 1961). The central questions are: *How is it possible to integrate all heterogeneous interests into the planning process? How can citizens' perceptions of urban spaces be measured? How can new technological approaches improve the entire process?* The Urban Emotions approach addresses these questions by using "human sensor" data, generated by social media, wearable sensor technology, and participatory sensing approaches, to develop a method set that creates a new point of view, viewing the "city as an organism" (Resch, Summa, Sagl, Zeile, & Exner, 2015). This approach is clearly influenced by the work of (Castells, 1996). Batty et al. (2012) state that, effectively, only citizens can make a city truly "intelligent" (in contrast to technologically driven understandings of Smart Cities), where "collective sensors" (e.g. social media channels or the cell phone network) are used to create a better understanding of humans' interactions and mobility in cities. Derived spatial, temporal, and spatio-temporal patterns help to identify urban processes and to characterise special social-cultural movements and developments.

2.2. Emotion Mapping

Emotion mapping is an emerging way of collecting and visualising citizens' feelings and perceptions. This field of research has its origins in the 1970s and tries to explain the relationship between the perception of the natural and the built environment (R. M. Downs & Meyer, 1978). In a cartographic representation, which is called "mental maps" or "cognitive maps", the subjective perception of people in (urban) space segments are visualised (R. Downs & Stea, 1974). The "Image of the City" describes the concepts of a cognitive representation of space: "We are not simply observers of this spectacle, but are ourselves a part of it, on the stage with the other participants. Nearly every sense is in operation, and the image is the composite of them all" (Lynch, 1960). "The steadily rising importance and the use of these maps in urban planning is addressed in the well-known 'Mappiness Project' or the work on Emotional Cartography" (Nold, 2009). A new approach, which is driven by the "quantified self" movement and the increasing availability of wearable sensors, is the analysis of physiological measurements to derive emotion information (Zeile, Resch, Loidl, & Petutschnig, 2016). However, the main goal of these efforts is to map and visualise emotion information, which is in contrast to our work where the aim is to extract emotions from social media for use in urban planning.

2.3. Extraction of Emotion Information from User-Generated Data

The field of "sentiment analysis" typically deals with a word's, sentence's, or document's polarity, i.e., whether it conveys a positive, negative, or neutral sentiment. Additionally, research has been conducted to determine the expressed sentiment's strength (Liu & Zhang, 2012). For our purpose we need a more sophisticated emotion model because knowing a tweet's polarity is not sufficient to convey the type of emotion, which is vital to understanding urban processes.

Detecting emotions from tweets focuses on classifying Twitter posts according to a number of distinct emotions. The two approaches by Roberts, Roach, Johnson, Guthrie and Harabagiu (2012) and Bollen, Mao and Pepe (2011) analyse the results of large-scale events and their influence on Twitter traffic for one or more days. In doing so, singular small-scale variations of Twitter traffic might be overseen. These smaller events may be important for urban planning as they affect smaller, local areas. Additionally, both approaches lack the geographic component, which is essential to our approach. Also, these previous efforts neglect the possibilities that arise for emotion detection from emoticon analysis.

Another approach by Hauthal and Burghardt (2013) aims to detect emotions in VGI, and to map emotional hot spots in a city. However, the approach works on the basis of a simple syntactical word-matching algorithm that is not able to cope with the complexity of unstructured text data like Twitter tweets and other social media posts. The same applies to López-Ornelas

and Morales Zaragoza (2015) and McGuire and Kampf (2015) who only analysed Twitter hashtags, and Do, Lim, Kim and Choi (2016) who pursue a lexicon-based approach that aggregates the weighted tweet-frequency values of words.

Strapparava and Mihalcea (2008) evaluate different algorithms that work in an unsupervised manner or with automatically obtained training data. Although the news headlines analysed by the authors share certain properties with tweets, such as brevity and partially incomplete sentences, they cannot be directly compared. While newspaper headlines are a source for short but edited text, tweets are not. Although news headlines can be considered as having a spatio-temporal dimension because they generally refer to current events, tweets are explicitly georeferenced and tagged with a timestamp. Thus, their approach inadequately takes the spatio-temporal dimension into account.

2.4. Linguistic Similarity

As mentioned in Section 1, our approach uses spatial, temporal, and linguistic similarity. Agirre, Cer, Diab and Gonzalez-Agirre (2012) define semantic textual similarity as "the degree of semantic equivalence between two texts". The main difference between their approach and ours is that they define a similarity rating of zero between two texts as "on different topics". In contrast to this, the topic does not influence the similarity score in the similarity metric proposed in this paper. In other words, the two definitions of similarity serve a different purpose and therefore the definition above is not applicable to the task at hand. To emphasise this contrast, we do not define textual similarity in terms of semantics, but with respect to a text's linguistic properties.

Many common similarity metrics for documents are defined on the documents' vector representations (Baba et al., 2015; Hill, Reichart, & Korhonen, 2015). However, because the data in our approach are not represented as vectors, geometric metrics are not applicable. This is due to the fact that the representation as a vector encodes the use of one dimension per feature, which would undermine the idea of a three-dimensional analysis based on the different dimensions of Twitter data. Furthermore, a vector would result in a bias towards the linguistic dimension as we define numerous linguistic features, but only a single parameter for the temporal and spatial dimensions, respectively. Consequently, a completely new approach to similarity computation is necessary in order to leverage the multiple dimensions.

3. Method for Extracting Emotions from Unedited Text

This section introduces our method for extracting emotions from unedited text like tweets. Figure 1 illustrates the stepwise method overview, in which we produce a set of labelled tweets from raw tweets, i.e., tweets are assigned a distinct emotion class. This is achieved by a semi-supervised learning approach, which labels tweets on the basis of a subset of the gold standard ("seeds"). The following subsections describe the single steps in more detail.

3.1. Data Preprocessing

The preprocessing step removes all tweets from our dataset that are not useful for our actual research goals. Furthermore, we apply a part-of-speech (POS) tagger (Owoputi et al., 2013) and a lemmatisation method (Manning et al., 2014) to optimise the dataset that is used for the subsequent analysis.

To eliminate non-relevant tweets, we first delete URLs and mentions of other users. This needs to be done because URLs are oftentimes abbreviated through services like bit.ly or goo.gl, and mentions, i.e., user names, are not unambiguous carriers of a tweet's emotion. If tweets are found to be empty after this first step, they are excluded from the dataset.

Second, we remove all tweets that do not contain any English words. According to Lui and Baldwin (2014), language identification of tweets is a complex problem, for which no perfectly accurate solution exists so far. To account for this shortcoming, we assessed the implications of wrongly classified tweets for the gold standard production (see Subsection 3.2) and similarity computation (see Subsection 3.3). Two cases have to

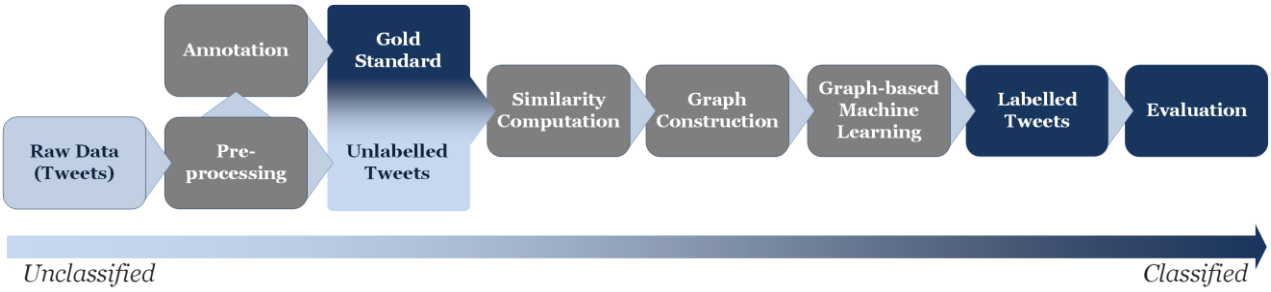

Figure 1. Workflow of the method for extracting emotions from unedited text.

be distinguished when considering the case of English versus all other languages. i.) tweets written in a different language are wrongly classified as English, thus remaining in the dataset. Consequently, if tweets are written in different languages they are probably not linguistically similar to each other. ii.) tweets written in English are wrongly classified as any other language and consequently discarded from the dataset. To reach the goal of eliminating non-English tweets, we combined two freely available state-of-the-art language classification tools in a voting architecture as proposed by Lui and Baldwin (2014): lang-id.py and compact language detector (cld2).

3.2. Annotation—Producing the Gold Standard

The production of a gold standard is necessary as we use a semi-supervised learning (SSL) approach for labelling the tweets. The SSL method requires a subset of ground-truth data to train the system and to evaluate the results. We base our annotation procedure on the work of Roberts et al. (2012), but adapted their approach to our environment and goals. Concretely, we employ more annotators (5 rather than 3) and we select annotators with little pre-knowledge in computational linguistics to avoid biases. In a related approach, Balabantaray, Mohammad and Sharma (2012) use seven emotion categories (Ekman and Friesen's six categories plus "no emotion"). We leverage new research results from emotion psychology, which defines four emotions, as stated in the introduction (see Section 1). Furthermore, the authors of the related study download tweets and user information randomly. In contrast, we use tweets originating from a particular place and time, which makes our data more homogenous in space, time, and user base, and hence easier to annotate.

In the actual annotation follows a three-phase process. The first annotation phase is the initial instruction phase to achieve a general agreement on the standard of the annotation procedure. Before the actual annotation, all annotators receive an annotation manual, providing all participants with the necessary information to understand the task and to standardise the resulting annotations, including a tree-like emotion structure that contains "sub-emotions" to each of the basic emotions (Shaver et al., 1987). The second phase constitutes a test phase in which the annotators individually label the same set of tweets. The results are then evaluated using the kappa metric for measuring the inter-annotator agreement. Concretely, we use the Fleiss Kappa index (Fleiss, 1971), which generalises the original Cohen's Kappa (Cohen, 1960) to more than two annotators and classes. The basic idea behind the kappa metric is to not simply measure the percentile agreement between two or more annotators, but to normalise this value by the expected agreement (produced by chance). The basic formula is $(p_o-p_e)/(1-p_e)$, wherein p_e is the expected (chance) agreement and p_o is the observed agreement. The larger the kappa value is, the higher the probability that the result was not produced by chance (Bortz, Lienert, & Boehnke, 1990). The second phase is completed if the kappa results are sufficiently high (at least 0.68). It shall be noted that the value of 0.68 represents broad agreement in the domain of computational linguistics, but further investigation is needed as to whether a lower threshold can be used when annotating emotions in tweets.

The third phase is the main annotation phase, in which all annotators individually annotate a different large set of tweets. This procedure provides a good compromise between ensuring high-quality annotations and reducing the load on each annotator. For the actual annotation we used the Crowdcrafting platform, which was chosen because of its free and open source nature and because of the promise of handling the task distribution correctly. Annotators were asked to label the same 400 randomly chosen tweets, which is a significant number to be used as a gold standard for the semi-supervised learning method, as widely agreed upon in existing literature.

3.3. Similarity Computation and Graph Construction

As a basis for classifying tweets according to contained emotions in a graph-based approach (see Subsection 3.4) we first need to perform a similarity computation. In our case, the concept of similarity is defined as the likelihood that two tweets contain the same emotion. The similarity computation comprises three dimensions, namely linguistic similarity, spatial similarity, and temporal similarity, which are combined to a single similarity value. As the details of the similarity computation are described in a separate research paper (Summa, Resch, & Strube, 2016) and because of its complexity, the following paragraph only provides a basic description of this process.

The similarity in the spatial and temporal dimensions are both formulated in exponential decay functions (see Subsection 4.1), in accordance with existing literature (Li, Goodchild, & Xu, 2013; Sakaki, Okazaki, & Matsuo, 2010). The computation of linguistic similarity uses the following "feature groups": hashtags, POS tags, word properties (word length, uni-, bi- and trigrams), emojis, spelling (e.g., recurring letters), and punctuation (e.g., several exclamation marks). More details on how these feature groups were defined can be found in Summa, Resch, & Strube (2016).

After the similarity between two tweets has been computed, the graph, which constitutes the input for the semi-supervised learning approach, is constructed. The graph is defined by the tweets (nodes) and the pairwise similarity values between tweets (weighted edges). If two tweets are not considered similar at all they receive an overall similarity score of zero, and no edge between

the two respective nodes is established in the graph. It is evident that the graph's density strongly depends on the edge weight threshold, which defines how many edges the graph contains. This is important because the graph's density (number of edges) clearly influences the SSL algorithm's running time and memory consumption. Finally, a node without any edges will not be part of the graph. This property is relevant for the evaluation because there is no guarantee that all seed and test tweets will actually be included in the graph.

3.4. Graph-Based Machine Learning for Labelling Tweets

For labelling the tweets (assigning an emotion to every tweet), we use the graph-based SSL algorithm Modified Adsorption (MAD), which has been found to be the most suitable method because "MAD is most effective for graphs with a high average degree, that is, graphs where nodes tend to connect to many other nodes" (Talukdar & Pereira, 2010). This is the case in our experiments as we initially calculate connections between all tweets. Generally speaking, graph-based SSL algorithms operate on a graph that is formally defined as $G = (V, E, W)$. The entities that are to be classified (tweets) are represented as nodes V, whereas possible connections between them are represented as edges E. Additionally, a matrix W stores the edges' weights (Bengio, Delalleau, & Le Roux, 2006). MAD is an example of transductive learning, i.e., no distinct training and test phases are conducted, but all instances are instantly labelled (Zhu & Ghahramani, 2002).

3.5. Evaluation Procedure

The evaluation of our results is difficult as no comparable approaches (integrated space-time-linguistics methods) exist so far, which in turn means that no standardised evaluation procedures have been defined yet. Furthermore, traditional evaluation metrics can only be applied conditionally, as they only work for single-disciplinary approaches from geoinformatics or computational linguistics, not for integrated methods. Thus, we propose a two-step evaluation setup to evaluate our results: i.) measuring a single feature combination's results, and ii.) comparing it to the other results. Furthermore, we selected a significance test that compares our results against two baselines. These procedures are described in more detail in the following paragraphs.

First, we chose a suitable evaluation measure: In general, classification tasks are evaluated by an approach to record correctly (true positive, true negative) and wrongly (false positive, false negative) labelled instances per class and to construct a confusion matrix (CM) accordingly. From the CM, precision, recall, and f-score values are calculated. After this has been completed for all classes, micro and macro averages are computed (see the next paragraph for details).

We chose to use precision, recall, and f-score as they are invariant towards a "change of true negative counts" and thus do not require a well-defined negative class (Sokolova & Lapalme, 2009), which is the case in our approach as laid out in Subsection 5.1. Additionally, our evaluation measure is applicable to a multi-class environment like our emotion classes, which we achieve by taking two averages over multiple classes: micro and macro averages (van Asch, 2013). The micro average constructs a confusion matrix from all classes' results and evaluates it like a single class. In contrast, the macro average evaluates all classes individually and the results' average is taken. Consequently, micro averaging is highly influenced by the larger classes' results, which contribute a larger fraction of the confusion matrix's counts. In contrast, macro averaging treats all classes alike, not being dependent on the number of test samples.

Next, we selected a suitable significance testing method, which computes the difference between our results and i.) a random baseline (assigns emotions randomly to tweets), and ii.) a majority baseline (assigns most frequent class label to all tweets). In our case, McNemar's test is the best choice for testing the significance, as underpinned by Dietterich (1998): "Given two classifiers C_A and C_B and enough data for a separate test set, determine which classifier will be more accurate on new test examples". This definition encompasses our setting (enough datasets and several runs). Furthermore, McNemar's test has an acceptable type 1 error ("false positives") while being computationally inexpensive, and it can handle data that are not normally distributed (Japkowicz, 2012), as in our case.

4. Case Study and Results

To test our approach presented in Section 3 we applied it in a case study, analysing about 200.000 Twitter tweets for the greater Boston area. The time period covers one week before and one week after the Boston marathon bombing. Table 1 summarises the properties of our dataset.

4.1. Experiments

In conducting our experiments, we followed three main steps in accordance with the methodological setup described in Section 3. The label distribution of the gold standard (manually annotated tweets) is illustrated in Table 2. It shows a strongly skewed class layout as the negative class ("none") is much larger than the other ones – this is discussed in more detail in Subsection 5.2. The first column indicates the number of agreements, i.e., how many annotators labelled a tweet with the same emotion.

Table 1. Dataset description summary.

Data Description Summary	
Geographic Bounding Box (WGS84)	-71.21°, 42.29°, -70.95°, 42.45°
Time Period (UTC)	08 April 2013–22 April 2013
Number of Georeferenced Tweets before Pre-processing	222,089
Number of Georeferenced Tweets after Pre-processing	195,380
Number of Unique Users	16,099

Table 2. Gold standard: Emotion labels and number of agreements.

	Emotion Labels					
Number of Agreements	Anger/disgust	Fear	Sadness	Happiness	None	Total
3	21	5	20	37	64	147
4	21	1	19	50	90	181
5	24	2	4	57	231	318
Total	66	8	43	144	385	646

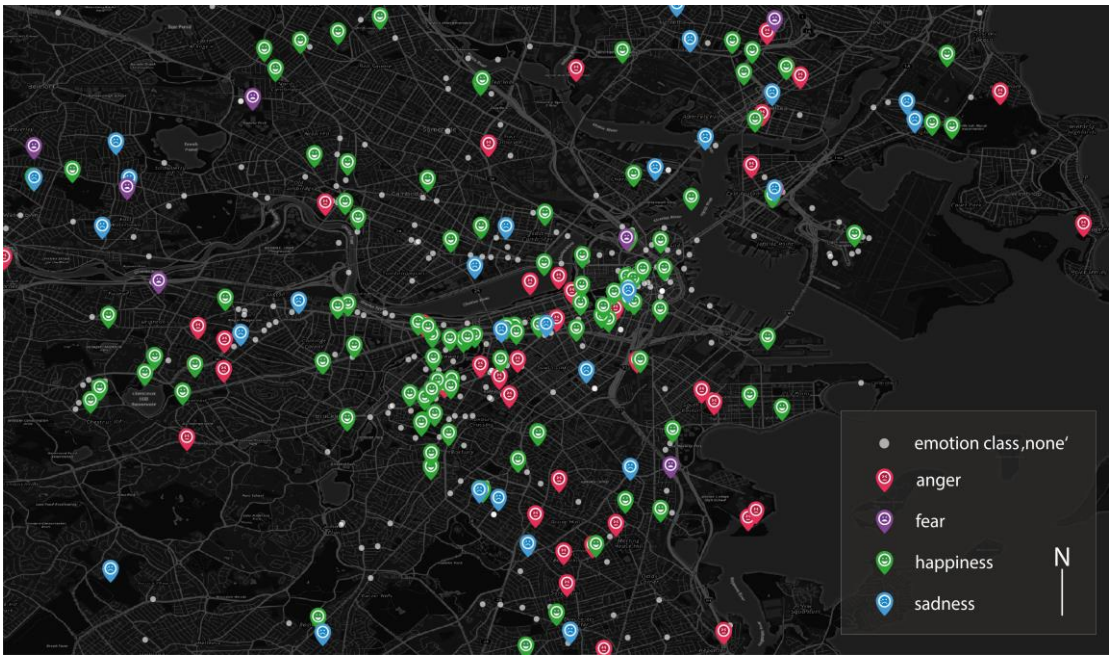

Figure 2. Spatial distribution of the gold standard.

First, we performed a seed selection, i.e., a random selection of seeds from the gold standard, to make sure that all runs in one experiment use the exact same training and test dataset. Second, our algorithm calculated the optimal combination of feature groups. The linguistic features are selected in an iterative approach, i.e., feature groups are added one by one, and the best combination between feature groups is finally applied, where "best" means the highest evaluation result (see Subsection 3.5). Then, the optimal parameter settings for spatial and temporal similarity are computed analogously. Third, we carried out the labelling procedure with the following parameters: amount of unlabelled data (5,000), the number and distribution of seeds (happiness: 70, sadness: 20, anger/disgust: 30, fear: 4, none: 70), the edge weight threshold (0.5), and the weighting parameters for the three dimensions linguistic (1.0), temporal (5.0), and spatial (5.0) similarity. These parameter values turned out to be the best ones with respect to the evaluation results (see Section 4.3), which were obtained by comparing different parameter combinations in our empirical optimisation.

4.2. Results

Figure 2 shows the spatial distribution of the tweets contained in the gold standard, where the icons indicate each tweet's emotion label. It can be seen that the gold standard is randomly distributed over space, where the tweet density correlates with the population density (e.g., higher density in the inner city). This is expected as we drew a random sample from the entire Twitter dataset for the annotation procedure. The map displays the gold standard tweets that we used for our analysis.

Applying our developed method to the Twitter dataset using the parameters described in Subsection 4.1 produced a result that is characterised by a high concentration of the labels for the emotion classes of

"happiness" and "none". This is not surprising as a consequence from the skewed dataset (see Subsection 5.2 for a thorough discussion). The maps shown in Figure 3 reveal strongly clustered "happiness" tweets (left) while "none" tweets (right) are more evenly dispersed over space—again with an apparent correlation between tweet density and population density. These results are discussed in more detail in Subsection 5.2. The number of unique users in our dataset is high enough to avoid clusters that are generated by only one person, which has oftentimes been a shortcoming in previous research.

4.3. Evaluation Results

The statistical evaluation results, which were obtained according to the procedure described in 0, show that we can reliably detect the emotion classes "happiness" and "none", and that our approach performs better than the baselines. Table 3 summarises the evaluation results for each of the measures. It shall be noted that precision, recall, and f-score are computed for every emotion class, whereas micro and macro averages are an integrated measure that consider all emotion classes, as described above. Following this rationale, it is evident that the averages are lower than the individual numbers of the "happiness" and "none" classes as no tweets have been labelled with the other three emotion classes. These evaluation results are discussed in Subsection 5.2.

The most significant method to evaluate the difference in the performance of our approach versus the baselines is to compare *macro averages*. The majority baseline by definition scores low according to the macro-averaged evaluation metrics because the macro-averaged metrics give equal weight to all evaluated classes. The fact that *TwEmLab* outperforms the majority baseline with respect to macro-averaged metrics is satisfying, but the comparison with the random baseline is even more meaningful in this case. This shows that the results are not produced by chance, but that meaningful similarities have been found between pairs of tweets.

5. Discussion of the Approach and the Results

This section discusses *TwEmLab* in two ways: First, in terms of the strengths and weaknesses of the methodological approach (Subsection 5.1), and second with respect to the obtained results (Subsection 5.2).

5.1. Discussion of the Approach

One central advantage of our approach is that it is virtually **language-independent**. It does not depend on specialised language resources and can work with languages other than English—given that a gold standard is available and the dataset exclusively contains text in one single language. This characteristic is important because it makes our approach transferrable to other study sites, as only 40% of tweets are written in English language. Furthermore, our approach is generic enough to be applicable to georeferenced posts from **other social media networks** like Flickr, Instagram, Panoramio, Facebook, etc.

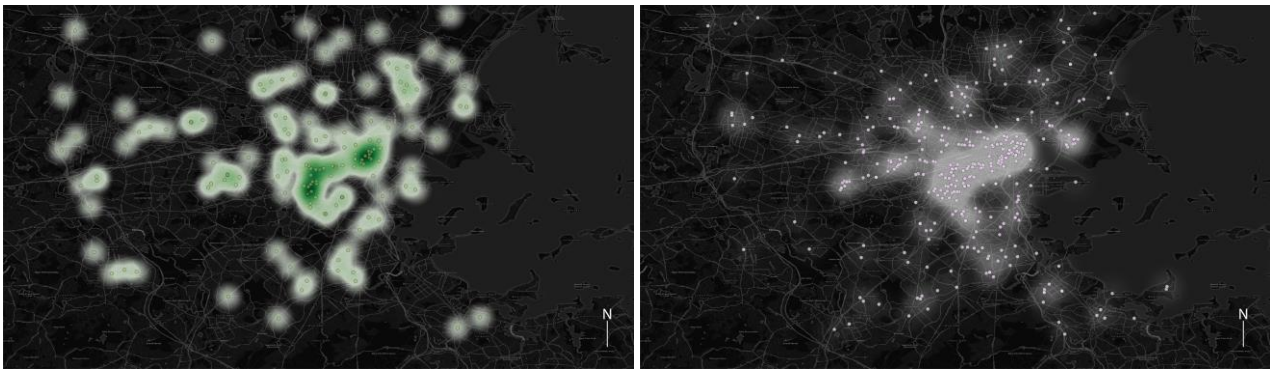

Figure 3. Spatial distribution and density of the tweets labelled with "happiness" (left) and "none" (right).

Table 3. Statistical evaluation results.

Evaluation Measure	Happiness	None	Random Baseline	Majority Baseline
Precision	0.65	0.68	n/a	n/a
Recall	0.24	0.98	n/a	n/a
F-score	0.35	0.80	n/a	n/a
Micro average precision	0.68		0.22	0.64
Macro average precision	0.27		0.23	0.13
Micro average recall	0.68		0.22	0.64
Macro average recall	0.24		0.14	0.20
Micro average f-score	0.68		0.22	0.64
Macro average f-score	0.25		0.18	0.16

Even though the evaluation of our approach shows promising results (see Subsection 3.5), a number of simplifying **assumptions** had to be made. First, we assume that exactly one emotion is present in a tweet, or none at all. From a psychological viewpoint, this may not necessarily be true, although the brevity of tweets makes them less vulnerable to changes in emotion compared to longer texts. Second, we assume that the tweets' textual content actually refers to the time and place from which they were sent, which has been a known restriction for most previous research efforts using Twitter data. Finally, we assume that there is a causal relationship between the expressed emotion and the user's environment.

A major challenge is the **construction of the gold standard**. Although our approach using human laymen annotators is scientifically justified, the resulting data set is still not unambiguous. This is rooted in two causes: i.) the way tweets are written, which makes them difficult to understand for other people; and ii.) the implications in Twitter users' abbreviated way of expressing themselves through 140 characters long messages, which are hardly suitable for conveying clear and unambiguous messages. Thus, labelling tweets with an emotion from a given set is a highly subjective task with considerable uncertainty. For instance, tweets may be understood and interpreted differently because of ironical language or use of slang; they may contain more than one prevalent emotion.

Another challenge that needs further research is the process of **computing similarity between tweets**. Here, we face a number of critical factors like defining the edge weights and finding appropriate thresholds for the weights. This is a particular and use case-dependent research challenge as no generic weights can be defined. The value of the weights of the single dimensions (space, time, and linguistics) obviously influences the results, which needs to be accounted for in the interpretation.

5.2. Discussion of the Results

Our results show that we can generally **detect emotions in tweets** using our approach in an integrated space-time-linguistics method. In Figure 3, the "happiness" tweet map is characterised by two **clusters** in the inner city. When looking at these particular clusters we can see that most of these tweets are related to the Boston marathon bombing, as they were written in the days after the marathon event. From a semantic viewpoint, it is interesting to observe that these tweets are classified as "happy", which results from a particularity in the characterisation of the dataset. Many of these tweets contain words like "proud", "supportive", "thanks", "love", "strong", "pride", etc., which are sub-emotions of happiness in the model by Shaver et al. (1987).

This special characteristic in our results arises from the **skewed nature of the dataset**: The "none" and "happiness" classes are dominant (see above), and in many cases only "none" and "happiness" labels occur in our results, depending on the parameter settings. This is a specificity of Twitter tweets, as confirmed by a number of psychological and sociological studies (Dodds et al., 2015; Wojcik, Hovasapian, Graham, Motyl, & Ditto, 2015). This effect arises because a high fraction of sad statements are oftentimes expressed as positive thoughts, as shown above. This distorts the input dataset and the results for emotion extraction.

Furthermore, the emotion classes of "none" and "happiness" can be more easily distinguished from each other compared to the other emotion classes because they are "more different" from each other. This may be due to the fact that "happiness" is the **only positive emotion class** in the used emotion model and thus in the gold standard annotations.

These distortions can probably be mitigated by a larger gold standard (allowing for the use of more seeds in all emotion classes) and by defining an appropriate number of seeds for each emotion class. Furthermore, while the emotion classes themselves are clearly specified, the **"none" class captures different phenomena**, such as "no emotion", "I didn't understand the tweet", and "I cannot decide". Consequently, the negative class is not of much interest for the project's evaluation.

From a purely quantitative viewpoint, our results prove that *TwEmLab* performs **better than the baselines**. This is a remarkable output given that our research constitutes the first approach towards a **joint metric** of computing similarity with respect to two tweets' emotional content along three dimensions (linguistic, temporal, and spatial), which clearly advances the state-of-the-art compared to previous single-disciplinary approaches or sequential methods.

Furthermore, our results show that it is possible to **generate a gold standard** through manual annotation of tweets, where the actual annotation is a subjective interpretation of a tweet's emotion by the annotators. Just like previous approaches, we assume that a high inter-annotator agreement (in our case 5 agreements among the 5 annotators) is considered a valid output. Furthermore, the actual annotation procedure is laborious and a high kappa index can only be achieved through distinct and unambiguous communication of the annotation task. Here, one essential research question will be the definition of a threshold for a sufficiently high kappa index, as current agreements (0.68) have been defined for edited text analysis, not for social media posts. Furthermore, the annotated tweets could then be used in a semi-supervised learning algorithm to label all tweets. We have shown that our trans-disciplinary similarity metric is not only theoretically possible, but also proven to be suitable for

emotion classification. Yet, it is still to be proven how the results of our method can be applied to spatial planning processes.

6. Use in Urban Planning

We are confident that our results are directly usable in urban planning processes. Apart from proving the general ability of our approach to detect emotions that are associated with places, we investigated a number of concrete examples. The first one is traffic-related. Here, we observed a number of tweets carrying different emotions, including "Traffic awful today (@ Kendall Square)", "Tourist traffic at Fenway already terrible", "Holy shit the traffic on Comm Ave is ridiculous. Thanks to those goddamn shit sox or whatever that soccer team is called #fenway", "So. Much. Traffic. #fuck", "At this rate, I might never make it to MNSB….I hate Red Sox traffic", "traffic on Mass Ave from Central Sq into Boston is gridlock. Avoid!", and many more. As all of these messages are associated with a geolocation and a timestamp, concrete traffic hot spots can be identified.

Another example is related to the Boston marathon bombing event itself. After the marathon bombing, the hashtag #BostonStrong was heavily used and oftentimes infused with emotion. Interestingly, we observed two different kinds of emotions. First, citizens expressed their sadness and sorrow in their tweets: "All of the aftermath from last week is still heavy, still brings tears.", "A crowded T of heavy hearts and sad faces, it hurts to see how shaken we are.", or "Still digesting the events from yesterday. Will be sad for the victims, their families & loved ones and our city for some time." Second, we observed a large number of positive tweets, which appear in the emotion class of "happiness". This is due to the fact that terms like pride, hope, love, optimism, and others are subsumed under happiness according to the emotion model by Shaver et al. (1987). Examples of such tweets include "A week ago our lives here in Boston changed forever. Always be thankful for the love in all of our lives. #blessed #bostonstrong", "Moment Of Silence In The Quad Was Amazing, Thank You EC #bostonstrong", or "I absolutely love all the #BostonStrong support around town!". These different ways of expressing one's emotions towards a tragic event need to be accounted for when interpreting the results of our research. Additionally, we were able to attribute emotional tweets to a wide variety of concrete planning issues like dog faeces on the streets, damaged pavements, or dangerous bicycle lanes.

This shows that social media constitute a valuable, open source of information for urban planning. This is particularly so as urban planning is oftentimes still a closed communication process between local governmental actors, and not an open, transparent procedure that integrates, discusses, and considers the requirements of citizens and civic interest groups. In an ideal planning workflow, all arguments should be collected, weighed against each other, and discussed in workshops, charrettes or other open formats to gather opinions and needs from citizens. However, in current deductive processes, which are typically initiated and installed by the government, citizens often do not feel that their requirements are heard and considered enough. This may be due to the fact that sectoral interests, diffuse goals, and unrealistic demands characterise the process (Olk, Somborski, & Stimbel, 2011). In contrast, public participation is increasingly promoted by politicians because it encourages democracy, increases acceptance through higher transparency, creates a more accurate repository of wishes and suggestions concerning the planning topic, delivers better results, can produce a legitimation of a specific planning approach, and reduces the costs of a planning process (Fürst & Scholles, 2008; Senatsverwaltung für Stadtentwicklung und Umwelt Berlin, 2012).

This is specifically important in emerging discussions about how stakeholders and politicians can foster participation and integrate the public into decision-making processes. The main question is how more people can be engaged in these processes and how new target groups can be involved in alternatives to traditional means of participation. The results of the research presented in this paper, i.e., a reliable method for extracting emotions from social media and correlating them with precise urban planning issues, will be a helpful mood sensor in future, to complement traditional surveys with a dynamic layer in planning processes. We clearly see the possibility of creating daily snapshots of citizens' remarks on planning aspects in cities. As an example, the growing problems of railway project "Stuttgart 21" were reflected in social media before 2010, but politicians and planners did not realise this at an early stage.

In addition, it will be helpful in the future to compare the results of the Twitter maps with government expenditures. As a result of citizens' protests, the State of Baden-Württemberg installed a State Counsellor for Civil Society and Civic Participation whose duty it is to improve civic participation on every level in the state and to integrate it into administrative processes (Erler, 2015). Against this background, the *big data* source of social media can be seen as an invaluable complement to traditional planning and participation processes as they contain plenty of potentially useful remarks concerning urban planning issues.

In this regard, our approach can deliver new insights into peoples' thoughts and expectations concerning their city. In contrast to top-down processes, *TwEmLab* pursues a bottom-up and inductive approach. The advantages are obvious: Bottom-up processes are self-organising, where the data acquisition of urban phenomena is done by interested people,

mostly laypersons, and not by institutions (Streich, 2011). If the citizens as "gatherers" of urban phenomena are not only data producers, but also provide an impetus for new planning issues, our approach is at the core of such self-organising processes of assessing phenomena in urban spaces. A simple example for such an inductive process is the hashtag used in Twitter messages and other social media posts. People use these hashtags to mark a specific annotation to a special event. This dynamic, together with more latent patterns like punctuation, spelling, words' properties, and others, allows us to gain up-to-date information about citizens' emotions and thoughts. Through this it is possible to obtain citizens' direct feedback for urban planning and as a supplementary decision support tool for ongoing planning processes using contextual emotion information.

One particularity of this approach is that it is not understood as a general tool of solving all planning issues, but that it can help to create another view and a more accurate understanding of the city as an organism. From a current viewpoint it would be beneficial if this new knowledge could be integrated as indicative information in official planning processes (Zeile et al., 2015). From a planning perspective, the annotations of the emotion labels "anger", "fear", and "sadness" seem to be a valuable information source of the future. Our experiences show that explicit comments concerning problems of the urban environment, like traffic jams or pollution, can be detected in tweets carrying this emotion. At this point, this kind of information helps to filter and identify planning-relevant tweets. In the future, we expect more accurate and reliable results using our method that can, for instance, be used in special planning processes or in combination with large-scale projects like Boston's "big dig"—the Central Artery/Tunnel Project (CA/T)—or "Stuttgart21", Stuttgart's controversial re-design of the area around the main station.

7. Conclusions

This paper presents the innovative, interdisciplinary method "*TwEmLab*" for identifying emotions in social media posts such as Twitter tweets, constituting a new approach towards jointly analysing the linguistic, spatial, and temporal dimensions of the data. To this end, we constructed a set of gold standard annotations for tweets with a set of discrete emotion labels. *TwEmLab* assigns similarity scores to pairs of tweets according to their three dimensions. It constructs a graph with the tweets serving as nodes and the similarity scores as edge weights between the respective tweets. After performing graph-based semi-supervised learning in order to label all tweets with their appropriate emotion classes, it evaluates the results through precision, recall, and f-score, as well as micro and macro averages.

Our results show that *TwEmLab* is able to generally detect emotions in tweets with some restrictions (see Section 5). Although this work is the first attempt to combine tweets' textual and spatio-temporal dimensions into a single metric for emotion detection and classification, its performance is better than the baseline's. A central challenge revealed by our results is that the "happiness" and "none" labels are disproportionately overrepresented. While this is not surprising, as several studies from the fields of sociology and psychology confirm, it still poses a significant challenge in identifying the other (negative) emotion classes in tweets.

Concluding from the discussion of our approach and our results (see Section 5), we identified a number of open future research issues: the development of a structured method for defining spatial, temporal and linguistic weights; the definition a formal method for determining the required size of the gold standard tweets used as seeds; the exact influence of the skewed dataset; the derivation of a suitable kappa index threshold for the inter-annotator agreement when annotating tweets; and research on how to improve the macro-averaged f-score to increase reliability of the results. Finally, the influence of the dataset size on our results needs to be further investigated, as it is not yet clear if and how larger datasets correlate with better results.

Conflicts of Interest

The authors declare no conflict of interests.

Acknowledgements

We would like to express our gratitude to the German Research Foundation (DFG—Deutsche Forschungsgemeinschaft) for supporting the project "Urban Emotions", reference numbers ZE 1018/1-1 and RE 3612/1-1. 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. Furthermore, we would like to thank Günther Sagl, Enrico Steiger, René Westerholt, Clemens Jacobs and Sebastian Döring for supporting the annotation procedure. Finally, we would like to thank the Doctoral College GIScience (DK W 1237-N23) at the Department of Geoinformatics—Z_GIS, University of Salzburg, Austria, funded by the Austrian Science Fund (FWF) and University of Salzburg's Open Access Publication Fund for their support.

References

Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). SemEval-201W2 Task 6: A Pilot on Semantic Textual Similarity. In *{*SEM 2012}: The First Joint Conference*

on Lexical and Computational Semantics (Vols 1 and 2, Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 385–393). Montréal, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S12-1051

Baba, S., Toriumi, F., Sakaki, T., Shinoda, K., Kurihara, S., Kazama, K., & Noda, I. (2015). Classification method for shared information on twitter without text data. *Proceedings of the 24th international conference on world wide web: WWW '15 companion* (pp. 1173-1178). New York, USA: ACM Press.

Balabantaray, R. C., Mohammad, M., & Sharma, N. (2012). Multi-class Twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, *4*(1), 48-53.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M. . . . Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, *214*(1), 481-518.

Bengio, Y., Delalleau, O., & Le Roux, N. (2006). Label propagation and quadratic criterion. In O. Chapelle, B. Schölkopf, & A. Zien (Eds.), *Semi-supervised learning* (pp. 193-216). Cambridge, MA: MIT Press.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of the fifth international AAAI conference on weblogs and social media* (pp. 450-453). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Bortz, J., Lienert, G. A., & Boehnke, K. (1990). *Verteilungsfreie methoden in der biostatistik*. Berlin, Heidelberg and New York: Springer-Verlag.

Brenner, N., Marcuse, P., & Mayer, M. (2012). *Cities for people, not for profit: Critical urban theory and the right to the city*. London and New York: Routledge.

Capdevila, J., Arias, M., & Arratia, A. (2016). GeoSRS: A hybrid social recommender system for geolocated data. *Information Systems*, *57*(April 2016), 111-128.

Castells, M. (1996). *The rise of the network society. Volume I: The information age. Economy, society, and culture*. Hoboken, NJ: Blackwell Publishers.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *XX*(1), 37-46.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classication learning algorithms. *Neural Computation*, *10*(7), 1895-1923.

Do, H. J., Lim, C.-G., Kim, Y. J., & Choi, H.-J. (2016). Analyzing emotions in Twitter during a crisis: A case study of the 2015 Middle East respiratory syndrome outbreak in Korea. *2016 international conference on big data and smart computing (BigComp)* (pp. 415-418). Hong Kong, China: IEEE.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R. . . . Megerdoomian, K. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*(8), 2389-2394.

Downs, R. M., & Meyer, J. T. (1978). Geography and the mind an exploration of perceptual geography. *The American Behavioral Scientist (Pre-1986)*, *22*(1), 59-77.

Downs, R., & Stea, D. (1974). *Image and environment: Cognitive mapping and spatial behavior*. London: AldineTransaction. Retrieved from http://books.google.com/books?hl=de&lr=&id=dnwkmQ1fuUAC&oi=fnd&pg=PR8&dq=Image+and+environment:+Cognitive+mapping+and+spatial+behavior+&ots=xoY70vTWGJ&sig=k0_X66Q58ypo8gvlu40pZGE1WLo

Eisenstein, J. (2013). What to do about bad language on the internet. *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp. 359-369). Atlanta, GA: Association for Computational Linguistics.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129.

Erler, G. (2015). Demokratie-Monitoring Baden-Württemberg. In Baden-Württemberg Stiftung (Ed.), *Demokratie-Monitoring Baden-Württemberg 2013/2014: Studien zu Demokratie und Partizipation* (pp. 11-15). VS Verlag für Sozialwissenschaften: Springer.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Fürst, D., & Scholles, F. (2008). *Handbuch theorien und methoden der raum und umweltplanung*. Dortmund, Germany: Verlag Dorothea Rohn.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Hauthal, E., & Burghardt, D. (2013). Extraction of location-based emotions from photo platforms. In J. Krisp (Ed.), *Progress in location-based services* (Vol. 49, pp. 1-20). Heidelberg, Germany: Springer.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187-192.

Jacobs, J. (1961). *The death and life of great American cities*. New York: Random House LLC.

Japkowicz, N. (2012). Performance evaluation for learning algorithms. *International conference on machine learning*. Edinburgh, UK: ICML Retrieved from http://www.mohakshah.com/tutorials/icml2012/Tutorial-ICML2012/Tutorial_at_ICML_2012_files/ICML2012-Tutorial.pdf

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG!

Proceedings of the fifth international AAAI conference on weblogs and social media (pp. 538-541). Barcelona, Spain: The AAAI Press.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61-77.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal & C. X. Zhai (Eds.), *Mining text data* (pp. 415-463). New York: Springer US.

López-Ornelas, E., & Morales Zaragoza, N. (2015). Social media participation: A narrative way to help urban planners. In G. Meiselwitz (Ed.), *Social computing and social media* (Vol. 9182, pp. 48-54). Cham, Switzwerland: Springer International Publishing.

Lui, M., & Baldwin, T. (2014). Accurate language identification of Twitter messages. *Proceedings of the 5th workshop on language analysis for social media (LASM)@EACL 2014* (pp. 17-25). Gothenburg, Sweden: Association for Computational Linguistics.

Lynch, K. (1960). *The image of the city*. Cambridge MA: MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55-60). Baltimore, MD: Association for Computational Linguistics.

McGuire, M., & Kampf, C. (2015). Using social media sentiment analysis for interaction design choices: An exploratory framework. *Proceedings of the 33rd annual international conference on the design of communication: SIGDOC '15* (pp. 1-7). New York: ACM Press.

Nold, C. (2009). *Emotional cartography: Technologies of the self*. Retrieved from http://emotionalcartography.net

Olk, T., Somborski, I., & Stimbel, T. (2011). Stadtgesellschaft macht Bildung. *Forum Wohnen Und Stadtentwicklung: Verbandsorgan Des Vhw*, *3*(3), 155-160.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. *Proceedings of NAACL-HLT* (pp. 380-390). Atlanta, GA: NAACL-HLT.

Pahl-Weber, E., Ohlenburg, H., Seelig, S., von Bergmann, N., & Schäfer, R. (2013). *Urban challenges and urban design approaches for resource-efficient and climate-sensitive urban design in the MENA region* (Vol. 5). Berlin, Germany: Universitätsverlag der TU Berlin.

Resch, B., Summa, A., Sagl, G., Zeile, P., & Exner, J.-P. (2015). Urban emotions: Geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data. In G. Gartner & H. Huang (Eds.), *Progress in location-based services 2014* (pp. 199-212). Cham, Switzwerland: Springer International Publishing.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. *Proceedings of the eighth international conference on language resources and evaluation (LREC-2012)* (Vol. 12, pp. 3806-3813). Istambul: Turkey: European Language Resources Association (ELRA).

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web* (pp. 851–860). Raleigh, NC, USA.

Senatsverwaltung für Stadtentwicklung und Umwelt Berlin. (2012). *Handbuch zur partizipation* (2nd ed.). Berlin, Germany: Kulturbuch-Verlag GmbH.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061-1086.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427-437.

Steiger, E., Resch, B., & Zipf, A. (2015). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, *30*(9), 1694-1716

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. *Proceedings of the 2008 ACM symposium on applied computing: SAC '08*, 1556-1560. Fortaleza, Brasil: ACM.

Streich, B. (2011). *Stadtplanung in der Wissensgesellschaft: Ein Handbuch* (2nd ed.). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

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

van Asch, V. (2013). *Macro and micro-averaged evaluation measures*. Retrieved from http://www.cnts.ua.ac.be/~vincent/pdf/microaverage.pdf

Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M., & Ditto, P. H. (2015). Conservatives report, but liberals display, greater happiness. *Science*, *347*(6227), 1243-1246.

Zeile, P., Resch, B., Exner, J.-P., & Sagl, G. (2015). Urban emotions: Benefits and risks in using human sensory assessment for extraction of contextual emotion information. In S. Geertman, J. Ferreira, R. Goodspeed, & J. Stillwell (Eds.), *Planning support systems and smart cities* (pp. 209-225). Cham, Switzerland:

Springer International Publishing.
Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation* (CMU-CALD-02-107). Pittsburgh, PA: Carnegie Mellon University.

About the Authors

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, Associated Faculty Member of the doctoral college "GIScience", and Executive Board member of Spatial Services GmbH.

Anja Summa holds a degree in Computational Linguistics from Heidelberg University. Her interests are semi-supervised learning methods for text analysis, and text mining of unedited documents. Anja is currently working as a software engineer.

Peter Zeile is research group leader at the University of Kaiserslautern in Computer Aided Design. He graduated in Spatial and Environmental Planning in 2003 and received his Ph.D. degree in 2010 at University of Kaiserslautern ("Real-time Planning"). Current research project "Urban Emotions", financed by German Research Foundation (Deutsche Forschungsgesellschaft—DFG). Previously, he was a lecturer at the HS Rapperswil (CH) and since 2011 at the University of Kaiserslautern. Peter is a member of the SRL and National Delegate of ISOCARP.

Michael Strube leads the Natural Language Processing (NLP) Group at HITS gGmbH, Heidelberg, Germany. There, he is involved in NLP related projects, works with the computational linguists at HITS, and supervises PhD students. In addition, he is a *Honorarprofessor* in the Computational Linguistics Department at Heidelberg University.