International Journal of
Geo-Information

MDPI

Article

Evaluating Geo-Tagged Twitter Data to Analyze Tourist Flows in Styria, Austria

Johannes Scholz * and Janja Jeznik

Research Group Geoinformation, Institute of Geodesy, Graz University of Technology, 8010 Graz, Austria; janja.jeznik@gmail.com
* Correspondence: johannes.scholz@tugraz.at

Received: 30 September 2020; Accepted: 10 November 2020; Published: 15 November 2020

Abstract: The research focuses on detecting tourist flows in the Province of Styria in Austria based on crowdsourced data. Twitter data were collected in the time range from 2008 until August 2018. Extracted tweets were submitted to an extensive filtering process within non-relational database MongoDB. Hotspot Analysis and Kernel Density Estimation methods were applied, to investigate spatial distribution of tourism relevant tweets under temporal variations. Furthermore, employing the VADER method an integrated semantic analysis provides sentiments of extracted tweets. Spatial analyses showed that detected Hotspots correspond to typical Styrian touristic areas. Apart from mainly successful sentiment analysis, it pointed out also a problematic aspect of working with multilingual data. For evaluation purposes, the official tourism data from the Province of Styria and federal Statistical Office of Austria played a role of ground truth data. An evaluation with Pearson's correlation coefficient was employed, which proves a statistically significant correlation between Twitter data and reference data. In particular, the paper shows that crowdsourced data on a regional level can serve as accurate indicator for the behaviour and movement of users.

Keywords: Twitter; crowdsourcing; spatial analysis; sentiment analysis; tourism; Styria

1. Introduction

Social media such as Twitter enables user communication and the sharing of their state of mind, behavior or activities. In addition, there is also the possibility to provide a current position or location of each tweet. From a geographical aspect, Twitter conveniently provides real time geo-data directly from the users, unlike the official data collections with postponed availability. Various spatial analyses may be performed with collected geographical data, such as identifying trends within the data to obtain information on locations with most Twitter users or analyzing their movement. Crowdsourced (geographic) data are data that are data that are voluntary and involuntary provided by citizens [1,2]. Furthermore, [1] among others refer to the fact that any data from social media are regarded as crowdsourced data In this research crowdsourced geographic data [1,2], from the platform Twitter, are collected to analyze the tourist flows in Styria, a state in the country of Austria. In this particular paper, we regard the influx of tourists as "flow" and subsequently try to analyze the spatio-temporal distribution of this tourist influx. As such, in contrast to many publications focusing on a greater scale [3,4], this study focuses on the reliability of Twitter data on a regional scale. The investigation of obtained data focused on the recognition of tourist spatial and temporal patterns as well as the evaluation of geo-tagged Twitter data reliability and adequacy in contrast to the public data of the statistical bureau of Austria.

The paper elaborates on the development of a methodology for spatial and semantic analysis that is applied in the test area—the Province of Styria in Austria. The plausibility and accuracy of results obtained with Twitter data are evaluated against official statistics on tourism [2,3]. The results and evaluations are interpreted in the form of maps, graphs and statistical evaluations. The detailed research questions of the paper are as follows:

- Do crowdsourced data on a regional scale accurately represent the touristic behavior of users?
- Are touristic-relevant crowdsourced data correlated to the official tourism statistics?
- Are touristic-relevant Twitter data on a regional level sufficient to draw conclusions on the topics covered and concerning the sentiment?

The paper is organized as follows. Section 2 discusses the relevant literature, and Section 3 elaborates on the methodology applied to the test data collected. The experiment conducted is described in Section 4 and the results obtained are listed in Section 5. A discussion of the results and conclusion is given in Section 6.

2. Relevant Literature

Literature in the field social sensing for tourism purposes has been published in different scientific fields. Most notably the advances in Geographic Information Science have most impact on the paper. In particular, publications dealing with Volunteered Geographic Information, social sensing, and Geospatial Artificial Intelligence are of particular interest.

The term social sensing combines machine learning approaches and artificial intelligence in general [5–7]. According to [8] it is defined as the usage of user-generated data to understand human dynamics. The methodologies can be utilized to understand human mobility patterns, social network patterns, or even support urban planning [3,9,10]. Hawelka et al. [11] have investigated the global mobility patterns of different countries in comparison to the Twitter market penetration. An evaluation of their results with the help of tourism statistics revealed that there is a correlation between the number of Twitter users, economic prosperity and mobility behavior of a country. Hence, Twitter is regarded as proxy for global mobility patterns.

The articles [12,13] report on migration patterns based on Twitter data. [12] analyzes migration patterns from Middle East and North Africa to Europe. The authors use spatial and semantic analyses (topic clusters along migration routes), utilizing the OPTICS method. The authors in [13] developed an estimator for migration patterns based on Twitter data. Other movement patterns involve the detection of trajectories using hot spot analysis and subsequently characterizing them via drift analysis [14]. The paper is using hot spot cluster analysis and Kernel Density Estimation to derive spatial trajectories. The methodology has been applied to the detection of a concert route of a pop singer.

User-generated data have been used in emergency situations as well. In literature, the Boston marathon bombing is used as an example where social media messages may be a tool for an early recognition of emergency situations [15]. In [16], the authors developed an application for earthquake detection and notification, based on social media. The study uses Kalman and particle filtering for location estimation and a semantic analysis based on keywords, number of words and their context, respectively.

Spatial analysis for tourism purposes has been covered in [17], whereas methodologies for semantic text analysis have been published in [18]. In [19] Flickr is used as basis for spatial mobility patterns. Of particular interest for this paper are the works [20], a geo-location prediction based on Twitter data [21], as well as [22] dealing with mobility patterns in cities located in Australia. More recent papers deal with the development of travel planning tools based on user-generated content [23], digital footprints of crowdsourced data for the management of protected areas [24], or an analysis of patterns of user-generated content with a focus on Flickr and Panoramio [25]. The papers [26–28] elaborate on data from Location-based social networks (Twitter, Flickr, etc.) to research on urban phenomena, and human activities in general and recreation patterns. On a more general side [29] looked into the role of Big Data for open innovation strategies in SMEs and large companies, which could be valuable for tourism as well. In particular, [30] elaborate on the future challenges and action for innovation research in the tourism industry. Social media usage of luxury hospitality facilities in Turkey is analyzed by [31]. In order to trigger the engagement of users for touristic events and destinations [32] the authors evaluated two events and their social media outreach different platforms. A contemporary research agenda for Tourism Geographies and Big Data in Tourism Geography has been provided in [33,34].

3. Methodology

The research did not follow any existent predetermined methodological workflow. Based on the existing literature we selected the most appropriate methods to perform the spatio-temporal and semantic analysis. The developed methodology and workflow in this paper includes methods for data acquisition, processing and filtering, followed by methods for spatial and semantic analysis as well as for visualization, interpretation and evaluation. The methodology is based on six pillars that are depicted in Figure 1, consisting of data acquisition, data processing and filtering, spatial analysis, semantic text analysis, visualization & interpretation and evaluation.

Data acquisition seeks to collect data from the social medial platform Twitter using the Twitterscraper [35] python package. A basic comparison with the existing Twitter REST API shows that the Twitter REST API has certain limitations in terms of data acquisition quantity (only 180 requests per 15 min, and a maximum of 100 tweets per request are returned) [36]. Hence, we use Twitterscraper to gather tweets published between 16 February 2008 and 22 August 2018.

Data preprocessing took place in a non-relational (NoSQL) document database. Before the data are entered in the database, a geocoding procedure takes place, where tweets not having a precise location (i.e., coordinates) associated, were analyzed regarding their municipality—and the name and centroid of the municipality are added as object to each data set. The filtering process is intended to do the following:

- sort out tweets of organizations
- filter out tweets with non-tourism keywords
- filter out tweets published several times by the same user and by/for same municipality

This filtering process is intended to sort the remaining tweets into three categories:

1. documents (i.e., tweets) with a strong relevance to tourism (with the help of defined keywords)
2. documents (i.e., tweets) that reveal the location of the author—similar to “I am at . . . ” or “I am in . . . ”.
3. documents (i.e., tweets) that do not fit the categories 1 or 2.

Text analysis is closely related to the filtering process of tweets, mentioned above. In order to perform topic modeling and word frequency analysis, the data have to be preprocessed. The preprocessing step involves a tokenization of the tweets, which breaks text into smaller units (but we preserved hashtags, emoticons or other symbols). In addition, we eliminated English and German stopwords, punctuation and brackets from the data. After the data preprocessing step, we erased all personal information (i.e., username) of the data set in order not to reveal any personal data.

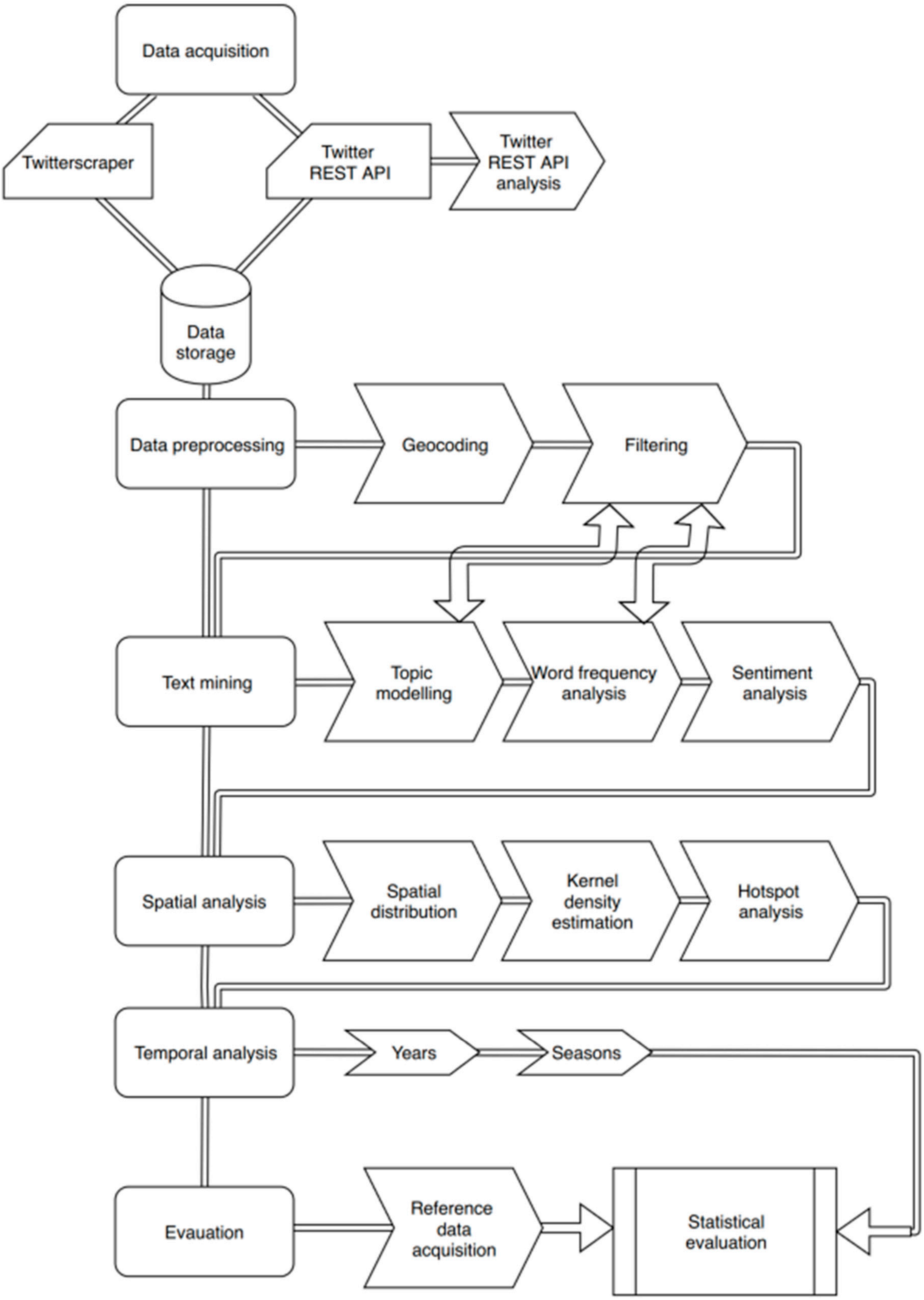


Figure 1. The overall methodology followed in this paper. The process starts with a data acquisition using Twitterscraper [35], based on location and tourism-related keywords on a municipal level. The data processing and filtering is done using a NoSQL document database. The Twitter data are analyzed in terms of spatial properties, whereas the text of the tweets is analyzed with a semantic text analysis. An evaluation seeks to analyze the correlation between the Twitter data and official statistical data on tourism.

Topic modeling is intended to determine groups of topics in accordance with the words used in the text of the tweets [37]. We used Latent Semantic Indexing and Latent Dirichlet Allocation [38–40] for this purpose. Frequency analysis serves the purpose to identify tourism-related keywords, that can be used for filtering. The frequency analysis implemented in R programming language [41]. The tourism related keywords are based on existing vocabularies [42–44]. The sentiment analysis reveals sentiment orientation, which classifies tweets into polarity classes, such as negative, positive or neutral [45]. In order to detect the sentiment of the tweets we utilized the VADER algorithm [46,47]. VADER is a valence-based approach for sentiment analysis, taking into account both the sentiment itself and its intensity. It the table below examples of words and the degree of intensity of their sentiment ranking are displayed. More positive words have higher and more negative words have lower ratings. As a result, Vader provides three metrics annotating the proportion of the text, that falls into the positive, neutral and negative categories. The fourth metric, called the compound score, is the sum of all of the lexicon ratings standardized between −1 and 1 [47]. The closer to 1, the more positive the sentiment is and the closer to −1 more negative the sentiment is. A score of 0 represents a completely neutral text. The approach has proven to be a good fit to social media data since it includes a selection of social media related terms or informal writing such as emotions, acronyms or multiple punctuation marks. Even acronyms "omg" (oh my god), or "smh" (shaking my head) are included in the lexicon. Acronyms help to determine the intensity of positivity and negativity. The paper [46] shows that Vader even outperforms human individual assessors.

The spatial and temporal analysis of the obtained data focuses on the evaluation of spatial distribution and temporal patterns within test region. Our objectives are to statistically analyze the growth of the tweets in our dataset through the years and to compare them with regard to touristic seasons. All tweets that are related to a single municipality have exactly the same coordinates—the municipality's centroid. To be able to perform further spatial analyses at both municipality and state level, we randomly distribute the tweets within each corresponding municipality. Based on this spatial distribution, further cluster-related analyses are performed. The tweets' distribution accuracy within the municipality is therefore not absolutely correct, but this approach is an appropriate solution for visualizing the tweets and obtaining answers of relative distribution within the whole state of Styria through further spatial analyses. Hot Spot analysis is a method of cluster identification and visualization that we took advantage of within our research. Hot Spot indicates statistically significant spatial clusters of high values as hot spots and of low values as cold spots. It is based on Getis-Ord Gi* statistic and optionally also on a set of weighted features [48,49]. Using Hot Spot analysis for area measurement purposes, an area of high quantity or intensity of observed feature values can be identified [50,51]. An alternative to Hot Spot analysis with manual settings, is the Optimized Hot Spot Analysis. With the use of Optimized Hot Spot Analysis, the characteristics of the input data are evaluated automatically, in order to extract the parameters that are needed for an optimal result. After aggregation of the incident data into weighted features an appropriate scale of analysis is determined. The statistically significant values of the end result are adjusted multiple times. Adjustment of the values is performed according to the testing and spatial dependency with the help of the False Discovery Rate (FDR) correction method [52]. The Hot Spot analysis is implemented with the tool Optimized Hot Spot Analysis within the desktop GIS "ArcGIS". Kernel Density Estimation (KDE) of Twitter data is a promising technique of density estimation, as it belongs to a non-parametric analysis with no fixed structure and depends on the point data [3]. With KDE, a magnitude-per-unit area is calculated from point features, hence in this Study from Twitter point data—meaning from positions of the Tweets. By calculating the density, we are spreading the input values over a raster cell. With KDE method in turn, the known quantity value (e.g., a population field) of a feature is spread out from the point location based on a quadratic kernel function [53] originally described by [54]. Results of the KDE can be visualized by a heat map. Heat map refers to feature concentration from a geographic, namely spatial, perspective. From point or line data an interpolated surface is created in order to indicate the density of occurrence. It is a way of visualizing a density surface by colored gradient in order to easily identify locations of higher densities or clusters of geographic features [55].

A temporal analysis is done with respect to the annual and seasonal tweet variability. In addition, the authors investigate the temporal pattern of tweet activity on a district level.

The evaluation seeks to underpin the results revealed from the social media data. Hence, a comparison between the tweets extracted through the Twitterscraper package and reference data is carried out. The reference data originates from official authorities for statistics—e.g., federal Statistical Office of Austria, and the statistics of the Province of Styria. The evaluation approach was based on correlation analysis between twitter data and official data, representing two scale variables. Although graphs may visually demonstrate the relationship between social media and the reference data, the correlation coefficient needs to be determined to confirm whether the correlation is statistically significant or insignificant. We used the Pearson's coefficient to determine the level of correlation of social media and reference data from 2008 till 2017, at both a province and district level. The seasonal aspect is also taken into consideration. Hence, the Pearson's coefficient concerning for twitter and reference data for summer and winter season, for each district is calculated as well.

4. Experiment

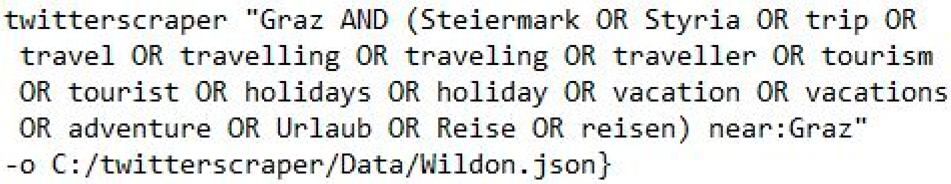
The experiment is, according to the research questions, carried out using a geographically small test area with a low number of tweets (compared to other touristic hotspots—like London or New York). Hence, our test area is the Province of Styria, Austria. The experiment follows the methodology described in Section 3, and analyzes tweets published between 16 February 2008 and 22 August 2018.

4.1. Data Acquisition

The processes started with the data acquisition with the help of Twitterscraper Python package [35]. The data query was performed with the help of the municipality names, exploiting geographic nearness, and used 12 tourism related keywords (see Table 1) based on [42–44]. We queried each community name separately—which resulted in 287 queries in total. The query blueprint used to acquire the data is shown in Figure 2. In Figure 2 the Twitterscraper query for the city of Graz is depicted.

Table 1. The tourism related keywords to acquire Twitter data with the help of the Twitterscraper package.

English Keywords	German Keywords
Styria	Steiermark
trip	Urlaub
travel	Reise
travelling, traveling	reisen
traveller	
Tourism	
Tourist	
holidays, holiday	
vacation, vacations	
Adventure	



```
twitterscraper "Graz AND (Steiermark OR Styria OR trip OR
 travel OR travelling OR traveling OR traveller OR tourism
 OR tourist OR holidays OR holiday OR vacation OR vacations
 OR adventure OR Urlaub OR Reise OR reisen) near:Graz"
-o C:/twitterscraper/Data/Wildon.json}
```

Figure 2. Query syntax for the municipality of Styria, Graz.

We acquired 35,234 tweets in total, from 287 municipalities for the time span 2 February 2008 until 22 August 2018. From the 287 queries, 80 queries returned no tweets at all, whereas we received tweets for 207 municipalities of Styria.

4.2. Data Preprocessing & Filtering

The acquired data is stored in a MongoDB database where each document (each tweet) gets geocoded—i.e., it was updated with the coordinates of corresponding municipality—in the coordinate reference system WGS84. The filtering process results in 3 different groups of tweets:

1. documents (i.e., tweets) with a strong relevance to tourism (with the help of defined keywords)
2. documents (i.e., tweets) that reveal the location of the author—similar to "I am at . . . " or "I am in . . . ".
3. documents (i.e., tweets) that do not fit the categories 1 or 2.

Based on the methodology described in Section 3, we applied term frequency analysis to determine the most frequent terms in conjunction with tourism contained in the tweets. This is implemented in the package R. For this purpose, the texts were transformed to lowercase letters, punctuation removed, extra whitespaces removed, numbers and stopwords (English & German) removed. Finally, URLs and ASCII signs are eliminated as well. This dataset is the basis for the document-term matrix that is used for word frequency determination. As a result, we obtained 229 words which were used for filtering purposes—i.e., to assign tweets to the groups 1, 2 or 3.

The word cloud of the first 150 most frequent words is depicted in Figure 3. During filtering we applied text mining in order to choose the final selection of tourism relevant words and then returned to filtering in order to extract the final dataset of our tweets. In the final dataset 6,953 documents/tweets are present. The quantity of tweets and their according groups (1–3) are depicted in Table 2.

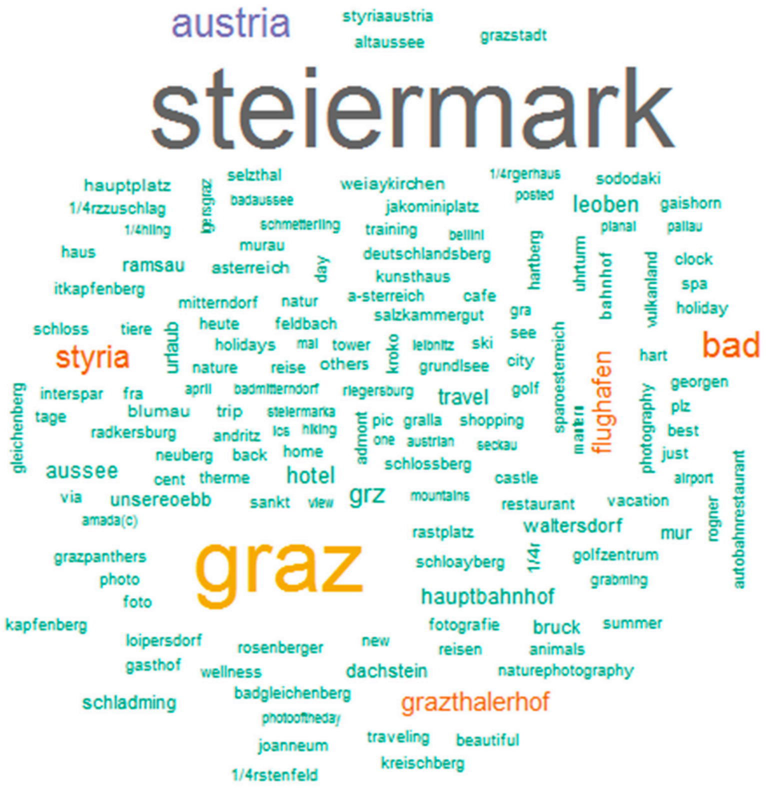


Figure 3. Word cloud of the 150 most frequent words in the twitter dataset.

Table 2. Quantity of tweets in each tweet group (1–3).

Group	Quantity of Tweets
1. predefined tourism keywords	1391
2. "I am at . . . " structure	2717
3. rest of tweets	2845
Total	6953

4.3. Sentiment Analysis

A sentiment analysis and topic modelling analysis were performed utilizing Orange Data Mining software [38,56]. Orange Data Mining software is an open-source software. To create the sentiment analysis, we applied the VADER algorithm. Sentiment was categorized according to the compound analysis result with more than 0 and up till 1 classified as positive, 0 as neutral and between less than 0 and −1 as negative sentiment. Any spatial analyses and geographical visualizations of the results are performed in ArcGIS.

4.4. Kernel Density Estimation & Hot Spot Analysis

The Hot Spot analysis is implemented using the Optimized Hot Spot Analysis tool provided by ArcGIS. For aggregation purposes a fishnet cell size of 1 km. Kernel Density Estimation was applied with a search radius of 5 km and a cell size of 500 m. Such search radius and output cell size were determined after analyzing results of different settings ranging from a 1 to 10 km search radius and a 100 up to 5000 m output cell size. Selected values were determined as suitable because these settings enable the creation of a convincing visualization with meaningful values. Due to big differences in resulting densities, we used a Jenks Natural Breaks classification method, which arranges values into different classes, minimizing the difference within a class and maximizing it between classes.

4.5. Temporal Analysis & Evaluation

The temporal analysis in the experiment considers classical annual variations as well as seasonal changes. Similar to the reference data, we used the summer and winter tourism season. Tweets from 1 May till 31 October are assigned to the summer season, while tweets from 1 November till 30 April of the following year to the winter season. For evaluation purposes we used the software SPSS in order to perform a correlation analysis between twitter data and reference data.

5. Results

This section contains the results achieved with the experiment and methodology described in the prior sections. The results section covers the evaluation of the results as well.

Basically, the resulting set of tourism-related tweets selected from all tweets in the time period between 16 February 2008 and 22 August 2018 contains 6953 tweets. The number of the tourist-related tweets at a district level is given in Figure 4. Obviously, the city of Graz—as capital of Styria—has most tweets followed by the district of Liezen and Hartberg/Fürstenfeld. In Figure 5 we depict the spatial distribution of the tourism-related tweets according to the districts of the Province of Styria.

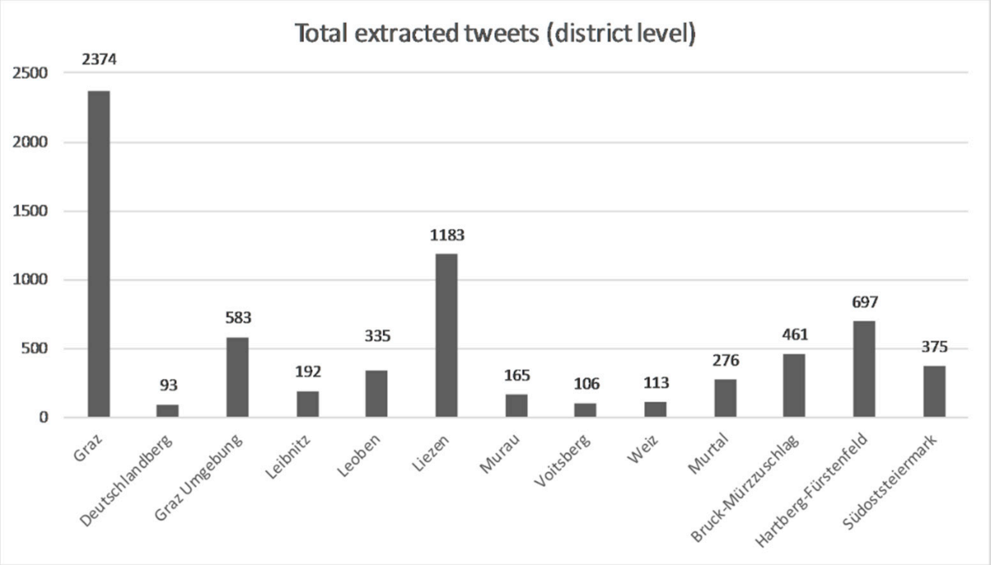


Figure 4. Number of extracted tweets for each district of the Province of Styria. On the *y*-axis the number of total tweets and on the *x*-axis the districts are depicted.

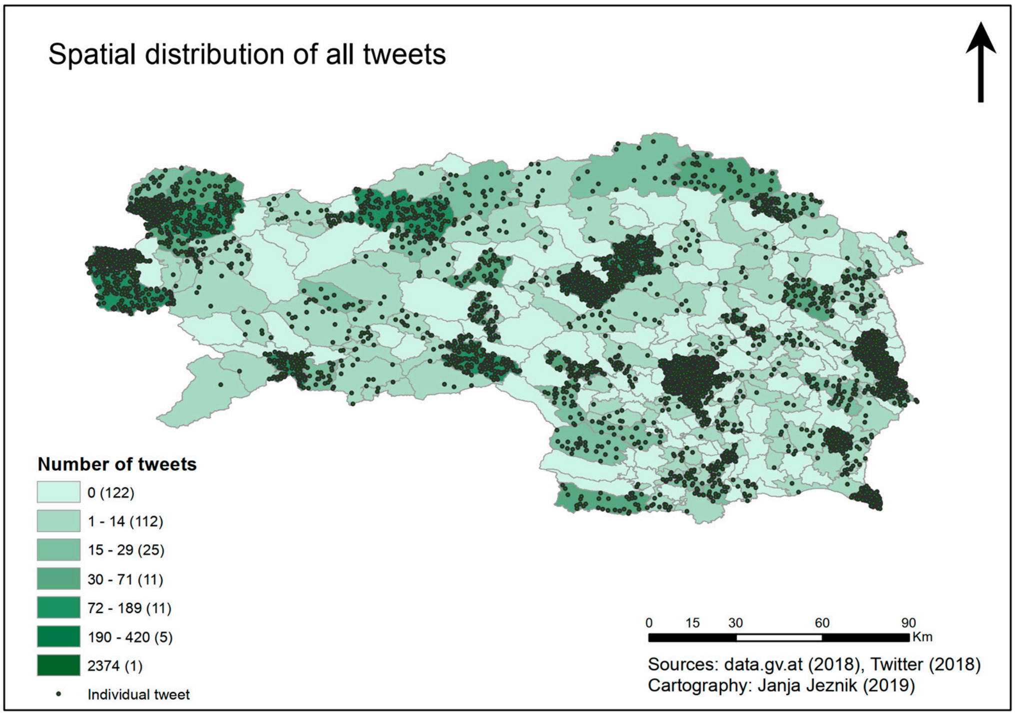


Figure 5. Spatial distribution of the tourism-related tweets according to the districts of the Province of Styria. Each district is colored according to the absolute number of tweets in the evaluation period and each individual tweet is represented by a green dot in the map.

Figure 6 shows the distribution of tweets according to the three categories, described in Section 4.2. In addition, the hot spot analysis (see Figure 7) in conjunction with Figure 6 shows six clear and evident hot spots with a confidence of 99%. As an obvious and expected result we can observe a hot spot of the state's capital city of Graz and its surroundings in a central part of southeastern Styria. Further city tourism-related hot spot is the urban agglomeration of the municipalities of Leoben, Bruck an der Mur and Kapfenberg. A hot spot in the shape of a semicircle in the northwest belongs to the tourism intensive municipalities of Bad Aussee and Ramsau am Dachstein, offering a selection of nature- and mountains-related tourism activities, e.g., hiking, skiing or bathing in natural lakes. The last three significant hot spots with 99% confidence belong to the municipalities of health and beauty thermal tourism, namely Bad Radkersburg in the extreme south-east, Bad Gleichenberg further north and Loipersdorf bei Fürstenfeld in the east. Statistically significant cold spots represent areas confronting low tourist visits and are, with the exception of the whole district of Deutschlandsberg located on the edges of more populated areas such as in the mountains or in the countryside. Cold spot clusters with 99% confidence are located in the western and northern mountainous parts of the state, mostly in the districts of Liezen and Bruck-Mürzzuschlag. However, significant cold spots are located also on the peripheral edges of all other districts (except Leibnitz).

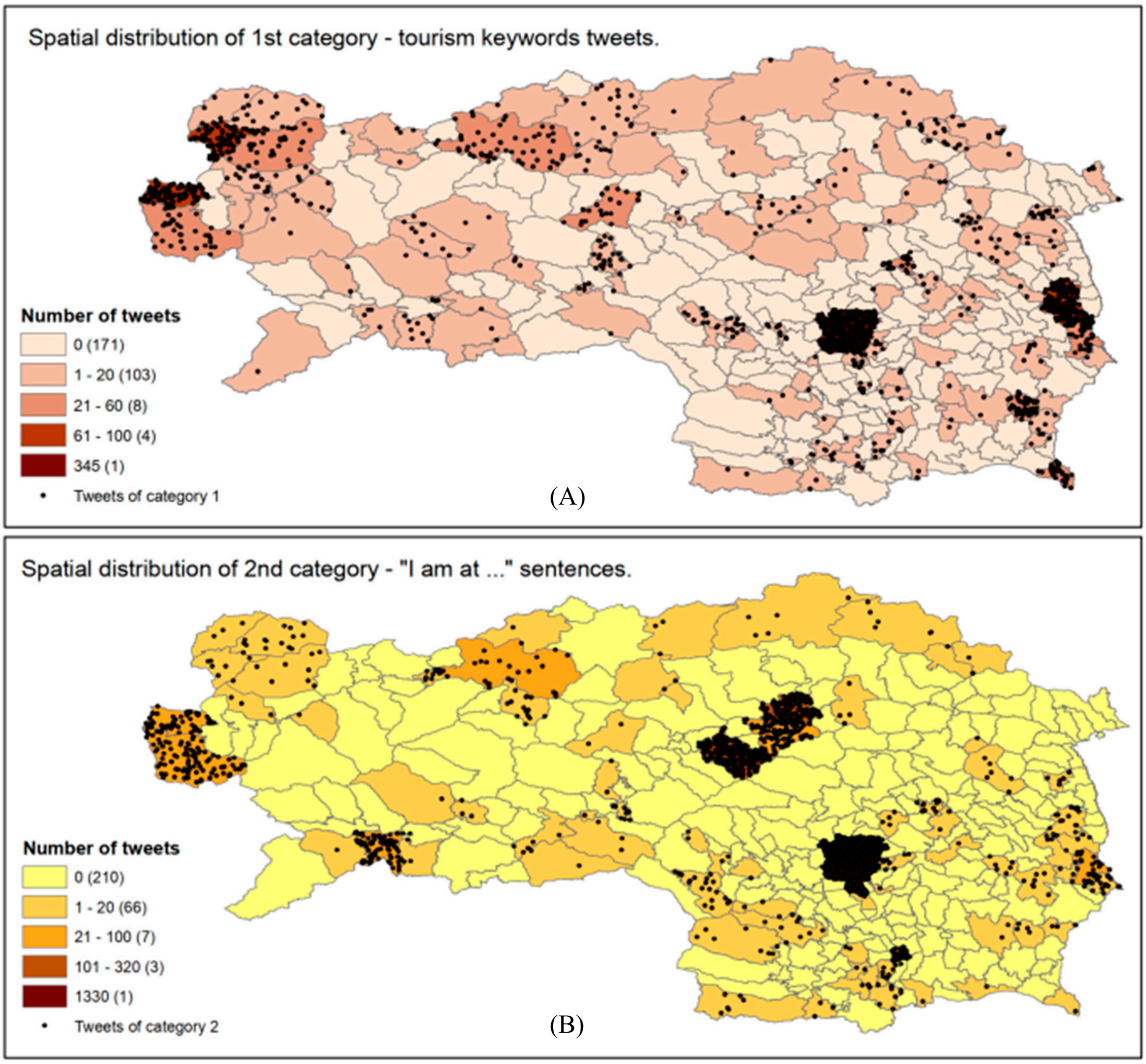


Figure 6. *Cont.*

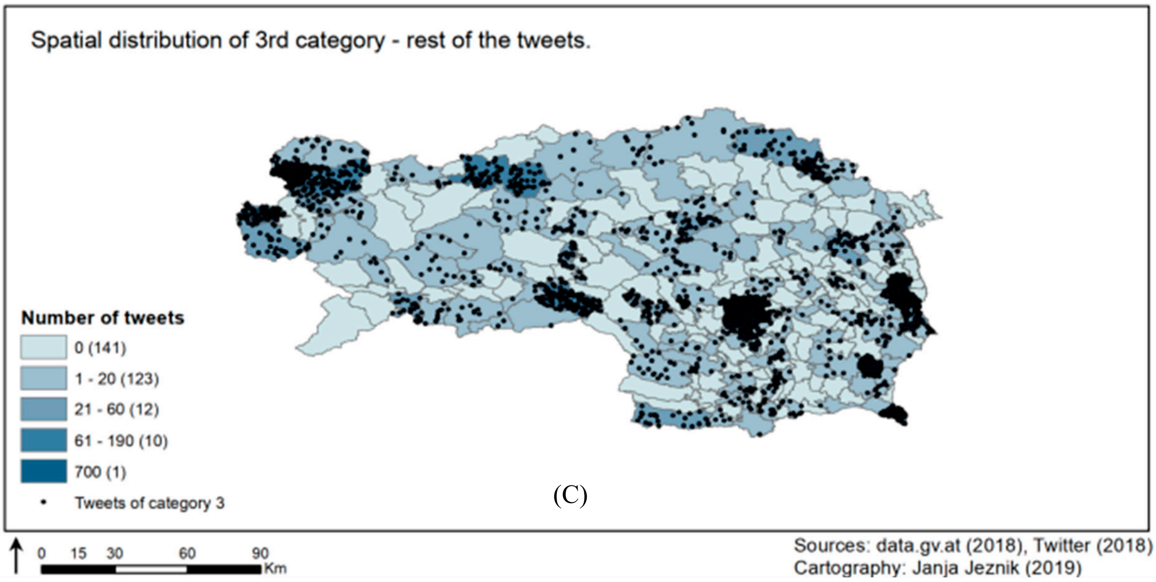

Spatial distribution of 3rd category - rest of the tweets.
Number of tweets
0 (141)
1 - 20 (123)
21 - 60 (12)
61 - 190 (10)
700 (1)
Tweets of category 3
(C)
0 15 30 60 90 Km
Sources: data.gv.at (2018), Twitter (2018)
Cartography: Janja Jeznik (2019)

Figure 6. Spatial distribution of the tourism-related tweets according their subgroups. Map (**A**) depicts the number of tweets of category 1, having tourist related keywords. Map (**B**) shows the tweets of category 2—tweets that are of the form "I am at . . . ". Map (**C**) shows the rest of tourist related tweets—i.e., category 3.

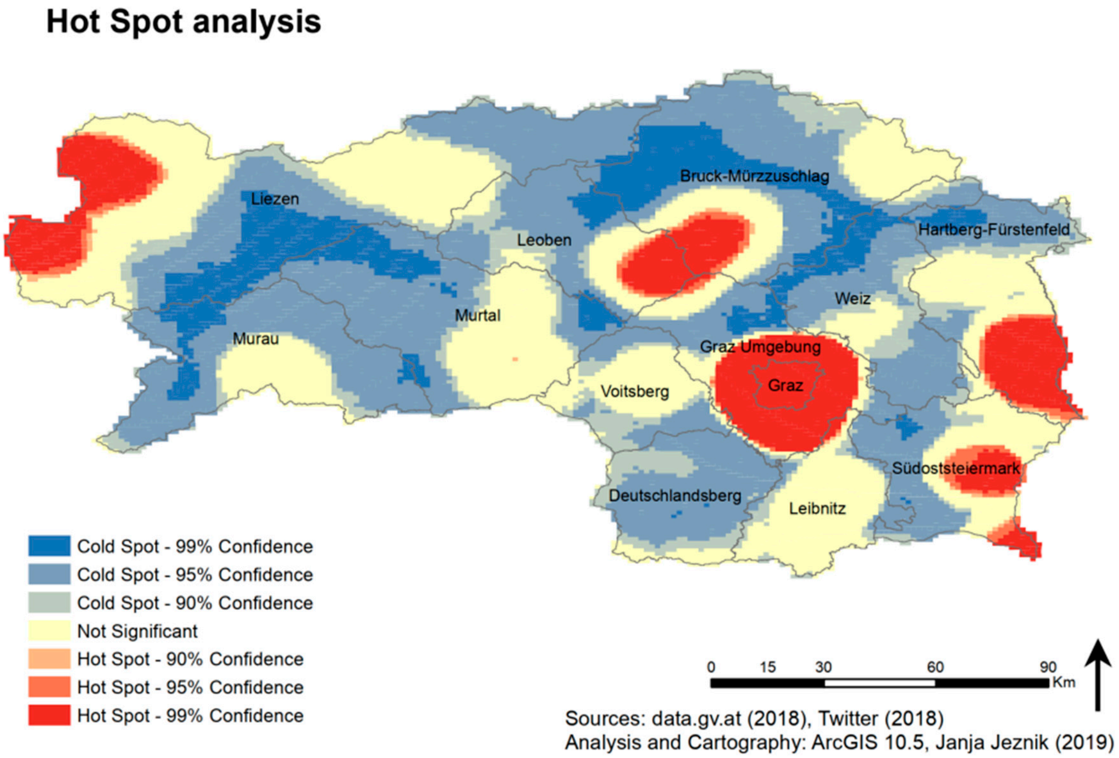
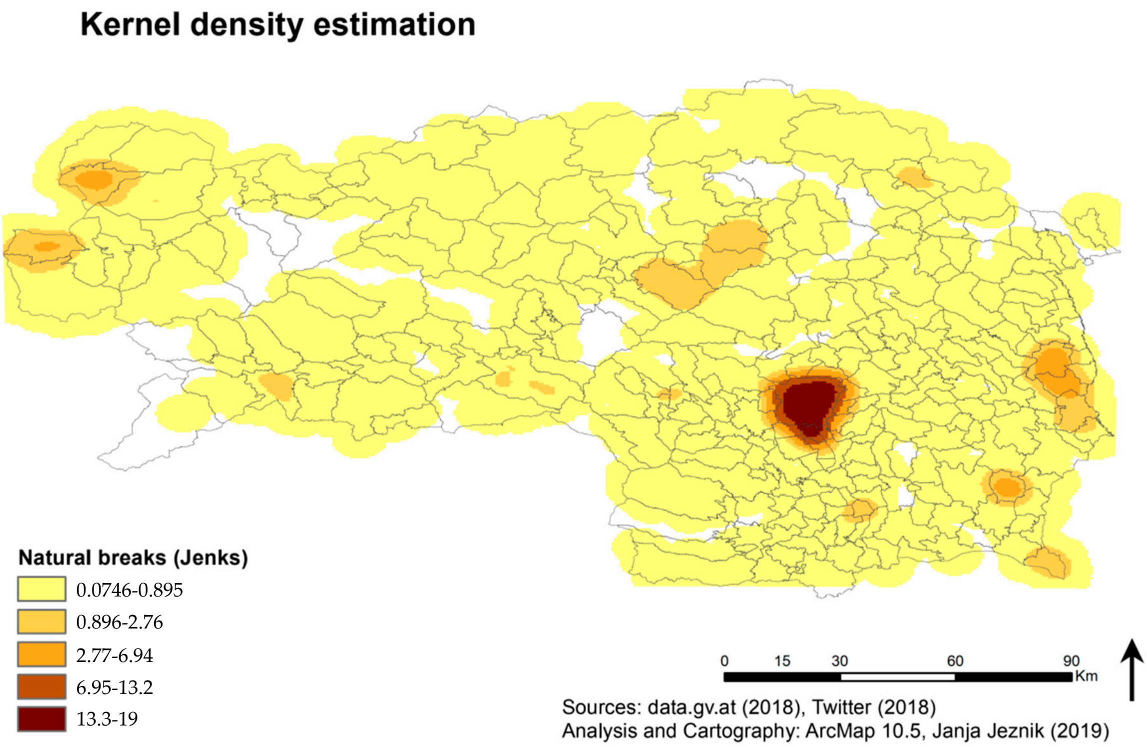

Hot Spot analysis
Liezen
Bruck-Mürzzuschlag
Leoben
Hartberg-Fürstenfeld
Murtal
Murau
Weiz
Graz Umgebung
Graz
Voitsberg
Südoststeiermark
Deutschlandsberg
Leibnitz
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
0 15 30 60 90 Km
Sources: data.gv.at (2018), Twitter (2018)
Analysis and Cartography: ArcGIS 10.5, Janja Jeznik (2019)

Figure 7. Optimized Hot Spot analysis of the tweets and the districts of the Province of Styria. The tweets are aggregated, based on a fishnet with 1 km cell size. Regions having a red color represent a hot spot, whereas regions colored in blue are regarded as cold spots.

The Kernel Density Estimation results confirm significant changes in values, especially between the state's capital Graz and other, more rural, areas, as mentioned above (see Figure 8). Most of the study area corresponds to a density of less than 1 Tweet per search radius. White values are locations with the kernel density of 0 and were excluded from the visualization in order to stress the difference to other areas.



Sources: data.gv.at (2018), Twitter (2018)
Analysis and Cartography: ArcMap 10.5, Janja Jeznik (2019)

Figure 8. Results of the Kernel Density Estimation with a search radius of 5 km and an output cell size of 500 m. The resulting densities are classified using Jenk's Natural Breaks method. The over-representation of the city of Graz, in comparison with the mostly rural area is visible.

Spatial patterns of positive, negative and neutral sentiments are represented in the map in Figure 9. However, it is important to take a notice, that there are some municipalities with the word "Bad" in their name, such as Bad Radkersburg or Bad Aussee. In these cases, although the user was referring to the name of the municipality, the tweet was categorized as negative due to the word "Bad" which refers to thermal or swimming areas in German, but is understood by the algorithm as the English word referring to something negative. As in Figure 10, six municipalities are marked with a cross hatch pattern, in order to point out their irrelevance to the category of very high negative sentiment percentage.

As can be seen in Figure 11, there is a strong relationship between the crowdsourced data and reference data from 2011 onwards. In order to achieve relevant results, we applied correlation analysis to two separate datasets—one to the whole dataset from 2008 till 2018 (see Figure 12), whereas in 2018 only the data of the first half year (and also only first half year of reference data) was used. There is a significant correlation at the 0.05 level with Pearson's coefficient of 0.650. The Sig(2-tailed) value under 0.05 confirms statistical significance of the correlation. On the other hand, considering only the full years between 2011 and 2017 (see Figure 13), the Pearson's coefficient increases to 0.772, which, according to the definitions in our theoretical section, is already a strong correlation. The correlation for years between 2011 and 2017 is also significant at the 0.05 level.

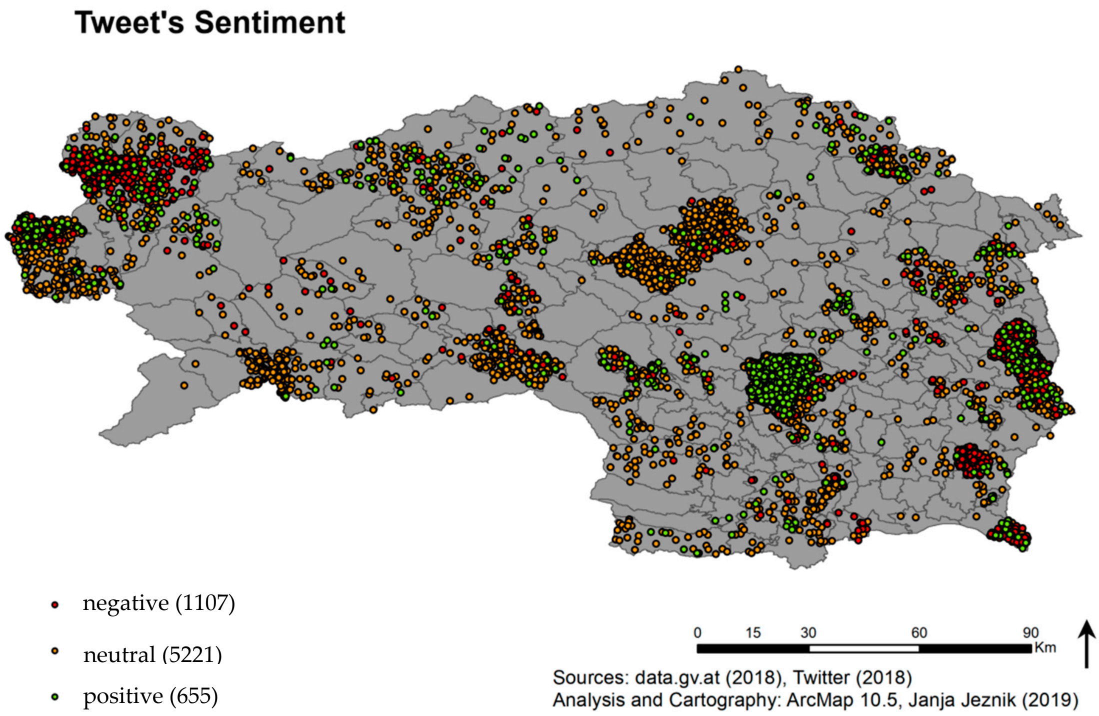

Tweet's Sentiment
● negative (1107)
● neutral (5221)
● positive (655)
0 15 30 60 90
Km
Sources: data.gv.at (2018), Twitter (2018)
Analysis and Cartography: ArcMap 10.5, Janja Jeznik (2019)

Figure 9. Results of the sentiment analysis. Basically, there are 1107 negative tweets (marked as red dots), 5221 neutral tweets (orange dots) and 655 positive tweets (green dots). The tweets with negative sentiment are also a result of German place names that start with "Bad ... ", denoting spa towns. The algorithm flags the German "Bad ... " with the meaning of the English word "bad".

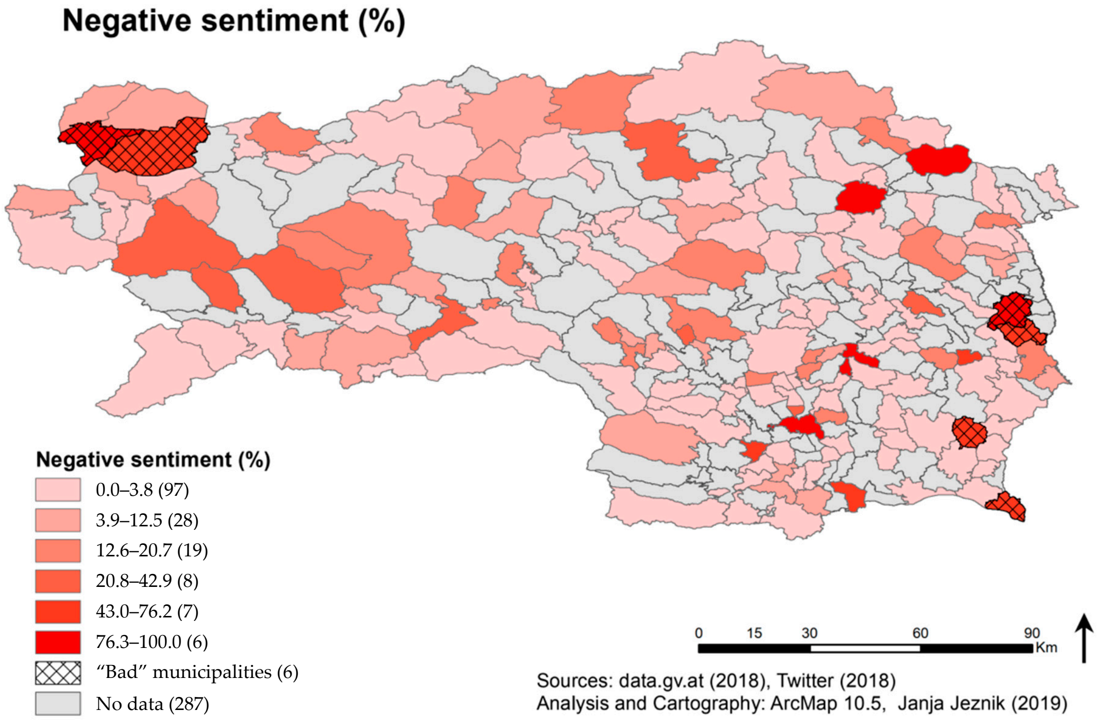


Figure 10. Results of the sentiment analysis. The percentage of the negative Tweets is depicted in this map in different classes. The tweets with negative sentiment are also a result of German place names that start with "Bad ... ", denoting spa towns. The algorithm flags the German "Bad ... " with the meaning of the English word "bad". These six towns are marked with a cross hatch pattern.

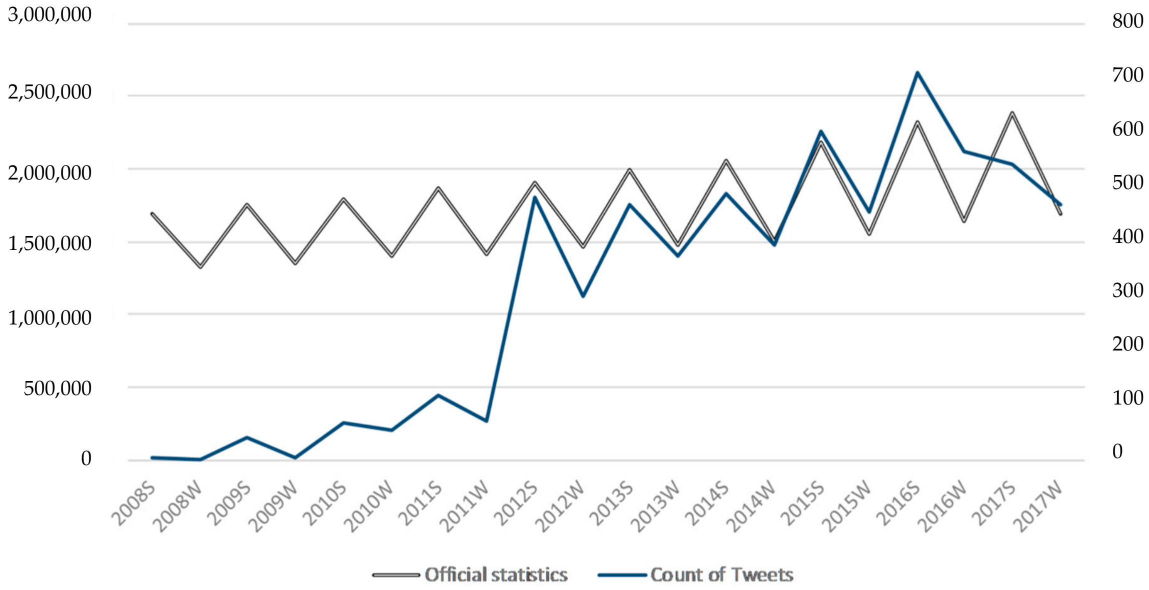


Figure 11. Comparison of absolute numbers of arrivals (from official statistics) and crowdsourced tweets for the evaluation period—for winter and summer season. Summer season is denoted with the suffix "S" and winter season with the suffix "W". The number of arrivals is denoted as grey line, whereas the absolute number of tweets is depicted as blue line. The x-axis on the left shows the number of arrivals from the official statistics, whereas the right x-axis denotes the absolute number of tweets.

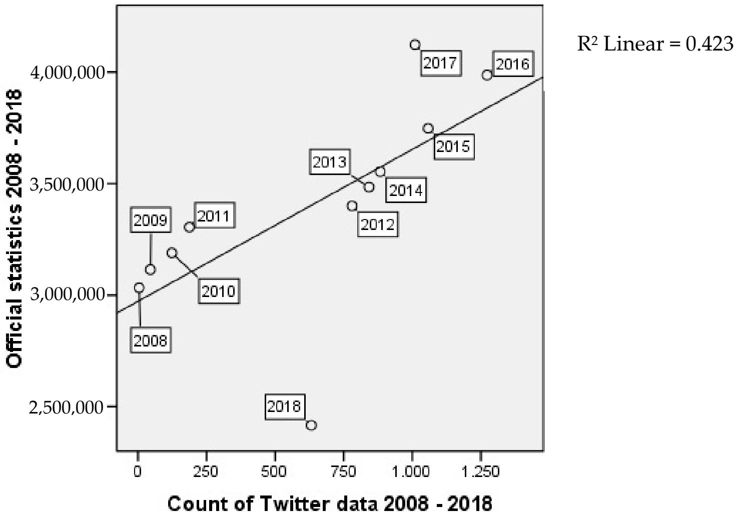


Figure 12. Scatter plot of official tourism and Twitter data distribution for the period 2008–2018. The x-axis denotes the absolute number of twitter data twitter data and the y-axis the absolute number of arrivals from the official statistics. R^2 is the coefficient of determination of the regression line.

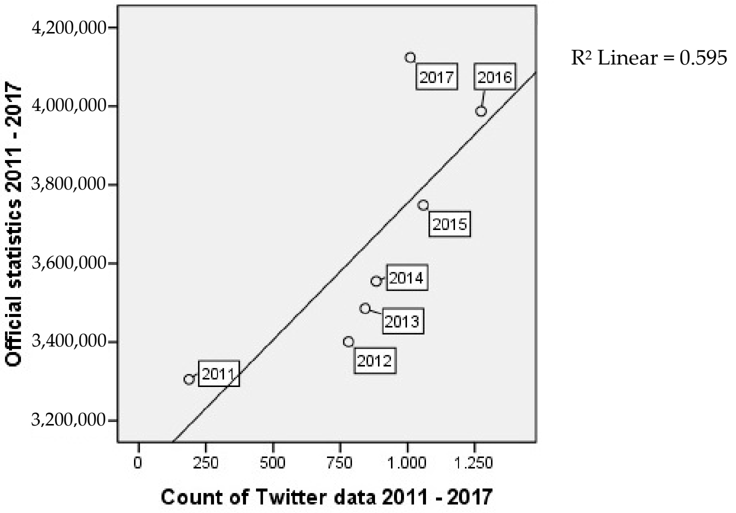


Figure 13. Scatter plot of official tourism and Twitter data distribution for the period 2011–2017. The *x*-axis denotes the absolute number of twitter data twitter data and the *y*-axis the absolute number of arrivals from the official statistics. R^2 is the coefficient of determination of the regression line.

Further correlations were also determined at the district level for the time period from 2011–2017, since we believe that the low count of tweets in the years 2008, 2009 and 2010 does not reflect such low tourist visitors but rather the low number of Twitter users in general. As can be seen in Table 3, there are 5 districts with positive correlation over 0.8 and are statistically significant at the 0.001 level. Murau, Liezen and Graz Umgebung have a strong correlation over 0.85 and Graz and Südoststeiermark even a very strong correlation over 0.9. With 0.694 there is also Leoben with significant correlation on 0.001 level. However, at half of the districts there is no significant correlation between our and reference data. One district, Voitsberg, even confronts negative correlation at 0.005 significance level. In addition, Table 3 shows the Pearson's coefficients for the summer and winter seasons. Districts are arranged according to the correlation strength in the summer season. It can be seen that correlations are, in general, higher in the summer compared to the winter. Graz, Südoststeiermark, Graz Umgebung and Liezen are on the top with the highest values for both seasons. The other districts fluctuate—for instance in Murau there is a significant positive correlation in summer but a very weak negative relationship with no significant correlation in winter. In Bruck-Mürzzuschlag, Deutschlandberg, Hartberg-Fürstenfeld and Voitsberg there are positive relationships in one season and negative relationships in another season.

Table 3. Correlation coefficients at a district level for tourism years 2011–2017 and for winter and summer season. The ** denote statistical significance at the 0.001 level, and * denotes statistical significance at the 0.005 level.

District	Pearson's Coefficient Overall	Pearson's Coefficient Summer	Pearson's Coefficient Winter
Graz	0.94 **	0.916 **	0.948 **
Südoststeiermark	0.937 **	0.892 **	0.790 **
Graz Umgebung	0.874 **	0.835 **	0.890 **
Liezen	0.888 **	0.839 **	0.913 **
Murau	0.894 **	0.736 *	0.557
Murtal	0.081	0.720 *	−0.168
Leibnitz	0.694 *	0.704 *	0.599
Leoben	0.623	0.683 *	0.582
Bruck-Mürzzuschlag	0.523	0.682 *	−0.715 *
Weiz	0.42	0.368	0.246
Deutschlandsberg	0.327	0.310	0.280
Hartberg-Fürstenfeld	0.288	−0.258	0.681 *
Voitsberg	−0.681	−0.564	−0.422

6. Discussion and Conclusions

The paper evaluates crowdsourced data of the platform Twitter in several "dimensions" in order to analyze the suitability to act as proxy for tourist related questions and tourist flows. In particular, we are interested in the validity of the results gained from small crowdsourced data sets—at a regional level. Therefore, we developed a methodology to analyze the tweets in a spatio-temporal and semantic dimension and tried to evaluate the findings with the help of official tourism data. The methodology was applied to the province of Styria in Austria, for the time period from 2008 until 2018.

First, the absolute annual number of tourist related Tweets is of interest in order to quantify the size of the user basis. From 2008 until 2010 the absolute number is far below 100 for each winter and summer season. Hence, any statistical evaluation of such a low number is questionable due to the low sample size. Thus, we decided to have a detailed look at the years 2011 until 2018—as there are several hundred tourist related Tweets to be analyzed for each season. Comparing the crowdsourced dataset with the official tourism statistic with the help of Pearson's correlation coefficient, results in half of the districts being significantly correlated. The same applies to the seasonal evaluation, where half of the districts relate better to the reference data in summer than in winter. This reveals that the spatial scale of a state is still an adequate one, while results at a district level must be critically reviewed.

Despite the small size of the crowdsourced tourism-related data set, it is visible that the spatial and temporal distribution of the Tweets is similar to the official tourism data. Spatial clusters of crowdsourced data coincide with the municipalities that have a strong tourism industry (i.e., high number of overnight stays). These touristic regions are located in the North-West (e.g., Bad Aussee, Ramsau am Dachstein) and in the thermal region in the South-East (e.g., Bad Radkersburg, Bad Gleichenberg, Loipersdorf) and the central region (Graz, Bruck/Mur, Leoben). The topic modeling for such a small data set using Latent Ditrichlet Allocation results in topics that are useless, whereas the term frequency analysis resulted in an accurate word cloud that represents touristic subjects in Styria. Text mining approaches for sentiment analysis applied to Twitter data are of great importance, as text in social media is specific and written without any general rules—mostly in an everyday language including abbreviations, typing mistakes, hashtags and emotion tags. Due to multilingual datasets, mistakes may happen during translation steps. In the paper, we highlighted the example of municipalities with the name "Bad" in it. These social media data are regarded to having a negative sentiment, due to the translation of the place name "Bad". Hence, improvements in analyses with multilingual data could help to improve the accuracy of the results.

In addition, data filtering is crucial for this kind of analysis—as only clean datasets lead to clear and objective results and information. The approach taken in this paper is a combination of an automatic and manual determination of tourism-related tweets—which is based on their content. We strongly believe that a machine learning approach could increase the quality of the filtering approach and could thus be one of the next steps in this research topic. Text mining approaches and sentiment analysis of crowdsourced datasets can help to determine and track user's opinions. Hence, these data can be a valuable proxy for the opinion of tourists. In particular, in our paper the Tweets—especially from 2011—2018)—can be used as proxy to determine the demand and occupancy in the tourism industry.

Author Contributions: Conceptualization, Johannes Scholz; data curation, Janja Jeznik; investigation, Janja Jeznik; methodology, Johannes Scholz and Janja Jeznik; project administration, Johannes Scholz; visualization, Janja Jeznik; Author—original draft, Johannes Scholz, Author revisions: Johannes Scholz. All authors have read and agreed to the published version of the manuscript.

Funding: Open Access Funding by the Graz University of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. See, L.; Mooney, P.; Foody, G.; Bastin, L.; Comber, A.; Estima, J.; Fritz, S.; Kerle, N.; Jiang, B.; Laakso, M.; et al. Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS Int. J. Geo Inf.* **2016**, *5*, 55. [CrossRef]
2. Papapesios, N.; Ellul, C.; Shakir, A.; Hart, G. Exploring the use of crowdsourced geographic information in defence: Challenges and opportunities. *J. Geogr. Syst.* **2019**, *21*, 133–160. [CrossRef]
3. Kovacs-Gyori, A.; Cabrera-Barona, P.; Ristea, A.; Havas, C.; Resch, B. # London2012: Towards Citizen-Contributed Urban Planning Through Sentiment Analysis of Twitter Data. *Urban Plan.* **2018**, *3*, 75–99.
4. Capineri, C.; Haklay, M.; Huang, H.; Antoniou, V.; Kettunen, J.; Ostermann, F.; Purves, R. *European Handbook of Crowdsourced Geographic Information*; Ubiquity Press: London, UK, 2016; ISBN 9781909188792.
5. Aggarwal, C.C.; Abdelzaher, T. Social sensing. In *Managing and Mining Sensor Data*; Springer: Boston, MA, US, 2013; pp. 237–297. ISBN 9781461463092.
6. Janowicz, K.; McKenzie, G.; Hu, Y.; Zhu, R.; Gao, S. Using Semantic Signatures for Social Sensing in Urban Environments. In *Mobility Patterns, Big Data and Transport Analytics*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 31–54.
7. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
8. Janowicz, K.; Gao, S.; McKenzie, G.; Hu, Y.; Bhaduri, B. GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 625–636. [CrossRef]
9. Zeile, P.; Resch, B. Combining Biosensing Technology and Virtual Environments for Improved Urban Planning. *GI_Forum* **2018**, *1*, 344–357. [CrossRef]
10. Zheng, Y.; Liu, T.; Wang, Y.; Zhu, Y.; Liu, Y.; Chang, E. Diagnosing New York city's noises with ubiquitous data. In *Proceedings of the UbiComp 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*; Association for Computing Machinery, Inc.: New York, NY, USA, 2014; pp. 715–725.
11. Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; Ratti, C. Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 260–271. [CrossRef]
12. Hübl, F.; Cvetojevic, S.; Hochmair, H.; Paulus, G. Analyzing Refugee Migration Patterns Using Geo-tagged Tweets. *ISPRS Int. J. Geo Inf.* **2017**, *6*, 302–325. [CrossRef]
13. Zagheni, E.; Garimella, V.R.K.; Weber, I. Inferring international and internal migration patterns from twitter data. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 439–444.
14. Senaratne, H.; Bröring, A.; Schreck, T.; Lehle, D. Moving on Twitter: Using episodic hotspot and drift analysis to detect and characterise spatial trajectories. In *Proceedings of the 7th ACM Sigspatial International Workshop on Location-Based Social Networks—LBSN '14*; ACM Press: New York, NY, USA, 2014; pp. 23–30.
15. Cassa, C.A.; Chunara, R.; Mandl, K.; Brownstein, J.S. Twitter as a sentinel in emergency situations: Lessons from the Boston marathon explosions. *PLoS Curr.* **2013**, 5. [CrossRef]
16. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860.
17. Hall, C.M. Spatial analysis: A critical tool for tourism geographies. *Routledge Handb. Tour. Geogr.* **2012**, *1*, 163–173.
18. Claster, W.; Pardo, P.; Cooper, M.; Tajeddini, K. Tourism, travel and tweets: Algorithmic text analysis methodologies in tourism. *Middle East J. Manag.* **2013**, *1*, 81–99. [CrossRef]
19. Beiró, M.G.; Panisson, A.; Tizzoni, M.; Cattuto, C. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Sci.* **2016**, *5*, 30. [CrossRef]
20. Bassolas, A.; Lenormand, M.; Tugores, A.; Gonçalves, B.; Ramasco, J.J. Touristic site attractiveness seen through Twitter. *EPJ Data Sci.* **2016**, *5*, 12. [CrossRef]
21. Han, B.; Cook, P.; Baldwin, T. Text-based Twitter user geolocation prediction. *J. Artif. Intell. Res.* **2014**, *49*, 451–500. [CrossRef]

22. Khan, S.F.; Bergmann, N.; Jurdak, R.; Kusy, B.; Cameron, M. Mobility in cities: Comparative analysis of mobility models using Geo-tagged tweets in Australia. In Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), Beijing, China, 10–12 March 2017; pp. 816–822.
23. Zhou, X.; Wang, M.; Li, D. From stay to play—A travel planning tool based on crowdsourcing user-generated contents. *Appl. Geogr.* **2017**, *78*, 1–11. [CrossRef]
24. Walden-Schreiner, C.; Leung, Y.-F.; Tateosian, L. Digital footprints: Incorporating crowdsourced geographic information for protected area management. *Appl. Geogr.* **2018**, *90*, 44–54. [CrossRef]
25. Alivand, M.; Hochmair, H.H. Spatiotemporal analysis of photo contribution patterns to Panoramio and Flickr. *Cartogr. Geogr. Inf. Sci.* **2017**, *44*, 170–184. [CrossRef]
26. Martí, P.; Serrano-Estrada, L.; Nolasco-Cirugeda, A. Social Media data: Challenges, opportunities and limitations in urban studies. *Comput. Environ. Urban Syst.* **2019**, *74*, 161–174. [CrossRef]
27. Jiang, B.; Ren, Z. Geographic space as a living structure for predicting human activities using big data. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 764–779. [CrossRef]
28. Sinclair, M.; Mayer, M.; Woltering, M.; Ghermandi, A. Using social media to estimate visitor provenance and patterns of recreation in Germany's national parks. *J. Environ. Manag.* **2020**, 263. [CrossRef]
29. Del Vecchio, P.; Di Minin, A.; Petruzzelli, A.M.; Panniello, U.; Pirri, S. Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. *Creat. Innov. Manag.* **2018**, *27*, 6–22. [CrossRef]
30. Pikkemaat, B.; Peters, M.; Bichler, B.F. Innovation research in tourism: Research streams and actions for the future. *J. Hosp. Tour. Manag.* **2019**, *41*, 184–196. [CrossRef]
31. Aydin, G. Social media engagement and organic post effectiveness: A roadmap for increasing the effectiveness of social media use in hospitality industry. *J. Hosp. Mark. Manag.* **2020**, *29*, 1–21. [CrossRef]
32. Pino, G.; Peluso, A.M.; Del Vecchio, P.; Ndou, V.; Passiante, G.; Guido, G. A methodological framework to assess social media strategies of event and destination management organizations. *J. Hosp. Mark. Manag.* **2019**, *28*, 189–216. [CrossRef]
33. Müller, D. *A Research Agenda for Tourism Geographies*; Edward Elgar Publishing: Cheltenham, UK; Northampton, MA, USA, 2019; ISBN 9781786439314.
34. Bauder, M. Engage! A research agenda for Big Data in tourism geography. In *A Research Agenda for Tourism Geographies*; Edward Elgar Publishing: Cheltenham, UK; Northampton, MA, USA, 2019; pp. 149–158.
35. Taspinar, A. Twitterscraper. Available online: https://github.com/taspinar/twitterscraper (accessed on 16 June 2020).
36. Steinert-Threlkeld, Z.C. *Twitter as Data*; Steinert-Threlkeld, Z.C., Ed.; Cambridge University Press: Los Angeles, CA, USA, 2018; ISBN 9781108529327.
37. Van Kessel, P. An Intro to Topic Models for Text Analysis. Available online: https://medium.com/pew-research-center-decoded/an-intro-to-topic-models-for-text-analysis-de5aa3e72bdb (accessed on 16 June 2020).
38. Orange, L.B.U. Topic Modelling. Available online: https://orange3-text.readthedocs.io/en/latest/widgets/topicmodelling.html (accessed on 5 May 2020).
39. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
40. Dumais, T. Latent Semantic Analysis. *Annu. Rev. Inf. Sci. Technol.* **2005**, *38*, 188–230. [CrossRef]
41. Wickham, H.; Grolemund, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016; ISBN 9781491910399.
42. Twinword Ideas Dictionary. Available online: https://www.twinword.com/ideas/graph/dictionary/ (accessed on 5 May 2020).
43. Beaver, A. (Ed.) *A Dictionary of Travel and Tourism*; Oxford University Press: Harlow, UK, 2012.
44. Tourismus. Available online: http://webterm.term-portal.de/DEUTERM/tourismus/tourismus_e.htm (accessed on 5 May 2020).
45. Gupta, V.; Rattikorn, H. Harnessing of Power of Hashtags in Tweet Analytics. In Proceedings of the 2017 IEEE International Conference on Big Data, Boston, MA, USA, 11–14 December 2017.
46. Hutto, C.J.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1–4 June 2014.

47. Burchell, J. Using VADER to Handle Sentiment Analysis with Social Media Text. 2017. Available online: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf (accessed on 13 November 2020).
48. ESRI. Using Proportional Symbols. Available online: http://desktop.arcgis.com/en/arcmap/10.3/map/working-with-layers/using-proportional-symbols.htm (accessed on 5 May 2020).
49. ArcGIS Pro. An Overview of the Mapping Clusters Toolset. Available online: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/an-overview-of-the-spatial-statistics-toolbox.htm (accessed on 5 May 2020).
50. Lu, Y. Spatial Cluster Analysis for Point Data: Location Quotients versus Kernel Density. In Proceedings of the University Consortium of Geographic Information Science Summer Assembly, Portland, Oregon, 21 June 2000.
51. Attaway, D. GIS Analysis Workshop. In *Proceedings of the 2016 GIS for a Sustainable World Conference*; ESRI, Ed.; ESRI: Geneva, Swtizerland, 2016; p. 51.
52. ESRI. Optimized Hotspot Analysis. Available online: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/optimized-hot-spot-analysis.htm (accessed on 5 May 2020).
53. ESRI. Differences between Point, Line, and Kernel Density. Available online: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/differences-between-point-line-and-kernel-density.htm (accessed on 5 May 2020).
54. Silverman, B.W. (Ed.) *Density Estimation for Statistics and Data Analysis*; University of Bath: London, UK; Chapman and Hall/CRC: New York, NY, USA, 1986.
55. Dempsey, C. Heat Maps in GIS. Available online: https://www.gislounge.com/heat-maps-in-gis/ (accessed on 5 May 2020).
56. Orange3. Text Mining Preprocess Text. Available online: https://orange3-text.readthedocs.io/en/latest/widgets/preprocesstext.html (accessed on 5 May 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.