sustainability

MDPI

Article

Using Volunteered Geographic Information and Nighttime Light Remote Sensing Data to Identify Tourism Areas of Interest

Bidur Devkota [1,*], Hiroyuki Miyazaki [1,2], Apichon Witayangkurn [1,2] and Sohee Minsun Kim [3]

1 School of Engineering and Technology, Department of Information and Communication Technologies, Asian Institute of Technology, Post Box No 4, Pathumthani 12120, Thailand
2 Center for Spatial Information Science, Tokyo University, Chiba 277-8568, Japan
3 School of Environment, Resources, and Development, Department of Development and Sustainability, Asian Institute of Technology, Post Box No 4, Pathumthani 12120, Thailand
* Correspondence: devkota.npl@gmail.com

Received: 25 July 2019; Accepted: 23 August 2019; Published: 29 August 2019

Abstract: Easy, economical, and near-real-time identification of tourism areas of interest is useful for tourism planning and management. Numerous studies have been accomplished to analyze and evaluate the tourism conditions of a place using free and near-real-time data sources such as social media. This study demonstrates the potential of volunteered geographic information, mainly Twitter and OpenStreetMap, for discovering tourism areas of interest. Active tweet clusters generated using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm and building footprint information are used to identify touristic places that ensure the availability of basic essential facilities for travelers. Furthermore, an investigation is made to examine the usefulness of nighttime light remotely sensed data to recognize such tourism areas. The study successfully discovered important tourism areas in urban and remote regions in Nepal which have relatively low social media penetration. The effectiveness of the proposed framework is examined using the F1 measure. The accuracy assessment showed F1 score of 0.72 and 0.74 in the selected regions. Hence, the outcomes of this study can provide a valuable reference for various stakeholders such as tourism planners, urban planners, and so on.

Keywords: social media; Twitter; tourism; volunteered geographic information; OpenStreetMap; nighttime light remote sensing

1. Introduction

Tourism is a key economic sector and plays a crucial role in expanding economic opportunities for a place. The United Nations reports that the tourism industry contributes up to 40% of the Gross Domestic Product (GDP) of developing countries [1]. World Travel and Tourism Council states that the tourism sector contributed to 10.4% of global GDP and created employment for 313 million people (i.e., 9.9% of global employment) in 2017 [2]. Also, the United Nations Sustainable Development Goals have identified tourism as one of the important tools to attain sustainable economic growth [3]. With the increase in the use of modern information and communication technologies, the global promotion of remote destinations has become easier resulting in the growth of international tourism. To raise the tourism industry even further, cost efficient and up-to-date identification and dissemination of information on tourist attractions is necessary. This will help the industry stakeholders update their knowledge of popular as well as obscure attractions. Hence, informed service providers can

better manage available resources and informed visitors can optimize their itineraries for better travel experience.

Over recent years, various problems in tourism have stimulated researchers. The traditional approach uses conventional techniques such as interviews, focus groups and questionnaire surveys [4,5]. Even if the various knowledge, methodologies, and techniques are available these methods are often slow, costly, and limited spatiotemporally. Previous studies have highlighted potentially serious measurement errors while using traditional approaches, particularly in developing and emerging economies [6]. In recent days, the increasing availability and accessibility of new cutting-edge technology and data are of great use for day-to-day decision-making processes. As a boon from the current "Information Age", a massive amount of data is now available via digital technologies and novel data sources [7]. Recent studies have demonstrated that new data sources such as social media can provide more insights in addition to the results from traditional surveys [8]. Paid and free data from online social media sources and remote sensing cater near-real-time data. Such data have been increasingly used for a huge range of applications including human settlement mapping [9,10], land use and land cover mapping [11], understanding socio-economic conditions [12], tourism studies [13–15], etc. In recent years, many studies have proposed various ways to discover tourism destinations and activities. Geo-tagged data from online social media have been extensively used in studies locating popular touristic sites [16–20], map tourist behaviors [21,22], comparing domestic and foreign tourists [18,21] and discovering obscure sightseeing locations [23]. This is possible because more and more data, i.e., digital footprint, is being generated by the unprecedented use of GPS-equipped smart devices by the users (both residents and tourists) and uploaded to the Internet as geo-tagged information. Researchers can access such geo-tagged contents from social media sites, analyze them, and discover interesting spatiotemporal patterns.

Data scarcity is a common problem in many studies. Often, it is advised to use more data than reasonably required while applying statistical methods. For machine learning problems, low quantities of data can negatively affect the performance of algorithms and their strength to generate useful outcomes. Many machine learning practitioners emphasize the need to get and use as much data as possible. For instance, Ester et al. [24], explained that data clusters and noise points can be identified by observing the high contrast in data density i.e., within cluster density is much higher than outside density. A larger amount of input data will generate denser clusters than fewer data. This implies that the use of large input data sets ensures better accuracy than smaller data sets. Much of the existing literature discusses exploring Area of Interests (AOIs) and places with abundant data availability, i.e., over tens of thousands of data points [16,17,19,25]. For a case in point, Yingjie et al., performed spatial clustering of geo-tagged social media data from cities such as New York [25]. A minimum point density threshold of 2% was used to distinguish clusters and noise. However, in remote areas with less social media penetration, this threshold may have to be lowered due to the data scarcity issue.

Essentially all the popular clustering methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) require input parameters (e.g., cluster density, inter-data distance) which influence the formation of clusters but estimating the optimal values for such parameters is a non-trivial task. Moreover, it may not be logical to use a single global value for such parameters to define the intrinsic cluster structures, as different local parameters may be required to identify clusters in different regions. For instance, while identifying social media clusters, a densely populated area with better infrastructure will possibly have relatively denser data points than other less populated areas. Hence, it is indispensable to tailor approaches to cope with such problems and estimate optimal (or multiple) parameter values as needed.

Most of the literature focuses on identifying tourism areas of interest, but it is indispensable to know if those areas cater minimum essential facilities to travelers such as food and shelter. Without such knowledge, chances may arise when visitors get stranded in uninhabited locations at odd hours. This study focuses on discovering tourism areas of interest (TAOI) while ensuring the accessibility to minimum essential facilities such as a bed and breakfast. This requirement is fulfilled

by spatial modeling of different types of VGI (Volunteered Geographic Information), i.e., tweets and OpenStreetMap (OSM) data, and nighttime light (NTL) remote sensing data. First, we automate the process of estimating the optimal combinations of clustering parameters. Next, relevant tweet clusters related to tourism are discerned from other merely popular clusters by analyzing foreigner participation, building footprint, and NTL data. Finally, the usefulness of the building footprint data and NTL data in discovering TAOIs evaluated in urban and rural places of Nepal. Despite the studied data set being sparser, we still managed to achieve good results. Also, the proposed framework works in an unsupervised manner without relying on any knowledge of tourism spots from external data sources such as Lonely Planet (https://www.lonelyplanet.com/) and TripAdvisor (https://www.tripadvisor.com).

2. Related Works

2.1. Volunteered Geographic Information

Volunteered Geographic Information is user-generated digital traces including both text and multimedia, about user's geographical information. VGI has emerged expeditiously with the advancement in information and communication technology. It contains rich spatiotemporal information generated by the human sensors and provides a way to explore and understand the socio-economic conditions of a place [26,27]. Though there have been concerns regarding information bias and lack of standard quality control, several studies have empirically revealed that VGI is of equally good quality as authoritative data [8,28,29]. Several researchers have used VGI resources including geo-tagged contents (e.g., tweets and photos) [16–18,30], check-in data [18,30], OSM [31] and so on. VGI has been used in a variety of applications such as tourism studies [16–18,30], urban studies [19,25,32,33], land use and land cover detection [11,34] and mobility analysis [35].

2.2. Extracting Interesting Regions from VGI

Visitors are very choosy in selecting the locations to visit as it is almost impossible for them to explore the whole area on an average 2–3-day visit to a place [36]. Hence, it is necessary for them to choose some attractions to visit and others to skip. This results in the formation of typical spatial patterns of tourism sites. Studies exploring such patterns reveal that visitors tend to be attracted to limited areas which have the main tourist attractions (e.g., museums, parks, historical buildings, etc.), leisure, shopping and lodging services [37]. The conventional method to elicit such information was through surveys. However, surveys are inherently limited in capturing visitor spatiotemporal behavior. Online social media data makes available a huge amount of spatial and temporal data which not only compliments but also overcomes the limits of traditional data sources. Hence, visitors' spatial footprints captured by various social media sites can be investigated to study tourism patterns.

Spatial clustering has proved to be an indispensable technique for data aggregation. Current literature has a rich set of methods ranging from the classic K-means [38] clustering algorithm to the popular DBSCAN algorithm [24]. K-means is a popular clustering algorithm which have many variants such as K-medoids [39]. K-means and K-medoids are used to solve spatial clustering problems such as Multivariate Clustering (implemented in ARCGIS (https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-multivariate-clustering-works.htm)). K-means clustering algorithm requires the number of clusters as input, which is an important factor in cluster quality. However, the number of clusters cannot be determined a priori in many scenarios. DBSCAN has emerged to be a better alternative as it is based on data density and not the number of clusters. It can move out sparsely distributed data and group data in arbitrary shape and arbitrary size. Furthermore, it can identify outliers as noise, unlike K-means which just dumps them into some cluster even if the data shows contrasting characteristics. The DBSCAN algorithm requires two parameters: a minimum number of points (*minPts*) in a cluster and the search radius (*eps*) that a data point can influence. Many algorithms improve and extend the DBSCAN algorithm. A new method called ST-DBSCAN

was proposed for discovering clusters from spatiotemporal data [40]. It is based on three attributes of data i.e., spatial, non-spatial, and temporal attributes. The original DBSCAN algorithm uses only one distance parameter to cluster similar data, but ST-DBSCAN requires one additional distance parameter. One distance parameter determines the closeness of the points in spatial scale while the other indicates the similarity of non-spatial attributes. Also, a concept of *density factor* is used to deal with the fixed density problem in the original DBSCAN algorithm. P-DBSCAN is an extension of DBSCAN algorithm using geo-tagged photos for detecting attractive areas based on the photo owner's density in the region [41]. This method is adaptive and flexible but focuses only on finding interesting points of interest (POIs). HDBSCAN extends DBSCAN by modifying it into a hierarchical clustering algorithm and then devising a way to obtain flat clusters [42]. It is more data-driven and hence automates the distance parameter. It requires only one parameter i.e., (*minPts*) and works well with clusters of different densities but compromises performance. Moreover, spatial point pattern methods, such as the Local Moran [43] (used for Cluster and Outlier Analysis in ArcGIS (https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm)) and Getis–Ord Gi [44] (used for Hot Spot Analysis in ArcGIS (https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/hot-spot-analysis.htm)) has also been used for detecting spatial clusters. However, generic clustering algorithms such as DBSCAN perform better in delineating aggregated data and shaping generated clusters [45]. Also, Dehuri et al. showed that if the clusters are of arbitrary shape, DBSCAN algorithm performs better than self-organizing map [46].

With the global availability of geo-information from different VGI sources, numerous studies have been done to discover and understand hotspots and areas of interest. Yang et al., recommended a self-tuning clustering method which can automatically determine the number of clusters without any subjective parameters such as the number of POIs, the shape of POIs and size of POIs [47]. Important tourist locations were assumed to be highly photographed. Hence, such interesting POIs were identified from the collection of geo-tagged photos by applying spectral clustering by self-tuning the necessary parameters. Laptev et al. suggested a parameter-free method for recommending AOIs by analyzing the density of geo-tagged photos in the selected region [48]. The algorithm required a single input parameter i.e., available time for travel. The quality of the result (i.e., recommended AOIs) was judged based on the number of POIs that can be observed within a given time. Hu. et al., proposed a coherent framework for discovering urban AOI based on geo-tagged Flickr photos [25]. Three well developed and three fast developing cities were examined to identify AOI using DBSCAN clustering algorithm and describe them using the textual tags and suitable photos. Korakakis et al. applied HDBSCAN for extracting AOIs [49]. They aimed to solve the "Tourist Trip Design Problem" and automatically extract popular POIs and AOIs to construct travel routes by using social knowledge embedded in the Flickr photos and its metadata. The extracted route aims to maximize user travel experience based on the user input such as start point, endpoint, and available time budget. Chen et al. extracted and visualized the dynamics of urban area of interest using the metadata from geo-tagged Flickr images in London [19]. A variant of DBSCAN clustering, i.e., HDBSCAN clustering, was applied to cluster the spatial data and explore their spatial and temporal insights. Sun et al., proposed a way to identify city centers using traveler flows obtained from check-in data [32]. Three algorithms (DBSCAN, Local Getis–Ord and Grivan–Newman) were implemented to get the best estimation of city centers. DBSCAN demonstrated better performance in delineating geometrically regular boundaries. Koutras et al. studied the top spots in an urban area and the tourists' attitude towards visiting those spots [33]. Density-based spatial clustering was applied on geo-tagged Flickr photos inside Athens. Lee et al. mapped the visitors' trajectory and hotspots in the Grand Canyon National Park using Kernel Density Estimation and Dynamic Time Warping to the geo-tagged flicker images [22]. Hasnat et al. discovered the spatial patterns of tourist destination choices by applying K-means, Mean-Shift and DBSCAN algorithm on the Twitter data collected inside Florida [50]. Encalada et al. analyzed the spatial distribution of the geo-tagged photos in Lisbon City and extracted places of tourist attractions and important tourism variables such as monuments [17]. They applied Local Moran clustering and

multiple linear regression for the purpose. Kuo et al. proposed a new method, i.e., Spatial Overlap algorithm, which considers the spatiotemporal properties and other metadata of the Flickr photos to extract and understand POIs/ROIs by efficiently eliminating noises [51]. Naming and merging of the clusters based on the photo attributes were done for obtaining better results. Moreover, this proposed algorithm applied local maximum to extract attractive footprints to deal with the problem in dense areas. Maeda et al. extracted the location of tourist destinations in Japan using geo-tagged tweets via attractiveness and originality of each place of interest [18]. Attractiveness was computed using DBSCAN clustering and gravity modeling. Originality was estimated by applying text analysis. Also, preference comparison of the foreigner and domestic tourists was done by relating FourSquare places characteristics.

Our research differs from previous works in several aspects. Many studies have used social media data from sources such as Flickr, Panoramio, FourSquare, Gowalla, Sina Weibo, etc. in different domains. However, to date most research has concentrated on attempting to derive the important AOIs and attractions of given regions, assuming the availability of social media data in abundance. Few of the works have used geo-tagged tweets and mainly focus on exploring urban AOIs and places. Limited work has been done to explore the area of interest in remote places and places with relatively less social media reach. Also, we came across many studies that use some sort of spatial clustering algorithm that requires some hyperparameters such as search distance. Though some of them have proposed ways to estimate or automate such parameters, they rely on some specific assumptions which are not suitable for our case (to be discussed in upcoming Section 3.2.1) [40]. To our knowledge, none of them suggest an easy way to automate the estimation of such parameter values which can work across different scarce data regions. However, we proposed a way to estimate such parameters for identifying TAOIs and prime locations. Furthermore, the recommended approach not just discovers TAOIs but also examines the accessibility to minimum essential facilities. We examined the proposed approach in sparse data regions in Nepal and further refined the results by using the freely available data from OSM and NTL. Hence, the proposed method is novel, and it encourages the discovery of TAOIs by fusing freely available data from VGI and remote sensing.

3. Methodology

In this section, we propose a framework devised to identify and map TAOIs from the subset of geo-tagged tweets. Our methodology consists of three main parts: data acquisition, cluster detection and TAOI identification as shown in Figure 1. Nighttime light satellite images, geo-tagged tweets, OSM data, and other ancillary data sets such as shapefiles are collected and used to identify TAOIs.

3.1. Tourism Area of Interest (TAOI)

A TAOI is defined as a center of attraction for many people and contains interesting tourist attractions such as restaurants, hotels, landmarks, leisure zones, and so on. Many people visit such areas and share their ideas and experiences in social media. Hence, many social media contents (e.g., geo-tagged tweets) are generated in such areas. Some TAOIs cater the maximum tourism facilities and become prime attractions while others may not act as a prime attraction but still provide basic essential facilities to the visitors within accessible locations. It is important to distinguish between such prime and non-prime regions within the span of a TAOI because non-prime areas are generally less crowded and provide economical tourism facilities.

Clustering of geo-tagged tweets is employed to extract dense regions of tweet activity as candidates for TAOIs. Also, it is essential that a minimum number of tweets are published by foreigners. Upon selecting such clusters, the final set of TAOIs is identified by ensuring the presence of basic infrastructure and services in the vicinity by examining freely accessible geo-information from OSM and NTL.

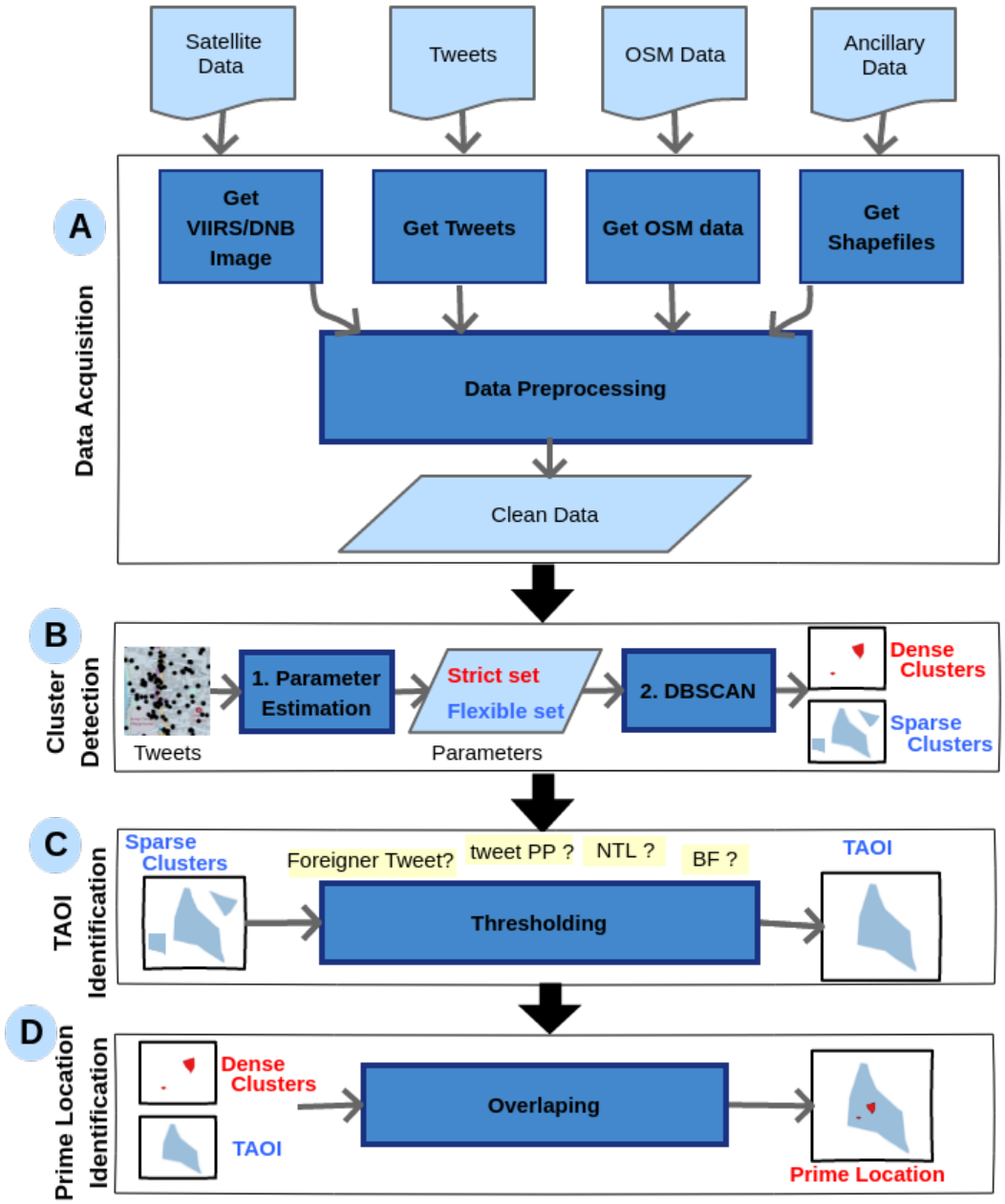

Figure 1. Overall Methodology.

3.2. TAOI Identification Algorithm

Geo-tagged tweets are aggregated using the clustering algorithm to identify and delineate the extent of active regions in social media. Clustering is a data mining technique of exploratory data analysis for discovering interesting patterns in data. It has been successfully used for finding dense regions such as the spatial hotspots. A plethora of clustering algorithms exist in the literature; however, we assume the following requirements for our choice:

- The algorithm should work in an unsupervised way as the exact number of clusters are not known beforehand.
- Clusters can have an arbitrary number, shape, or size (point percent) depending on the context (e.g., study area, data source, etc.)
- Method should detect high-intensity gathering of tweets all over the study area.
- The algorithm must aggregate significant data while discarding any outliers and noise.

DBSCAN algorithm adheres to the above specifications. Also, the comparative summary of different clustering algorithms in Table 1 indicates DBSCAN algorithm as a suitable candidate. Hence, DBSCAN algorithm is selected for generating prominent tweet agglomerations. After the determination of active tweet clusters, TAOIs are selected by examining building footprints, nighttime light intensity as well as non-local user presence within the extent of the clusters.

Lately, Yan et al. proposed a model based on the probabilistic model to explore the geographic distribution of visitors by examining the underlying mechanisms behind it [52]. They suggested to couple maximum entropy modeling with geo-tagged photos to determine the location of tourists in relation to different environmental factors. They quantified the correlations between tourism presence and various environmental factors which seem very useful to envision the impact of abnormal scenarios such as a disaster and other crisis. Differing from our study, this approach modeled tourist distributions in a probabilistic way and determined touristic regions based on the probabilities. Our proposed approach outputs TAOIs in a binary form; however, it does not declare the presence or absence of tourist but tries to identify interesting places of interest (i.e., areas with tourism facilities such as hotels, viewpoints, landmarks, etc.) based on the digital footprint of the travelers. The underlying method accepts the fact that such tourism traces may be present at other places also (even if the proposed method do not highlight them as interesting TAOI), for example, the DBSCAN clustering algorithm identifies them as outliers i.e., isolated points which are not qualified as an interesting cluster. Though expressing AOIs with probabilities seem more informative, it may complicate the understanding. Moreover, Yan et al. accomplished their study in data rich region, i.e., San Diego city, by collecting over hundreds of thousands of data points because probabilistic models commonly require very large dataset to detect statistically significant relationships. Hence, the proposed method highlights the interesting tourist hotspots in scarce data regions in an easy and unambiguous way.

Table 1. Comparison of DBSCAN with other clustering methods.

Algorithm	Comments
DBSCAN [24]	- Works in an unsupervised way as the exact number of clusters are not known beforehand. - Clusters can have an arbitrary number, shape, or size. - Detects high-intensity gathering of points all over the study area. - Aggregates significant data while discarding any outliers and noise. - Require minimum domain knowledge to determine the input parameters.
Classic clustering method such as K-means [38] and K-medoids [39]	- Require pre-knowledge of the number of clusters to be generated. - Cannot identify outliers as noise. - Final result is sensitive to initial starting values. - Assumes the true underlying clusters are globular.
Spatial Point Processing methods such as Local Moran[43] and Getic-ord Gi[44]	- Cannot outperform generic clustering algorithms (e.g., DBSCAN) in delineating aggregated data and shaping generated clusters.
Self-Organizing Maps [46]	- If the clusters are of arbitrary shape, DBSCAN algorithm performs better than the self-organizing map.
Mean-Shift Algorithm [53]	- Cannot identify outliers as noise.
Kernel Density Estimation [54]	- Does not generate a clear hard-lined definitions between points in different clusters.
Affinity Propagation [55]	- Assumes the true underlying clusters are globular.
Spectral clustering[56]	- Require pre-knowledge of the number of clusters to be generated.

3.2.1. Clustering Terminology and Mechanism

As discussed in the preceding section, DBSCAN algorithm is chosen for generating tweet clusters. It is an extensively used density-based clustering algorithm. Figure 2 shows a basic representation of a

DBSCAN cluster where each point represents a geo-tagged tweet. The fundamental terminologies are described as follows:

(i) The set of tweet points to be clustered is D, $D = \{p : p(latitude, longitude)\}$ where p denotes the location of any tweet.
(ii) The tweet density of a point p is determined by the number of neighboring tweets within distance eps from the point p.
(iii) The eps-Neighborhood of p is represented by $N(p) = \{q : q \in D, dist(p,q) < eps\}$, where $dist$ gives the distance between the two points p and q.
(iv) A point p is a core point if it has a minimum of $minPts$ points within its neighborhood such that: $|N(p)| \geq minPts$.
(v) A point q is a border point if it has fewer than $minPts$ points within its neighborhood but lies within the neighborhood of core point p.
(vi) Any tweet point r which is neither a core nor a border point is considered to be a noise point.
(vii) A point p is directly density reachable from another points q if p is within eps-Neighborhood of q and q is a core point such that:

    (a) $p \in N(q)$.
    (b) $|N(q)| \geq minPts$.

(viii) A point p is density reachable from another points q if there is a series of core points leading p to q such that: $p_1 = p, p_2, ..., p_n = q$, where p_i denote core points.
(ix) A cluster C is a non-empty subset of D where each point is density reachable such that:

    (a) $\forall p, q$ if $q \in C$ and p is density reachable from q then $p \in C$.
    (b) $\forall p, q \in C, \exists r \in C$ so that both p and q are density reachable from r.

DBSCAN is simple and efficient algorithm for clustering large-scale data. The basic mechanism for identifying clusters follows the steps listed below:

(i) Start with an arbitrary point

    (a) Determine the neighborhood points adhering to eps and $minPts$ requirements.

        (1) Recursively apply step (a) for all new neighboring points.

    (b) Density reachable and density connected points are used to create a new cluster. Any other points are marked as noise. If the noise point satisfies eps and $minPts$ for a different point in later iterations then it can still be a cluster point.
    (c) All points within the cluster are marked visited.

(ii) Repeat step (i) with new unvisited points until all the points are marked visited or noise.

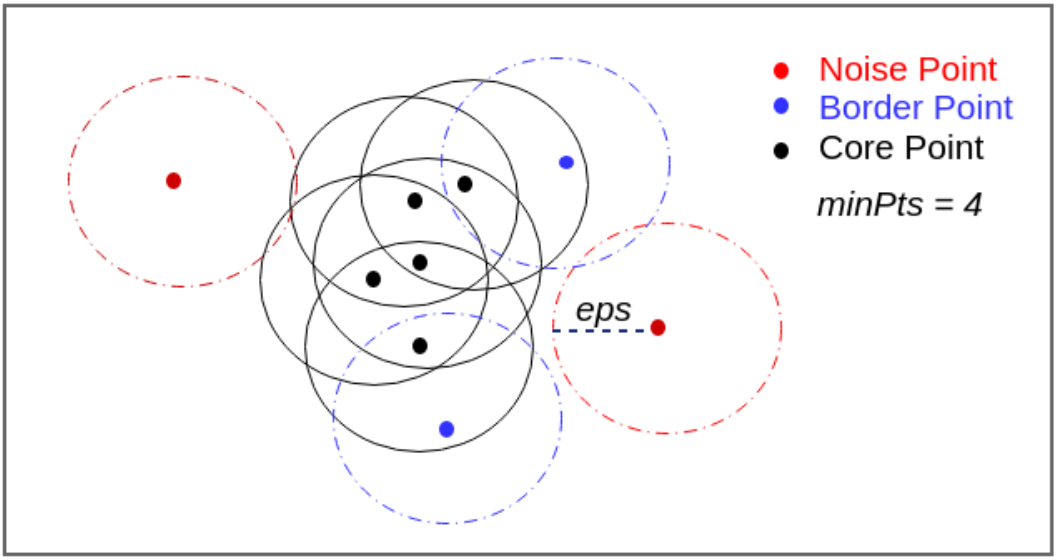

Figure 2. A single DBSCAN cluster with Core, Border, and Noise Points.

3.2.2. Tuning of Clustering Parameter

DBSCAN clustering outcomes are determined by two parameters: *eps* (the search distance or radius) and *minPts*(the minimum data count within the search radius). These parameters define a minimum density threshold, such that clusters are discovered at locations where the data density exceeds the minimum threshold. To use DBSCAN proper values of *eps* and *minPts* must be chosen. The *eps* parameter is associated with the geographic scale of the study area. A larger value for *eps* results in broader clusters, while a smaller value establishes narrower clusters. The *minPts* parameter specifies the minimum number of points (i.e., cluster members) required to produce a new cluster. A larger value of *minPts* assures a more robust cluster but may exclude some potentially smaller areas as it attempts to merge them in a larger one. On the other hand, a smaller value extracts many clusters, but the resultant clusters may include noise as well.

Generally, a minimum *minPts* can be obtained from the number of *dimensions* in the data set, as $minPts \geq dimensions + 1$. The extremely low values of $minPts \leq 2$ do not provide a significant gain of using DBSCAN clustering and hence *minPts* must be set to be at least 3. Larger values are considered more robust to noise and yield more significant clusters. In general, $minPts = 2 * dimensions$ is suggested while larger values are encouraged for larger data sets containing noise and duplicates [57]. A single absolute value for *minPts* may not be appropriate (e.g., comparing different places with varying numbers of tweets) therefore a percentage of the tweets is taken as a general approach [25]. The distance parameter *eps* for DBSCAN is comparatively more difficult to set. DBSCAN algorithm inherently encourages the smallest possible value for *eps*. It is important to consider the application domain knowledge and the target geography while setting this parameter. Also, researchers advocated that if *eps* is based on domain knowledge then *minPts* may vary, to get different insights into the data. Many studies in the literature follow a trial-and-error approach with various values and compare the clustering results with the background knowledge of the study area in order to select some absolute values as the parameters (e.g., *eps* = 100 and *minPts* = 10) [19,25,49]. Others such as Briant et al. use some heuristics (i.e., $minPts = ln(n)$) to determine such clustering parameters [40]. Such tricks may not be suitable because this study requires the formation of maximum numbers of clusters as possible. Apart from being sparse data region, the number of tweets and geographic characteristics of the selected study areas differ significantly. Hence, it is necessary to test the different possible combinations of parameters to estimate reasonable values. To address this problem of parameter estimation, we proposed two objective functions:

(i) *StrictParameters* (i.e., lowest *eps* and highest *minPts*): The strict version of parameters identifies groups with the smallest possible extent and utmost tweet density (i.e., *Dense Clusters*). The Dense Clusters are supposed to have the highest user attention and are the potential regions for prime tourism locations.

(ii) *MaxParameters* (i.e., parameter combination yielding maximum clusters): A maximum number of clusters may be obtained with relaxed DBSCAN parameters and such accumulations (i.e., *Sparse Clusters*) are the candidate TAOI locations.

3.2.3. Cluster Detection

The input data points along with the parameters estimated by *MaxParameters* and *StrictParameters* are provided to DBSCAN clustering algorithm which examines the input and generates relevant clusters. The best part of DBSCAN clustering is that it can identify and avoid irrelevant noise points from appearing in the output clusters. Once DBSCAN clustering identifies the Dense and Sparse clusters, the next step is to build a perceptual polygon to enclose the extracted cluster points. This helps in distinguishing the area of interest along with the sets of important points. Various researchers have used convex hulls for the construction of boundaries that envelop all geo-tagged data within a cluster [58]. However, convex hulls may contain unnecessary empty areas unoccupied by the selected cluster points [59]. Hence, to delineate the cluster shape more accurately, we adopted a chi-shape concave hull algorithm by Duckham et al. [60] as described in previous study [25].

3.2.4. Identifying TAOIs by Cluster Pruning

While active cluster regions have been extracted based on tweet density, the TAOIs of each region remains vague. This section aims to explore the extracted clusters and characterize them as TAOIs. Related studies have used geo-tagged social media data for mining tourist locations without separating the data from local and non-local users [20,49]. Also, previous studies in the proposed area (i.e., Nepal) illustrated that approximately one third portion of the tweets was contributed by non-local users [61]. Hence, to avail more input for the underlying clustering algorithm, the proposed approach makes use of the tweets posted by both local and non-local users in the beginning. Later, a strict requirement is imposed so that foreign user tweet must be present in the selected TAOIs. Twitter user profile information is examined to confirm if the tweets are published by local or foreign users. False user profile information can degrade the location inference; however, several past studies have used such data as a source for ground truth in their studies [62,63]. Jurgens et al., experimentally illustrated that there are not so many false location information [64]. Furthermore, recent studies have indicated that NTL data is highly correlated with urban infrastructures and socio-economic activities such as tourism [65,66]. Therefore, regions with either enough BF (or enough NTL intensity) are assumed to ensure minimum infrastructure necessary for tourism activities. Also, it has been observed that higher tweet percentage is associated with active tourism areas. Hence, the following criteria are defined to distinguish TAOIs from ordinary tweet clusters:

1. *MaxParameters* must be used to generate Sparse Clusters from geo-tagged tweets.
2. Sparse Clusters must satisfy sufficient amount of foreigner tweets (FT) and tweet point percentage (PP).
3. Lastly, BF and NTL thresholds must be fulfilled. Based on the contribution of BF and NTL, three different methods are defined to identify TAOIs as follows:

   (a) $tClust\_B$: Clusters must adhere to minimum BF threshold.
   (b) $tClust\_N$: Clusters must maintain minimum NTL threshold.
   (c) $tClust\_NB$: Clusters must adhere to minimum NTL threshold as well as BF threshold.

Any clusters which does not meet these requirements are pruned out from the list of TAOIs. Once TAOIs are discovered, prime tourism locations are identified as follows:

1. *StrictParameters* must be used to generate Dense Clusters from geo-tagged tweets.
2. Dense Clusters must satisfy sufficient amounts of NTL, foreigner tweets, tweet point percentage, and building footprint requirements. Any cluster which does not meet these requirements are pruned away.
3. The overlap of the selected Dense Clusters with TAOIs, if present, identifies the location of the prime tourism spots.

4. Experiment Setup

This section describes the experimental setup of the research which includes the details about the study area, data sources and implementation tools used.

4.1. Study Area

The spatial area of interest for this study is Nepal. Nepal is the land of ten UNESCO World Heritage sites (https://whc.unesco.org/en/statesparties/np). Seven are in Kathmandu Valley and one each in Lumbini (Birthplace of Lord Buddha), Chitwan National Park, and Sagarmatha National Park. Tourism is regarded as one of the most promising sectors for sustainable development, yet its total contribution to the economy is very low. The total contribution of tourism to the economy in 2017 was 7.8% and total employment including jobs indirectly supported by the industry was 6.6% [2] of the workforce.

This research is carried out by using geo-tagged tweets over two well-known tourism destinations in Nepal, i.e., Kaski district and Solukhumbu district. Nepal Government has categorized administrative sectors into urban municipalities and rural municipalities based on criteria such as infrastructure, population, and revenue. Kaski contains one urban municipality (i.e., Pokhara city) and four rural municipalities and Solukhumbu district consist of a total of eight rural municipalities (http://mofaga.gov.np/). Both Kaski and Solukhumbu district are popular tourist destinations in Nepal. Table 2 summarizes the population, area, location, and tourism importance and Figure 3 presents their location on the map of Nepal. Pokhara city of Kaski district is a popular destination for national and international tourists. Pokhara is the second largest city in the country and the capital of Province 4. Solukhumbu District of Province 1 lies in the eastern part of Nepal and is very rural compared to Kaski. Mount Everest, the highest peak on earth, lies in the northern part of this district, within the region of Sagarmatha National Park. Everest region is globally popular trekking destination.

Table 2. Geo-demographic summary of the selected regions [67,68].

	Kaski	**Solukhumbu**
Area (sq.km.)	2017	3312
Population	492,098	105,886
Bounding Box (degrees)	(83.70,28.08), (84.28,28.61)	(86.36,27.34), (87.01,28.11)
Features	- Pokhara, the tourism capital of Nepal. - Pokhara ranked 7 in "Top Experiences in Nepal" by Lonely Planet [69]. - A part of Mount Annapurna and range.	- Mount Everest. - World heritage site: Sagarmatha National Park. - 'Everest Base Camp Trek' ranked 2 in "Top Experiences in Nepal" by Lonely Planet [69].

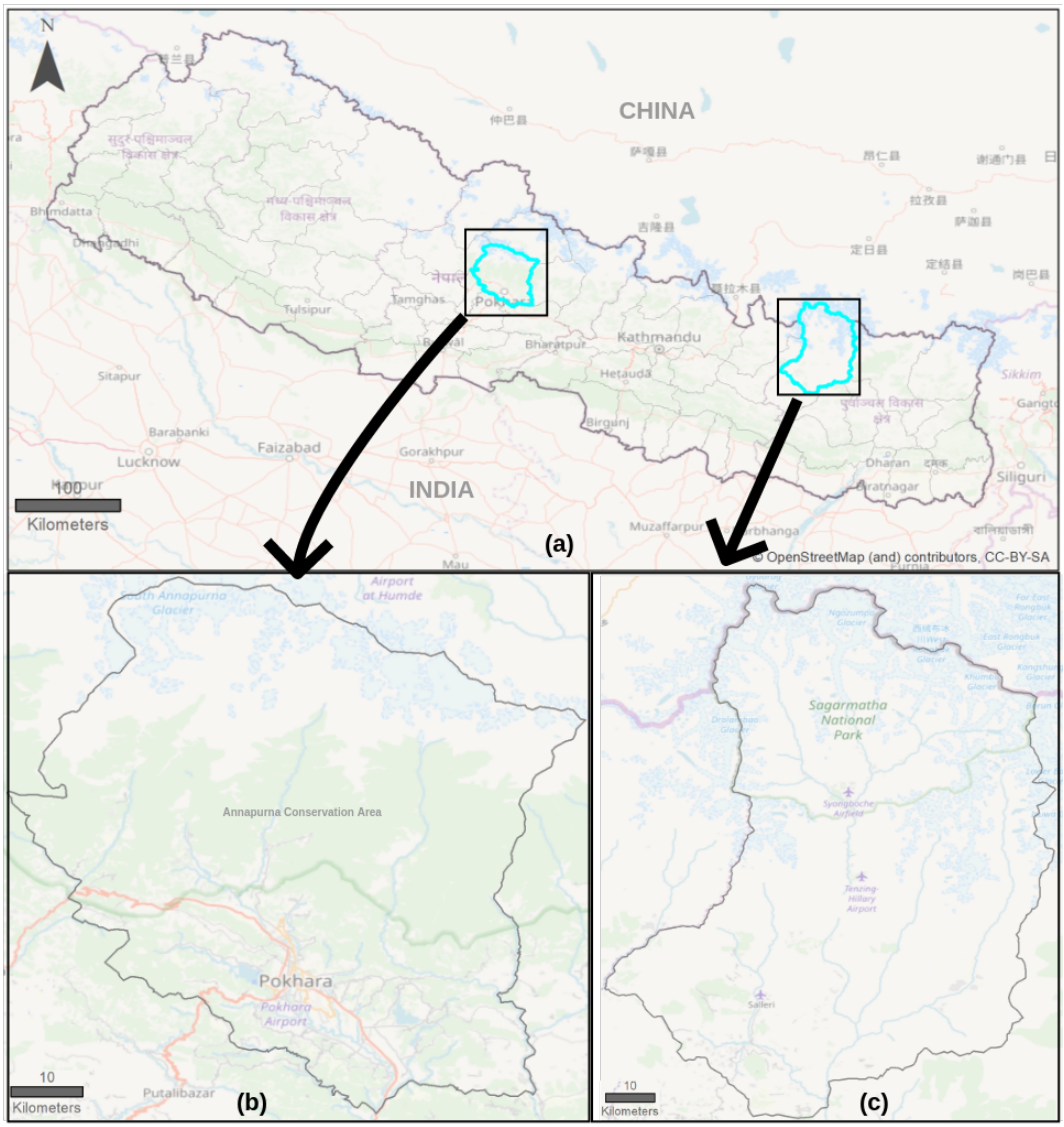

Figure 3. Map highlighting selected study areas in Nepal: (**a**) Nepal map, (**b**) Kaski district map and (**c**) Solukhumbu district map).

4.2. Data Acquisition

Freely available geospatial data from various sources such as Twitter, OpenStreetMap, and earth observation satellites were collected and made available for further processing. A comprehensive description for each of them is provided in this section.

4.2.1. Twitter

We chose Twitter for this study as it has a simple and well-defined public interface for extracting data. Twitter is one of the most popular social networking sites as well as a micro-blogging site. Twitter provides a platform for online users to share a text message up to 280 characters which is popularly known as a tweet. An average of 500 million tweets containing rich data such as texts, images, links, and videos are posted every day on Twitter [70]. Most of its content is public in nature, unlike other popular social media sites such as Facebook. Cesare et al., [71] investigated 60 pieces of existing research literature for social media sites to review user demographic traits. More than half, i.e., thirty-nine (i.e., 65%) of the studies, focused on Twitter, two on Facebook, and so on. The popularity of Twitter in the research and academic community is attributed to various characteristics such as message size, metadata, availability and accessibility [72]. Furthermore, Puschmann et al.,

pointed out various reasons for considering Twitter as a reasonable source for research data [73]. It is considered a global phenomenon with a growing number of users and posts every day. Twitter is deeply rooted in our media ecology and it is used by the public as well as politicians, journalists, and marketers. An investigation on the social media usage in Nepal reported that Facebook, Twitter, and YouTube were the most popular social media sites during the first quarter of 2017 [74]. Twitter is used by government offices, celebrities, authorities, politicians, and the general public for sharing and accessing information. Nepal Police (https://twitter.com/NepalPoliceHQ) started its Twitter activities on April 2015, i.e., during Nepal Earthquake 2015.

Using the free Streaming Twitter API, we collected the geo-tagged tweets within Nepal and stored only the tweets whose location data is recorded via GPS on mobile devices as they provide the best location accuracy [75]. Sometimes, the sample data made available via the free streaming service has been questioned for its representativeness. Morstatter et al., conducted various examinations and conveyed that the Streaming API provides the approximate set of the geo-tagged tweets even if the amount of geo-tagged tweets is relatively very small (i.e., around 1%) [76,77].

In this study, the original Twitter data set contains 89,228 geo-tagged tweets within Nepal covering the study area for 685 days between March 2017 and June 2019. Data collection was interrupted for some period in the middle due to problems in the data collection server. Not all the collected tweets are usable since some of them may have been generated by spammers. Frias-Martinez et al. assumed that some mobile advertising agents may publish a huge number of tweets on a daily basis and proposed a way to filter such tweets so that any GPS location which publishes more than 20 tweets per user in one day should be removed [78]. Furthermore, Zhao et al. defined spam tweets as any geolocated tweets that are published in more than one far-away (>500 km) locations within a very short time interval (<1 h) by the same user. Such tweets are supposed to be generated either by the hackers or due to errors in the Twitter system or GPS devices [79]. Hence, from the total collection of tweets, such tweets were filtered away. Table 3 provides a summary of the total tweets collected in the whole of Nepal and the selected study areas during the study period. Figure 4 shows the plot of geo-tagged tweets over the map of Nepal indicating its distribution in different parts of the country.

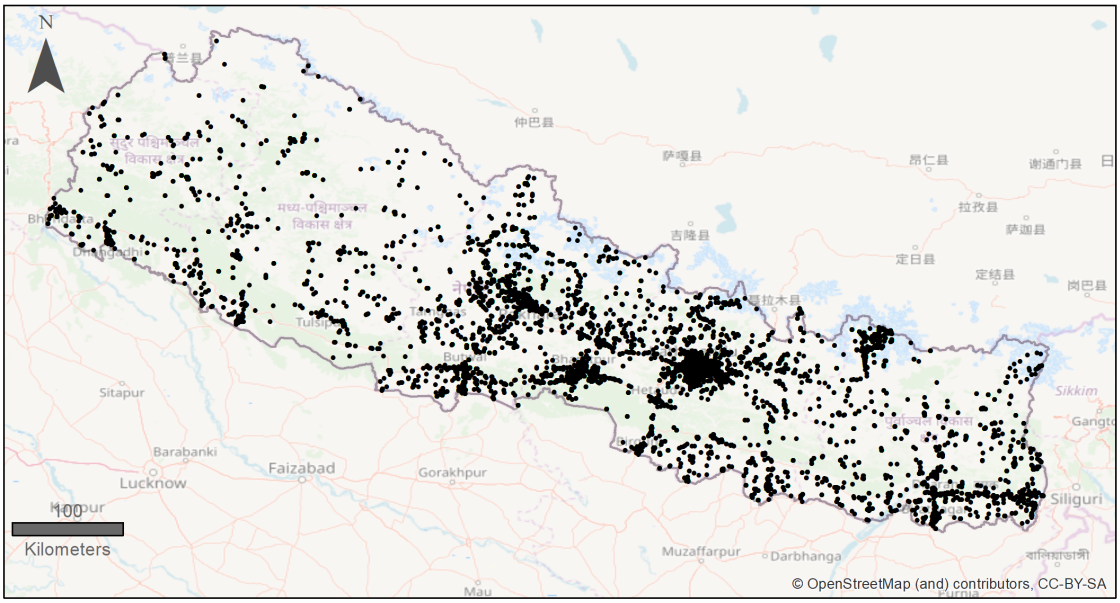

Figure 4. Map of Nepal showing the locations of geo-tagged tweets.

Table 3. Summary statistics of collected tweets.

Place	Total Tweets	Total Users	Average Tweets Per User
Kaski	8787	2150	4
Solukhumbu	5472	718	7
Nepal	89,228	14,216	6

4.2.2. Building Footprint

OpenStreetMap project provides an easily accessible platform that enables free access to geo-information across the world. It is a crowdsourcing platform where volunteers from around the world can contribute in the generation of the geographic data. Though OSM has no strict quality control mechanism, studies have indicated that data obtained from OSM are good enough and comparable to authoritative data to some extent [28]. OpenStreetMap encodes data in different formats such as points, polylines, and polygons. Points usually indicate the point of interests such as shops, post office, and bus stops. Polylines store information such as roadways, waterways, and railways. Polygons represent features such as buildings, parks, or forests. At present, several research communities use this free global dataset for investigating and solving problems in different domains such as spatial computing, urban planning, and geographic information system, ecology, etc. This free dataset has opened boundless possibilities in commercial as well as academic research.

The proposed study uses OSM building footprint data. Building footprint data indicates an area on a site that is covered by the building structure. The perimeter of the building plan defines the extent of the built-up structure. The building footprint differs from the general built-up areas because it excludes any other regions used by other natural and man-made structures such as landscapes, parking areas, and other non-building facilities. The OSM data is freely downloadable from geofabrik website (http://download.geofabrik.de/asia/nepal.html). The count of building footprint data for Solukhumbu district and Kaski district were 35,348 and 137,945 respectively as of 23 June 2019.

4.2.3. Nighttime Light

Nighttime Light remote sensing is a technique using satellite sensors to acquire city lights, fishing vessel lights, gas flares, and burning biomass. It is important for studying social issues such as poverty, environment, and ecology because NTL reflects real human activities. NTL remote sensing is used as an important supplementary dataset to the census in decision-making processes. It is observed that the increase in NTL intensity and extent tends to be correlated with economic growth. NTL can be used to investigate different types of development issues related to human activities, including ecological pressures, the degree of country-level economic activities, the rate of city-level urbanization, and light pollutions [80].

One of the main drivers for the growth in NTL research is the development of satellite sensors which can capture imagery data during the nighttime. The Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) sensor onboard the Suomi National Polar-orbiting Partnership (NPP) satellite platform became operational in 2012. Using advanced processing schemes (e.g., excluding/correcting data impacted by stray light), NOAA-NGDC https://ngdc.noaa.gov/eog/viirs/download_viirs_ntl.html is producing global composite products featuring average radiance values at 15 arc-sec (about 450 m) spatial resolution [81]. The radiance values are measured in floating point and its unit is $nanoWatts/(cm^2.sr)$. We use the annual composite which excludes any data impacted by stray light, ephemeral lights and background (non-lights). It has been extensively used in understanding the various anthropogenic phenomenon. Also, correlating the NTL brightness with other variables such as buildings and settlement areas, bridges and communication routes, and demographic and socio-economic activities have generated good results which have proved useful for decision making [65,66,82,83].

The annual composite average radiance images are available for the years 2015 and 2016 only (as of June 2019). The radiance images are dis-aggregated to obtain the average radiance values for each district of Nepal in 2015 and 2016, respectively. The Pearson correlation coefficient measure of 0.995 is obtained for 2015 and 2016 (Table 4). This hints that the annual composite nighttime light images have a strong association. Figure 5 shows the map of the annual composite average radiance image in Nepal during 2016. It can be observed that the overall nighttime light intensity in Nepal is very low. Close observation in Kaski district reveals very faint lighting around Pokhara city. Solukhumbu district seems to be the least illuminated region. Recent studies have suggested that nighttime light data can act as a good indicator for an urban area at national or regional scale [83]. Therefore, the Kaski district is considered more urbanized than the Solukhumbu district.

Table 4. Pearson correlation of nighttime light intensity of different years; NTL 2015: Nighttime lightintensity for year 2015; NTL 2016: Nighttime light intensity for year 2016.

	NTL 2015	NTL 2016
NTL 2015	1	0.995
NTL 2016	0.995	1

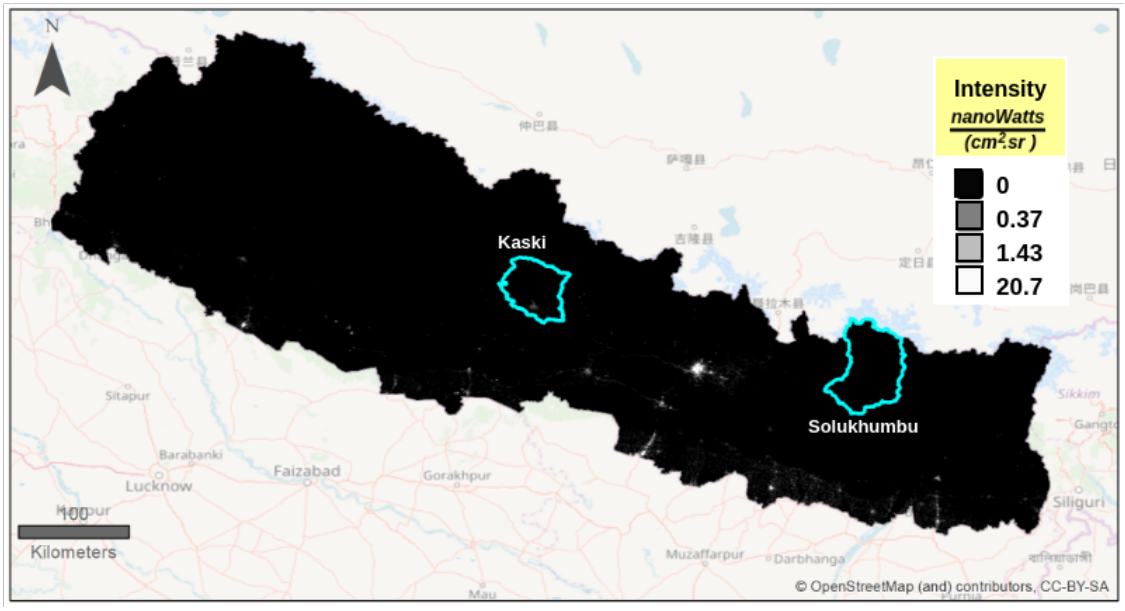

Figure 5. Average Radiance NTL Map highlighting Kaski and Solukhumbu district in Nepal (2016).

4.3. Softwares

The system implementation is done using free and open source resources. Programming languages and modules in Python and Java were used. Tweepy (https://www.tweepy.org/) Python library was used for collecting tweets. QuickOSM (https://plugins.qgis.org/plugins/QuickOSM/) Python module for QGIS was used for collecting data from OSM. PyProj (https://pypi.org/project/pyproj/) module was used for performing cartographic transformations between different projection systems. Shapely (https://pypi.org/project/Shapely/) module was used for manipulation and analysis of geometric objects. Java implementation of DBSCAN [25] was used for clustering and extent extraction. QGIS was used for visualization of spatial data and results. Additionally, other Python modules like pandas (https://pandas.pydata.org/), JSON (https://docs.python.org/2/library/json.html), haversine (https://pypi.org/project/haversine/), functools (https://docs.python.org/3/library/functools.html) were also used.

5. Results

The proposed technique is applied to discover TAOIs using geo-tagged tweets, OSM, and NTL data. This section illustrates the experimental results of applying the proposed TAOI identification method in two different districts of Nepal.

5.1. Selection of Clustering Parameters

The objective functions, *StrictParameters* and *MaxParameters*, help in the selection of appropriate *eps* and *minPts* values required by the DBSCAN algorithm. The knowledge of the study area can guide in determining better values for maximum and minimum *eps* and *minPts*. We examined the geographic features of the study area as a guide to estimate a set of minimum and maximum values for the DBSCAN parameters. Minimum *eps* was set based on the average diameter of tourism features in the study area. Maximum *eps* was set based on the nearest neighbor distance among the tourism features in the study area. Maximum *minPts* value was taken referring to the values used by similar studies in more data dense regions. We examined 12,100 sets of *eps* and *minPts*. A total of 22 different values for *eps* from 25 m to 550 m with an increment of 25 meters were examined. For *minPts*, 550 different values from 0.01% to 5.5% with increments of 0.01% were examined. The set of *StrictParameters* and *MaxParameters* and the number of Dense Clusters and Sparse Clusters obtained are listed in Table 5. Optionally, in short of the knowledge of the target geography, the desired values for *eps* and *minPts* can be estimated without the need for any user input. The objective functions choose appropriate *eps* by examining a range of possible values from minimum (i.e., zero) to maximum (i.e., length of the study area) distance values. Similarly, *minPts* is also selected by scanning minimum data points (i.e., 3 data points as defined by DBSCAN algorithm) to maximum possible data points in the input. In this way, the objective functions, i.e., *StrictParameters* and *MaxParameters*, estimate the required values for generating Dense and Sparse clusters.

Table 5. Tweet Clusters obtained by applying *StrictParameters* and *MaxParameters*.

Place	*StrictParameters*			*MaxParameters*		
	***eps* (m)**	***minPts* (%)**	**Clusters**	***eps* (m)**	***minPts* (%)**	**Clusters**
Kaski	25	0.11	1	175	0.01	75
Solukhumbu	25	0.51	1	350	0.01	44

5.2. TAOI Identification

Table 6 illustrates the significance of different input variables in distinguishing TAOIs. The selected thresholds for foreign tweet presence, NTL, and building footprint is one or more. Similarly, the minimum requirement for tweet point percentage is 0.06% for Kaski and 0.09% for Solukhumbu district. The threshold values are estimated based on the knowledge gathered from various tourist maps and trekking guides published by recommended agencies such as the Nepal Tourism Board and Milestone Guides [84,85]. Furthermore, seven local tourism experts (i.e., tourist guides, backpackers and hotel staffs and owners) were also consulted while setting the threshold values. Local experts verified whether the extents of the identified TAOIs contain touristic spots or not. For Kaski, DBSCAN with *MaxParameters* discovered a total of 75 tweet clusters. Out of the 75 clusters, only 47 of them contain tweets from foreigners, 60 clusters have sufficient NTL values, 68 clusters pass BF threshold and 68 clusters confirm the minimum tweet PP requirement. Much of the tweet clusters in Solukhumbu have foreigner tweet presence while almost half of the clusters in Kaski do not have a significant contribution from foreigners. Inversely, many clusters in Kaski seem to be formed around built-up areas while the cluster presence around built-up regions in Solukhumbu is much low. Also, NTL acted as a stronger discriminator in Solukhumbu district. This is because Solukhumbu is extremely rural in

comparison with Kaski district. Almost one third of the clusters in Kaski and less than a quarter of the clusters in Solukhumbu were below the PP threshold.

Table 6. Effect of different data sources on cluster selection; NTL: Nighttime light intensity; PP: Tweet Point Percentage; FT: Foreigner Tweet; BF: Building Footprint.

Place	Total Clusters	Clusters Confirming			
		FT	NTL	BF	PP
Kaski	75	47	60	68	68
Solukhumbu	44	39	7	26	40

Table 7 summarizes the result obtained by applying different TAOI identification methods. The first row shows the outcomes of the TAOI selection methods for Kaski. Out of the 75 tweet clusters, *tClust_B* selected 41 of them as TAOIs and eliminated 34 clusters. Similarly, *tClust_N* elected 33 clusters as TAOIs and disqualified 42 clusters. In addition, *tClust_NB* nominated 28 clusters as relevant ones and excluded 47 of them from the result. For each study area, single prime locations were identified. The TAOI determination methods *tClust_N* and *tClust_NB* elected comparatively few TAOIs in the rural region (i.e., Solukhumbu district) because most of the touristic spots in the district are located in and around the Sagarmatha National Park which does not have enough NTL illumination. A closer inspection revealed some level of human settlement in few of the un-lit regions. Hence, the mandatory requirement of NTL threshold seems biased for such areas. However, the method based on the building footprints, i.e., *tClust_B*, behaved in a much unbiased manner while selecting TAOIs. Further exploration of the *tClust_B* TAOIs and OSM tourism features was made. Table 8 provides a summary of the overlapping between the TAOIs and the polygons labeled as tourism in OpenStreetMap. Out of the 41 TAOIs in Kaski, only 19 of them overlapped with tourism polygons in OSM. Accordingly, about half of the TAOIs in Solukhumbu do not find any related features in OSM. Surprisingly enough, none of the prime locations showed any overlay of OSM tourism features. Close inspection of the study areas uncovered the fact that OSM data is not complete and consistent in the regions which resulted in a low agreement between the discovered touristic spots and OSM data.

Table 7. Summary of TAOIs identified by tClust_B(FT & PP & BF), tClust_N(FT & PP & NTL) and tClust_NB(FT & PP & NTL & BF).

Place	Total Clusters	tClust_B		tClust_N		tClust_NB		Prime Location
		TAOI	Pruned	TAOI	Pruned	TAOI	Pruned	
Kaski	75	41	34	33	42	28	47	1
Solukhumbu	44	24	20	7	37	5	39	1

Table 8. Summary of TAOIs and OSM tourism feature overlap.

Place	tClust_B		Prime Locations	
	TAOIs	Overlap	Locations	Overlap
Kaski	41	19	1	0
Solukhumbu	24	13	1	0

The spatial distribution of identified TAOIs depicted the locations of notable tourism spots where prominent natural and man-made landmarks are located. This result supports our assumption that important landmarks receive more user attention in reality, and accordingly in the social media than regular areas. To illustrate this claim, concrete evidence from our study area are discussed. Figure 6 shows Lakeside, a part of Pokhara city, which is the most touristic region in the Kaski district. Detail investigation about Lakeside revels (i) high social media activity (ii) dense OSM building footprint and

(iii) comparatively high NTL in the region. Such a place should contain TAOIs as per the definition proposed in Section 3.1. Accordingly, the proposed TAOI mapping scheme identified Lakeside as a touristic region. Furthermore, a prime location was discovered in the core of Lakeside (i.e., a red polygon within). On a par with our assumption, this identified prime location provides better tourism services and facilities as well as active social media presence in the whole region. This is the most prominent tourist location which lies within 100 meters of 'Phewa Lake'. Phewa Lake is the main tourist attraction in Lakeside and is listed as the top attraction in Pokhara by travel website TripAdvisor (https://www.tripadvisor.com/Attractions-g293891-Activities-Pokhara_Gandaki_Zone_Western_Region.html). Next, we examined some of the tourism facilities at an obscure location, i.e., very far from the prime location, but within the extent of the TAOI. For example, 'Hotel Mountain View' is located at the bottom end of TAOI i.e., most distant from the prime tourism centers. Though the hotel is not at the main road (i.e., Lakeside Road), visitors find its location good enough to stay as its cost is economical and the hotel is situated away from crowded locations (expressed in the hotel reviews in TripAdvisor (https://www.tripadvisor.com/Hotel_Review-g293891-d5982537-Reviews-Hotel_Mountain_View-Pokhara_Gandaki_Zone_Western_Region.html).

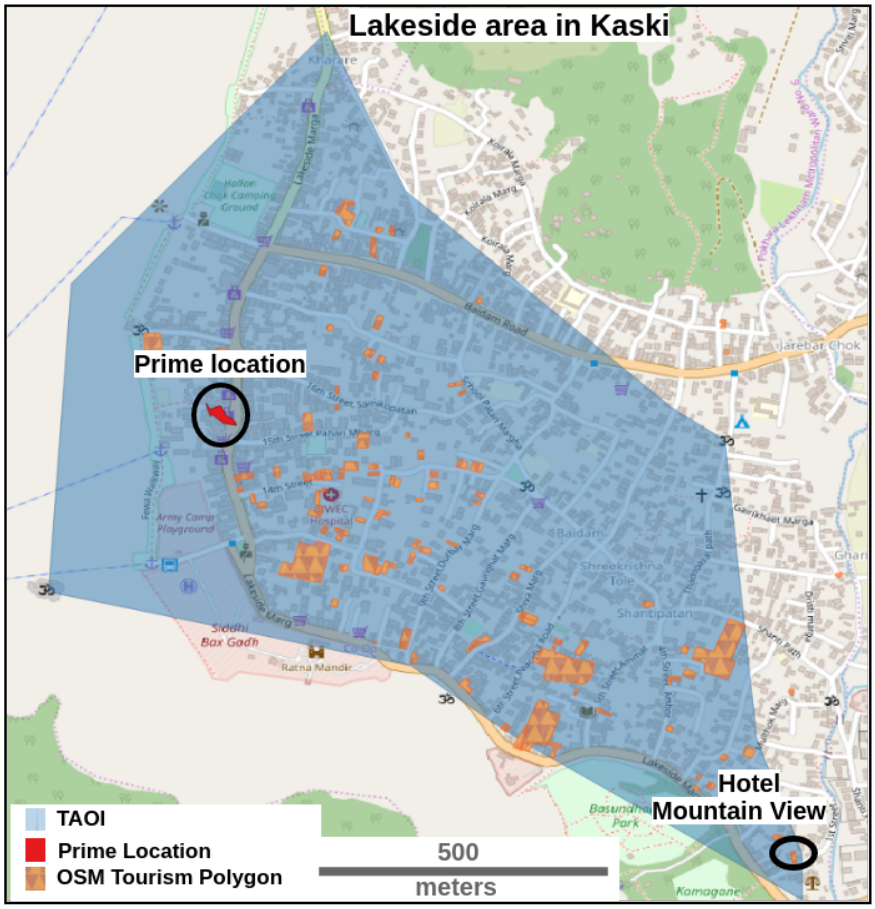

Figure 6. Lakeside region as an important TAOI in Kaski with a prime location.

Similarly, Figure 7 displays the spread of TAOI within Namche Bazaar region in the Solukhumbu district. Namche Bazaar is an important tourism hub located in the rural municipality of the district. It is situated inside the Sagarmatha National Park area, a world heritage site. The shadowed region indicates the stretch of TAOI, and the enclosed red polygon represents the prime location. An interesting pattern divulged while observing the spatial distribution of TAOIs in the Solukhumbu district. As evident in Figure 8a the distribution of TAOIs follow the trekking routes in the Everest Base Camp (EBC) region. In addition, relaxation of the thresholds resulted in the pattern becoming more

significant (e.g., Figure 8b–d). The TAOIs which conforms to the defined threshold is relatively less in number but are in areas that provide minimum infrastructures and services. Only five TAOIs in the whole district and three TAOIs in the EBC region were found to be illuminated during the night. In fact, these illuminated TAOIs represent district headquarter Salleri (*TAOI*_25), neighborhood along Phaplu airport (*TAOI*_31) and other important locations along the Everest trekking route that act as indispensable tourism centers in the region i.e., Namche Bazzar (*TAOI*_12), Lukla area (*TAOI*_7) and Tengboche region (*TAOI*_6).

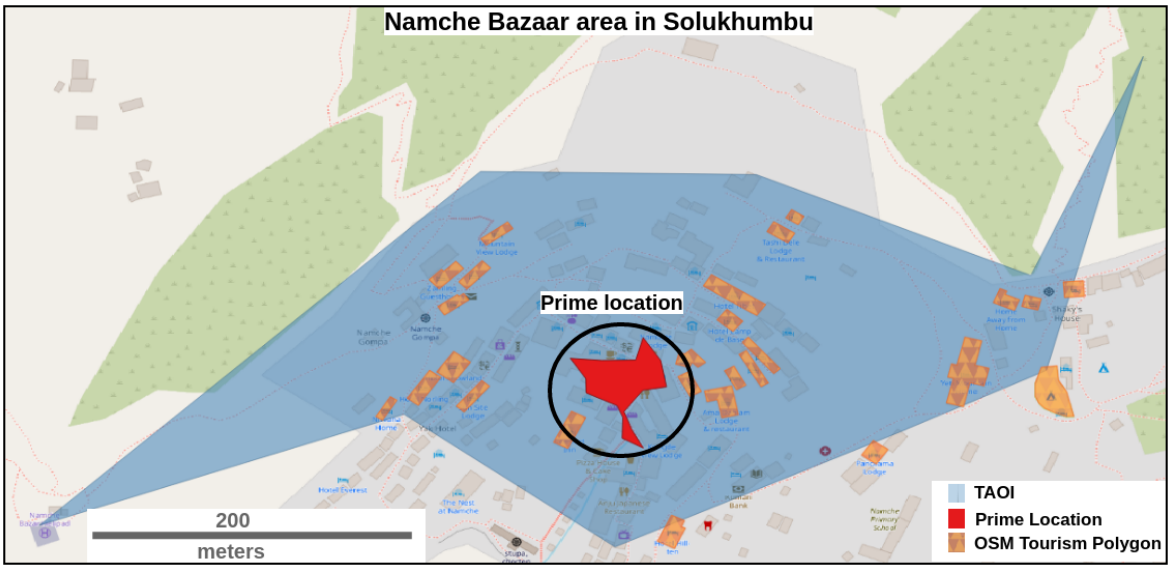

Figure 7. Namche Bazaar as an important TAOI in Soukhumbu with a prime location.

5.3. Validation

To evaluate and compare the performance of the proposed methods, standard measures of quality i.e., F1 score is used. The accuracy assessment is performed by selecting well-known tourism sites of Nepal. Pokhara city, one of the best urban tourism destinations in Nepal, is selected from the Kaski district. EBC region which is popular for remote area trekking is chosen from the Solukhumbu district. The choice of these areas represents well-known urban and remote tourist destinations in Nepal. The ground truth data were collected from tourist maps and trekking guides published by recommended agencies such as the Nepal Tourism Board and Milestone Guides [84,85]. Figure 9a,b respectively show the confusion matrices for the EBC region and Pokhara city by comparing ground truth data against the TAOIs recommended by the proposed *tClust_B* method. In the EBC region, 22 TAOIs are correctly recognized and 17 TAOIs were not identified by *tClust_B*. Similarly, in Pokhara 20 clusters were recognized as TAOI correctly, 9 TAOIs were missed, and 7 clusters were incorrectly identified as TAOIs. F1 scores of 0.72 and 0.74 were calculated for EBC region and Pokhara, respectively. For EBC, the model was able to determine the tourist regions correctly as indicated by the high precision score. However, the recall value hints that *tClust_B* sometimes make incorrect predictions by categorizing valid regions as invalid. In the case of Pokhara city, a decrease in precision and an increase in recall values were observed. However, the overall F1 score depicted the consistency of the method over different areas. Figure 10 provides a comparative overview of the three proposed methods for TAOI detection. It is determined that *tClust_B* method offers the best performance among all.

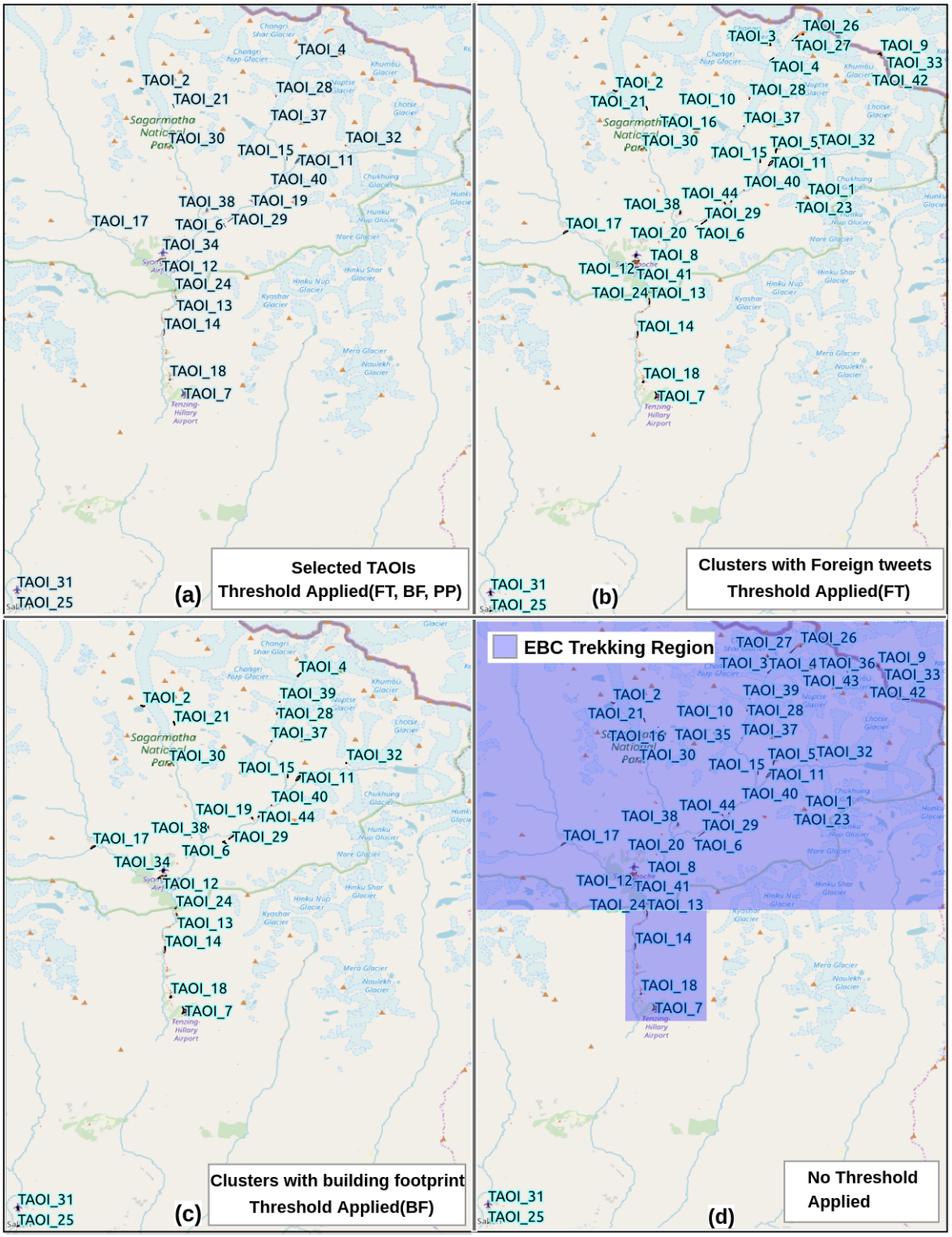

Figure 8. Distribution of TAOIs in Solukhumbu. (**a**) All selected TAOIs , (**b**) all tweet clusters with foreigner tweets, (**c**) all tweet clusters around building footprints, (**d**) all tweets clusters (no threshold).

		Predicted	
		True	*False*
Real	*True*	22	17
	False	0	0

(a)

		Predicted	
		True	*False*
Real	*True*	20	9
	False	7	0

(b)

Figure 9. Confusion Matrices for (**a**) EBC Region and (**b**) Pokahra City.

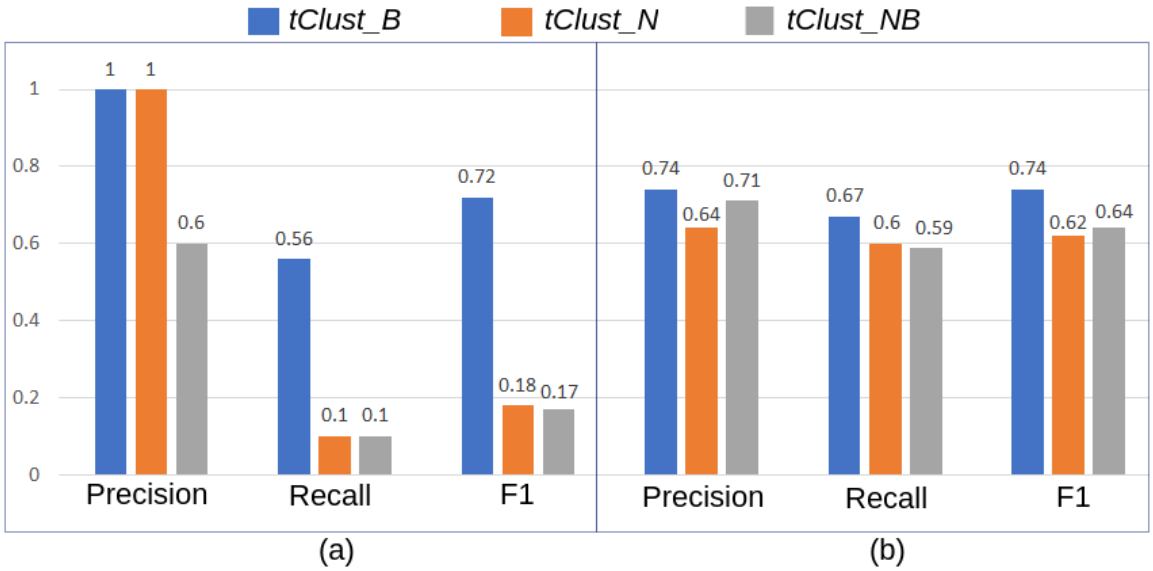

Figure 10. Performance comparison of TAOI detection using tClust_B(FT & PP & BF), tClust_N(FT & PP & NTL) and tClust_NB(FT & PP & NTL & BF): (**a**) EBC Region (**b**) Pokhara.

6. Discussion

The proposed approach automatically searches and identifies touristic places of interest based on collective knowledge. OpenStreetMap and Nighttime light remote sensing data are used to enhance the knowledge obtained from the social media data. Results show that our approach can identify major tourist attractions in different rural and urban tourism destinations in Nepal.

6.1. Significance of the Data Sources Used

Previous works implemented using density-based clustering for mining important tourist locations (such as [20,49]) do not examine if the social media clusters are merely popular or actually related to tourism. Our proposed framework uses multiple data sources such as geo-tagged tweets, NTL data, and building footprint data to discard non-touristic social media clusters. For example, in the proposed study, the presence of a foreigner tweet was essential for the tweet cluster to become touristic. Additional consideration of infrastructures such as buildings in the locale of the identified clusters ensured the availability of minimum essential services for the travelers. Tables 6 and 7 clearly shows that not all spots merely popular in the social media are relevant to tourism.

Furthermore, Table 6 indicates that rural region (i.e., Solukhumbu district) have far less tweet clusters than urban region(Kaski district). It is clear that most of the rural clusters have foreigner tweet presence but those clusters are poorly lit during the night. However, most of the clusters in the urban are located around built-up areas and are illuminated during the night. Furthermore, a detail investigation of the significance of different data sources and the *tClust_B* selected TAOIs in the rural region (i.e., EBC region) and urban region (i.e., Pokhara city) are examined. The contribution of each input data source in the final TAOI selection using *tClust_B* are presented in Figure 11. All the built-up

areas with enough tweet density in the rural region are touristic. In contrast, only half of the clusters with building footprints in the urban region are touristic. This is because more local users publish tweets in urban than in rural. Moreover, almost all clusters with NTL exist around areas with buildings. Though most urban TAOIs have noticeable NTL, it is not the case in the rural TAOIs. However, NTL presence ensures the existence of better infrastructure and facilities in the neighborhood [83,86]. Hence, NTL availability indicates crucial information useful for identifying prime touristic locations.

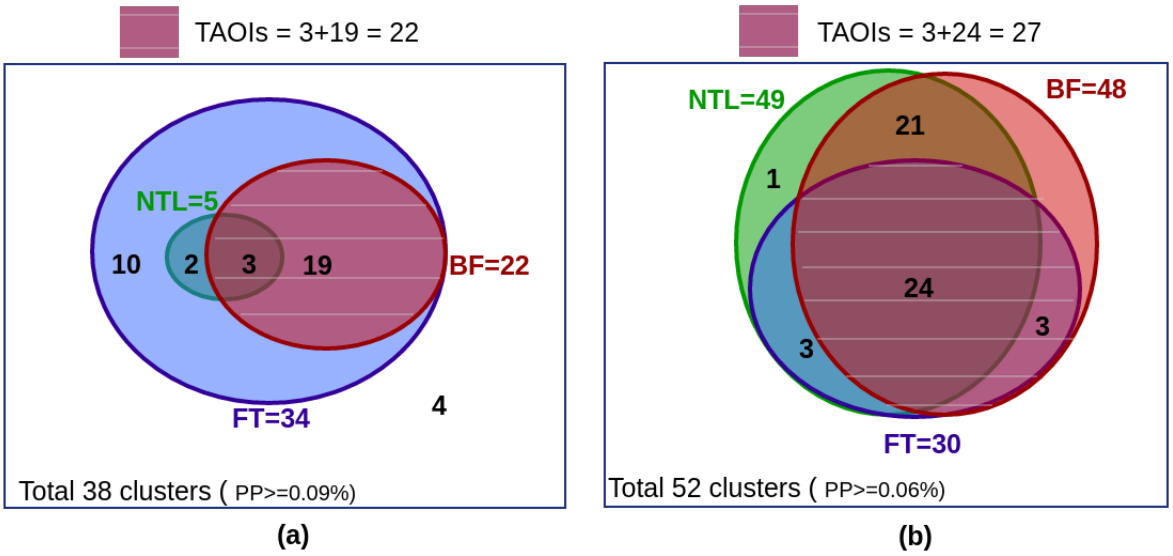

Figure 11. *tClust_B* TAOIs selection from clusters with sufficient tweet PP in (**a**) rural region (EBC, Solukhumbu) and (**b**) urban region (Pokhara, Kaski).

6.2. Significance of Cluster Pruning

It is important to recognize the desired tourism areas and eliminate any non-tourism clusters from the result. However, it is equally important that the selected TAOIs provide minimum essential services (e.g., overnight accommodation and breakfast) to the travelers. Korakakis et al. identified tourist locations without guaranteeing such facilities [49]. Although Majid et al., examined the context of travelers such as time and whether but no consideration was made for such essential requirements [20]. An example in our study clarifies the need to ensure the availability of minimum shelter requirements in the tourism spots. For instance, 'Kala Patthar' (i.e., *TAOI*_3 in Figure 8b), located in Solukhumbu district, is a very popular spot among travelers and caters with an amazing view of the snow-capped mountains including the Mount Everest. Though this place has extreme geography and harsh climatic conditions, a lot of geo-tagged tweets, from foreigners, were found forming a significant cluster; however, it was not included in the final list of TAOIs because this place does not have any infrastructure support for serving the travelers. The BF check helped to prune this spot from the final result.

Similarly, one of the active tweet clusters in Kaski district was identified around a collage area which was not selected as a final TAOI because the area does not fulfill the FT requirement. Field examination revealed that this particular spot does not contain any tourism facilities. Similarly, many other clusters were pruned out of the final result due to the absence of sufficient tweet points or foreigner tweets or building footprint data. Table 7 provides the number of clusters filtered out as non-tourism clusters while identifying interesting TAOIs in the different regions. Hence, the thresholding step is crucial in filtering out irrelevant tweet clusters from the final selection.

Krikigianni et al. suggested the high NTL areas are relatively more urbanized and maintain more tourism infrastructure and services than the low-lit areas [65]. Also, a relatively small number of TAOIs were identified in the rural district than in urban districts. In our investigation, the NTL threshold acted as a stronger discriminator in the rural regions than in the urban districts. It was evident that

NTL (i.e., *tClust_N* and *tClust_NB*) could not act as a good proxy for monitoring TAOIs in rural regions where there is very low or even no nighttime light illumination. Hence, the model suggested by Krikigianni et al. may not provide good insights in such places. However, our proposed model (i.e., *tClust_B* which uses the building footprint information served as a better alternative to NTL by providing a fair decision while discovering TAOIs in both rural and urban region. Furthermore, thresholding may be improved by including other socio-economic and geographical features such as accessibility, population, etc. Exploration in this direction will be done in the future.

6.3. TAOI Categorization

This study focuses on identifying important touristic spots which can fulfill the basic needs of a travelers. This indicates that the vicinity of the discovered TAOIs must serve, at least, a basic bed and breakfast facility. Also, an approach for examining the existence of prime attractions within the extents of a TAOI has been explored in this research. Guidelines for identifying less crowded and economical alternatives for those prime locations have also been discussed. However, naming and categorizing the TAOIs based on their inherent features would be more informative. Analyzing the textural data shared by the users in social media can be helpful in labeling the TAOIs with categories, for example, TAOI_7 (Location Name: Lukla, Details: airport, hotels, mountain scene).

6.4. Cost Effective and Quicker Alternative to Traditional Methods

Buntain et al. suggested that social media data such as tweets can provide better insights in understanding social phenomena more rapidly and at lower cost than ground surveys [87]. For instance, Figure 12 shows newly emerged touristic hubs in the Kaski district. These spots which lie along the coast of Phewa lake, used to be non-touristic in the past. The proposed framework successfully discovered these newly formed tourism sites although most of the traditional data repositories are still outdated. Hence, an exhaustive investigation of the false positive results generated by our system may give more insights. For instance, the presence of foreigner tweets in non-touristic regions may signify an emerging tourist attraction.

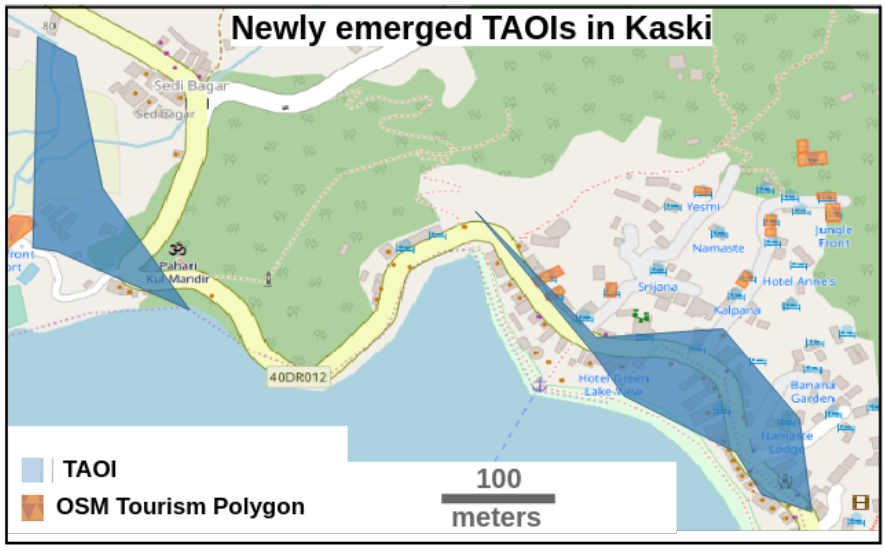

Figure 12. Newly emerged TAOI in Kaski.

6.5. Comparison with OSM

OSM is a popular crowdsourced project that acts as a free source for geodata. In OSM, polygons (or closed ways) are used to depict the boundaries of areas such as parks, buildings, or forests. The quality and coverage of OSM data are not consistent globally [28,88]. For instance, it is surprising that many TAOIs identified do not have any such OSM tourism features. Furthermore, the prime

locations discovered in Kaski and Solukhumbu do not have any tourism polygon feature mappings in the OSM. However, field survey uncovered the fact that though there are several physical structures (e.g., buildings) on site, many of them do not have a proper mapping in the OSM. Many tourism facilities such as hotels, restaurants, viewpoints, etc. were observed on the ground and most of them were present in the OSM but proper labeling was not used. This means OSM lacks the representation of such features, or even if some features are represented they are not annotated correctly with the 'tourism' tag. Hence, social media data can act as a more up-to-date and closer to real-time tool for identification of active tourism centers.

6.6. Impact of Location Accuracy of Geo-Tagged Tweets

Our assumption of geo-tagged location is based on the capacity of smartphone which may compromise location accuracy up to 13 meters [89,90]. The underlying DBSCAN clustering algorithm generates social media clusters with reference to the location of geo-tagged tweets. The search distance parameter (i.e., *eps*) of the algorithm controls the formation of the clusters based on the horizontal distance among the geo-tweets. Therefore, the positional changes in the tweet points may affect the formation of the clusters. Many previous studies (e.g., [17,18,50,58]) which use manually selected static values for the clustering parameters may not adapt well in such cases. However, our proposed approach works well even in such scenarios because the parameters for clustering algorithm are estimated by using objective functions *StrictParameters* and *MaxParameters*. Moreover, the positional accuracy of the tweets may affect the geometry of the cluster formed and if the cluster size becomes smaller then fewer TAOI may be detected in regions with sparse building footprints.

6.7. Limitations in Coverage

Though we were able to get good results, it should be considered in the light of limitations of the input data and method used. The quality of OSM data has often been questioned for consistency and completeness. Though nighttime light images are used as a proxy in various research studies, it is of low resolution and it cannot capture activities unrelated to light. Furthermore, Twitter usage does not include all online activities because not all Internet users participate in this social network. However, the remotely sensed data and VGI data used are publicly accessible and consistently available across the world. Therefore, we assume nighttime light remote sensed data and VGI data can still act as a proxy for our purpose. Also, there is some room for improvements in the method. The DBSCAN clustering algorithm used for identifying TAOIs is not density adaptive. The process of NTL, BF, and tweet percentage threshold estimation is manual. Tuning of the threshold values is expected to improve results.

7. Conclusions

Twitter provides freely accessible geo-tagged volunteered geographic information which can be investigated across various domains of human activities. This study demonstrated that tweets hold great potential for mapping touristic areas even in remote and data deficient regions. Freely available OSM data and remotely sensed nighttime light data helped in enriching the TAOI discovery process. In this study, we collected and cleaned the geo-tagged tweets, and then applied clustering to get the spatial distribution of the social media clusters. Next, tourism related clusters were distinguished from normal clusters by exploring different Twitter metadata along with OSM and NTL data. The final list of selected TAOIs were not only popular among the travelers but also ensured the availability of essential tourism infrastructure in the vicinity. The proposed framework was examined in two different tourism districts in Nepal with satisfactory results. This framework not only identified popular TAOIs but also helped in spotting prime tourism locations in the neighborhood.

Our proposed analytic framework will enhance the state-of-the-art tourism studies in several directions. First, by discovering tourism areas of interest in data deficient remote areas by fusing free social media data and other free data sources such as OSM and NTL. Secondly, by devising a

method to estimate clustering parameters values which can be applied across different scarce data regions. Thirdly, by ensuring the availability of minimum essential facilities in the recommended touristic sites. Moreover, useful guidelines were prescribed for identifying obscure alternatives apart from prime touristic locations. Most importantly, the approach outlined in this paper presents a novel methodology previously unused in tourism attraction studies.

The proposed approach examined the discovery of TAOIs from social media clusters which were located mainly in built-up regions (or well-lit regions). Some of the TAOIs with fewer media attention or located in non-built (or non-lit) regions may be discarded. A simple way to include such regions in the final selection is by softening the threshold values. However, optimal threshold selection and method universalization will be explored in future iterations of our work. Further improvements are to be made in other aspects such as the elimination of bias arising from the inherent characteristics of selected data sources (Twitter, OSM, remote sensing images).

Author Contributions: B.D. and H.M. contributed to the overall study design and supervised all research. B.D. completed the data collection and development of analysis tools. B.D. and H.M. contributed to the data analysis and the manuscripts. A.W. and S.M.K. reviewed and provided critical comments for the improvement of the work.

Funding: We would like to acknowledgment the Japanese Government Scholarship. This study is supported by the Japanese Government Scholarship at the Asian Institute of Technology (grant: August 2016).

Acknowledgments: We thank Dhiraj Pahari and Niraj Pahari for assisting us during the data collection and study area surveys. Their knowledge and network on Nepalese Tourism and helpful comments has greatly contributed in the study. The authors would also like to thank Michelle Han for her help in editing the manuscript.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ashley, C.; De Brine, P.; Lehr, A.; Wilde, H. *The Role of the Tourism Sector in Expanding Economic Opportunity*; John F. Kennedy School of Government, Harvard University Cambridge: Cambridge, MA, USA, 2007.
2. WTTC. World Travel and Tourism Report 2018 Report for Nepal. Available online: https://www.wttc.org/-/media/files/reports/economic-impact-research/countries-2018/nepal2018.pdf (accessed on 18 March 2019).
3. United Nations. Sustainable Tourism. Available online: https://sustainabledevelopment.un.org/topics/sustainabletourism (accessed on 27 June 2019).
4. Morrison-Saunders, A.; Hughes, M.; Pope, J.; Douglas, A.; Wessels, J.A. Understanding visitor expectations for responsible tourism in an iconic national park: Differences between local and international visitors. *J. Ecotourism* **2019**, 1–11. [CrossRef]
5. Martín Martín, J.M.; Guaita Martínez, J.M.; Molina Moreno, V.; Sartal Rodríguez, A. An Analysis of the Tourist Mobility in the Island of Lanzarote: Car Rental Versus More Sustainable Transportation Alternatives. *Sustainability* **2019**, *11*, 739. doi:10.3390/su11030739. [CrossRef]
6. Henderson, J.V.; Storeygard, A.; Weil, D.N. Measuring economic growth from outer space. *Am. Econ. Rev.* **2012**, *102*, 994–1028. [CrossRef] [PubMed]
7. Bennett, P.; Giles, L.; Halevy, A.; Han, J.; Hearst, M.; Leskovec, J. Channeling the deluge: Research challenges for big data and information systems. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 2537–2538.
8. Heikinheimo, V.; Minin, E.D.; Tenkanen, H.; Hausmann, A.; Erkkonen, J.; Toivonen, T. User-generated geographic information for visitor monitoring in a national park: A comparison of social media data and visitor survey. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 85. [CrossRef]
9. Jendryke, M.; Balz, T.; McClure, S.C.; Liao, M. Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. *Comput. Environ. Urban Syst.* **2017**, *62*, 99–112. [CrossRef]
10. Miyazaki, H.; Nagai, M.; Shibasaki, R. Development of Time-Series Human Settlement Mapping System using Historical Landsat Archive. *Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2016**, *41*, 1385. [CrossRef]

11. Sitthi, A.; Nagai, M.; Dailey, M.; Ninsawat, S. Exploring land use and land cover of geotagged social-sensing images using naive bayes classifier. *Sustainability* **2016**, *8*, 921. [CrossRef]
12. Preoţiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; Aletras, N. Studying user income through language, behaviour and affect in social media. *PLoS ONE* **2015**, *10*, e0138717. [CrossRef] [PubMed]
13. Levin, N.; Lechner, A.M.; Brown, G. An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas. *Appl. Geogr.* **2017**, *79*, 115–126. [CrossRef]
14. Park, J.H.; Lee, C.; Yoo, C.; Nam, Y. An analysis of the utilization of Facebook by local Korean governments for tourism development and the network of smart tourism ecosystem. *Int. J. Inf. Manag.* **2016**, *36*, 1320–1327. [CrossRef]
15. Del Vecchio, P.; Mele, G.; Ndou, V.; Secundo, G. Creating value from social big data: Implications for smart tourism destinations. *Inf. Process. Manag.* **2018**, *54*, 847–860. [CrossRef]
16. García-Palomares, J.C.; Gutiérrez, J.; Mínguez, C. Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Appl. Geogr.* **2015**, *63*, 408–417. [CrossRef]
17. Encalada, L.; Boavida-Portugal, I.; Cardoso Ferreira, C.; Rocha, J. Identifying tourist places of interest based on digital imprints: Towards a sustainable smart city. *Sustainability* **2017**, *9*, 2317. [CrossRef]
18. Maeda, T.; Yoshida, M.; Toriumi, F.; Ohashi, H. Extraction of Tourist Destinations and Comparative Analysis of Preferences Between Foreign Tourists and Domestic Tourists on the Basis of Geotagged Social Media Data. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 99. doi:10.3390/ijgi7030099. [CrossRef]
19. Chen, M.; Arribas-Bel, D.; Singleton, A. Understanding the dynamics of urban areas of interest through volunteered geographic information. *J. Geogr. Syst.* **2018**. doi:10.1007/s10109-018-0284-3. [CrossRef]
20. Majid, A.; Chen, L.; Mirza, H.T.; Hussain, I.; Chen, G. A system for mining interesting tourist locations and travel sequences from public geo-tagged photos. *Data Knowl. Eng.* **2015**, *95*, 66–86. [CrossRef]
21. Vu, H.Q.; Li, G.; Law, R.; Ye, B.H. Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos. *Tour. Manag.* **2015**, *46*, 222–232. doi:10.1016/j.tourman.2014.07.003. [CrossRef]
22. Lee, J.Y.; Tsou, M.H. Mapping Spatiotemporal Tourist Behaviors and Hotspots Through Location-Based Photo-Sharing Service (Flickr) Data. In *Lecture Notes in Geoinformation and Cartography, Proceedings of the Progress in Location Based Services 2018, LBS 2018, Zurich, Switzerland, 15–17 January 2018*; Kiefer, P., Huang, H., Van de Weghe, N., Raubal, M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 315–334.
23. Zhuang, C.; Ma, Q.; Liang, X.; Yoshikawa, M. Discovering obscure sightseeing spots by analysis of geo-tagged social images. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 25–28 August 2015; pp. 590–595.
24. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, USA, 2–4 August 1996; Simoudis, E., Han, J., Fayyad, U., Eds.; AAAI Press: Palo Alto, CA, USA, 1996; Volume 96, pp. 226–231.
25. Hu, Y.; Gao, S.; Janowicz, K.; Yu, B.; Li, W.; Prasad, S. Extracting and understanding urban areas of interest using geotagged photos. *Comput. Environ. Urban Syst.* **2015**, *54*, 240–254. [CrossRef]
26. 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]
27. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
28. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. Plan. Des.* **2010**, *37*, 682–703. [CrossRef]
29. Li, L.; Goodchild, M.F. Constructing places from spatial footprints. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Redondo Beach, CA, USA, 6 November 2012; pp. 15–21.
30. Salas-Olmedo, M.H.; Moya-Gómez, B.; García-Palomares, J.C.; Gutiérrez, J. Tourists' digital footprint in cities: Comparing Big Data sources. *Tour. Manag.* **2018**, *66*, 13–25. doi:10.1016/j.tourman.2017.11.001. [CrossRef]
31. Estima, J.; Painho, M. Exploratory analysis of OpenStreetMap for land use classification. In Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Orlando, FL, USA, 5 November 2013; pp. 39–46.

32. Sun, Y.; Fan, H.; Li, M.; Zipf, A. Identifying the city center using human travel flows generated from location-based social networking data. *Environ. Plan. Plan. Des.* **2016**, *43*, 480–498. [CrossRef]
33. Koutras, A.; Nikas, I.A.; Panagopoulos, A. Towards Developing Smart Cities: Evidence from GIS Analysis on Tourists' Behavior Using Social Network Data in the City of Athens. In *Smart Tourism as a Driver for Culture and Sustainability*; Katsoni, V.; Segarra-Oña, M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 407–418.
34. Xing, H.; Meng, Y.; Hou, D.; Song, J.; Xu, H. Employing crowdsourced geographic information to classify land cover with spatial clustering and topic model. *Remote. Sens.* **2017**, *9*, 602. [CrossRef]
35. Li, Y.; Li, Q.; Shan, J. Discover patterns and mobility of Twitter users—A study of four US college cities. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 42. [CrossRef]
36. Mazanec, J. Segmenting city tourists into vacation styles. In *International City Tourism: Analysis and Strategy*; Pinter: London, UK, 1997; pp. 114–128.
37. Shoval, N.; Raveh, A. Categorization of tourist attractions and the modeling of tourist cities: Based on the co-plot method of multivariate analysis. *Tour. Manag.* **2004**, *25*, 741–750. [CrossRef]
38. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 21 June–18 July 1965; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
39. Rousseeuw, P.J.; Kaufman, L. Clustering by means of medoids. In *Statistical Data Analysis Based on the L1 Norm and Related Methods*; Dodge, Y., Ed.; North-Holland/Elsevier: Amsterdam, The Netherlands, 1987; pp. 405–416.
40. Birant, D.; Kut, A. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data Knowl. Eng.* **2007**, *60*, 208–221. [CrossRef]
41. Kisilevich, S.; Mansmann, F.; Keim, D. P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, Washington, DC, USA, 21–23 June 2010; p. 38.
42. Campello, R.J.; Moulavi, D.; Sander, J. Density-based clustering based on hierarchical density estimates. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Gold Coast, Australia, 14–17 April 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.
43. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
44. Getis, A.; Ord, J. The Analysis of Spatial Association by Use of Distance Statistics, Geographycal Analysis. In *Perspectives on Spatial Data Analysis*; Springer: Berlin/Heidelberg, Germany, 1992.
45. Wang, T.; Ren, C.; Luo, Y.; Tian, J. NS-DBSCAN: A Density-Based Clustering Algorithm in Network Space. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 218. [CrossRef]
46. Dehuri, S.; Mohapatra, C.; Ghosh, A.; Mall, R. Comparative study of clustering algorithms. *Inf. Technol. J.* **2006**, *5*, 551–559. [CrossRef]
47. Yang, Y.; Gong, Z. Identifying points of interest by self-tuning clustering. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 24–28 July 2011; pp. 883–892.
48. Laptev, D.; Tikhonov, A.; Serdyukov, P.; Gusev, G. Parameter-free discovery and recommendation of areas-of-interest. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, TX, USA, 4–7 November 2014; pp. 113–122.
49. Korakakis, M.; Spyrou, E.; Mylonas, P.; Perantonis, S.J. Exploiting social media information toward a context-aware recommendation system. *Soc. Netw. Anal. Min.* **2017**, *7*, 42. [CrossRef]
50. Hasnat, M.M.; Hasan, S. Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transp. Res. Part Emerg. Technol.* **2018**, *96*, 38–54. [CrossRef]
51. Kuo, C.L.; Chan, T.C.; Fan, I.; Zipf, A. Efficient Method for POI/ROI Discovery Using Flickr Geotagged Photos. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 121. [CrossRef]
52. Yan, Y.; Kuo, C.L.; Feng, C.C.; Huang, W.; Fan, H.; Zipf, A. Coupling maximum entropy modeling with geotagged social media data to determine the geographic distribution of tourists. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 1699–1736. [CrossRef]
53. Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799. [CrossRef]

54. Rosenblatt, M. Remarks on some nonparametric estimates of a density function. In *The Annals of Mathematical Statistics*; JSTOR: New York, NY, USA, 1956; pp. 832–837. doi:10.1214/aoms/1177728190.
55. Dueck, D. Affinity Propagation: Clustering Data by Passing Messages. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2009.
56. Ng, A.Y.; Jordan, M.I.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2001; pp. 849–856.
57. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst. (TODS)* **2017**, *42*, 19. [CrossRef]
58. Zhou, X.; Xu, C.; Kimmons, B. Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. *Comput. Environ. Urban Syst.* **2015**, *54*, 144–153. [CrossRef]
59. Akdag, F.; Eick, C.F.; Chen, G. Creating Polygon Models for Spatial Clusters. In *Foundations of Intelligent Systems*; Andreasen, T.; Christiansen, H.; Cubero, J.C.; Raś, Z.W., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 493–499.
60. Duckham, M.; Kulik, L.; Worboys, M.; Galton, A. Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. *Pattern Recognit.* **2008**, *41*, 3224–3236. [CrossRef]
61. Devkota, B.; Miyazaki, H. An Exploratory Study on the Generation and Distribution of Geotagged Tweets in Nepal. In Proceedings of the 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, Nepal, 25–27 October 2018; pp. 70–76.
62. Yamaguchi, Y.; Amagasa, T.; Kitagawa, H. Landmark-based user location inference in social media. In Proceedings of the First ACM Conference on Online Social Networks, Boston, MA, USA, 7–8 October 2013; pp. 223–234.
63. Chong, W.H.; Lim, E.P. Fine-grained Geolocation of Tweets in Temporal Proximity. *ACM Trans. Inf. Syst. (TOIS)* **2019**, *37*, 17. [CrossRef]
64. Jurgens, D. That's what friends are for: Inferring location in online social media platforms based on social relationships. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Atlanta, GA, USA, 28 June 2013.
65. Krikigianni, E.; Tsiakos, C.; Chalkias, C. Estimating the relationship between touristic activities and night light emissions. *Eur. J. Remote. Sens.* **2019**, *52*, 233–246. [CrossRef]
66. Checa, J. Urban Intensities. The Urbanization of the Iberian Mediterranean Coast in the Light of Nighttime Satellite Images of the Earth. *Urban Sci.* **2018**, *2*, 115. [CrossRef]
67. Nepal, S.K.; Kohler, T.; Banzhaf, B.R. *Great Himalaya: Tourism and the Dynamics of Change in Nepal*; Swiss Foundation for Alpine Research: Zurich, Switzerland, 2002.
68. Central Bureau of Statistics, Government of Nepal. National Population and Housing Census 2011. Available online: https://unstats.un.org/unsd/demographic/sources/census/wphc/Nepal/Nepal-Census-2011-Vol1.pdf (accessed on 18 March 2019).
69. LonelyPlanet. Top Experiences in Nepal. Available online: https://www.lonelyplanet.com/nepal (accessed on 7 April 2019).
70. InternetLiveStats. Twitter Usage Statistics. 2013. Available online: http://www.internetlivestats.com/twitter-statistics/ (accessed on 18 March 2019)
71. Cesare, N.; Grant, C.; Nsoesie, E.O. Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices. *arXiv* **2017**, arXiv:1702.01807.
72. Burghardt, M. Tools for the Analysis and Visualization of Twitter Language Data. Available online: https://epub.uni-regensburg.de/35669/ (accessed on 18 March 2019).
73. Puschmann, C.; Bruns, A.; Mahrt, M.; Weller, K.; Burgess, J. Epilogue: Why Study Twitter? In *Twitter and Society*; Peter Lang: New York, NY, USA, 2014; Volume 89, pp. 425–432.
74. SocialAves. Social Media Landscape Nepal. 2017. Available online: https://socialaves.com/social-media-landscape-nepal/ (accessed on 18 March 2019)
75. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77. [CrossRef]
76. Morstatter, F.; Pfeffer, J.; Liu, H.; Carley, K.M. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv* **2013**, arXiv:1306.5204.

77. Morstatter, F.; Pfeffer, J.; Liu, H. When is it biased? Assessing the representativeness of twitter's streaming API. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 555–556.
78. Frias-Martinez, V.; Soto, V.; Hohwald, H.; Frias-Martinez, E. Characterizing urban landscapes using geolocated tweets. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 239–248.
79. Zhao, N.; Cao, G.; Zhang, W.; Samson, E.L. Tweets or nighttime lights: Comparison for preeminence in estimating socioeconomic factors. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *146*, 1–10. [CrossRef]
80. Elvidge, C.D.; Safran, J.; Tuttle, B.; Sutton, P.; Cinzano, P.; Pettit, D.; Arvesen, J.; Small, C. Potential for global mapping of development via a nightsat mission. *GeoJournal* **2007**, *69*, 45–53. [CrossRef]
81. Baugh, K.; Hsu, F.C.; Elvidge, C.D.; Zhizhin, M. Nighttime lights compositing using the VIIRS day-night band: Preliminary results. *Proc. Asia-Pac. Adv. Netw.* **2013**, *35*, 70–86. [CrossRef]
82. Li, X.; Elvidge, C.; Zhou, Y.; Cao, C.; Warner, T. Remote sensing of night-time light. *Int. J. Remote. Sens.* **2017**, 38, 5855–5859. [CrossRef]
83. Mellander, C.; Lobo, J.; Stolarick, K.; Matheson, Z. Night-time light data: A good proxy measure for economic activity? *PLoS ONE* **2015**, *10*, e0139779. [CrossRef]
84. Board, N.T. Greater Pokhara Valley Lake side and City Map. In *Greater Pokhara Valley and City Map*; Nepal Map Publisher: Kathmandu, Nepal, 2011.
85. Banerjee, P.S. *Everest Trekking Maps and Complete Guide*; Milestone Himalayan Series; Milestone Books: Calcutta, India, 2017.
86. Nel, O.; López, J.; Martín, J.; Checa, J. Energy and urban form. The growth of European cities on the basis of night-time brightness. *Land Use Policy* **2017**, *61*, 103–112. [CrossRef]
87. Buntain, C.; McGrath, E.; Golbeck, J.; LaFree, G. Comparing Social Media and Traditional Surveys around the Boston Marathon Bombing. In Proceedings of the #Microposts: 6th Workshop on Making Sense of Microposts, Montréal, QC, Canada, 11–15 April 2016; pp. 34–41.
88. Zheng, S.; Zheng, J. Assessing the completeness and positional accuracy of OpenStreetMap in China. In *Thematic Cartography for the Society*; Springer: Cham, Switzerland, 2014; pp. 171–189.
89. Tomaštík, J.; Saloň, Š.; Piroh, R. Horizontal accuracy and applicability of smartphone GNSS positioning in forests. *For. Int. J. For. Res.* **2016**, *90*, 187–198. [CrossRef]
90. Merry, K.; Bettinger, P. Smartphone GPS accuracy study in an urban environment. *PLoS ONE* **2019**, *14*, e0219890. [CrossRef] [PubMed]