GEO-INFORMATION SCIENCE AND REMOTE SENSING

THESIS REPORT GIRS-2017-36

Citizen Perception of Nature on Social Media

M.C.S. Holtslag

09 December 2017

WAGENINGEN
UNIVERSITY & RESEARCH

Citizen Perception of Nature on Social Media

M.C.S. Holtslag

(940720-356-030)

Supervisors:

Dr. Ir. Ioannis N. Athanasiadis[1]

Dr. Ir. Arend Ligtenberg[2]

[1] Information Technology Group
Wageningen, The Netherlands

[2] Laboratory of Geo-Information Science and Remote Sensing
Wageningen, The Netherlands

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research,
The Netherlands.

09 December 2017

Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2017-36

Wageningen University and Research
Laboratory of Geo-Information Science and Remote Sensing

ABSTRACT

With the internet becoming available to the wide public, social media platforms have become an important part of people's everyday life worldwide. This has drawn interest in the academic world, using the possibility to extract and analyse people's opinions from social media. This research investigates whether data from social media can be used to determine citizen perception. A case study is carried out to find the citizen perception of nature, to assess cultural ecosystem services in the Pike-San Isabel National Forest, Colorado USA. The results are compared with the results from a study in the same area, using survey data to determine the social value of the park, in order to find whether the proposed method can replace the use of surveys. A large dataset has been created by harvesting data from different social media platforms. Relevant data from this dataset was selected using a Naive Bayes classification combined with Expectation Maximization. A sentiment analysis has been applied to extract the citizen perception from the messages. This data has been extrapolated towards social value maps using the Social Values for Ecosystem Services (SolVES) tool. After which a hotspot analysis was performed to identify hotspots with a high social value in the area. When these results were compared to the results from the previous study using surveys, it could be concluded that social media data can be used as a complement to this data to increase the value of knowledge about cultural ecosystem services.

Keywords: Cultural ecosystem services, Expectation Maximization, Hotspot analysis, Naive Bayes, Sentiment analysis, Social media, SolVES

ACKNOWLEDGEMENTS

The last couple of months have shown me that I am capable of developing my own research and performing every step of it myself. I learned how to solve problems, or formulate questions in order to be able to ask others for help. For this, I especially would like to thank my two supervisors: Ioannis Athanasiadis and Arend Ligtenberg, for providing me useful feedback, and always being willing to help me improve my work. Even when little time was available they were able to find moments to squeeze a meeting in.

In some areas of my research expertise from different sources was needed. Therefore, I would like to thank John Stuiver for helping me solve all issues regarding ArcMap and Add-Ins, by spending multiple hours trying different approaches. I would also like to thank Sytze de Bruin for spending time to help me understand the statistical methods used in this research. Whenever I had issues regarding the SolVES tool I got immediate response from Ben Sherrouse to help me solve my problems, therefore my gratitude. Special thanks to Kenneth Bagstad for performing the study that lead to my research questions, and providing me the maps needed for my validation.

Finally, I would like to thank my fellow students for the joyful moments in and outside the thesis room. Special thanks to Robbert-Jan Joling and Gijs Peters for proofreading my work and providing me useful feedback. And last, but not least, my gratitude to Jorn Habes for giving me the needed support and motivation, and willingness to proofread my work.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 SOCIAL MEDIA

The internet was first developed for military use in 1969, but from the 80's onwards it became more widely used by the non-military world as well. By 1992, there were many internet services available, of which the World Wide Web gained the most attention, and eventually grew to be the biggest and most widely used internet service (Marson, 1997). By the upcoming of the internet during the 21st century, the term Web 2.0 was mentioned by O'Reilly (2005). Web 2.0 offered many new possibilities for users on the internet, examples being: websites could be used as platforms, instead of only being standalone and static, it became possible to cross link between websites, and data driven web applications got the opportunity of being developed, since database systems evolved. These developments resulted in the evolution of social media (Kaplan and Haenlein, 2010). Social media platforms can be described as a group of applications that allow the exchange of user generated content and is based on Web 2.0 technologies (Khan et al., 2016).

Social media gives users the opportunity to share and discuss their personal experiences easily online. In the research of Samuel Fosso et al. (2016) it was concluded that 74 per cent of online adults in the USA are active on social media. Since people easily share their personal opinions on social media, the marketing world has developed a great interest in this content: i.e. online commercials could be targeted at a specific person. The large amount of opinion data available on social media cannot only be used for marketing purposes, but also for academic research (Khan et al., 2016). Researchers have been collecting, monitoring, analysing, summarising, and visualising data from social media to extract patterns and intelligence (Samuel Fosso et al., 2016).

One sort of information that can be extracted from social media data is the citizen perception regarding a certain topic. Existing social media analysis methods are able to extract topics from text data (Nie et al., 2013), after which opinion mining or sentiment analysis can be applied to find the citizen perception regarding these topics (Sobkowicz et al., 2012). A next step would be to extrapolate this information to a larger scale, by finding spatial patterns in this data (Wood et al., 2013). An interesting field to apply this approach to is by assessing the natural environment, by selecting messages regarding e.g. nature.

1.2 CULTURAL ECOSYSTEM SERVICES

As the human society we are dependent on the services provided to us by nature, which are referred to as ecosystem services. Ecosystems are known for their provision of supporting services (underlying natural processes), provisioning services (delivering of goods), regulating services (regulation of natural processes in the benefits of people), and cultural services (beneficial services for social well-being). Lately a decrease of almost two thirds of these services provided by nature is found worldwide (Millennium Ecosystem Assessment, 2005). Therefore more research is performed

to provide a quantification of ecosystem goods and services, in order to improve decision making in these areas (Ruhl et al., 2013).

The most widely used way to quantify ecosystem services is by creating biophysical models, which are developed based on hydrological, soil, climate, topographical, remote-sensed, and land-cover data (Martínez-Harms and Balvanera, 2012). The only ecosystem service difficult to quantify using biophysical models are the cultural services, which cannot be directly deduced from any of the sources of biophysical data (Martínez-Harms and Balvanera, 2012). Historically this has limited the influence of cultural services on decision making (Daniel et al., 2012). Nowadays cultural services are being mapped often through Public Participatory Geographic Information Systems (PPGIS) approaches, which are based on i.e. surveys of the publics values and attitudes towards ecosystem services (Sherrouse et al., 2011). PPGIS, or in other words social values mapping, can influence decision making by offering a means to quantify the cultural services, or social value, of an ecosystem (Brown, 2012).

An example of such a study is the research from Bagstad et al. (2016), who tried to find the relationship between the results from a social values mapping study and corresponding biophysically modelled ecosystem services. The research area that was used for this study is the Pike-San Isabel (PSI) National Forest in Colorado, USA. The social value of the area was determined by performing a survey among residents living close by the PSI. Bagstad et al. (2016) found that social values mapping is not able to replace biophysical modelling, since the relationship was found to be very limited. However, they concluded that social values mapping can work as a complement to biophysical models instead, increasing the value of ecosystem services maps by indicating areas of synergy or conflict.

1.3 Problem Description

Studies to analyse the cultural ecosystem services are survey-based, which means that the respondents answered pre-produced questions. However, as mentioned in the discussion of the paper from Bagstad et al. (2016), responses tend to be influenced by the wording and presentation of questions (Schwarz, 1999). Which might, in its turn, influence the results of a study, making them less valuable. Besides, the execution of a survey takes a substantial amount of time and could be accompanied by high costs, even though the effective response rate can be low: i.e. 19 per cent for the survey from Bagstad et al. (2016). To alleviate the bias introduced by pre-produced questions and reduce the costs and time, this thesis will explore the value of using data retrieved from various social media platforms to determine citizen perception. In this way data is voluntarily delivered by social media users, instead of surveys being executed.

A method using social media data to gain insight in the citizen perception does not yet exist, but this information can be obtained by combining various social media analytic methods, which will transform social media data into social value maps. Text analysis methods are required to find relevant messages within the large collection of social media data (Samuel Fosso et al., 2016). With a data mining method this data can be transformed into information regarding the citizen perception (Woo et al., 2015). By using the geographic information available in the social media

data, spatial patterns can be recognised and related to the information retrieved through data mining (Wood et al., 2013). Combining these methods into one research, instead of only using a single analysis method, increases the value of knowledge retrieved from the social media data (Samuel Fosso et al., 2016).

The term citizen perception, which is used in this study, describes the opinion of citizens regarding a certain topic (Choudri et al., 2017). When this term applies to an area, it can be described as the social value of this area. The social value can be seen as the emotional and psychological value people experience in an area (Yoo et al., 2014). As this value can be positive of negative it can also be seen as the sentiment of a person towards the given topic (Reuter and Spielhofer, 2017).

1.4 Research Objective and Questions

The main objective of this research is to analyse the citizen perception of nature using social media, by developing and demonstrating a method to assess cultural ecosystem services by identifying the sentiment of messages. In order to validate the results a case study will be executed covering the same area as the research from Bagstad et al. (2016): the Pike-San Isabel National Forest in the state of Colorado, USA.

Based on this objective the research questions are defined as follows:

1. Which social media platforms can be used for citizen perception analysis?
2. What is the best way to store the collected social media data?
3. How can information about citizen perception be retrieved from social media messages?
4. How do the results from the social media analysis compare to the results from Bagstad et al. (2016)?

1.5 Study Area

The area used for this research is the Pike-San Isabel National Forest (PSI). This national park is located in the state of Colorado, USA. In Figure 1.1 the study area of the PSI is shown. The PSI covers a mountainous area of around 950 000 ha (Donnegan et al., 2001; Taylor et al., 1984).

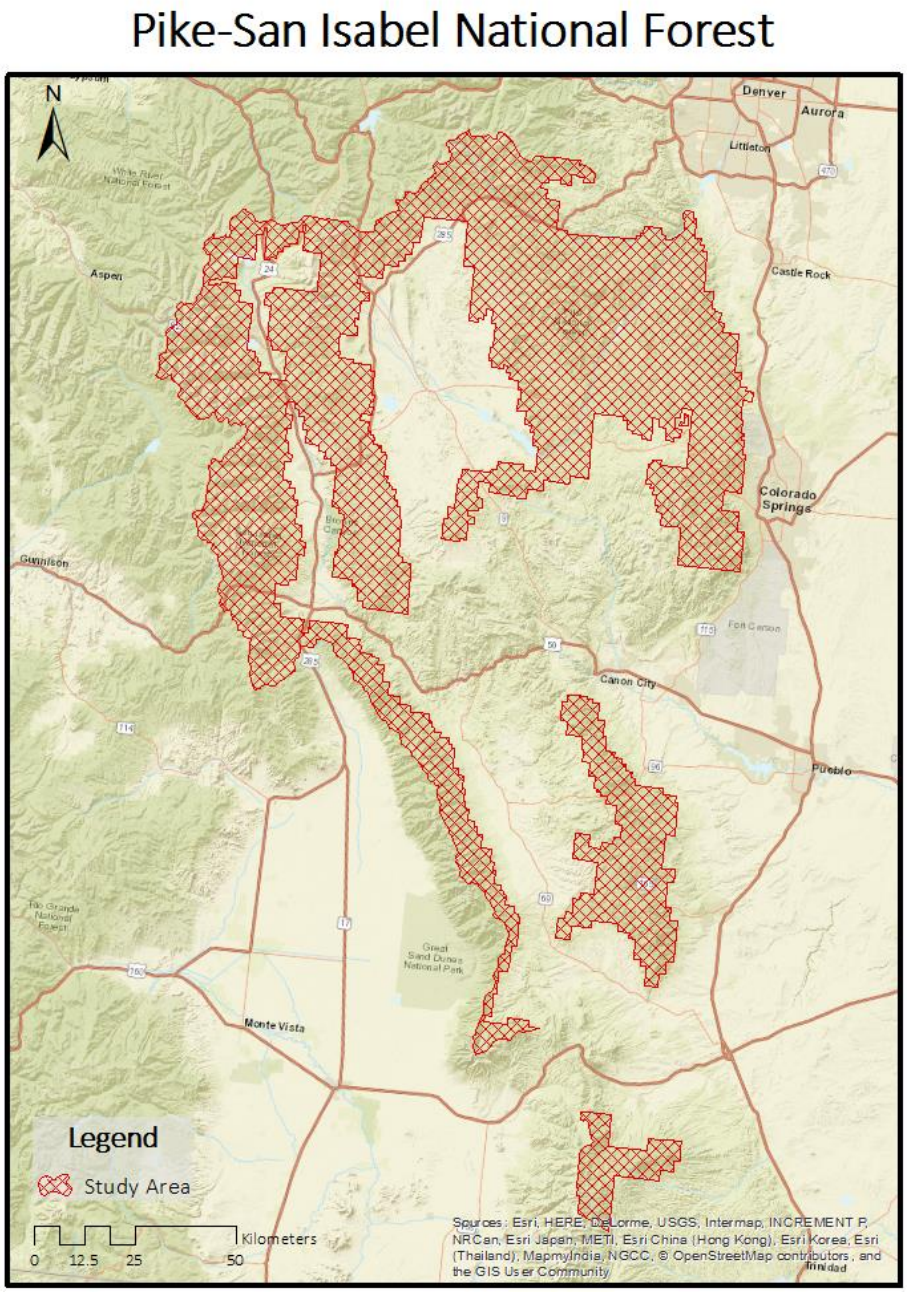

Figure 1.1: Study area of the Pike-San Isabel National Forest

Since the appearance of settlers in the region of the PSI forest, fires have been taking place, damaging over 75 per cent of the total area (Donnegan et al., 2001). The most recent large scale forest fire affecting the PSI is the Hayman Fire (Thompson et al., 2016). Being active from June 8 until July 18, 2002, it burned an area of 52 600 ha (Wang and Zhang, 2017). These forest fire could have a negative influence on the citizen perception of the area.

1.6 FRAMEWORK

The main process in this research is extraction of data from a large dataset. The process of turning data into information can be described by Knowledge Discovery in Databases (KDD), which emphasizes that knowledge is the end product of data-driven analysis. According to the KDD principles (Fayyad et al., 1996), a data-driven research can be performed by following some predesigned steps. These steps are not necessarily performed once, as multiple iterations can be made to optimise the results. The process of this research follows the same principle components as found in the KDD methodology, but the specific steps are adapted to fit the research.

The goal of this research is to determine the usefulness of social media data for analysing citizen perception and propose a method to perform this analysis. The proposed method will be tested to determine the citizen perception of nature, by performing a case study based on the research of Bagstad et al. (2016).

The KDD steps adapted for this research are shown in Figure 1.2. The dataset in the first step corresponds to the large amount of social media data available on the web on multiple social media platforms. In the data harvesting step potentially relevant data will be collected, resulting in a smaller dataset which can be used for further processing. With data selection a subset of this data is taken, based on the available information in the metadata. After classification the final subset is left, this dataset will contain all relevant and useful data. Next, a data mining method needs to be chosen to extract patterns from the dataset. Finally, the creation of maps using the extracted patterns will visualise the patterns, which will reveal knowledge about the citizen perception on social media.

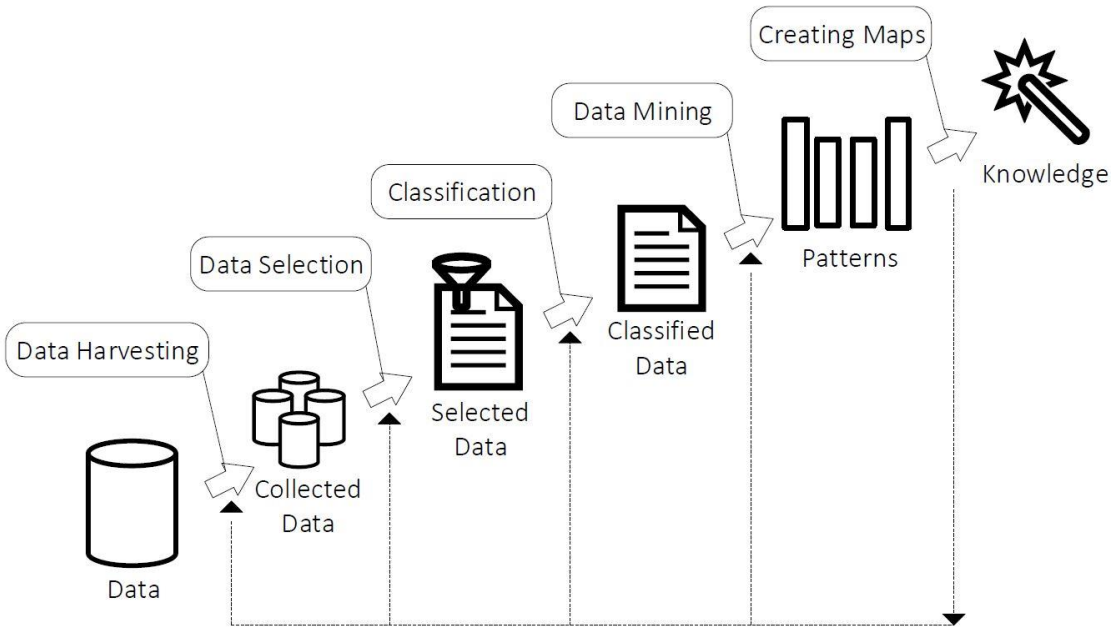

Figure 1.2: Steps performed during Knowledge Discovery in Databases (adapted from Fayyad et al. (1996))

1.7 READING GUIDE

The process in this thesis follows the steps described in Chapter 1.6, the chapters in this report follow this structure as well. An exact overview of the chosen steps is given in Chapter 3.1, these choices are made based on the information given in Chapter 2. Chapter 2.1 gives an overview of methods used to assess cultural ecosystem services.

For the data harvesting step various social media platforms are compared in Chapter 2.2, the methods performed to harvest data from these platforms are given in Chapter 3.2, in Chapter 4.1 some resulting numbers are given, and in Chapter 5.1 the used methods and obtained results of the data harvesting are discussed. To find a suitable database to store the collected data, different types of databases are compared in Chapter 2.3. How the data is stored in the chosen database is explained in Chapter 3.3.

An overview of various social media analytics is given in Chapter 2.4. How one of these methods can be used to filter the collected dataset to only keep the relevant data for this study is discussed in Chapter 2.4.1. In Chapter 3.4 the method used to perform the chosen classification is explained and the results are given in Chapter 4.2. A discussion about the chosen method and resulting classification is given in Chapter 5.2. Chapter 2.4.2 gives some examples of studies using social media data to determine citizen perception. The data mining method used in this thesis is explained in Chapter 3.5, the results are given in Chapter 4.3, and these are discussed in Chapter 5.3.

Chapter 2.5 shows how point data regarding citizen perception can be extrapolated to create overall citizen perception maps. How this method has been applied for this thesis research is explained in Chapter 3.6, the resulting maps are shown in Chapter 4.4, and the used method is discussed in Chapter 5.4. The method to validate these results is given in Chapter 3.7, of which the results are available in Chapter 4.5, and a discussion about these results in Chapter 5.5.

In Chapter 6 of this report an overall conclusion is given and the research questions from Chapter 1.4 are answered. Chapter 7 gives recommendations on how to improve the current research, and some ideas for further research.

2 RELATED WORK

The purpose of the case study in this research is to determine the citizen perception of nature, to assess cultural ecosystem services. Therefore, some previously performed researches that assess cultural ecosystem services are discussed in Chapter 2.1. For the data harvesting step it is important to know which social media platforms could be used for perception analysis. In Chapter 2.2 various social media platforms are discussed, to find suitable platforms. For the classification step a variety of classification methods are discussed in Chapter 2.3. To determine the data mining method, an overview of social media analyses is given in Chapter 2.4. Chapter 2.4.2 describes how some of these methods were already used for citizen perception analyses. In the mapping stage data has to be converted from point data to raster data, a method to do this is discussed in Chapter 2.5.

2.1 CULTURAL ECOSYSTEM SERVICES

In the Millennium Ecosystem Assessment (2005) is stated: "Everyone in the world depends on nature and ecosystem services to provide the conditions for a decent, healthy, and secure life." Following this assessment various researches have been performed to quantify these ecosystem services, of which generally used methods to assess cultural ecosystem services are explored below.

Daily et al. (2009) proposed a framework that considers a number of services simultaneously, and use this framework on a case study in Hawaii. In Figure 2.1 this framework is visualised, focussing on cultural ecosystem services in the bottom two ovals (Services and Values). They created the tool InVEST to perform integrated valuation of ecosystem services. In a part of their research they compared cultural ecosystem services to the economic value of ecosystem services. In economic valuation methods, the impact of changes in the ecosystem services is given in monetary terms (Daily et al., 2000). In some cases, the cultural value of areas can be used in addition to monetary values to increase the quality of evaluation of land-management decisions. They concluded that institutions should be developed to monitor the social values of ecosystem services.

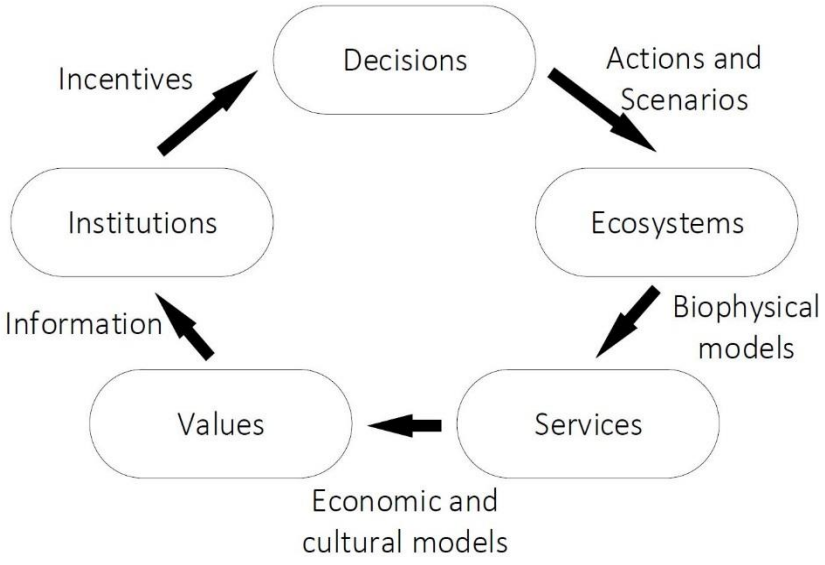

Figure 2.1: The framework proposed by Daily et al. (2009)

In their research Daniel et al. (2012) highlighted the importance of cultural ecosystem services as they are mostly left out of ecosystem services frameworks, which is mainly caused by their difficult integration. They provided examples of socioecological methods that could be adapted to improve the evaluation of cultural ecosystem services and focused on adapting methods proposed by Kumar (2012), who already indicated that his methods are limited in the fields of valuation and linking with other ecosystem services. Daniel et al. (2012) were able to improve the methods, however, they still indicated that a lot of progress needs to be made to fully integrate cultural ecosystem services into the broader framework.

Bagstad et al. (2016) performed a survey to assess the value of cultural ecosystem services in the PSI. This survey consisted of two parts: in the first part respondents answered questions about their opinion of the forest management in the area, in the second part they were asked to allocate a hypothetical 100 dollars to twelve value types and to mark locations in the PSI they found corresponding to these value types. Based on the collected survey data social value maps were generated using the Social Values for Ecosystem Services (SolVES) 2.0 tool. In this tool, the points marked by the respondents were related to six environmental data layers to determine the relationship between the value types and environmental conditions. This resulted in raster layers covering the whole PSI giving a value between 1 and 10 for each social value type, 1 indicating a low influence of the social value and 10 a high. For identifying hotspots and coldspots in these created raster layers the Getis-Ord Gi* statistic was used, to find local spatial autocorrelation between ecosystem services and social values. Regions of overlap between these hot- and coldspots related to the social values, and hot- and coldspots in biophysically modelled ecosystem services, could also be determined using this tool. Afterwards a regression analysis was performed to quantify the relationship between social values and ecosystem services.

2.2 Technical Overview of Social Media

Most social media platforms offer an API (Application Programming Interface) to developers, APIs allow a user to programmatically access software components, without accessing the component itself directly (De Souza et al., 2004). In principle, when a software component needs another component to work, changes made to the needed component can have influence on the collaboration between the two. However, by using an API for the needed component, instead of using it directly, changes to this component will not influence the collaboration (des Rivieres, 2004).

The two most popular social media platforms are Twitter and Facebook (Lago Vázquez, 2017). Twitter is a social media website on which users are able to share their opinions in real-time, and it is used by businesses as well to spread news all around the world (Palomino et al., 2016). Data from Twitter can be used to: spread information to a wide audience, track the real-time spreading of news or events, or, retrieve views from the general public about a certain topic. Twitter is a very suitable platform for research, since users share their ideas openly for the public to see (Palomino et al., 2016). From all academic articles using social media data, 60 per cent uses Twitter as the main deliverer of data (Lago Vázquez, 2017).

Lago Vázquez (2017) states that the second most used social media platform in academic research is Facebook (29 per cent). In December 2015 there were around 1.59 billion monthly active users on Facebook, 70 per cent of these users are active on a daily basis. Facebook is the most widely used social media platform at the moment, and its users are a good representation of the general population. As opposed to Twitter, data on Facebook is not openly available. On Facebook, users can choose to keep their messages private (Palomino et al., 2016). Facebook data can be harvested only when the user's profile is connected to the profile to harvest from. With the Public Feed API it is possible to harvest real-time posts from all Facebook users, however, this API is not freely accessible (Facebook, 2017).

A comparable platform to Facebook is Google+, this platform is owned by Google, and allows users to share information in the same way as on Facebook. A user's Google+ account is linked to its account for all Google products, like Gmail, Maps, Hangouts and YouTube (Miller, 2014). This combination of all Google products can be seen as the main reason for Google to enter the social media market, since it reveals personal information about a user (Landeweerd et al., 2013). When messages are uploaded publically, it is possible to retrieve them for free using the Google API (Google, 2017). The content of Google+ has been used in academic research, for example to determine the use of Google+ for sharing breaking news (Osborne and Dredze, 2014).

Facebook, Twitter, and Google+ are all examples of Social Networks, but there are more types of social media platforms, for example: Social News Aggregators. The difference between Social Networks and Social News Aggregators is mainly in the identity of users. On Social Networks the identity of a user is of great importance, since messages are shared to connections. Whereas on Social News Aggregators the identity of a user is of no importance for the content of the message or the reach (Glenski et al., 2017). One widely used Social News Aggregator is Reddit, this platform is open for anyone over the age of 13 (Reddit Inc., 2017). The content of Reddit is widespread, ranging from cute puppy pictures to breaking news, divided over numerous *subreddits*. Everything posted on Reddit is open for harvesting through the Reddit API (Reddit Inc., 2017). Reddit has been of interest in the academic world, for studies regarding the content of posted messages (Suran and Kilgo, 2017; Cole et al., 2017).

Another form of social media platforms are Image Publishing Sites, one of these platforms is Flickr. Flickr can be seen as a place to both store and share personal images. Images uploaded to Flickr can either be shared with personal connections of the user, or with the wide public (Angus and Thelwall, 2010). Flickr images can be retrieved from the web through the title, description, and tags given to it by the posting user (Angus and Thelwall, 2010). Since a large amount of the images on Flickr are geotagged, Flickr is commonly used in academic researches, using the locations of pictures as reference (Wood et al., 2013; Zheng et al., 2010).

A comparable platform to Flickr is Instagram, it has been available on the internet since October 2010, and can be called the most popular image capturing and sharing application. Compared to other platforms, such as Twitter, the use of Instagram data in academic researches is relatively sparse (Hu et al., 2014). It is possible for developers to use the Instagram API to develop their own applications (Instagram, 2017). The terms of service from the Instagram API are more restrictive

compared to other platforms. It is only possible to use the Instagram API for marketing and advertising. Other use cases, such as harvesting data or monitoring user activities are not allowed (BrightPlanet, 2017). Instagram has been used in academic research, for example to identify differences in social media usage between different cultures (Sheldon et al., 2017).

A very different kind of social media platform is Foursquare. This type of platform is called a Location-based Social Network. It is used by users to *check in* at locations, such as restaurants and museums, to write tips and share comments and photos about these locations (Li et al., 2018). All this information can be retrieved through the Foursquare API (Foursquare, 2016). Since all data on Foursquare is related to a location, researches have been performed using this geospatial information (Mueller et al., 2017; Arampatzis and Kalamatianos, 2017; Huguenin et al., 2017). These researches mainly use data from Foursquare to find relationships between locations and the type of people checking-in at these locations.

Another location-based social media platform is Strava, on this platform users upload information about cycling rides, walks, runs, and hikes across the world (Sun et al., 2017). All information uploaded to Strava can be accessed through the Strava Metro service. Information on Strava is only quantitative and not qualitative, since users are only able to upload routes and not give any information about their personal experiences (Strava, 2017). Examples of data from this service being used in researches are: to find relationships between environmental factors and cycling behaviour (Griffin and Jiao, 2015), to map cycling activities (Jestico et al., 2016), or to assess air pollution in cities (Sun et al., 2017).

2.3 DATABASE

To be able to store the collected social media data a unified geodatabase is required. In comparable researches various types of databases have been used. Chen et al. (2016) stored collected, georeferenced Tweets in an ArcSDE-based geodatabase, by creating a geo-tagged Tweet layer. This database was based on a Microsoft SQL Server database management system, with the geo-tagged Tweet layer as a feature class. This database was created for storing real-time Tweets and updates continuously as new Tweets were harvested. The Microsoft SQL Server database is a relational database. Relational databases store their information in tables, in which attributes can be related to attributes in other tables based on a common property (Varga et al., 2016). Another example of a relational geodatabase is the PostgreSQL database, with the PostGIS extension. This database was used by Gazaz et al. (2016) to store geo-tagged Tweets. They only stored a selection of attributes of the collected Tweets, removing all information that was not useful for their research. PostgreSQL is an open-source database, with many features, including spatial operations added by the PostGIS extension.

A different approach to storing social media data is to use a graph database. Graph databases store their attributes in the form of *nodes*, and declare relations between them as *edges*. A *node* is the main element in a graph database, information about this element is stored in its properties. To create sets of the same type of *nodes*, *labels* are used, which describe the type of *node*. It is possible to create directional relationships between *nodes* by adding *edges*, one *node* can have multiple

edges (Neo4j Inc., 2017). An example of a graph database is the Neo4j Platform, which was used by Palomino et al. (2016). The Neo4j Platform is open-source and it is possible to add a spatial extension to it. In Table 2.1 some of the differences between a relational database and a graph database are given.

Table 2.1: Differences between Relational Databases and Graph Databases (Neo4j Inc., 2017)

	Relational Databases	Graph Databases
Data Storage	Fixed storage; pre-defined tables with rows and columns	Graph storage structure with index-free *nodes*
Data Modelling	Model is translated from a logical model to a physical one	No mismatch between logical and physical model
Query Performance	Processing performance suffers with number of JOINs	Near zero latency and real-time performance
Query Language	SQL	Cypher

Since all data in graph databases is stored in n*odes* and relationships can be easily found through *edges*, queries on these relationships perform relatively fast. Relational databases first require a JOIN to be executed before relating properties can be matched, while in graph databases, in contrast, the relational information is stored in the properties of a *node* and can be directly accessed through the n*ode* itself. In studies using social media data, relationships between *nodes* are an important factor and queries should perform as fast as possible, especially when the volume of the data and the velocity of data collection increase.

2.4 Social Media Analysis

According to Samuel Fosso et al. (2016), there are eight different types of social media analytics, which are given in Table 2.2.

Table 2.2: Types of social media analytics (Samuel Fosso et al., 2016)

Type of Social Media Analytics	Purpose
Topic Modelling	Detecting themes
Opinion Mining	Extracting views, beliefs, and judgements
Sentiment Analysis	Extracting emotions (positive/negative)
Social Network Analysis	Analysing the network of relations
Trend Analysis	Predicting market trends
Popularity Prediction	Forecasting future demands of products
Customer Engagement	Determining the success of online activities
Visual Analytics	Visualising relationships

Some of these methods are mainly used in the marketing world: trend analysis, popularity prediction, and customer engagement, while the other methods are used in academic research as well (Goodchild and Glennon, 2010; Lago Vázquez, 2017; Palomino et al., 2016). Some examples of usage of these methods in academic research are briefly discussed below.

Topic Modelling

In the process of topic modelling a graph is created in which similarities between different documents are used to connect the corresponding information. By finding dense subgraphs in this constructed graph, topics can be extracted (Nie et al., 2013). This method of social media analysis can be used to, for example, assign topics to groups (Nie et al., 2013), or find messages related to a certain topic (Karami et al., 2018).

Opinion Mining

With the development of Web 2.0, the internet has become the place for people to share their opinion. It is interesting for governments and researchers to extract the general opinion of the public from social media. For this, opinion mining is used (Tian et al., 2018). Two approaches exist to detect the opinion in text messages: Natural Language Processing (NLP) and Semantic Web approaches (SW). Sobkowicz et al. (2012) proposed a method to combine the strength of both approaches. By creating a knowledgebase containing online opinions, they are able to process data using an NLP engine based on machine learning techniques. This engine detects which part of the data corresponds to an opinion, and on which topic.

Sentiment Analysis

A method close to opinion mining is sentiment analysis, but instead of finding an opinion in messages the emotion is extracted to determine the polarity (Sobkowicz et al., 2012). Sentiment analysis has been widely used to analyse social media data and a large variety of methods has been created. In most sentiment analysis methods, a lexicon is used, in which words and their corresponding sentiment are stored. All words in a sentence are compared to this lexicon to determine a sentiment for the complete sentence. The main difference between methods is in the lexicon used; e.g. some are based on machine learning, while others include emoticons in their lexicon (Gonçalves et al., 2013).

Social Network Analysis

Lago Vázquez (2017) compared a sample of articles about social media analysis published between 2010 and 2015 to analyse the methods used. In half of the studies in her sample, quantitative methods were used; using statistics to build their conclusions. She found that in most studies social media data is collected and analysed manually, since there is not yet a unified technique. One of the more recent developments in social sciences is SNA (Social Networks Analysis), which uses relationships between users to identify social structures. Joy (2010) described two tools (Netvizz and Gephi) which can be used to apply SNA to Twitter and Facebook.

Visual Analytics

Visual analytics of social media uses visualization techniques to find relationships between data (Wu et al., 2016). There is a wide variety of visual analytics methods, Chen et al. (2017) summarized them into six categories. Visual monitoring (1) gives a quick overview of the information, and is the basis for further identification of patterns. With feature extraction (2) a feature (any attribute

belonging to the data) is used as a characteristic to analyse the data. Event detection (3) can take place using four attributes; topic, time, people, and location, in order to identify groups of these attributes. Anomaly detection (4) depends on a normal pattern, outliers of this pattern can be used to find abnormal trajectories. Using historical data, it is possible to do a predictive analysis (5), as long as temporal patterns can be recognised. The last visual analytics method is situation awareness (6), which combines multiple of the previous methods to help the user in decision making. All these methods use a variety of visualization methods to solve complex social media data problems.

2.4.1 Classification of Messages

Data harvested from social media can be divided into many categories, of which not all are relevant for every research. Topic modelling can be used to create a selection of only the relevant categories. Various articles have been published about topic modelling or classification methods, either supervised, semi-supervised, or unsupervised. The first two being able to categorize text data into given categories, thus being relevant for this research, while unsupervised classification is mainly used for clustering knowledge (Xu et al., 2017).

One of the most widely used methods for text classification is the Naive Bayes classifier. Jiang et al. (2013) demonstrated and compared the performance of various forms of this supervised classifier. The standard Naive Bayes classifier is based on the assumption that all words in a document are independent of each other. Naive Bayes classification makes use of the multivariate Bernoulli model, in which the probability of each word contained in a document to belong to a certain class is multiplied to obtain the overall probability of the document belonging to that class. This method does not take into account the number of times a word appears in the document, to overcome this the Multinomial model (MNB) was proposed by Jiang et al. (2013). This model uses the same approach, but does use the frequency of a word as parameter. One shortcoming of this model is the fact that the number of training documents for each class influences the probabilities; a class with few training data automatically has lighter weights assigned. Another form of Naive Bayes is Complement Naive Bayes (CNB), which uses complement classes while classifying documents to balance for the amount of training documents. Combining MNB with CNB forms the one-versus-all-but-one (OVA) classifier, this ensemble classifier is based on the assumption that each document may belong to multiple classes (Rennie et al., 2003).

Classification accuracy can be further improved by using, next to labelled data (supervised classification), a large amount of unlabelled data (semi-supervised classification). Nigam et al. (2000) proposed an algorithm based on the combination of Expectation Maximization (EM) and a Naive Bayes classifier. First a standard Naive Bayes model is created with the labelled training data, using this classifier the probabilities of the document belonging to each class are determined. Next, a new Naive Bayes model is made using the labelled training data and the weighted class labels of the unlabelled documents. These last two steps are iterated until a stable model is created. This method significantly improves classification, compared to single Naive Bayes, when a small amount of labelled data is available.

Xu et al. (2017) introduced a semi-supervised framework using Convolutional Neural Networks (CNNs). This method integrates embedded small text regions from unlabelled data into a supervised CNN. First, they determined relevant categories and created concepts around these category names. Next, based on the similarity between these concepts and a selected set of unlabelled documents, the documents were labelled into these categories. From these newly labelled documents only the documents with the highest probability values were selected as training data. Subsequently two classifiers were built using two semi-supervised methods: Transductive Support Vector Machine (TSVM), and SVM based on Deterministic Annealing (DA). TSVM classifies data by building a hyperplane classifier, resulting in the margin between two classes to be maximal. SVM based on DA uses the same principle, however it is able to overcome local minimum issues present in TSVM (Sindhwani et al., 2006). SVM methods are only able to analyse data in numbers, therefore each document should be converted to a vector of numbers (Tripathy et al., 2016).

2.4.2 Social Media Analysis on Citizen Perception

As shown in the previous section there are many methods to analyse social media data, some of these have already been used to analyse citizen perception. One of the fields in which social media data is already used, is disaster management. By making use of real-time information the impact of a disaster on the citizens can be determined and management can be adjusted to this information. Chen et al. (2016) proposed a system using real-time harvested Twitter data to support mass evacuation and resource allocation. The novelty of their research was creating the possibility to store harvested data, to make it possible for different applications to access the data. However, the data they collect is not analysed automatically, the system is able to visualise the data, but analysing is still performed manually.

Woo et al. (2015), on the other hand, used Twitter data to determine how the public mood changed after a human-made disaster, in their case the 2014 Sewol ferry disaster in South Korea. By using natural-language processing and text-mining technologies, they were able to investigate the emotional reactions of citizens to the disaster. For this they used certain keywords related to disasters, combined with a topic based sentiment analysis. The authors proof that social media data can be used as an information source regarding public health and to monitor the public's emotions.

Understanding the public opinion on social media regarding a certain topic, is a very useful approach to characterise issues. Karami et al. (2018) analysed the characteristics of the public's opinion on diabetes, diet, exercise, and obesity. They collected data from Twitter and were able to discover topics within these data using a modelling approach to fuzzy cluster semantically related words. Afterwards they revealed thoughts, feelings, personality, and motivations regarding each topic from the collected texts. They found that Twitter data can be used to analyse public health issues, since a correlation is found between their results and known census data. However, since they do not use the geographical location of the messages, they are not able to find spatial relationships between these public health issues. They mention this as a good way to improve their analysis.

Not only the content of social media messages can be useful for research, but, for example, the locations and amount of data can be used as well. Wood et al. (2013) used geotagged photographs from Flickr to estimate visitation rates at recreational sites. They found a correlation between the amount of geotagged Flickr photographs and the amount of visitors to the area, making Flickr a good source to estimate visitation rates. An extra advantage they show of using Flickr data is the ability to determine the origin of visitors. Since the country of origin was available of all Flickr users posting photographs. A positive correlation was found within this data as well.

2.5 Citizen Perception Maps

The Social Values for Ecosystem Services (SolVES) tool is developed to incorporate spatial social value measures into ecosystem service assessments (Sherrouse and Semmens, 2015). It is able to quantify and map social ecosystem services across a study area, by relating available point data to characteristics of the underlying physical environment. The tool quantifies the social value of social ecosystem services on a 10-point value index, assigning values between 1 and 10 to the study area. These values are created using inserted point data, concerning the citizen perception of the area. In principle this point data is collected through surveys, but data collected in a different manner could be used as well. The locations and values of the points are then correlated to the underlying environment, such as land cover or the average distance to water, to create one output raster describing the social value of the area. This process is visualised in Figure 2.2.

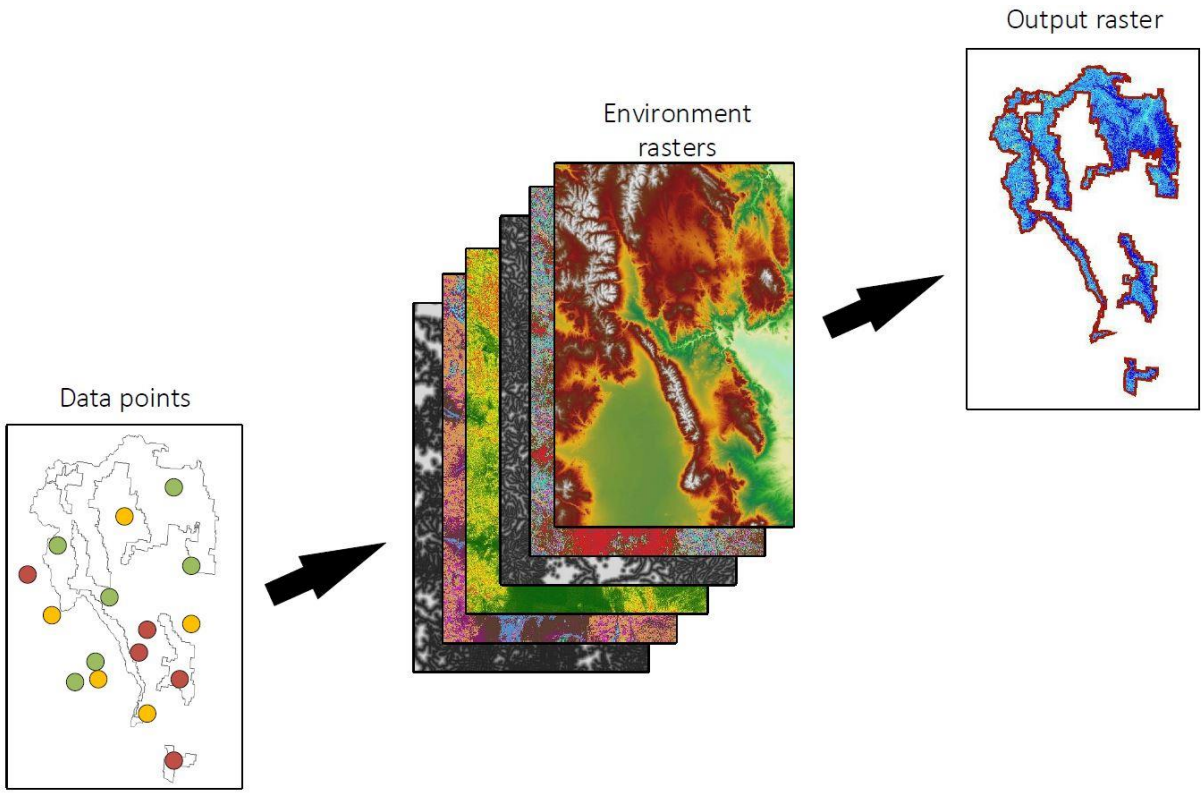

Figure 2.2: A simplified view of the proces performed by the SolVES tool

Since the tool is integrated with the MaxEnt maximum entropy software, it is able to offer statistical models describing the relationship between the environmental factors and the assigned values (Sherrouse and Semmens, 2015). The MaxEnt maximum entropy modelling software was originally developed to model the geographic distribution of species, by finding the probability distribution which has maximum entropy based on environmental factors (Phillips et al., 2017). The maximum entropy distribution was first mentioned by Jaynes (1957), according to him the most important property of the maximum entropy distribution is that mathematically no possibility is ignored. Or in other words: it assigns a positive probability to every situation, unless it is absolutely excluded by the given information.

The expression for entropy, as found in statistical mechanics, is given in Equation (2.1):

$$H(p_1 \cdots p_n) = -K \sum_i p_i \ln p_i , \qquad (2.1)$$

in which K is a positive constant, set equal to 1. This expression needs to be maximised, in order to find the probability distribution with maximum entropy. One property of a probability distribution is the fact that each probability is between 0 and 1 and the sum of all probabilities is equal to 1:

$$\sum p_i = 1. \qquad (2.2)$$

This is the first constraint used in maximum entropy, the second constraint is based on the fact that in principle entropy has its maximum value when all probabilities are equal. However, this result cannot be reached when additional information is available, since the result should meet this information as well. Therefore, it is assumed that the expected value for a quantity is known ($\tilde{F}$), resulting in Equation (2.3) (Massachusetts Institute of Technology, 2016):

$$\check{F} = \sum_{i=1}^{n} p_i f(x_i). \qquad (2.3)$$

By introducing Lagrangian multipliers λ and μ, which indicate the functions are proportional to each other, Equation (2.1) can be maximised to the constrains of Equations (2.2) and (2.3) giving:

$$p_i = e^{-\lambda - \mu f(x_i)}. \qquad (2.4)$$

The constants λ and μ can be determined by substituting into Equations (2.2) and (2.3):

$$\check{F} = -\frac{\delta}{\delta \mu} \ln Z(\mu), \qquad (2.5)$$

$$\lambda = \ln Z(\mu). \qquad (2.6)$$

In Equations (2.5) and (2.6) the partition function ($Z(\mu)$) is:

$$Z(\mu) = \sum_i e^{-\mu f(x_i)}. \quad (2.7)$$

To generalise these equations to any number of known quantities ($\check{F}$), giving the averages:

$$\check{F}_r = \sum_i p_i f_r(x_i), \quad (2.8)$$

the partition function from Equation (2.7) will become:

$$Z(\lambda_1, \cdots, \lambda_m) = \sum_i exp\{-[\lambda_1 f_1(x_1) + \cdots + \lambda_m f_m(x_m)]\}. \quad (2.9)$$

Using this equation for the partition function, the maximum entropy probability distribution becomes:

$$p_i = exp\{-[\lambda_0 + \lambda_1 f_1(x_1) + \cdots + \lambda_m f_m(x_m)]\}. \quad (2.10)$$

Therefore, the entropy of this distribution will be (Jaynes, 1957):

$$S_{max} = \lambda_0 + \lambda_1 \check{F}_1 + \cdots + \lambda_m \check{F}_m. \quad (2.11)$$

The SolVES model uses maximum entropy to find the best possible probability distribution, while satisfying constraints represented by the environmental variables. In the MaxEnt output of the SolVES model the assigned probability values (between 0 and 1) represent the relative intensity assigned by survey respondents to a chosen social value type. Combining these results with the kernel density method used by SolVES itself, more complete maps of the study area can be retrieved (Sherrouse and Semmens, 2015).

The process inside the SolVES model can be described by the following steps (Sherrouse and Semmens, 2015):

1. Using the inserted points and values assigned to these points, the model calculates a kernel density surface and average nearest neighbour statistics
2. The model identifies the most highly valued location, by comparing the kernel density surface, and assigns its value to the maximum value parameter
3. Using this maximum value, the model normalizes the kernel density surface to produce a kernel density-based value-index surface
4. By making use of the provided environmental data layers, MaxEnt produces a map output based on the relations between these layers and the provided points. It also generates a statistical model describing the relationship
5. Using all previous results, a final social value map is created

3 METHODOLOGY

The following chapter describes the methods used during this research for collecting, storing, filtering, analysing and visualising the social media data. In Figure 3.1 an overview is shown of the steps performed and the corresponding sections of this chapter.

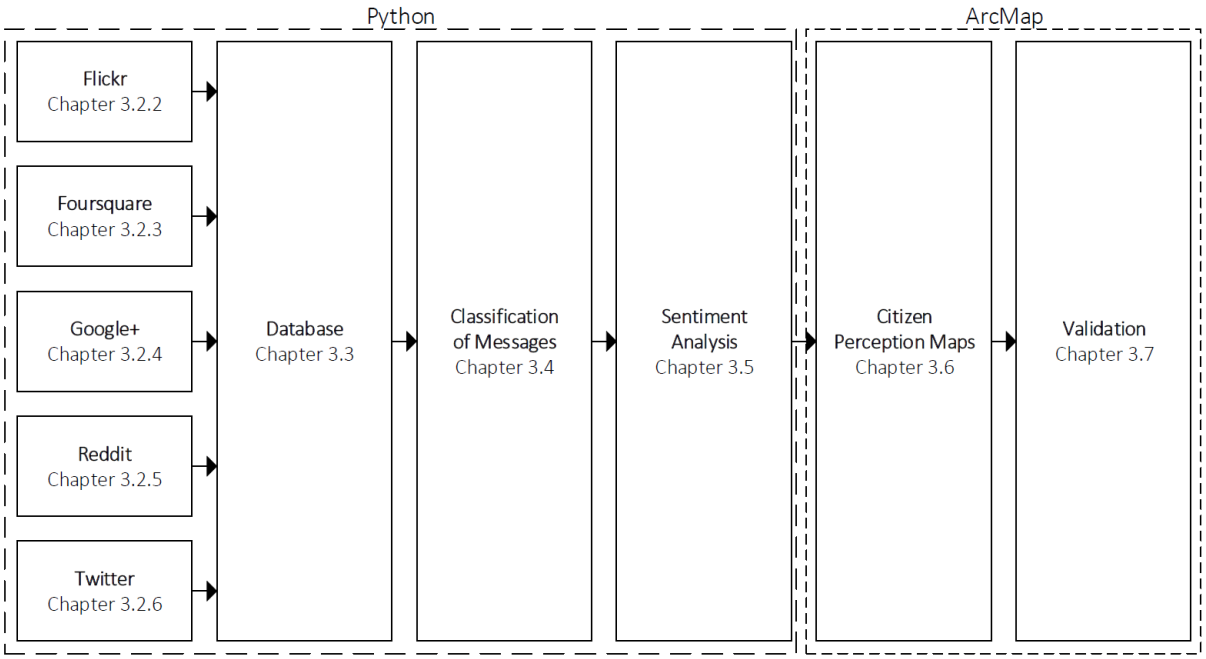

Figure 3.1: Flowchart of the methodology

The first four steps in this research are performed in the Python language. This language is chosen as the APIs used for social media data collection are readily available as Python libraries. The final steps of creating the citizen perception maps and validating the results are performed in ArcMap, since it offers the required tools.

The first research question has been answered in Chapter 3.1, indicating which social media platforms are suitable. In Chapter 3.2 it is elaborated for these social media platforms how and which types of data can be collected. For each platform a large amount of data will be collected, which requires storage in a geo-database. In Chapter 3.3 various databases are compared to find the best suitable database to store the collected data, answering research question 2.

Since a lot of data will be collected and not all data is relevant for this research, categorisation of the data is necessary. The method to filter the collected data is described in Chapter 3.4, this section contains three chapters, each describing the categorization of a different part of the data.

The methods described in Chapters 3.5 and 3.6 can be used to answer research question 3. Chapter 3.5 describes the analysis performed on the data. The results of this analysis are transformed into citizen perception maps, which is described in Chapter 3.6.

The final research question can be answered by the methods described in Chapter 3.7: the step of creating hot- and coldspot maps and validating these results by comparing them to the maps created by Bagstad et al. (2016).

3.1 Chosen Methods

The conclusions from the researches described in this section were used to propose the best suitable methods to fit the framework from Chapter 1.6. An adapted version of this framework is shown in Figure 3.2, all chosen methods are visualised.

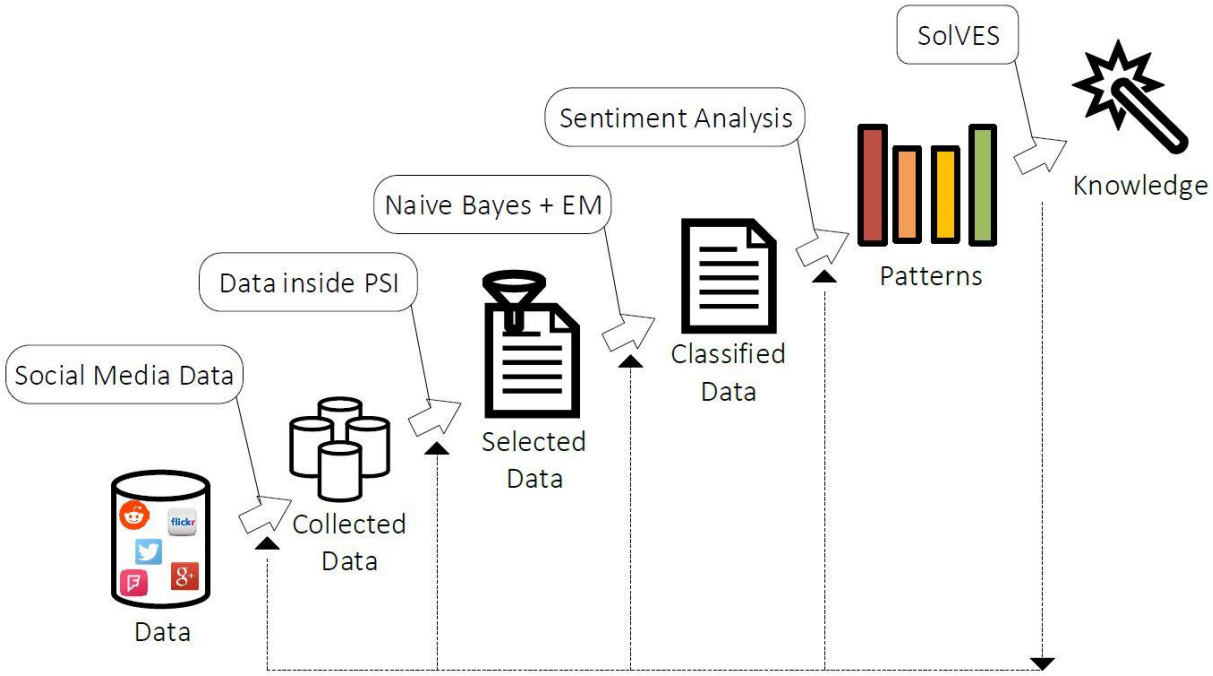

Figure 3.2: Framework of chosen methods

For the harvesting of data a selection was made of suitable social media platforms. This selection was based on whether both the content and the availability of the data are sufficient (Table 3.1). In the case of Strava, the content of the data is not sufficient for this research, as the data is not qualitative and does not provide information about the citizen perception. Data from Facebook and Instagram could be relevant for this research, however, this content is not openly available and therefore not usable. All the collected data is stored in a Neo4j database, as relevant queries perform better in graph databases and performance will not deteriorate when the data size increases.

Table 3.1: Suitability of analysed platforms.

Social Media Platform	Content	Availability	Suitable
Twitter	Useful	Available	Yes
Facebook	Useful	Unavailable	No
Google+	Useful	Available	Yes
Reddit	Useful	Available	Yes
Flickr	Useful	Available	Yes
Instagram	Useful	Unavailable	No
Foursquare	Useful	Available	Yes
Strava	Not useful	Available	No

In the data selection step three sets were created containing a certain part of the data based on the available metadata: (1) a set with all messages which are posted inside the PSI bounding box, (2) a set with all messages mentioning certain highlights in the PSI, and (3) all messages not belonging to either of the two. Messages belonging to set 1 could also belong to set 2. The first set was used for further processing. By classifying the messages, only relevant messages from this set remained. For classification it was chosen to use the same method as proposed by Nigam et al. (2000): the Naive Bayes method combined with Expectation Maximization. This method was chosen over single Naive Bayes, since a large amount of unclassified data was available and EM has proven to increase the classification result in this case.

The next step to analyse the data is data mining, in this step information is retrieved from the dataset. It is chosen to perform a sentiment analysis on the data, since these results are most comparable to the results from the survey used by Bagstad et al. (2016) and will give a good overview of how visitors experience the park.

In the final step of creating social value maps the tool used by Bagstad et al. (2016) was used: SolVES. Followed by a hotspot analysis using the Getis-Ord Gi* statistic, to create hot- and coldspot maps for validation. Using these tools makes the comparison of the results to Bagstad et al. (2016) more reliable, since only the source of the input data is different.

3.2 Data Collection

When using the APIs of all chosen social media platforms, the created application should be authenticated in order to gain access to the social media data, the method to perform this is elaborated in Chapter 3.2.1. In Chapters 3.2.2 through 3.2.6 it is described how data has been collected for each platform in alphabetic order. The syntax for all described methods is given in Appendix B: Code Snippets.

3.2.1 OAuth

For making use of the different APIs an account is needed, with a corresponding username and password. The simplest way to authenticate an application would be by using this username and password. However, when a user agrees to share these credentials, he automatically gives another person full access to his account. To overcome this security problem OAuth has been developed (OAuth, 2017). OAuth allows sites to give users access to their content without sharing their private credentials. Instead of asking the user for their credentials and using them to log into the preferred website, OAuth asks the user to give permission to the application by asking them to log in to the website themselves and authorizing the application. In that way the user can specify which content the application will be able to access and which should remain private (Bihis, 2015). In Figure 3.3 an overview is given of this difference. All platforms used in this research require authentication through OAuth.

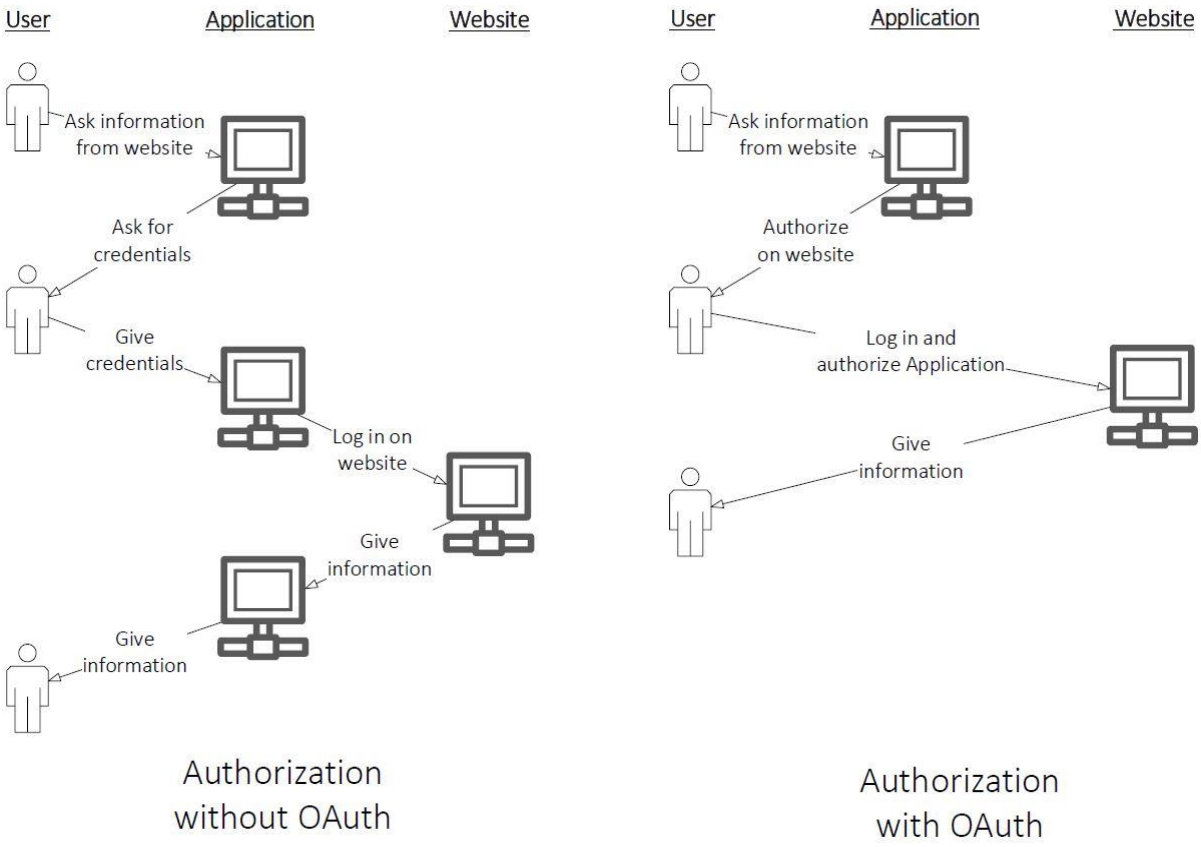

Figure 3.3: Overview of process OAuth (adapted from Bihis (2015))

3.2.2 Flickr

In 2013 more than 6 million Flickr users had shared over 6 billion photographs on the website. From these images around 197 million were assigned coordinates (Wood et al., 2013). Since 2013 this number has been increasing with up to 2 million photos per day (Figure 3.4).

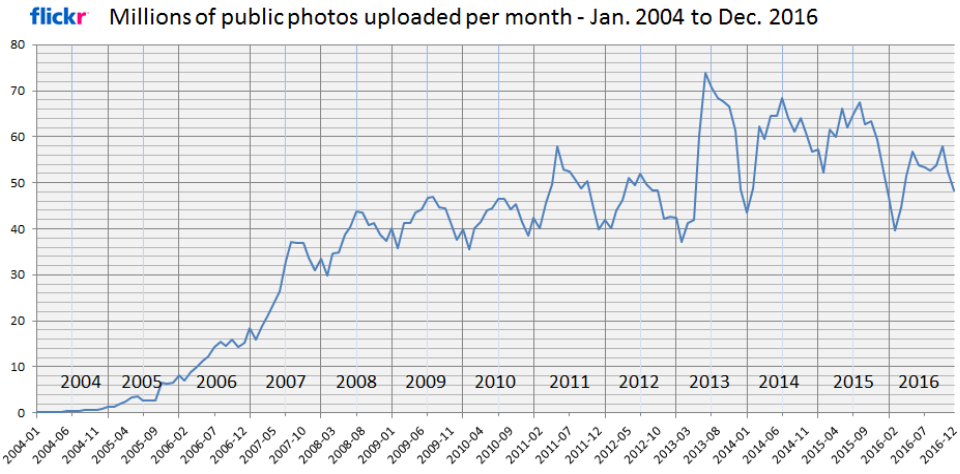

Figure 3.4: Graph showing number of photos uploaded to Flickr since 2004 (Michel, 2017)

The developers of Flickr have made it possible for people to freely experiment with their data. This data and all functionalities that run Flickr can be accessed using the Flickr API (Flickr, 2016). The Flickr API has a large number of search methods to find all sorts of data and metadata, in other words: data can be harvested using the Flickr API.

With the large number of geotagged photographs available finding relevant photos by location is the first logical method to apply. The method used for searching photos based on location is *flickr.photo.search*. This method returns a list of photos based on the entered parameters. These parameters are: (1) *bbox*: a bounding box in which the API will search for photos, (2) *per_page*: the number of photos to return per page, with a maximum of 500, (3) *extras*: extra information to be retrieved with the photos, and (4) *max_upload_date*: only photos uploaded before this data are returned, this variable can be used to search for more than 500 photos.

Besides searching by location another interesting search method for finding data on Flickr is by using search queries. This search uses one different parameter than the previous search. Instead of searching using a bounding box the *text* parameter is used, which is a search term to find relevant photos. The used queries can be found in Table A, Appendix A: Lists of Used Queries.

The third and final search method used is a follow-up of the previous two. The previous methods are able to collect most of the relevant data. However, sometimes photos about the park are not found by either of the previously described methods. Therefore, during collection of relevant photos, a list is created containing the user id of all users uploading those photos. Since these users once uploaded a picture inside or about the park it can be expected that they have done it more often. This will increase the chance of finding more relevant photos. For finding photos from a user the method *flickr.people.getPhotos* can be used, which uses the *id* of a user to find al photos posted by this user.

The Flickr API, however, has some limitations to prevent abuse. The main limitation is the rate limit. The maximum number of requests per hour is 3600 for each key, which is a unique identifier of each API user. There is no method available to check whether this rate limit is exceeded. Therefore a function is written to automatically save the program from exceeding this limit.

3.2.3 Foursquare

Since the development of Foursquare, over 10 billion check-ins have taken place. These check-ins are made by the more than 50 million visitors per month, together being responsible for around 9 million check-ins per day (Foursquare, 2017). The Foursquare API makes it possible for users to harvest information about venues from Foursquare. Resources can be accessed through a unique URL for each venue, aspects from these venues can be accessed by adding them to the URL (Foursquare, 2016).

Instead of using these URLs directly, endpoints are developed to access resources more easily, which call the URLs through previously defined functions. The most interesting endpoint for this resource is the venue search. The *Search Venues* method returns venues near the inserted location, depending on the *intent* used. There are four parameters involved in this search: (1) *ll*: the latitude and longitude around which to search, (2) *radius*: search radius around the given coordinate, (3) *limit*: the number of venues to return per search, and (4) *intent*: selected search method, *browse* returns all venues within in the given area.

By performing some tests with the radius parameter, the radius which returned the most venues was 5000 meter. Since this will not cover the whole park in once, the method is used multiple times in a row, each time increasing the latitude and longitude. This way overlapping search circles, like in Figure 3.5, are created, covering the whole area.

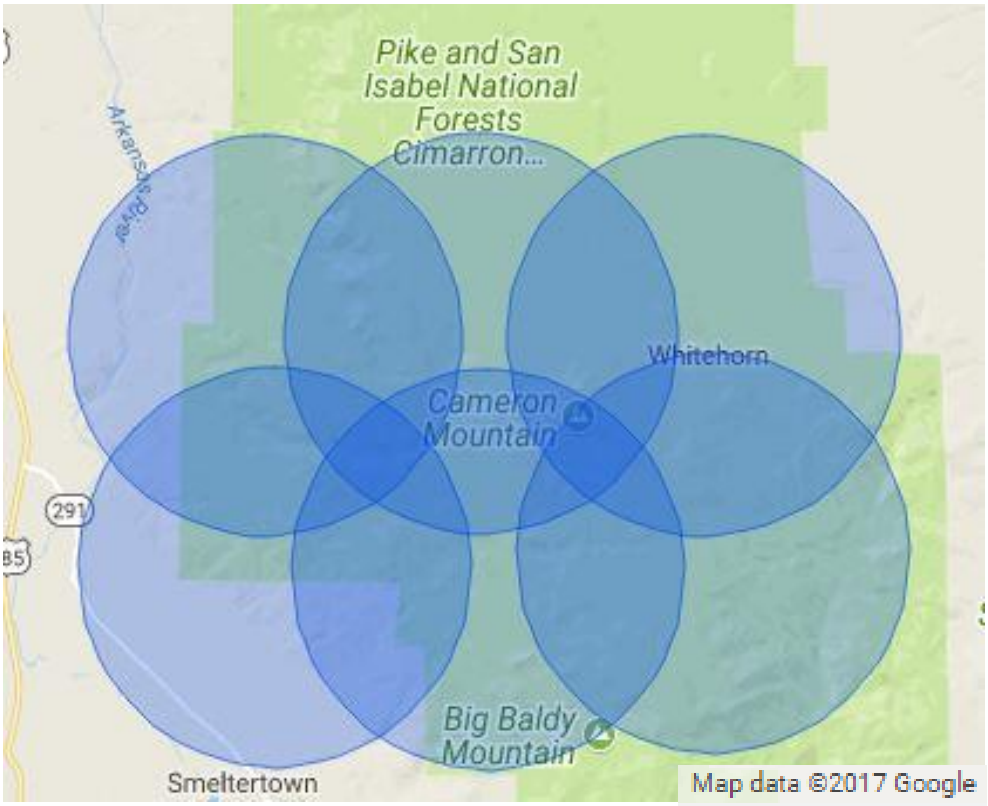

Figure 3.5: Example of circles used for searching

The *Search Venues* method only returns the users name and id, while more information about the users is preferred. For this the method *Find users* can be used. This method will find all information available about a user, which can be found using its name. Besides users checking in to the venue, relevant users are: users giving tips about the venue and users uploading photos from the venue.

The Foursquare API also has some limitations set for its use, in the form of rate limits. These rate limits are set for each top-level endpoint group separately. This means that once a user has exceeded the rate limit for one endpoint group, it is still possible to continue sending requests to another endpoint group. For the *venues* endpoint group a user is allowed to make up to 5000 requests per hour, for all other groups this is only 500. The Foursquare API has a variable indicating the current remaining rate (*X-RateLimit-Remaining*), which can be used to prevent an application from exceeding the rate limit (Foursquare, 2016).

3.2.4 Google+

Google makes it possible for users to create applications using their API, for this research the API was used for harvesting data from Google+. Google+ is a social networking site, enabling users to connect to other users worldwide. Since the development of the platform in 2011 the platform has had usage peaks of up to 540 million active users monthly (McGee, 2013). Google+ has been trying to compete with the social media platform Facebook, but it has not been able to reach the same

level of usage and importance (Miller, 2014). Google Sign-In is used to allow a user to connect to an application, this service makes it possible for users to use their own Google account to log into an application (Google, 2017). The user is asked to refresh its credentials every hour.

Users post activities on their Google+ feed, these activities can be retrieved through the Google+ API. Google+ does not assign locations to these activities, therefore it is not possible to search based on a location. The only possible relevant search is a search based on queries, returning all activities containing these search words. Activities can be retrieved by HTTP requests, which will return the result in a JSON data format. To perform these requests in Python the *google-api-python-client* is used, which converts the HTTP requests into Python functions (Google, 2017).

The first step to retrieve activities is to create an activity resource, making communication possible with the API. One of the methods combined to this activity resource is the search method: *activity_resource.search*. This method requires three input parameters: (1) *query*: a query to find relevant activities for, (2) *maxResults*: the number of activities to return per request, the maximum allowed value is 20, and (3) *pageToken*: the page to return the activities from, this token can be used to move through pages each containing twenty activities.

The search method only returns the id of the user posting the activity, to retrieve more information about this user the *people_resource.get* method can be used. This function finds all available information based on the given user *id*.

The number of API calls the application is allowed to make is restricted by a usage courtesy quota, which is set to 10,000 for the used application. Once this quota is exceeded an exception is thrown, which will set the program to sleep for a random amount of time.

3.2.5 Reddit

The Reddit API can be accessed by anyone who registers an application, which is free of charge as long as the application is not used commercially. For using the Reddit API with Python, the Python Reddit API Wrapper (PRAW) can be used. Three types of applications are supported by PRAW, the relevant one for this research is the Script Application. To use a Script Application, only four parameters are needed: *client_id*, *client_secret*, *password*, and *username*. The *client_id* and *client_secret* are obtained by registering the application, while the *password* and *username* correspond to the ones used to register the application (Reddit Inc., 2017; Boe, 2017).

With the endpoints delivered by the Reddit API it is not possible to search the whole of Reddit at once, instead it is only possible to search in subreddits. Therefore, at first all relevant subreddits are selected using the endpoint *Reddit.subreddits.search*, this method returns a list of subreddits, which can be used to search in. Parameters used in this method are: (1) *query*: search term to find subreddits for, the used list of queries is given in Table B, Appendix A: Lists of Used Queries, (2) *after*: the name of the last found subreddit, this parameter determines where to start searching, (3) *limit*: the number of subreddits to return per search, the maximum allowed value is 100, and (4) *show*: indicates what should be returned, when set to *all* every found subreddit will be returned, independent of internal filters.

Once all relevant subreddits are found, it is possible to search for posts inside these subreddits. The query list in Table A in Appendix A: Lists of Used Queries, is used to find posts, by executing the *Reddit.[subreddit_name].search* method. The parameters in this method are: (1) *query*: corresponds to the queries in the given query list, and (2) *sort*: determines the order in which posts are returned, when set to *comments* they are ordered based on the number of comments. These comments could also be useful for this research, therefore they are harvested as well, using the *Reddit.[submission].comments* method.

The rate limit set by the Reddit API is thirty requests per minute. To determine whether this rate limit is exceeded, the *Reddit.auth.limits* variable can be used. This variable exists of a dictionary containing three parameters of which two can be used to make sure this rate limit is not exceeded. The *remaining* parameter stores the number of request that are remaining for the current credentials, this parameter will reset every minute. The r*eset_timestamp* gives a Unix timestamp for the moment the rate limit will be reset by the API (Boe, 2017).

3.2.6 Twitter

The Twitter API platform makes it possible for users to programmatically harvest Tweets from their website. Three options are available to collect historical Tweets: The Search API, The 30-Day Search API, and the Full-Archive Search API. As the names of the last two imply the 30-Day Search API is able to return Tweets posted up to 30 days ago, while the Full-Archive Search API will return all historical Tweets. These two options are only available in the Enterprise edition of the API and therefore only accessible for premium operators. The Search API is freely available for all developers and will return Tweets posted up to 7 days ago (Twitter Inc., 2017).

The method used to collect all Tweets is *GET search/tweets*. The parameters used in this search are: (1) *geocode*: a latitude, longitude, and radius, which determine the search area, (2) *count*: the number of Tweets to return per request, the maximum allowed value is 100, and (3) *max_id*: Tweets posted before the Tweet with this *id* will be returned, this parameter can be used to return more than 100 Tweets. Since the maximum radius that can be used by the method is only 5 kilometres, the same strategy as with the Foursquare search (Figure 3.5) has been applied.

The *GET search/tweets* method could also be used to find Tweets using queries, instead of locations. The parameter for this search is *q*, which represents a search query of up to 500 characters to find relevant Tweets. For this research the previously defined query list in Appendix A: Lists of Used Queries (Table A) is used.

Unfortunately, the previous two methods only return Tweets posted up to 7 days before harvesting, however there is a free method that is able to harvest Tweets from longer ago. It is possible to retrieve the timeline of a selected user, by using method *GET statuses/user_timeline*. This method is able to harvest up to 3,200 Tweets from a user's timeline, regardless of the time they were posted. During collection of Tweets based either on location or a query, a list is created of all users posting these Tweets, which is used to retrieve all their timelines. A lot of these Tweets will be irrelevant for this research, but it can be assumed that someone that once posted something inside

or about the PSI will do this again or has done this before. The parameters used for this search are comparable to the previous parameters, but instead of a query or location a *user_id* is requested.

A special feature about Twitter, compared to the other platforms, is that it is possible for Google+ and Foursquare users to add their Twitter account to their Google+ or Foursquare account. This makes it possible to retrieve a larger set of users to retrieve timelines from.

The Twitter API is bound to a rate limit, which is defined per method. The *GET search/tweets* method has a rate limit of 180 requests per 15 minutes, while the *GET statuses/user_timeline* method has a limit of up to 900 requests per 15 minutes. The remaining rates at a certain moment can be retrieved from the Twitter API, since it is an attribute of the API instance. A method is built into the application to check these values frequently and to pause the program when the rate limits are almost reached (Twitter Inc., 2017).

3.3 DATABASE

The Cypher query language is developed especially for the Neo4j Platform, this language is intuitive and human-readable. When you have knowledge about SQL it is easy to learn the Cypher language as it is inspired by SQL. The spatial extension for Neo4j has its own version of queries, to, for example, create spatial n*odes*, import a shapefile, or select all n*odes* inside a defined bounding box (Neo4j Inc., 2017).

Each social media message was stored as a n*ode*, with the platform as l*abel*. All properties of the message were stored in these *nodes*, while separate *nodes* have been created for the users, storing the properties of the user. For the geocoded messages, a spatial *node* is created storing the coordinates of the message. During creation of the *nodes* they were immediately connected by *edges*. The structure of the graph database makes it easy to display relationships between different messages; for example, when a message is posted by a user that is already in the database, the new *node* will be linked to this user, instead of creating a new one. The same goes for messages posted on the same location.

In Figure 3.6 an example is given of some Tweets stored in the Neo4j database. The given Tweets are connected by their coordinate *node*. Each Tweet has a user connected to it, this is the user that posted the Tweet. When one user has posted more Tweets, the user's *node* has a relationship with all corresponding Tweet *nodes*. Some other platforms (e.g. Reddit, Flickr) can have more complex relationships, since on these platforms users can perform different actions, like commenting on a post.

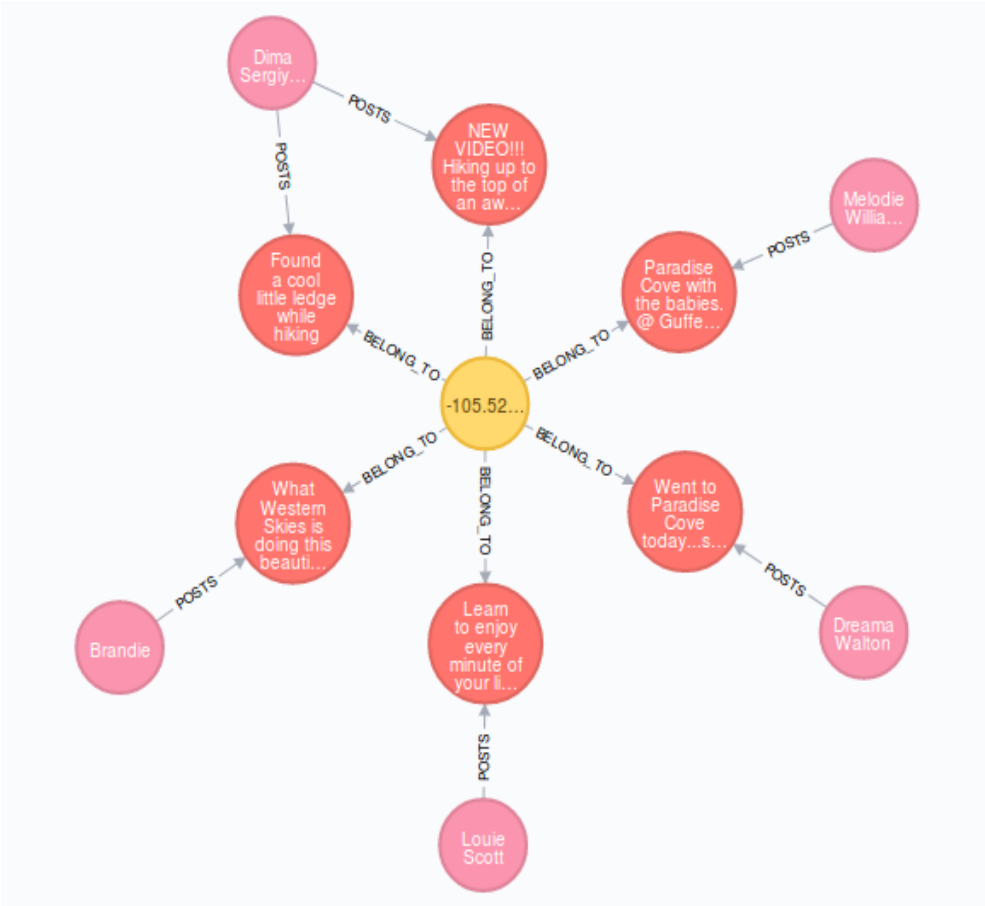

Figure 3.6: Example of Tweets stored in Neo4j database

3.4 Classification of Messages

Not all found social media data are relevant for this research, therefore it is classified as either relevant or not. At first the retrieved data can be divided into three categories, based on their metadata. These categories are:

1. Data inside a bounding box around the PSI
2. Data containing words out of a certain set of search words
3. All remaining data

The categories are visualised in Figure 3.7, it can be seen that category 1 and 2 overlap, as there is a part of the data that is inside the bounding box and containing a search word.

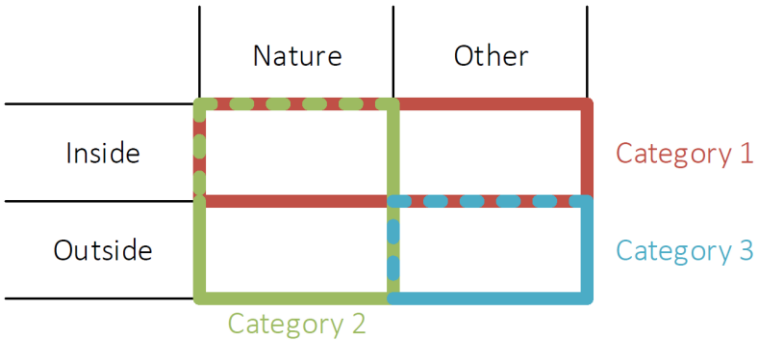

Figure 3.7: Division of categories

Each data set required a different method to analyse whether the data is relevant for this research. Data that can be used is data containing a relevant location and the content of the message should be about the PSI or nature.

The data in the first category, inside the bounding box, satisfied the first constraint: it has a relevant location. This dataset had to be filtered in order to remove all messages having a content other than about the PSI or nature. In Chapter 3.4.1 is explained how this data is retrieved from the database and how the dataset is filtered to only select the relevant messages. The data in the second category satisfied the second constraint: is has a relevant content. A part of this dataset has already been processed in the classification of category 1. The remaining part of this dataset, being relevant but not containing a location, can be processed to find the corresponding location. In Chapter 3.4.2 it is discussed how data from this category is retrieved from the database and further processing steps are discussed. Chapter 3.4.3 describes the third category, which contains all data not being assigned to one of the other two classes.4

3.4.1 Data inside bounding box

When the found social media data contained a location, this location has been stored as a *Coordinates node* in the database. Using the spatial extension of Neo4j it is possible to select data based on these coordinates. Within the *spatial.bbox* function the user can define a bounding box, the function will return all *Coordinates nodes* inside this bounding box. From these *nodes* related *nodes* can be found, in this case these *nodes* can be social media nodes from Twitter, Flickr, or Foursquare. In Code snippet 3.1 the Cypher query is shown which will retrieve all data from Twitter inside the bounding box around the PSI.

```
CALL spatial.bbox('geom',{lon:-106.6965,lat:37.3341},
{lon:-104.9881, lat:39.6353})
YIELD node
MATCH (node)-[:BELONGS_TO]-(tw)
WHERE 'Tweet' IN LABELS(tw)
RETURN tw;
```

Code snippet 3.1: Cypher query to find all Twitter nodes inside the bounding box

The collected data was assumed to all be posted inside or close by the PSI. As not all data placed inside or close by the park is relevant for this research the collected data set has been filtered. For this the machine learning Naive Bayes classification is used, combined with Expectation Maximization to increase accuracy. Naive Bayes is able to divide texts between different classes, in this case two classes are used: one containing all messages regarding nature and/or the PSI, and one containing all non-relevant messages.

Naive Bayes classification used a training set of data to create a Bag Of Words for each class containing all words present in the training data and their occurrence count. Using these Bags Of Words the unlabelled data could be classified, calculating the chance of the text belonging to each class. The Naive Bayes classifier is trained by estimating the probability parameters (θ) of the generative model, written as $\hat{\theta}$. The most probable value of θ can be calculated using the maximum a posteriori estimate: $\arg max_{\theta}\, P(\theta|X,Y)$. From all labelled training data it is assumed that they

were created using one mixture component for each class. The probability of a word w_t occurring in document x_i belonging to class y_i can be given by equation (3.1):

$$\hat{\theta}_{w_t|c_j} \equiv P(w_t|c_j;\hat{\theta}) = \frac{1 + \sum_{x_i \in X} \delta_{ij} x_{it}}{|X| + \sum_{s=1}^{|X|} \sum_{x_i \in X} \delta_{ij} x_{is}}, \quad (3.1)$$

in which δ_{ij} is 1, when $y_i = c_j$ and 0 when another class is selected. $|X|$ corresponds to the total vocabulary size.

The overall class probability of class c_j can be estimated using equation (3.2):

$$\hat{\theta}_{c_j} \equiv P(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|X|} \delta_{ij}}{M + |X|}, \quad (3.2)$$

where M corresponds to the total number of classes.

Using labelled training documents it is possible to determine the mixture components that would have created the documents, by turning the generative model backwards. This will result in an equation to calculate the probability that a certain mixture component generated a given document:

$$P(y_i = c_j|x_i;\hat{\theta}) = \frac{P(c_j|\hat{\theta})P(x_i|c_j;\hat{\theta})}{P(x_i|\hat{\theta})} = \frac{P(c_j|\hat{\theta}) \prod_{w_t \in X} P(w_t|c_j;\hat{\theta})^{x_{it}}}{\sum_{k=1}^{M} P(c_k|\hat{\theta}) \prod_{w_t \in X} P(w_t|c_k;\hat{\theta})^{x_{it}}}. \quad (3.3)$$

In the case of a simple Naive Bayes classification the class with the highest class probability, $arg\,max_j\, P(y_i = c_j|x_i;\hat{\theta}$, will be selected as the true class (Nigam et al., 2006).

The accuracy of this method can be questioned, especially when a small amount of training data is used. A method to increase this accuracy is Expectation Maximization (EM). EM can be seen as an advanced version of the Naive Bayes method. Instead of only classifying the unlabelled data using the training data set, once classified unlabelled data is used to further classify the unlabelled data. This method is especially useful when only a small amount of labelled data is available, which is the case in this study. Equation (3.4) shows the EM method:

$$l(\theta|X,Y) = \log(P(\theta)) + \sum_{x_i \in X_u} log \sum_{j \in [M]} P(c_j|\theta)P(x_i|c_j;\theta) + \sum_{x_i \in X_l} \log(P(y_i = c_j|\theta)P(x_i|y_i = c_j;\theta)), \quad (3.4)$$

this equation gives an iterative process, resulting in finding a local maximum probability.

In this study EM has been applied using two classes: (1) *Nature*: containing messages about nature or the PSI, the training set is created using the nature search words given in Table C, Appendix A: Lists of Used Queries, and (2) *None*: containing messages definitely not about nature or the PSI, the

training set is created using the non-nature search words given in Table D, Appendix A: Lists of Used Queries. All messages not belonging to one of the two training sets have been used as the *Unlabelled* dataset in EM.

The process can be explained by the following steps (Nigam et al., 2006):

1. A Naive Bayes classifier is created using a small set of labelled training data
2. The remaining unlabelled data is classified using this classifier
3. Instead of assigning the most probable class to the text, the probability of the text belonging to each class is assigned
4. A new Naive Bayes classifier is created using the labelled and unlabelled data, using the estimated class probabilities as real class labels
5. These steps are repeated, until a stable classifier and set of labels is created

Once a stable classifier is created all data belonging to the Nature class is selected for further processing, leaving out all non-relevant social media messages.

3.4.2 Data containing search words

Using cypher it is possible to select data in the database based on a certain search term. For this the attribute should be selected in which the search term should be found. For Flickr for example this attribute is the *title* of the post (Code snippet 3.2).

```
MATCH (flickr:Flickr)
WHERE lower(flickr.title)
CONTAINS lower('san isabel')
RETURN flickr;
```

Code snippet 3.2: Cypher query to find Flickr photos using a search term

It can be assumed that all data containing the selected search terms is relevant data for the analysis. However, not every message had a location assigned, which is required for further processing. First the part of the data that had a location assigned was selected. The next step was to determine whether this location was close enough to the PSI to be useful. All data inside the bounding box, thus falling in both category 1 and 2, is selected to be relevant for further processing. Messages containing coordinates outside the bounding box were not used any further.

The remaining part of this data does not have a location assigned, making it unusable for this research. However, since this data is collected based on search terms, which are most often certain highlights inside the park, it could still be possible to determine a location based on the content of the message. Several methods exist for retrieving a location from a text (e.g. Gazaz et al., 2016; Fatkulin et al., 2018; Inkpen et al., 2017). These methods, however, are outside the scope of this research, and will therefore not be used.

3.4.3 Remaining data

The remaining data do not contain any of the search words, and do not have a location assigned close by or inside the PSI. For determining the citizen perception of the PSI this data was considered as not relevant. Therefore it is chosen to not further use this data in the analysis.

3.5 Sentiment Analysis

To perform the sentiment analysis it was chosen to use the SentiStrength software, developed by Thelwall (2017). SentiStrength uses a large lexicon for classifying sentiment, this lexicon contains information obtained from the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker et al., 2003) and the General Inquirer list of sentiment terms (Stone et al., 1966). During the testing phase of the development this set has been expanded. The basic sentiment detection of SentiStrength returns both a positive and a negative score for the inserted text. The positive score ranges between 1 and 5; 1 indicating no positive sentiment and 5 indicating a highly positive sentiment. The same accounts for the negative sentiment; returning a score in the range of -1 to -5. This same score for positive and negative sentiment is assigned to each word in the lexicon, originally based on a corpus of 2600 comments on the social networking site MySpace (Thelwall, 2013).

Normally, SentiStrength only gives a positive and a negative score for the inserted text, however it is possible to select some other options; trinary, binary, and scale. The trinary option will return, next to the positive and negative score, a standardised classification score. This standardised score gives a value between -1 and 1, to indicate whether a text is positive, negative, or neutral. It does not distinguish between different levels of positivity or negativity. The binary option is comparable to the trinary one, but instead of returning a value of 0 for a neutral text it will always assign a positive or negative sentiment to the text. The scale option sums the positive and negative sentiment score to end up with one score to indicate the overall sentiment. This value can therefore be between -4 and 4 (Thelwall, 2017).

The algorithm used in the SentiStrength software has been designed for short texts, which is useful for most of the collected social media data, as can be seen in Figure 3.8. However, Flickr, Google+, and Reddit do have some longer texts than the other platforms, which might not be classified as short texts. To account for these texts the Java version of SentiStrength has a special mode for binary and trinary classification on longer texts (Thelwall, 2017). SentiStrength is especially useful for analysing social media data because of its linguistic rules; it corrects for misspellings and takes the sentiment indicated by emoticons into account (Pfitzner et al., 2012).

SentiStrength classifies a string based on the scores assigned to words in the lexicon. Each word inside the string, which is also present in the lexicon is assigned the corresponding score. The positive and negative sentiment scores of the string itself are determined by the word with the strongest corresponding sentiment. For example the string: *"The view was beautiful, but unfortunately it was raining badly."*, is classified as follows: *"The view was beautiful* [3]*, but unfortunately* [-2] *it was raining badly* [-2]*."*, resulting in a positive classification of 3 and a negative classification of -2. The program also takes into account *booster words*, these are words that strengthen the sentiment of the immediate following sentiment word; e.g. really, very (Thelwall, 2013). When a booster word is used, the sentiment of the corresponding word is increased or decreased by 1.

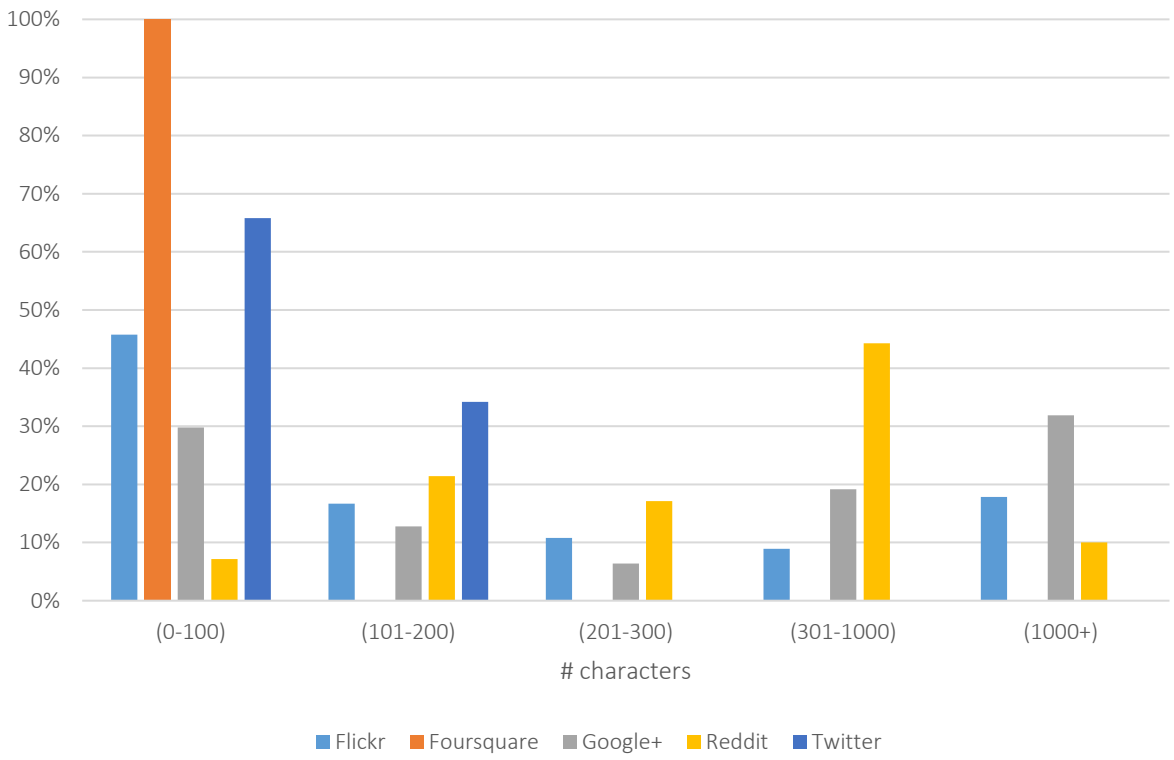

Figure 3.8: Division of text length for collected data per social media platform

A part of the data (7.1 per cent) consisted of long texts, in which only binary and trinary classification is allowed (Thelwall, 2017). Therefore it was chosen to use trinary classification, which was preferred over binary, since some texts do not have a sentiment at all and should be classified neutral. After performing the sentiment analysis on the text, the retrieved analysis was stored as attribute of the corresponding *node* in the database.

3.6 Citizen Perception Maps

The sentiments retrieved through the sentiment analysis were used to create overall citizen perception maps of the PSI. For this the SolVES tool was used, which is developed by the USGS Geosciences and Environmental Change Science Centre, and can be incorporated into ArcMap. In this research Version 3.0 (SolVES 3.0) has been used.

The previously created dataset is used as input for the SolVES model, using the locations of the relevant messages and their assigned sentiments. SolVES uses an attitude range (corresponding to sentiment) between 1 and 5, 1 indicating a strong positive and 5 a strong negative sentiment. The first step was therefore to normalize the collected sentiments to this scale. The sentiments were normalised using the functions from Equation (3.5):

$$sentiment = \begin{cases} 1\ if\ pos \geq abs(neg) * 2\ and\ pos > 2 \\ 2\ if\ pos > abs(neg) \\ 3\ if\ pos + neg = 0 \\ 4\ if\ abs(neg) > pos \\ 5\ if\ abs(neg) \geq pos * 2\ and\ neg < -2 \end{cases} \quad (3.5)$$

It was chosen to perform the normalisation this way, to make sure only really strong positive or negative messages were assigned a sentiment of respectively 1 or 5, all other positive or negative messages were assigned a sentiment of 2 or 4.

After conversion the sentiment data had to be inserted into the corresponding tables in the SolVES model. The points themselves were stored in a feature class, which was linked to an attitude table through the id. Other inputs for the SolVES model were environmental raster layers, to which the user wanted to correlate the social value points. Six of these layers were used, which are described in Table 3.2. The final input was the boundary of the study area, the model extrapolated to this area. The environmental layers and bounding box of the PSICC were available on the website of the USGS (US Geological Survey, 2017).

Table 3.2: Description of used environmental layers

Environmental Raster Layer	Description
DTR	Distance to the closest road
DTW	Distance to the closest water body
ELEV	Elevation of the area
LANDFORM	Classification of landforms
LULC	Classification of land cover
SLOPE	Slope of the area

Within the MaxEnt software, which is included in the SolVES tool, all input data points are correlated to the environmental layers. By applying Equation (2.3) to each point, the probabilities for the environmental values is determined. MaxEnt optimises these probabilities for each environmental layer, after which the results are combined into a single output raster. The MaxEnt software also calculates the relative influence of each environmental layer on the output. Jack-knife statistics visualise this influence, by showing the regularized training gain when only one environmental layer is selected, and when every layer except that layer is selected. The regularized training gain represents the performance of the model compare to a uniform distribution (Gormley et al., 2011). It can help the user to optimise their model by removing environmental layers that have a negative influence on the results, which are layers for which the training gain improves when that layer is removed.

To determine whether the output model of the MaxEnt software is useful a Receiver Operating Characteristic (ROC) curve is created. This curve shows the relationship between the false positive rate of the predicted class membership (x-axis) and the true positive rate (y-axis). The area under this curve (AUC) gives the usefulness of the created model, a value below 0.5 indicates that the model performs worse than a random model. Once a value is higher than 0.7, the model can be seen as useful.

To test the performance of the SolVES tool with different inputs, five scenarios were created. The difference between these scenarios were either in the environmental layers used, or in the points used as reference data.

3.7 VALIDATION

To identify hotspots and coldspots in these social value maps the Getis-Ord Gi* statistic was used, it determines which locations in an image are statistically significant hotspots and coldspots. It calculates a p-value (probability) and z-score (standard deviation) for each location in the image. These values are determined by calculating the local sum for a location and its nearest neighbours, which is then compared proportionally to the sum of all features. If the difference between the local sum and the expected local sum are too large to be the result of random change (i.e. the highest and lowest 5 percent), the z-score will be significant. A positive z-score relates to a local hotspot, and a negative z-score to a local coldspot (Environmental Systems Research Institute, 2016).

The Getis-Ord Gi* tool, which is incorporated in ArcMap, requires a feature class as input, however the created social value maps are raster layers. Therefore they are first converted to polygon feature classes using the Raster to Polygon tool. The *Conceptualization of Spatial Relationships* parameter in the Hot Spot Analysis tool was set to a fixed distance band, which ensures that all features have at least one neighbour, which were determined by their Euclidian Distance. The result of this tool is a map indicating areas of hot- and coldspots, classified by their confidence level (Environmental Systems Research Institute, 2016).

To compare the results from this research to the results from Bagstad et al. (2016), the created hot- and coldspot maps were used to find areas of synergy. The created maps were each overlaid by the map from Bagstad et al. (2016) and areas where both maps indicated a hotspot were assigned a value of 1 (red), areas where only the newly created map indicated a hotspot a value of 2 (yellow), areas where only the map from Bagstad et al. (2016) indicated a hotspot a value of 3 (green), and areas where neither of the maps showed a hotspot were given a value of 4 (blue).

A measurement of agreement was calculated using the Kappa Statistic in the Map Comparison Kit 3 (Geonamica, 2011). The Kappa Statistic value is calculated by equation (3.6):

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \qquad (3.6)$$

in which p_0 corresponds to the number of pixels in which both maps are classified the same, divided by the total number of pixels, and p_e is the agreement expected by chance. In other words, the Kappa value corresponds to the proportion of agreement between two rasters after removing any chance agreement (Flight and Julious, 2015). The scale from Altman (2006) was used to determine the goodness of the value: <0.20 represents poor, 0.21-0.40 represents fair, 0.41-0.60 represents moderate, 0.61-0.80 represents good, and 0.81-1.00 represents very good.

4 RESULTS AND VALIDATION

During this research various intermediate results have been collected, alongside the final results. In this chapter all these results are elaborated. In Chapter 4.1 facts and numbers about the collected data are given; in Chapter 4.2 the outcome of the classification is shown; in Chapter 4.3 the sentiment analysis is discussed; in Chapter 4.4 the created perception maps are shown. Validation of the maps is done in Chapter 4.5.

4.1 COLLECTED DATA

During a period of 5 weeks (June–July 2017) data has been collected extensively. Later in the research this dataset has been expanded with more recent data, by on-and-off collecting of data. Figure 4.1 shows the division of the used messages over the years.

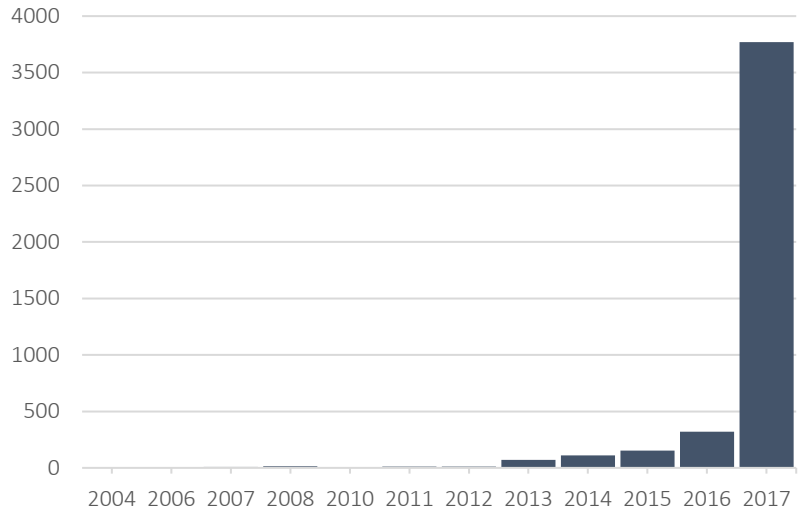

Figure 4.1: Number of collected messages per year

In Table 4.1 the total number of collected messages per social media platform is given, including the number of users connected to these messages. Users can be connected either by posting the message itself, or commenting on the message, depending on the actions allowed by the social media platform. In the table can be seen that most messages are collected from Twitter, followed by Flickr, Reddit, Foursquare and the least from Google+. On Twitter and Flickr more messages are posted than the number of users connected to them. While on Foursquare, Reddit and Google+ more users are connected to the messages than the number of messages itself.

Table 4.1: Number of collected messages and users per social media platform

	Flickr	Foursquare	Google+	Reddit	Twitter
# of messages	8,726	735	68	890	1,709,838
# of users	1,728	1313	89	1,015	131,014

4.2 Classification of Messages

After division of the retrieved data into the three subsets, defined in Chapter 3.4, relevant data could be selected. Since the geocoding of messages is outside of the scope of this research, only data with coordinates inside the bounding box around the PSI was used. Data from Twitter, Flickr, and Foursquare contained spatial information, and could therefore be retrieved from the database using a bounding box. The number of collected messages per platform inside the bounding box around the PSI can be seen in the bottom *Total* row in Table 4.2. Using these messages two training sets were created: *Nature* and *None*, and the remaining data was used as *Unlabelled* data.

Table 4.2: Number of messages used for classification, for each class and platform

	Twitter	Flickr	Foursquare	Total
Nature	706	291	18	*1,015*
None	2,250	8	6	*2,264*
Unlabelled	6,904	1,986	397	*9,287*
Total	*9,860*	*2,285*	*421*	*12,566*

The results after the Naive Bayes and Expectation Maximization classification are visible in Table 4.3, all messages were significantly classified after 70 iterations. A total of 4,694 social media messages could be used for further processing, this correspond to around 37 per cent of the total messages posted inside the park.

Table 4.3: Number of messages assigned to each class per social media platform

	Twitter	Flickr	Foursquare	Total
Nature	3,173	1,308	213	*4,694*
None	6,687	977	208	*7,872*
Total	*9,860*	*2,285*	*421*	*12,566*

To determine whether the messages classified as Nature are indeed relevant, a manual inspection was done of 50 messages. The queries used to create the training sets where adapted based on the findings during this inspection, after which the classification was performed again. This process has been repeated until none of the 50 manually checked messages were assigned wrongfully. Figure 4.2 shows two of the Twitter messages assigned to the Nature class during the last iteration.

Figure 4.2: Examples of Twitter messages assigned to the nature class

4.3 Sentiment Analysis

The sentiment analysis has been applied to all Twitter and Flickr messages that were classified as Nature in the previous step. The classified Foursquare messages were not included for further classification, since these messages are only descriptive and do not contain a sentiment.

The SentiStrength tool assigned a positive, negative, and trinary (-1, 0 or 1) score to the messages. Table 4.4 gives the assigned sentiments for some example Twitter and Flickr messages. In this table the first sentiment value corresponds to the positive sentiment found in the text, the second value corresponds to the negative value, and the third value gives the trinary classification.

Table 4.4: Sentiment assigned by SentiStrength to example messages

Twitter	Sentiment
"The view is breathtaking.....literally because it's hard to breathe up here"	*[4,0,1]*
"Went for a 9-mile run on the Ring in the Peak trail after 8 hours of trail work on the Ring the Peak trail. Good day!"	*[4,0,1]*
Flickr	**Sentiment**
"A late spring snow storm clears as the sun sets. Maloit Park near Minturn, Colorado. We're used to these spring snowstorms, but this one dropped a little more than usual and hung around a little longer than usual. Still made for a pretty scene!"	*[4,-4,-1]*
"The view from John's place – just amazing"	*[2,-1,1]*

To gain insight in the division of the Flickr and Twitter messages over the different sentiment classes, the histogram in Figure 4.3 has been constructed, in this histogram all positive and negative scores are merged: ('-', 4 & 5), ('+/- ', 3), ('+', 1 & 2). This shows that 60 per cent of the messages is assigned a neutral sentiment, and that the rest of the messages are overall more positive than negative.

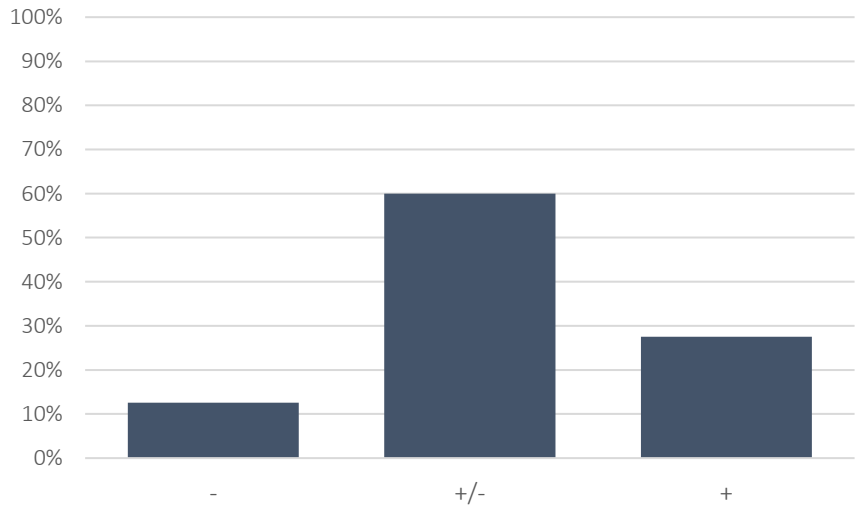

Figure 4.3: Division of assigned sentiments

To see the difference between sentiments on different social media platforms, the sentiment analysis has also been applied to texts from Google+ and Reddit. In Figure 4.4 histograms are shown of the division of sentiments in Flickr, Google+, Reddit, and Twitter messages. Reddit and Twitter show a larger number of positive messages, while on the other hand messages on Google+ seem more negative and messages from Flickr are distributed evenly over the positive and negative sentiments.

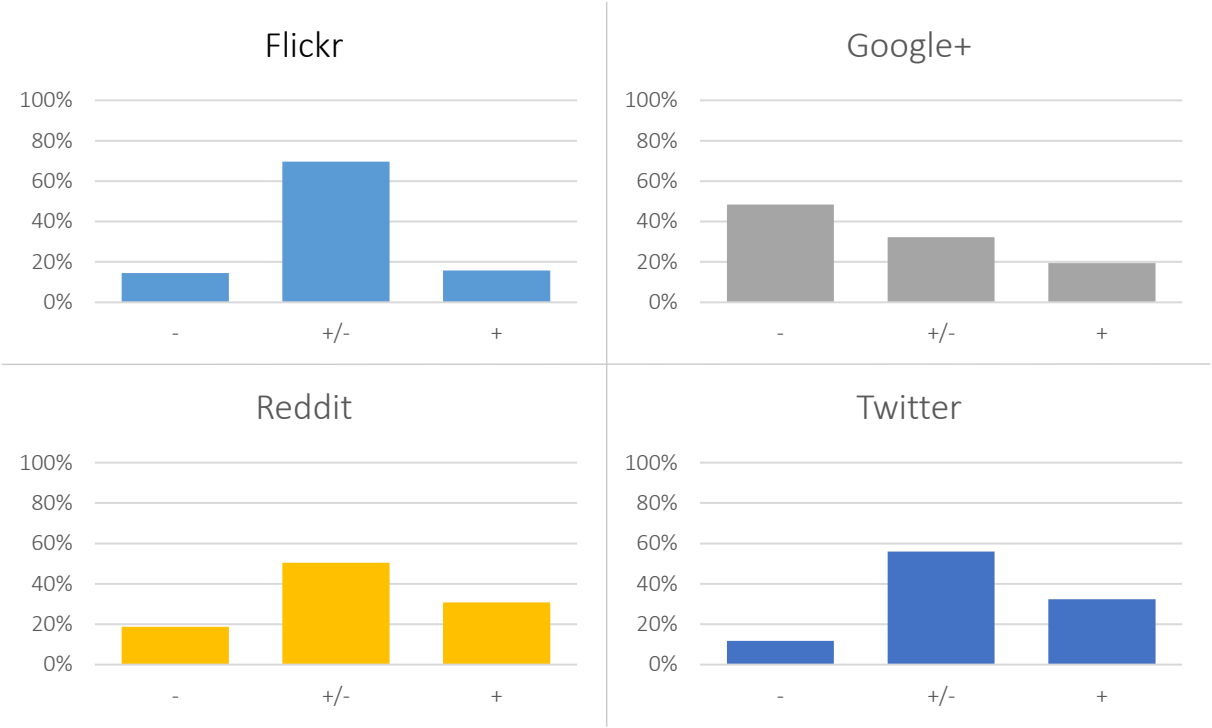

Figure 4.4: Division of assigned sentiments per platform

Using the available information about the users posting the messages, other comparisons on this information could be made. For example the comparison in Figure 4.5, where the sentiments of messages posted by residents of Colorado are compared to the sentiments of people with a different origin. In general it can be seen that residents from Colorado show a more positive sentiment towards the park than people from a different origin.

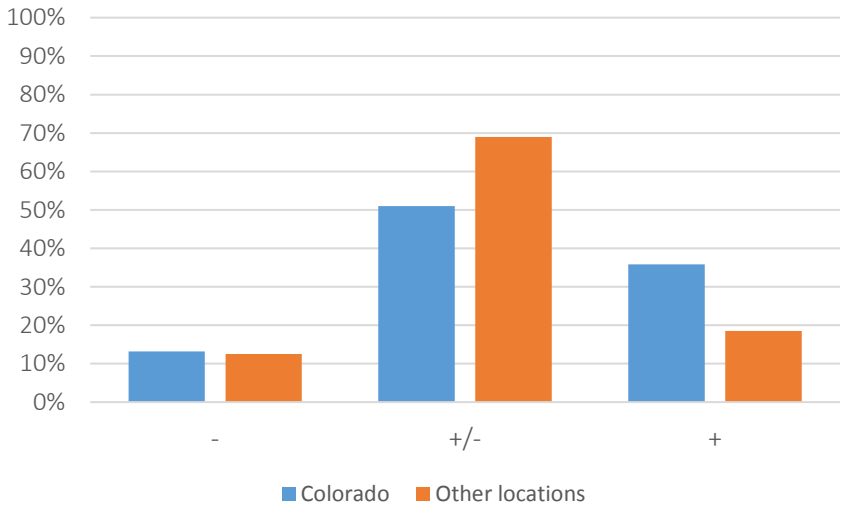

Figure 4.5: Division of sentiments for residents of Colorado and other locations

4.4 Citizen Perception Maps

After the sentiments have been determined for each message and these values have been normalised to the required input for the SolVES tool, a map showing the messages and their corresponding sentiments could be formed (Figure 4.6). Already some clusters of positive or negative points are visible.

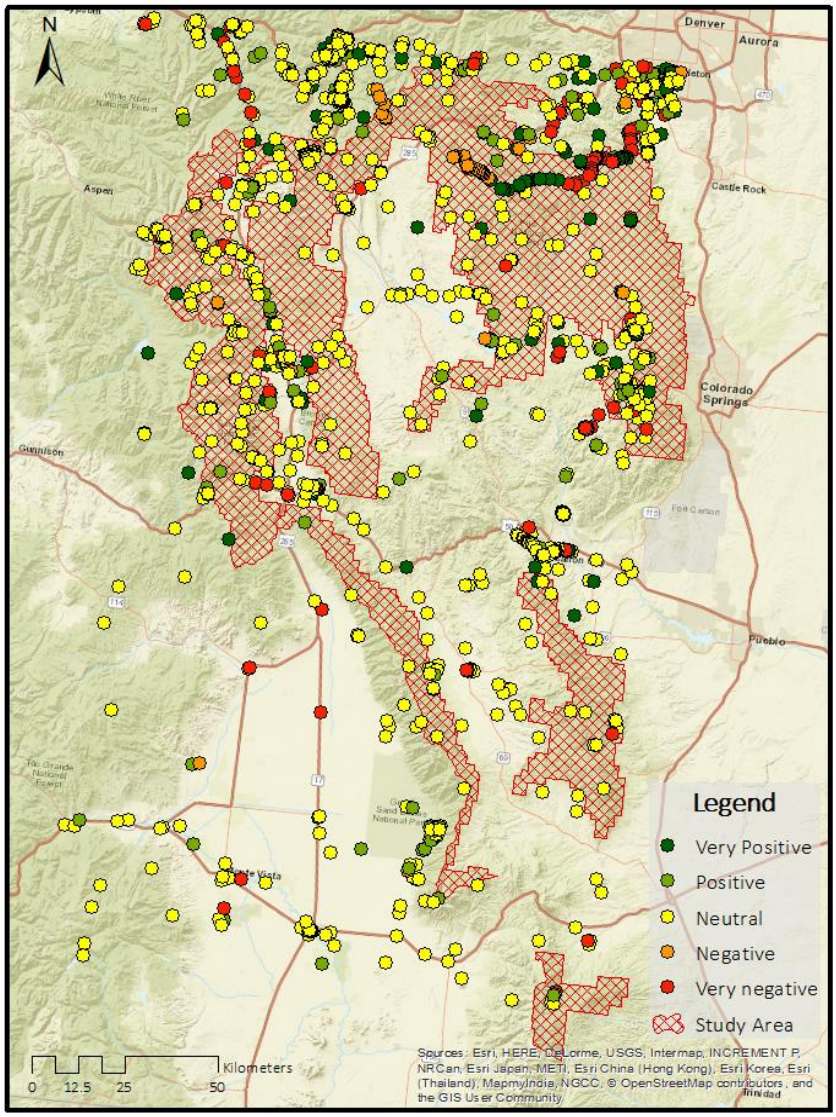

Figure 4.6: Spatial distribution of sentiments

The different scenarios described in this section, used (a subset of) these points as reference data for their analysis. These scenarios are shown in Table 4.5.

Table 4.5: Overview of used scenarios

Scenario	Name	Environmental layers used	Point data used
1	*Total*	All	All
2	*Optimal*	Optimal	All
3	*Positive*	Optimal	Positive and Neutral
4	*Negative*	Optimal	Negative and Neutral
5	*No Neutral*	Optimal	Positive and Negative

The first scenario used all points and all environmental layers as input. In Figure 4.7 the social value map created using this scenario is shown. The assigned values are normalised to a 1-10 scale. The corresponding jack-knife is given in Figure 4.8. Both a visual and common sense comparison were done to find the best suitable environmental layers.

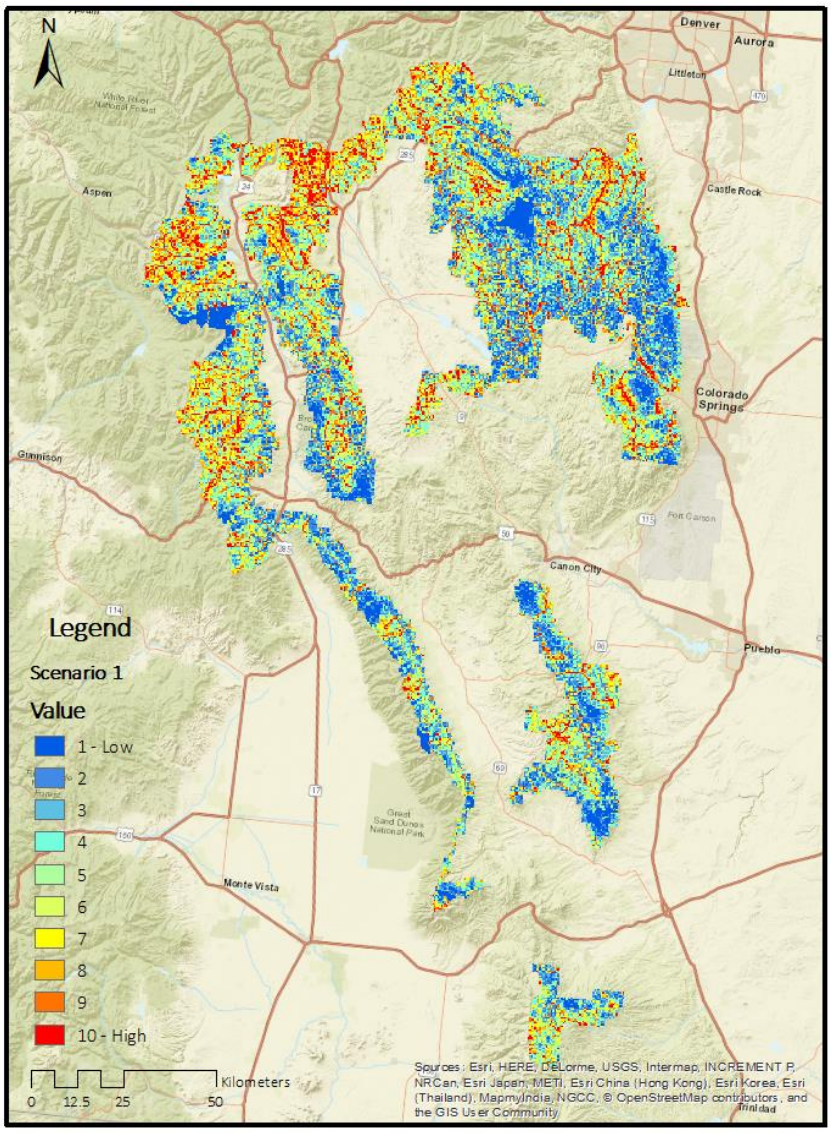

Figure 4.7: Resulting social value map from scenario 1

In the jack-knife it can be seen that all environmental layers do influence the outcome. Of all layers the elevation has the biggest influence, which is emphasised by the fact that the training gain is already higher than 0.7 when only this variable is used. Based on the jack-knife no environmental layers have to be removed to increase the training gain. However, when looking at the map it can be seen that the values show a linear pattern throughout the park, especially in the North-Eastern part. These patterns correspond to the walkable roads in the park, and therefore to the DTR (Distance To Roads) layer. The used locations are based on social media data and social media users are generally using these roads, therefore this relationship seems logical, but it is not desired. The DTR layer was removed for further processing to minimise this influence, even though the layer shows a positive influence in the jack-knife.

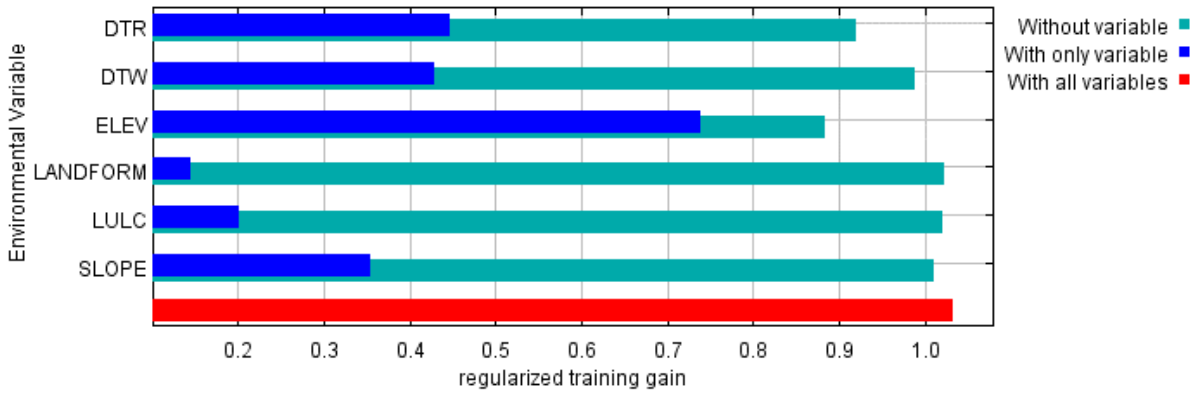

Figure 4.8: Jack-knife of scenario 1

The second scenario used the results from scenario 1 to retrieve the optimal results by removing the DTR layer. Figure 4.10-A gives the resulting social value map for this scenario. Compared to the map in Figure 4.7 the values seem more spread, instead of following the narrow linear structures.

The graph from Figure 4.9 shows the Receiver Operating Characteristic (ROC) curve and corresponding Area Under Curve (AUC) statistics for scenario 2. The ROC curve shows the relationship between the true positive ratio (y-axis) and the false positive ratio (x-axis), the closer the curve is to the upper left corner, the better the model fit. Two values for AUC were calculated for two different sets: the training data was used to create the model and the test data was used to test the performance of the model. As expected, the AUC of the training set is larger than the AUC of the test set, because this data has been used to create the model itself. The AUC of the test set gives an indication of the suitability of the model to be used for other areas as well. Both AUC values are higher than 0.7: 0.826 for the training data and 0.801 for the test data.

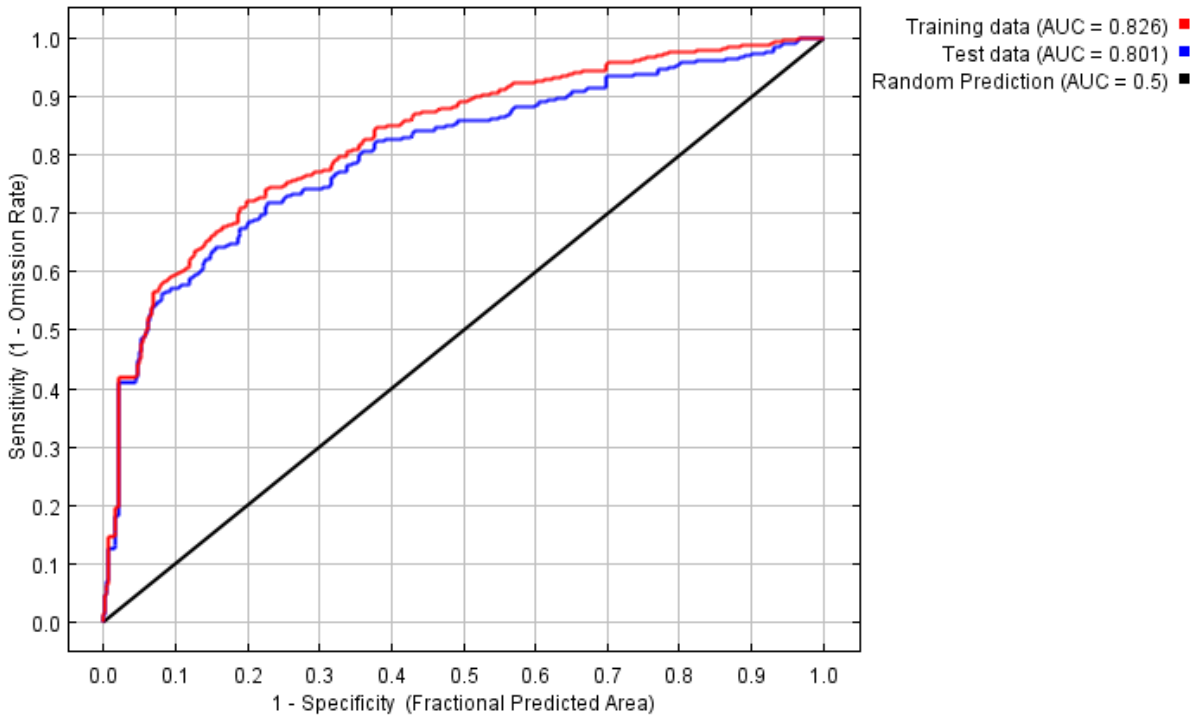

Figure 4.9: Receiver Operating Characteristic (ROC) curve for scenario 2

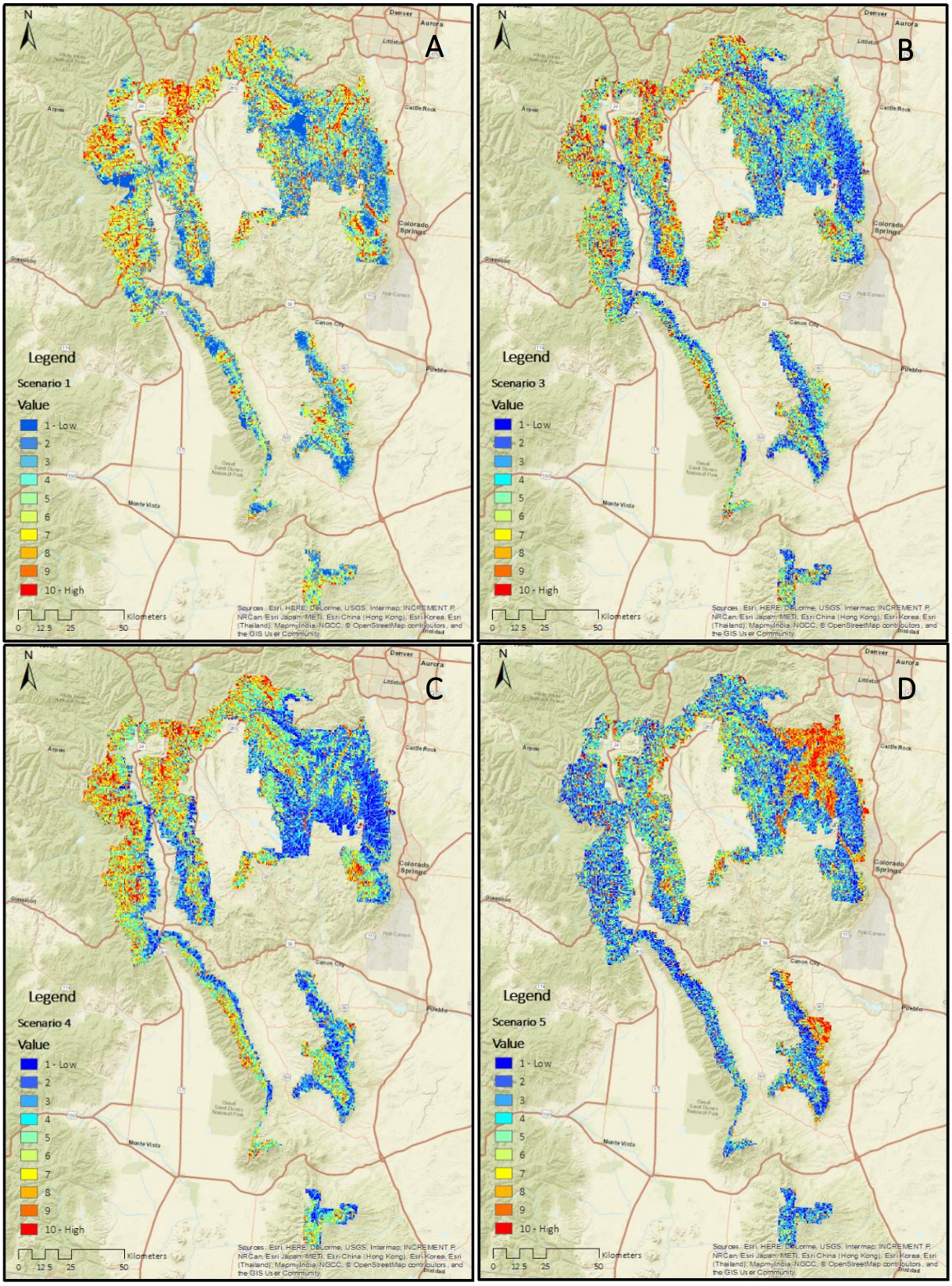

Figure 4.10: Resulting social value maps for scenarios 2-5. A: Scenario 2, B: Scenario 3, C: Scenario 4, D: Scenario 5. The calculated social values normalised to a 1-10 scale for each image separately, therefore the images cannot be directly compared.

The third scenario only used positive and neutral points from the sentiment data. This resulted in the map shown in Figure 4.10-B. By only selecting these points and leaving the negative ones out, the AUCs increased: for the training data to 0.833 and for the test data to 0.8194. In this map the blue colours correspond to areas where messages were neutral, as in the input data the neutral messages had the lowest values. The red colour corresponds to the most positive messages, which seem to be most present in the north-western part of the PSI.

The fourth scenario used, in contrast to the positive scenario, only the negative and neutral points. Using only these points, the map from Figure 4.10-C was created. In this map the red colours correspond to a neutral sentiment, as these messages had the highest values. The blue colours indicate areas with the most negative messages, which seem to be on the eastern side of the PSI.

Since the largest part of the data is assigned a neutral sentiment (Figure 4.3) and this data does not give any information about the citizen perception, the fifth scenario was created in which these data points are left out. The results from this scenario are visualised in Figure 4.10-D. By leaving the neutral points out the AUCs increased: for the training set to 0.8928 and for the test set to 0.8378. This scenario shows a more extreme result, as the neutral messages were left out only the positive and negative messages influenced the outcome. In this map a distinction can be made between very positively rated and very negatively rated areas.

4.5 Validation

Using the created social value maps, hot- and coldspot maps were constructed for all scenarios. To validate these results the maps have been compared to hotspot maps from the social value maps created by Bagstad et al. (2016), as explained in Chapter 2.1. The comparison map is given in Figure 4.11, the yellow colour indicates that there is a hotspot in the map from Bagstad et al. (2016), but not in the map from scenario 1. The same counts for the green colour, but the other way around. A similar comparison between the maps from scenarios 2-5 and the map from Bagstad et al. (2016) is visible in Figure 4.12. In Table 4.6 confusion matrices are given describing all comparison maps, it can be seen that the highest overlap in hotspots is found using scenario 4 (4.6 per cent), while the highest overlap in total is found using scenario 3 (87.8 per cent).

Table 4.6: Confusion matrices of hot- and coldspots for scenario 1-5 and the social value maps created by Bagstad et al. (2016)

Maps from Bagstad et al. (2016) ↓		*Scenario 1*		*Scenario 2*		*Scenario 3*		*Scenario 4*		*Scenario 5*	
		High	Low	High	Low	High	Low	High	Low	High	Low
Social value maps	High	3.2%	7.8%	2.9%	8.1%	2.8%	8.2%	4.6%	6.4%	1.6%	9.4%
	Low	13.5%	75.5%	4.3%	84.7%	4.0%	85.0%	8.6%	80.4%	7.3%	81.7%

Table 4.7: Calculated Kappa Statistics

Scenario	Kappa Statistic
1	0.112
2	0.257
3	0.256
4	0.295
5	0.072

The calculated Kappa Statistics for all scenarios are given in Table 4.7. The highest value is reached by scenario 4, which also had the highest overlap in hotspots. Scenario 1 and 5 have the lowest Kappa Statistic value, indicating a poor fit between the two maps, while the fit of the other three scenarios can be seen as fair.

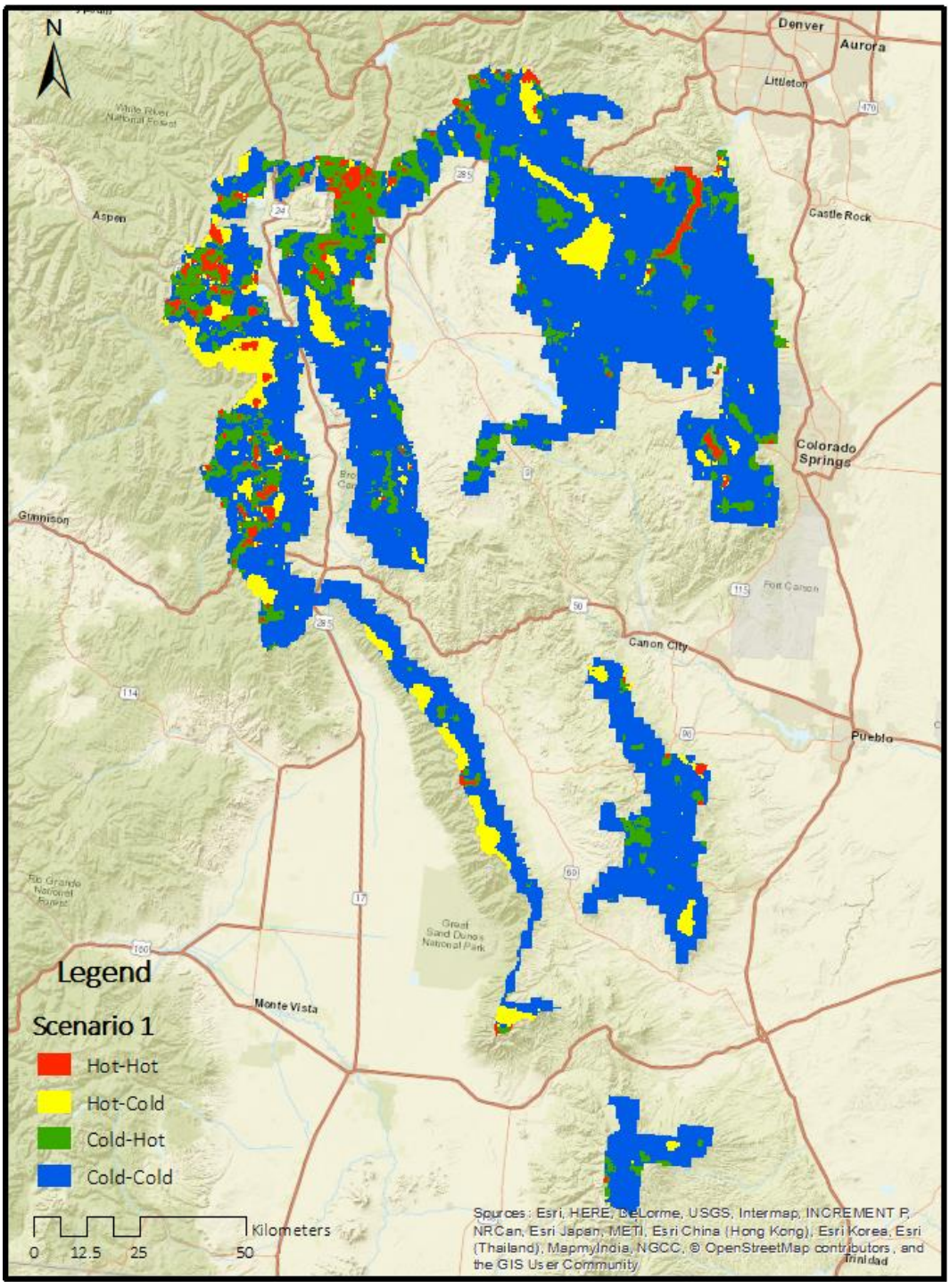

Figure 4.11: Comparison between the social value map from Bagstad et al. (2016) and scenario 1

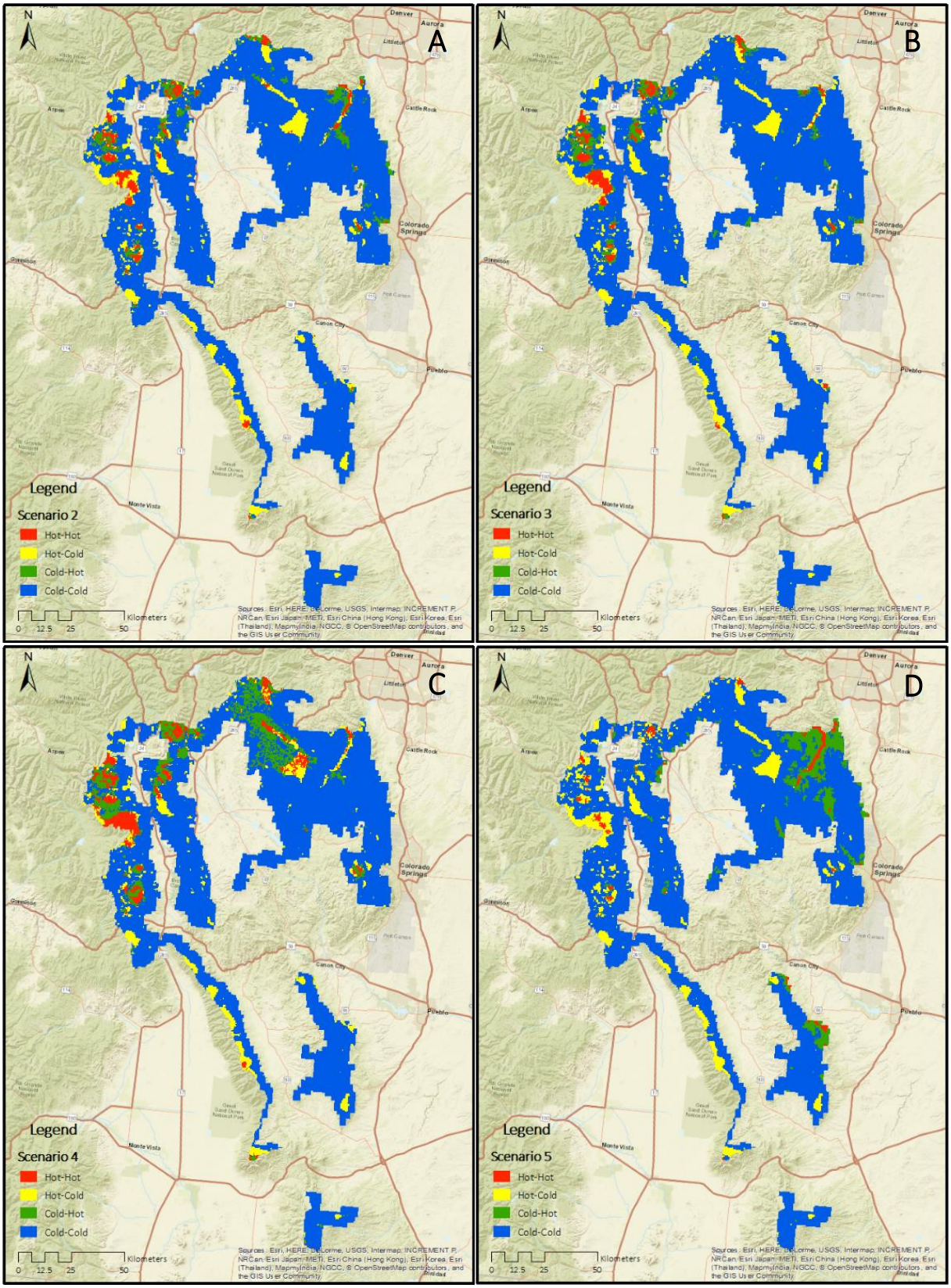

Figure 4.12: Maps showing overlapping hotspot areas between the different scenarios and the social value map from Bagstad et al. (2016). (A) scenario 1, (B) scenario 2, (C) scenario 3, (D) scenario 4

To determine whether the combination of social media data and survey data could increase the knowledge about cultural ecosystem services the created hotspot map from scenario 1 has been combined with the social value hotspot map from Bagstad et al. (2016). In this map the highest value from each map has been kept as true, this resulted in the hotspot map from Figure 4.13. As the highest of two values on one location is chosen, more hotspots exist than coldspots. The red areas in this map correspond to social value hotspots according to the social media analysis, the survey based analysis, or both.

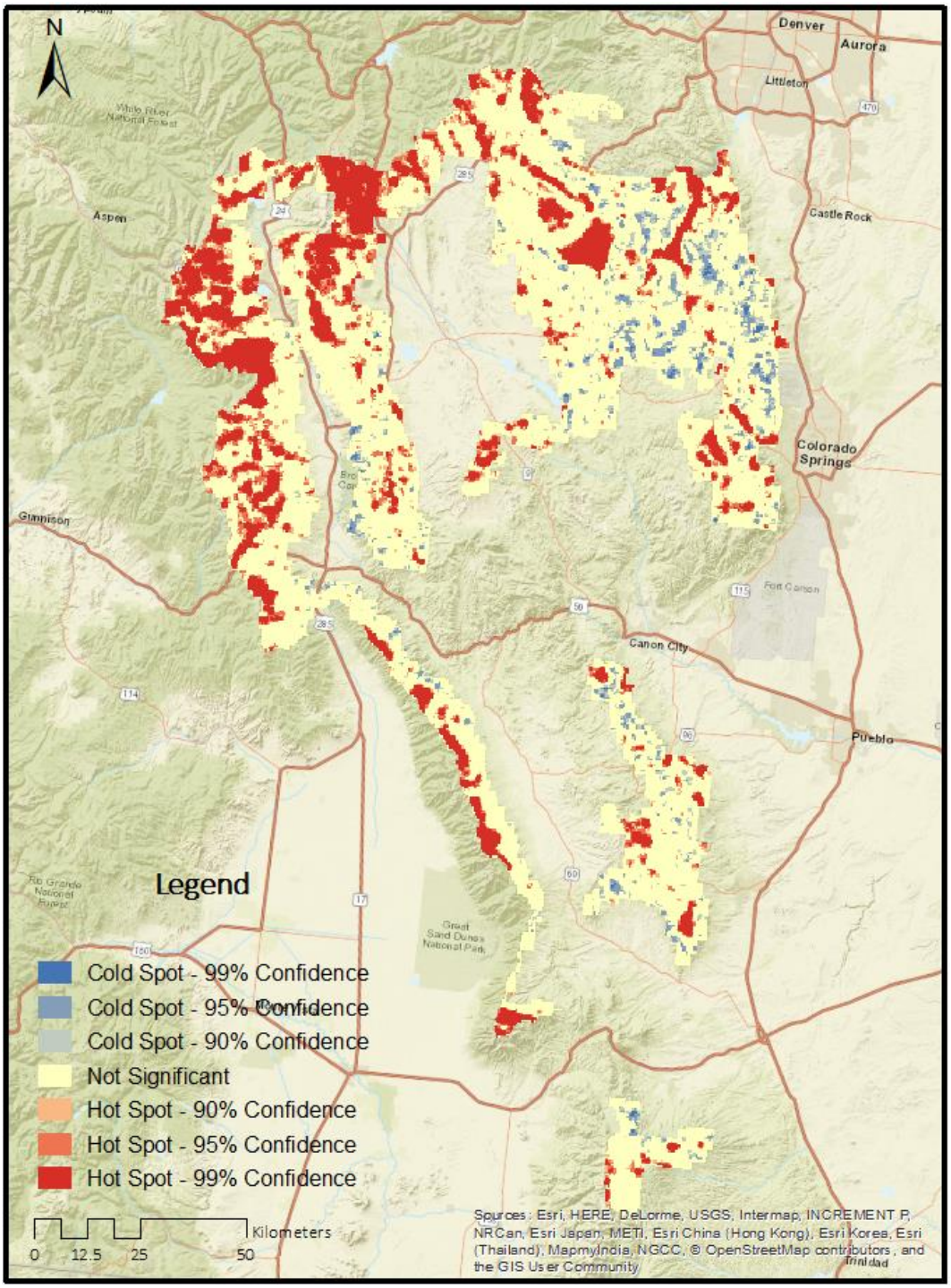

Figure 4.13: Combined hotspot map of scenario 1 and the social value map from Bagstad et al. (2016)

In the study from Bagstad et al. (2016), their social value hotspot map was compared to a hotspot map created using a biophysical model, in order to see whether social values could explain the value of ecosystem services. To determine the added value of using social media data in the analysis of social values, the newly created combined hotspot map has been compared to the same biophysical hotspot map. The map showing this comparison is given in Figure 4.14, in this map yellow indicates a hotspot in the hotspot map created using the biophysical model from Bagstad et al. (2016) and green indicates a hotspot in the combined social value hotspot map from Figure 4.12.

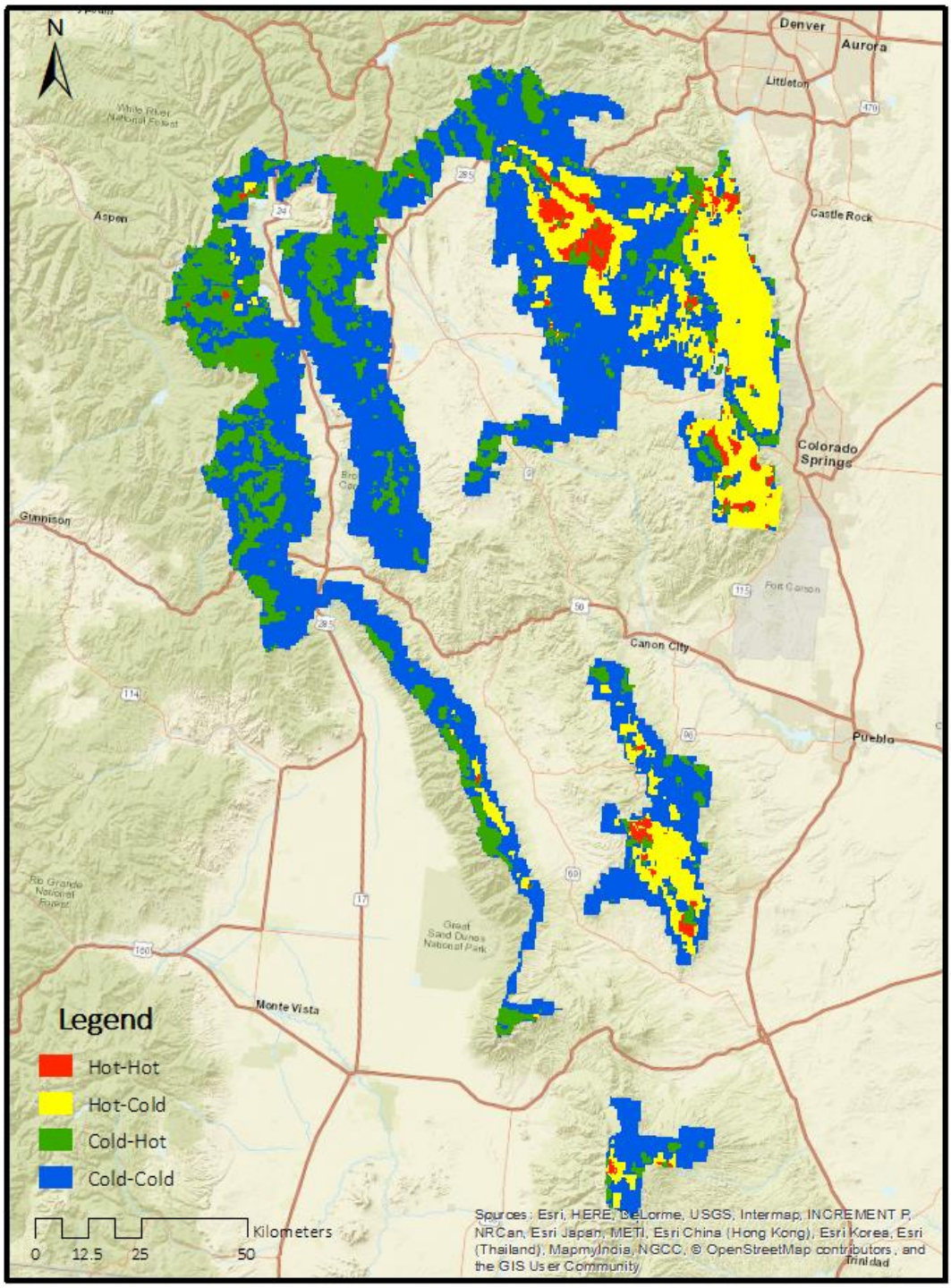

Figure 4.14: Comparison between the hotspot map from Figure 4.13 and the map created using a biophysical model

The Kappa Statistic value corresponding to this comparison is 0.552, which indicates a moderate equality between the two maps. Two other Kappa variables are the Kappa Histo, which refers to the similarity of quantity, and the Kappa Location, which refers to the similarity of location (Visser and De Nijs, 2006). These values are respectively 0.552 and 1, the Kappa Location indicates that the categories are spatially equally distributed over the map. The Kappa Histo, on the other hand, tells that the correspondence between the histograms of the two maps is a little more than 50 per cent. The overlap between the hotspots is 2.8% of the area of the PSI.

In Table 4.8 the confusion matrix of this comparison is shown, it shows that the overlap between the social value map and the biophysical map is 61.8 per cent. As the confusion matrix shows 16.9 per cent of the area is seen as a hotspot in the biophysically modelled map, which corresponds to the eastern side of the PSI as seen in the comparison map. Hotspots in the social value map are more present on the western side of the PSI, indicated by the green colour on the map, these hotspots account for 21.3 per cent of the area.

Table 4.8: Confusion matrix of combined social value map and biophysically modelled map

		Social value map	
		High	Low
Biophysical map	High	2.8%	16.9%
	Low	21.3%	59.0%

5 DISCUSSION

This study tried to find the added value of using social media data to determine the citizen perception of nature. A process to extract this information from the large amount of available data on social media platforms has been proposed, in this chapter the steps taken are discussed.

5.1 SOCIAL MEDIA

Data has been collected from Flickr, Foursquare, Google+, Reddit, and Twitter in order to be used to determine the citizen perception of the Pike-San Isabel National Forest (PSI). The total amount of collected messages has been given in Table 4.1. The largest amount of data has been harvested from Twitter: 99 per cent, of which only 0.6 per cent was useful for classification, as these messages had a posting location in the given bounding box. The number of collected messages, the number of messages used for Naive Bayes and EM classification, and the number of messages used for the sentiment analysis are given in Table 5.1. This table shows that only messages from Flickr and Twitter have been used to create the final result, even though it was concluded from the literature study that all chosen platforms could be useful for this research. Also, the part of the data from Flickr and Twitter that was useful is rather small.

Table 5.1: Number of messages used for each processing step

Platform	Collected	Classification	Sentiment Analysis
Flickr	8,726	2,285 (26.2%)	1,308 (15.0%)
Foursquare	735	421 (57.3%)	0 (0.0%)
Google+	68	0 (0.0%)	0 (0.0%)
Reddit	890	0 (0.0%)	0 (0.0%)
Twitter	1,709,838	9,860 (0.6%)	3,173 (0.2%)

During the classification step it was chosen to not use the data from Google+ and Reddit. As argued before, this data can be relevant for the citizen perception of the area (Miller, 2014; Reddit Inc., 2017), however, it does not contain any spatial information (Google, 2017; Reddit Inc., 2017). Therefore, when sampling on a bounding box, no data from Google+ or Reddit was returned. To overcome this limitation, a method could be developed to geocode messages, using highlights mentioned in the texts, and assign the found locations to the corresponding messages (Fatkulin et al., 2018; Gazaz et al., 2016; Inkpen et al., 2017). This method of extracting a location from text data could also be used to increase the number of usable Flickr and Twitter messages.

Table 5.1 also shows that Foursquare messages have been used for classification, but were left out of the further analysis. This is due to the fact that, even though users are able to give information in the form of text about venues (Li et al., 2018), hardly any of this text data was available for the selected venues. Data from Foursquare could therefore not be used for the sentiment analysis, but the data could be very useful for other types of social media analytics, such as finding relationships between locations and the type of people checking-in or liking these locations (Arampatzis and Kalamatianos, 2017; Huguenin et al., 2017; Mueller et al., 2017). Besides, some users have linked

their Twitter account to their Foursquare account, making it possible to find relationships between locations these people visit and their Twitter posts (Foursquare, 2016).

As seen in Chapter 2, there are other suitable social media platforms for citizen perception analysis. However, most of them have a restricted API, which can only be used when enough funding is available (Facebook, 2017; Kaplan and Haenlein, 2010; BrightPlanet, 2017; Instagram, 2017). These platforms were not used in this research, since it intends to develop a method to map citizen perception using openly available data sources.

Another question that arises is whether the collected data gives a sufficient overview of the opinion of the general public, as a bias might exist in the used data. It is questionable whether social media users are a good representation of people visiting the PSI. Pew Research Center (2017) has been monitoring social media usage of American adults. They have seen a rise in usage over to past years towards 69 per cent of American adults using a form of social media, which corresponds to 74 per cent of the online adults in the USA (Samuel Fosso et al., 2016). With the growth in social media usage, the social media user base has increased its representation of the general public, which can be seen in Figure 5.1. Even though the percentages of social media users between different age groups are not equal, still every age group of American adults in represented on social media.

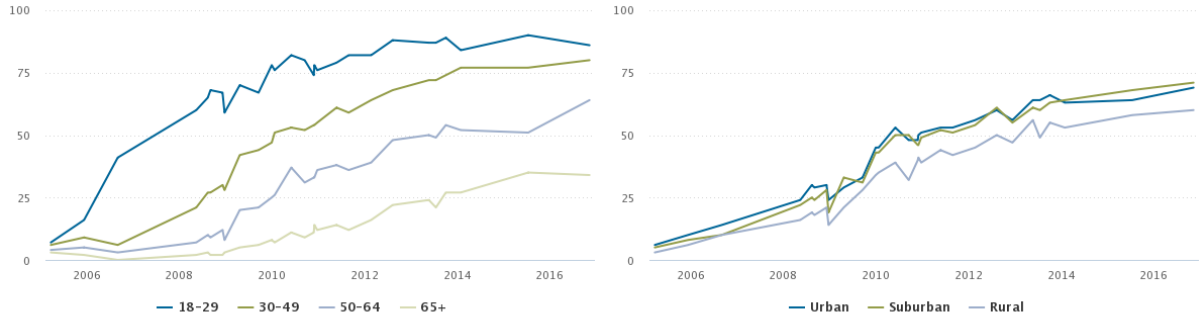

Figure 5.1: Percentage of social media users among American adults. Left: by age group, Right: by community type (Pew Research Center, 2017)

Besides the general bias on social media data, the dataset used for this research might also be biased. Since data has only been harvested for five consecutive weeks, this probably does not give a good overview of the general opinion. To test whether this bias is indeed present a longer harvesting time is needed. When the method is executed multiple times, using different subsets of the harvested data, differences between the created maps can be compared. If the differences are not significant it can be concluded that an unbiased sample has been used and the time of harvesting does not influence the outcome (Toepke, 2017).

5.2 Classification of Messages

During the classification step three categories of data were distinguished: (1) data inside a bounding box around the PSI, (2) data containing search words about the PSI, and (3) all remaining data. The third category, which contained all remaining data, was not used further, since this category existed of non-relevant data without coordinates, or with a location outside the defined bounding box. The

second created category contained messages without spatial information, but with a relevant content. These messages were not used further as well, but by using the method discussed before (Chapter 5.1) these messages could be geocoded, creating the possibility to use them for spatial analysis. All messages not used during the current study could be used in further research, by, for example, performing a network analysis to find the spatial reach of messages (Lago Vázquez, 2017).

The category of messages that was used to perform the classification, contained messages with a known location inside the given bounding box. The method used to classify these messages was Naive Bayes, combined with Expectation Maximization. For the purpose of this research this method was determined to have the best performance (Nigam et al., 2000). However, there are several ways in which this method could be improved, which will be discussed below.

In this research it was chosen to divide the messages between two subsets: one with messages regarding nature and one with all other messages. The Bag Of Words, created by the Naive Bayes classifier, for the nature class consisted of words like: *view*, *hiking*, *forest*, which are all clearly words regarding nature. The Bag Of Words of the second class, on the other hand, had a large variety in words, since there was no single identifier for this class. This makes it difficult for the classifier to find a distinct meaning for this class and messages are never 100 per cent certain assigned to it (Jiang et al., 2013; Rennie et al., 2003). To increase classification accuracy the number of used classes should be increased, which can be achieved by substituting the non-nature class into various subclasses, depending on the content of the messages. To determine these subclasses a sample should be taken from the dataset in order to visually find reoccurring topics. The increased number of classes will make it more certain for a message to belong to its assigned class, also decreasing the chance of wrongly assigning messages to the nature class (Rennie et al., 2003).

Another way of improving classification is by not only looking at the text of a message, but also to corresponding images. Advanced methods exist which can identify the context of a photograph, either by making use of the metadata of the used camera, or by looking at the content of the image itself (Yang and Ro, 2006). Amornpashara et al. (2015) developed a method which is able to identify landscapes on Flickr photos. They used three identifiers for photos taken from landscapes: (1) using tags regarding landscapes, (2) using the metadata from the images, since most photographers use different camera settings for photographing landscapes, and (3) using face detection techniques to find images containing people, and therefore they classify them as no landscape. Their method is able to classify images with landscapes quite accurately and could therefore be used in this study to increase the number of relevant social media messages which could be used for the sentiment analysis.

5.3 Sentiment Analysis

The sentiment analysis was performed using the SentiStrength software, with the standard data provided with this tool. To statistically test the performance of the SentiStrength two methods could be used. The first of these methods requires a sample of human classified texts, which can be compared to the classification of SentiStrength for the same texts (Thelwall et al., 2010). Another method would be to apply a 10-fold cross-validation procedure, which uses 9/10 of the given data

to train the term weights, and uses the other 1/10 to assess the accuracy. This is repeated ten times, until every part of the dataset is used to assess the accuracy (Thelwall, 2017).

In Figure 4.3 it can be seen that the largest part of the collected messages (60 per cent) was assigned a neutral sentiment. These texts either have no sentiment at all (40 per cent of these messages), or are equally positive as negative (60 per cent of these messages). That such a large amount of the messages does not have a sentiment at all, can be explained by the fact that not all users use social media to share their perception. Some of the posted messages are just informative, and therefore do not contain any sentiment, in fact only 26 per cent of messages on social media are regarding a person's opinion (GO-Gulf, 2014). Leaving out the neutral messages it can be seen that from the collected messages more have a positive sentiment than a negative sentiment. Which is not only the case in this dataset, but this trend can be seen over the general social media content: more than 70 per cent of the sentiment containing messages posted on social media this sentiment is positive (GO-Gulf, 2014).

This trend of people posting more positive messages than negative messages can also be seen in the collected datasets from Reddit and Twitter (Figure 4.4). However, the dataset from Google+ shows an opposite trend, having more negative messages than positive ones. There is no explanation for this difference, other than the fact that the dataset harvested from Google+ on which the sentiment analysis was applied was very small: only 31 messages. Therefore, no hard conclusions can be drawn from the observed difference. For Flickr, on the other hand, the dataset used for the sentiment analysis had a larger size. In this dataset 70 per cent of the messages was neutral, and the amount of messages assigned a positive sentiment was equal to the amount of messages assigned a negative sentiment. That such a large amount of the messages was neutral is most likely due to the fact that the messages are only titles and descriptions of photos, which in many cases are only descriptive of the content of the photo (Angus and Thelwall, 2010). It would be interesting to use methods to find sentiment in images, to see if there is a sentiment connected to the collected images (Ko et al., 2016)

Another comparison made using the sentiment of the data, has been visualised in Figure 4.5. This figure shows that residents from Colorado are more positive about the PSI on social media than people of a different origin. This might be due to the fact that people living close by the area, and potentially visiting it a lot, have a higher attachment to the area itself and are therefore more positive about it (Anton and Lawrence, 2014).

There are several ways in which the performance of the sentiment analysis could be perfected. One way is by improving the lexicon, which is used for classifying the sentiments. The currently used lexicon contained 2310 words and their corresponding sentiment, classified using text from the social network site MySpace (Thelwall, 2013). This lexicon can be adapted manually, by either adding new words or changing the current sentiment assigned to a word. In this way the lexicon can be adapted to fit the collected dataset. Another way of optimising the lexicon is by using a classified text collection. This text collection should contain at least 500 texts from the collected dataset, which are manually assigned a positive and a negative sentiment by humans. The

SentiStrength tool is able to create a new lexicon using the texts and the assigned sentiments, this lexicon can be used to replace the old one (Thelwall, 2017).

The texts used for the sentiment analysis were chosen based on their content, in this case all texts are regarding nature. However, the used lexicon only consists of general words containing a sentiment, while there are certain words that only contain a sentiment when they are regarding nature. To adjust for this bias in the sentiment analysis a topic based sentiment analysis could be performed, by changing the lexicon and adding these specific words. This could be done manually, but SentiStrength also offers the option to find a selection of relevant words for your specific texts automatically. These words are found using a dataset with human coded texts, which is compared to the classification SentiStrength assigns to it, using the default lexicon (Thelwall, 2017).

5.4 Citizen Perception Maps

The citizen perception maps for this study were created using the SolVES tool, which used the MaxEnt software to find a correlation between the inserted points and selected environmental layers. The MaxEnt software has proven to be valuable in comparable research, Lee et al. (2018) used the MaxEnt software on a dataset containing social media messages on which a sentiment analysis was applied. By using the MaxEnt classifier, next to a set of other comparable classifiers, to find relationships between sentiments and lexical properties, they found that MaxEnt outperformed all other used classifiers. Yoshimura and Hiura (2017) used the MaxEnt software to estimate the aesthetic demand of an area using photos uploaded on Flickr. They found that the prediction accuracy of MaxEnt does not decrease when the sample size decreases, making it a useful tool for mapping ecosystem services using social media data.

Five different scenarios were created to perform using the SolVES tool, which were used to test the performance of the tool on the collected data. The differences between the results from the different scenarios are explained below.

The first scenario contained all environmental layers used by Bagstad et al. (2016) and all collected social media points. In this scenario the map shows linear patterns, corresponding to the roads in the PSI. In fact, all collected social media points are on locations that are accessible by humans, as people are not able to easily access, for example, the high peaks in wilderness areas. This will have an influence on the final outcome of this research, because it results in the fact that the messages only give information about accessible areas, while the other areas are of importance as well.

The social value maps of the four other scenarios are given in Figure 4.10, in these maps, however, the scales are not numerically comparable. The SolVES tool calculates a kernel density surface for the environmental layers, which will always have different values, depending on the relationship between the environmental layers and the inserted point data (Sherrouse and Semmens, 2015). In the social value maps the ten used categories were created using quantiles to determine the borders, this means that every category represents $1/10^{th}$ of the data. The maps thus give an overview of the division of social values across the park for each scenario and can be used to find areas of accordance.

The map created by scenario 3 shows low values on the eastern side of the PSI, while in scenario 2 this area contains higher values. Scenario 3 only contains positive and neutral data, removing the influence of negative data. This results in the neutral messages being assigned the lowest values, moving the scale of the categories more towards the higher values. In this map positive areas are exaggerated and negative areas are smoothened. While in the map from scenario 4 the positive values are removed, resulting in the neutral messages being assigned the highest values. This almost completely removes the positive area in the north-eastern part of the PSI, which is visible in the map from scenario 2. Indicating that this area was assigned a high value because of the large amount of positive messages in this area. This is emphasised by the map from scenario 5, which does not contain any neutral messages. In this map the north-eastern area is assigned the highest value, corresponding to the large amount of positive messages in this area.

The training AUCs for all scenarios (Table 4.5) indicate that it is possible to create a reliable model using the given information, which means that a correlation exists between the sentiments and environmental layers. Besides the training AUCs, the test AUCs also gave an indication that the created models are potentially valuable to be transferred onto other areas (Elith et al., 2011).

5.5 VALIDATION

To create the hotspot maps for the different scenarios the Getis-Ord Gi* tool in ArcMap was used, these maps contained information about the statistics at an $\alpha = 0.10$ and an $\alpha = 0.05$ significance level. The study from Bagstad et al. (2016) used the hotspot map with an $\alpha = 0.05$ significance level for comparison, therefore this map was used in this research as well. However, other methods to create hotspot maps exist, which were evaluated by Bagstad et al. (2017) using the same social value and biophysical maps. The other methods proposed were: (1) top and bottom 33% of values, (2) top and bottom 10% of values, (3) top and bottom values covering 33% of the park, (4) top and bottom values covering 10% of the park. These proposed methods could be used for validation of the results from this research as well, in order to improve the validation, as these methods are proven to give significant differences in hotspot extent, clustering, and number {Bagstad, 2017 #213}.

As discussed before, the created hotspot maps from scenarios 2-5 do not contain the Distance To Roads (DTR) layer, this will have an influence on the comparison to the results from Bagstad et al. (2016), as the DTR layer was used to create their map. Therefore, only scenario 1 can be directly compared to these results, even though the Kappa Statistic (Table 4.7) for this scenario is lower than for some other scenarios.

The Kappa Statistic from scenario 1 indicates that the fit between the social value map created using social media data and the social value map creates using survey data is poor (0.112). Based on this information, it can be concluded that the value of social media data is different from the value of survey data and can therefore not work as a replacement. To see if social media data is able to increase the knowledge on citizen perception instead, the created hotspot map was combined with the social value hotspot map from Bagstad et al. (2016). Next, this map was compared to the biophysical hotspot map, using the same comparison methods. The resulting map

can be seen in Figure 4.13, which shows that the biophysical model assigns a higher value to the eastern side of the PSI, while the social value method gives more value to the western part. The Kappa Statistic for this comparison is 0.552, which indicates a moderate equality between the two maps. Besides, the overlap between the hotspots on the social value map and the biophysical map is 2.8%, which is considerably higher than the overlap found by Bagstad et al. (2016) when only using the survey data, which is 1.8%. This indicates that the accuracy of social value maps increases by using social media data next to survey data to determine the citizen perception.

The added value of social media data to survey data can be explained by the content of these two data sources. When using surveys to determine the citizen perception of nature, pre-produced questions are used, which are asked to the respondents after they have visited the area. The moment these surveys are taken has an influence on the answers respondents give, as they might not remember every exact detail (Schwarz, 1999). Social media gives users the opportunity to share their opinion on the spot, an opinion they might have forgotten a few minutes later. This will result in people sharing opinions on social media, which they might not have remembered when questioned for a survey. Besides, the respondents answers can be influenced by the way the questions are formulated (Schwarz, 1999), while on social media a user is unbiased. The fact that social media users share their messages on the spot, results in a spatial difference between the locations from social media messages and the locations assigned by survey respondents. In the survey, respondents mainly indicated highlights in the PSI, at which they indicated an important social value (Bagstad et al., 2016). While the messages on social media were mainly posted from points where people could see these highlights, corresponding to viewpoints and roads (Figure 4.6). Combining these points will give a better coverage of the entire area and a better coverage of people's opinions. The spread of social media locations can be improved when geocoding of messages is applied, as this will lead to highlights in the PSI being indicated, which will correspond to locations indicated by survey respondents.

6 CONCLUSIONS

Based on the executed research and its results the following conclusions could be made, answering the research questions.

1. *Which social media platforms can be used for citizen perception analysis?*

   Flickr and Twitter currently are the most suitable platforms for citizen perception analysis. The content of messages posted on these platforms gives relevant information about citizen opinions, and a large amount of the data is geotagged, making the information spatially relevant as well. The content of Google+ and Reddit messages is relevant for citizen perception as well, however there is no spatial information available about these messages, therefore it cannot be used directly. Foursquare, on the other hand, has spatially interesting information, but it contains no text data that gives information about the citizen perception.

2. *What is the best way to store the collected social media data?*

   Within social media data one of the most important information sources are the relationships between messages, users, and locations. In relational databases these can be stored, however, once these relationships become more advanced, querying will become increasingly difficult. A better way of storing this data is therefore by using a graph database, in which relationships between different data *nodes* are easily created, visualised, and queried. For this study Neo4j is the best suitable graph database to store the collected social media data as it has a spatial extension making it possible to apply spatial queries.

3. *How can information about citizen perception be retrieved from social media messages?*

   To turn data harvested from social media into information about citizen perception, some steps have to be taken. Since citizen perception is concerning a specific topic, the first step is to find data regarding this topic. By using Naive Bayes and Expectation Maximization data can be found similar to a given training set and corresponding topic. Sentiment analysis can give information about the polarity of a message and therefore a person's feelings towards the given topic. This information can be used to create social value maps visualising the citizen perception of a certain area regarding the selected topic. The performed case study has proven that this method is able to give insight in the citizen perception.

4. *How do the results from the social media analysis compare to the results from Bagstad et al. (2016)?*

   Based on the found Kappa Statistic (0.112) and visual validation between the created social value map and the social value map from Bagstad et al. (2016), it can be concluded that social media data cannot work as a direct replacement for survey data. However, the use of social media data, together with survey data, has proven to increase the knowledge of citizen perception, as the similarities between the combined hotspot map, from the social

media data and survey data, and the biophysical hotspot map are significantly: $\kappa = 0.552$. This can be substantiated by emphasising the spatial and temporal differences between social media data and survey data, which are complementary to each other.

This study proposed a method to analyse the citizen perception of nature using social media, and implemented this method on a case study of the Pike-San Isabel National Forest (PSI). A dataset was created which could be used to analyse the citizen perception of nature, by collecting messages from various social media platforms and performing a classification on them to create a subset containing messages inside the PSI and relevant for nature. Identifying the sentiments of these messages gave insight in the social value of these locations, as sentiment was used as a proxy for social value. To gain insight in the overall citizen perception of the PSI, social value hotspot maps were created, indicating areas of high or low social value. After comparison of these maps with maps created during a previous study from Bagstad et al. (2016), it could be concluded that using social media data can increase the knowledge about citizen perception of nature.

7 RECOMMENDATIONS

The method proposed in this research has proven to be valuable, some implementations to further improve the results are discussed below.

Use connections between platforms: One of the strengths of this research is making use of data from different social media platforms. An interesting attribute in the data from some of these platforms is a link to a user's accounts on other social media platforms. This attribute is currently only used to find more Twitter users by their Foursquare and Google+ accounts, but it could also be used to visualise relationships between messages on different platform (Veiga and Eickhoff, 2016).

Harvest more social media data: A larger dataset will increase the reliability of the results, as the influence of single messages decreases. When a larger dataset is available it is also possible to perform tests on the bias produced by a small harvesting window. Overall the size of the dataset determines the accuracy of the results.

Geocode messages and use images for classification: As seen in Chapter 5.1 a large part of the collected data has not been used to create the final results. To increase this number two methods are proposed. The first method is to geocode relevant messages, as a large part of the collected data does not contain spatial information. By performing geocoding methods on these messages a location could be found, making the messages useful for this research (Fatkulin et al., 2018; Gazaz et al., 2016; Inkpen et al., 2017). Another method to increase the number of useful messages is by using images in the classification step. Amornpashara et al. (2015) proposed a method to identify topics in images, making it possible to select images regarding nature, increasing the size of the usable dataset.

Improve sentiment analysis lexicon: Currently, the default lexicon was used for the sentiment analysis, but, since all messages contain data regarding nature, this lexicon could be improved to fit the right topic. By creating a lexicon especially for the collected data, the accuracy of the sentiment analysis could be improved. Besides, a classified collection of texts could be used to automatically improve the lexicon.

Use images for sentiment analysis: Most collected Flickr messages were assigned a neutral sentiment, this was mainly due to the fact that these messages did not contain a lot of text. To improve the classification of sentiment for these messages it is possible to use the corresponding image as well. Ko et al. (2016) developed a method to identify sentiments in images, which could be used to find the sentiments in the available Flickr images.

Use more environmental layers: The social value maps were created using six environmental layers, but there could be more environmental factors which influence the citizen perception. To see the influence of other environmental factors these should be used in the SolVES tool as well. The corresponding jack-knife will give information about the influence of these layers. In the current research only the environmental layers used by Bagstad et al. (2016) were used, to be able to directly compare the results. If other environmental layers are being used, these layers should also

be used to create new social value maps using the data from their research to be able to validate the outcome.

Use other methods to create hotspot maps: The hotspot maps used for the validation were created using the Getis-Ord Gi* statistic with an $\alpha = 0.05$ significance level. To further validate the results different hotspot methods could be used, such as the ones proposed by Bagstad et al. (2017) and discussed in Chapter 5.5. It could be compared whether different hotspot methods result in a higher correlation between biophysical models and social values mapping.

Perform social network analytics: In the performed research the collected messages have been used to determine the citizen perception. A next step would be to use the retrieved results to find spatial patterns. This could be done by performing a social network analysis, to see, for example, the reach of the messages (Lago Vázquez, 2017).

Perform method on other area: As it is proven that this method is able to increase the value created by survey data regarding citizen perception. Therefore, it is interesting to see how this method performs on a different area, to determine whether the method is applicable in other situations as well.

8 BIBLIOGRAPHY

ALTMAN, D. G. 2006. *Practical Statistics for Medical Research*, Chapman \& Hall/CRC.

AMORNPASHARA, N., ARAKAWA, Y., TAMAI, M. & YASUMOTO, K. Landscape photo classification mechanism for context-aware photography support system. 2015 IEEE International Conference on Consumer Electronics, ICCE 2015, 2015. 663-666.

ANGUS, E. & THELWALL, M. Motivations for image publishing and tagging on flickr. ELPUB 2010 - Publishing in the Networked World: Transforming the Nature of Communication, 14th International Conference on Electronic Publishing, 2010. 189-204.

ANTON, C. E. & LAWRENCE, C. 2014. Home is where the heart is: The effect of place of residence on place attachment and community participation. *Journal of Environmental Psychology,* 40**,** 451-461.

ARAMPATZIS, A. & KALAMATIANOS, G. 2017. Suggesting points-of-interest via content-based, collaborative, and hybrid fusion methods in mobile devices. *ACM Transactions on Information Systems,* 36.

BAGSTAD, K. J., REED, J. M., SEMMENS, D. J., SHERROUSE, B. C. & TROY, A. 2016. Linking biophysical models and public preferences for ecosystem service assessments: a case study for the Southern Rocky Mountains. *Regional Environmental Change,* 16**,** 2005-2018.

BAGSTAD, K. J., SEMMENS, D. J., ANCONA, Z. H. & SHERROUSE, B. C. 2017. Evaluating alternative methods for biophysical and cultural ecosystem services hotspot mapping in natural resource planning. *Landscape Ecology,* 32**,** 77-97.

BIHIS, C. 2015. *Mastering OAuth 2.0,* Birmingham, Packt Publishing Ltd.

BOE, B. 2017. *PRAW: The Python Reddit API Wrapper* [Online]. Available: http://praw.readthedocs.io/en/latest/index.html [Accessed 10 October 2017].

BRIGHTPLANET. 2017. *Social Media Data - Instagram Pulls Back on API Access* [Online]. Available: https://brightplanet.com/2017/01/instagram-data/ [Accessed 13 October 2017].

BROWN, G. 2012. Public participation GIS (PPGIS) for regional and environmental planning: Reflections on a decade of empirical research. *URISA Journal,* 24**,** 7-18.

CHEN, S., LIN, L. & YUAN, X. 2017. Social Media Visual Analytics. *Computer Graphics Forum,* 36**,** 563-587.

CHEN, X., ELMES, G., YE, X. & CHANG, J. 2016. Implementing a real-time Twitter-based system for resource dispatch in disaster management. *GeoJournal,* 81**,** 863-873.

CHOUDRI, B. S., BAAWAIN, M., AL-ZEIDI, K., AL-NOFLI, H., AL-BUSAIDI, R. & AL-FAZARI, K. 2017. Citizen perception on environmental responsibility of the corporate sector in rural areas. *Environment, Development and Sustainability,* 19**,** 2565-2576.

COLE, J. R., GHAFURIAN, M. & REITTER, D. 2017. Is word adoption a grassroots process? An analysis of Reddit communities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

DAILY, G. C., POLASKY, S., GOLDSTEIN, J., KAREIVA, P. M., MOONEY, H. A., PEJCHAR, L., RICKETTS, T. H., SALZMAN, J. & SHALLENBERGER, R. 2009. Ecosystem services in decision making: time to deliver. *Frontiers in Ecology and the Environment,* 7**,** 21-28.

DAILY, G. C., SÖDERQVIST, T., ANIYAR, S., ARROW, K., DASGUPTA, P., EHRLICH, P. R., FOLKE, C., JANSSON, A., JANSSON, B. O., KAUTSKY, N., LEVIN, S., LUBCHENCO, J., MÄLER, K. G., SIMPSON, D., STARRETT, D., TILMAN, D. & WALKER, B. 2000. Value of nature and the nature of value. *Science,* 289**,** 395-396.

DANIEL, T. C., MUHAR, A., ARNBERGER, A., AZNAR, O., BOYD, J. W., CHAN, K. M. A., COSTANZA, R., ELMQVIST, T., FLINT, C. G., GOBSTER, P. H., GRÊT-REGAMEY, A., LAVE, R., MUHAR, S., PENKER, M., RIBE, R. G., SCHAUPPENLEHNER, T., SIKOR, T., SOLOVIY, I., SPIERENBURG, M., TACZANOWSKA, K., TAM, J. & VON DER DUNK, A. 2012. Contributions of cultural services to the ecosystem services agenda. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 8812-8819.

DE SOUZA, C. R. B., REDMILES, D., CHENG, L. T., MILLEN, D. & PATTERSON, J. Sometimes you need to see through walls - A field study of application programming interfaces. Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 2004. 63-71.

DES RIVIERES, J. 2004. Eclipse APIs: Lines in the Sand. *EclipseCon,* 2004.

DONNEGAN, J. A., VEBLEN, T. T. & SIBOLD, J. S. 2001. Climatic and human influences on fire history in Pike National Forest, central Colorado. *Canadian Journal of Forest Research,* 31**,** 1526-1539.

ELITH, J., PHILLIPS, S. J., HASTIE, T., DUDÍK, M., CHEE, Y. E. & YATES, C. J. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions,* 17**,** 43-57.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE, I. 2016. *How Hot Spot Analysis (Getis-Ord Gi*) works* [Online]. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm [Accessed 23 November 2017].

FACEBOOK. 2017. *Public Feed API* [Online]. Available: https://developers.facebook.com/docs/public_feed/ [Accessed 10 October 2017].

FATKULIN, T., BUTAKOV, N., DZHAFAROV, B., PETROV, M. & VOLOSHIN, D. 2018. An approach to location extraction from russian online social networks: Road accidents use case. *Advances in Intelligent Systems and Computing.*

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. 1996. Knowledge discovery and data mining: towards a unifying framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* Portland, Oregon: AAAI Press.

FLICKR. 2016. *The Flickr Developer Guide* [Online]. Available: https://www.flickr.com/services/developer/ [Accessed 18 Jan 2017].

FLIGHT, L. & JULIOUS, S. A. 2015. The disagreeable behaviour of the kappa statistic. *Pharmaceutical Statistics,* 14**,** 74-78.

FOURSQUARE. 2016. *The Foursquare API* [Online]. Available: https://developer.foursquare.com/overview/ [Accessed 18 Jan 2017].

FOURSQUARE. 2017. *About Foursquare* [Online]. Available: https://foursquare.com/about [Accessed 1 Sept 2017].

GAZAZ, H., CROITORU, A., DELAMATER, P. L. & PFOSER, D. 2016. Geo-fingerprinting social media content. *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data.* San Francisco, California: ACM.

GEONAMICA. 2011. *Map Comparison Kit* [Online]. Available: http://mck.riks.nl/ [Accessed 29 November 2017].

GLENSKI, M., PENNYCUFF, C. & WENINGER, T. 2017. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems*.

GO-GULF. 2014. *What People Share on Social Networks - Statistics and Trends* [Online]. Available: https://www.go-gulf.ae/blog/what-people-share-on-social-networks/ [Accessed 30 November 2017].

GONÇALVES, P., ARAÚJO, M., BENEVENUTO, F. & CHA, M. 2013. Comparing and combining sentiment analysis methods. *Proceedings of the first ACM conference on Online social networks.* Boston, Massachusetts, USA: ACM.

GOODCHILD, M. F. & GLENNON, J. A. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth,* 3**,** 231-241.

GOOGLE. 2017. *Google+ Platform* [Online]. Available: https://developers.google.com/+/ [Accessed 26 September 2017].

GORMLEY, A. M., FORSYTH, D. M., GRIFFIOEN, P., LINDEMAN, M., RAMSEY, D. S., SCROGGIE, M. P. & WOODFORD, L. 2011. Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology,* 48**,** 25-34.

GRIFFIN, G. P. & JIAO, J. 2015. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport & Health,* 2**,** 238-247.

HU, Y., MANIKONDA, L. & KAMBHAMPATI, S. What We Instagram: A First Analysis of Instagram Photo Content and User Types. ICWSM, 2014.

HUGUENIN, K., BILOGREVIC, I., MACHADO, J. S., MIHAILA, S., SHOKRI, R., DACOSTA, I. & HUBAUX, J. 2017. A Predictive Model for User Motivation and Utility Implications of Privacy Protection Mechanisms in Location Check-Ins. *IEEE Transactions on Mobile Computing*.

INKPEN, D., LIU, J., FARZINDAR, A., KAZEMI, F. & GHAZI, D. 2017. Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems,* 49**,** 237-253.

INSTAGRAM. 2017. *The Instagram API Platform* [Online]. Available: https://www.instagram.com/developer/ [Accessed 18 Jan 2017].

JAYNES, E. T. 1957. Information theory and statistical mechanics. *Physical review,* 106**,** 620.

JESTICO, B., NELSON, T. & WINTERS, M. 2016. Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography,* 52**,** 90-97.

JIANG, L., CAI, Z., ZHANG, H. & WANG, D. 2013. Naive Bayes text classifiers: A locally weighted learning approach. *Journal of Experimental and Theoretical Artificial Intelligence,* 25**,** 273-286.

JOY, S. 2010. Using Netvizz & Gephi to Analyze a Facebook Network. *Persuasion* [Online]. Available from: https://persuasionradio.wordpress.com/2010/05/06/using-netvizz-gephi-to-analyze-a-facebook-network/ [Accessed 31 October 2017].

KAPLAN, A. M. & HAENLEIN, M. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons,* 53**,** 59-68.

KARAMI, A., DAHL, A. A., TURNER-MCGRIEVY, G., KHARRAZI, H. & SHAW, G., JR. 2018. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management,* 38**,** 1-6.

KHAN, M. M., GHANI, I., JEONG, S. R., IBRAHIM, R. & HABIB-UR-REHMAN 2016. Social media usage in academic research. *Journal of Theoretical and Applied Information Technology,* 87**,** 7.

KO, E., YOON, C. & KIM, E. Y. Discovering visual features for recognizing user's sentiments in social images. 2016 International Conference on Big Data and Smart Computing, BigComp 2016, 2016. 378-381.

KUMAR, P. 2012. *The economics of ecosystems and biodiversity: Ecological and economic foundations*.

LAGO VÁZQUEZ, D. 2017. The Usefulness of Social Networks as Research Tools for the Media. *Media and Metamedia Management.* Cham: Springer International Publishing.

LANDEWEERD, M., SPIL, T. & KLEIN, R. 2013. The success of google search, the failure of google health and the future of google plus. *IFIP Advances in Information and Communication Technology.*

LEE, H., SEO, H., LEE, N. & SONG, M. 2018. Exploring a Supervised Learning Based Social Media Business Sentiment Index. *Lecture Notes in Electrical Engineering.*

LI, H., DENG, K., CUI, J., DONG, Z., MA, J. & HUANG, J. 2018. Hidden community identification in location-based social network via probabilistic venue sequences. *Information Sciences,* 422**,** 188-203.

MARSON, S. M. 1997. A selective history of internet technology and social work. *Journal of Technology in Human Services,* 14**,** 35-46.

MARTÍNEZ-HARMS, M. J. & BALVANERA, P. 2012. Methods for mapping ecosystem service supply: A review. *International Journal of Biodiversity Science, Ecosystems Services and Management,* 8**,** 17-25.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY 2016. Chapter 9 - Principle of Maximum Entropy. *Information, Entropy, and Computation.*

MCGEE, M. 2013. Google+ Hits 300 Million Active Monthly "In-Stream" Users, 540 Million Across Google. *MarketingLand*, 29 October.
MICHEL, F. 2017. How many photos are uploaded to Flickr every day, month, year? *In:* PHOTOS_UPLOADED_FLICKR.PNG (ed.).
MILLENNIUM ECOSYSTEM ASSESSMENT 2005. *Living Beyond Our Means: Natural Assets and Human Well-being : Statement from the Board*, Millennium Ecosystem Assessment Board.
MILLER, C. C. 2014. The Plus in Google Plus? It's Mostly for Google. *The New York Times*, 14 February.
MUELLER, W., SILVA, T. H., ALMEIDA, J. M. & LOUREIRO, A. A. 2017. Gender matters! Analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science,* 6.
NEO4J INC. 2017. *Neo4j* [Online]. Available: https://neo4j.com/ [Accessed 4 October 2017].
NIE, W., WANG, X., ZHAO, Y. L., GAO, Y., SU, Y. & CHUA, T. S. 2013. Venue semantics: Multimedia topic modeling of social media contents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*
NIGAM, K., MCCALLUM, A. & MITCHELL, T. 2006. Semi-supervised text classification using EM. *Semi-Supervised Learning*, 33-56.
NIGAM, K., MCCALLUM, A. K., THRUN, S. & MITCHELL, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning,* 39, 103-134.
O'REILLY, T. 2005. *What Is Web 2.0* [Online]. Available: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html [Accessed 5 October 2017].
OAUTH. 2017. *OAuth* [Online]. Available: https://oauth.net/ [Accessed 30 Aug 2017].
OSBORNE, M. & DREDZE, M. Facebook, twitter and google plus for breaking news: Is there awinner? Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, 2014. 611-614.
PALOMINO, M., TAYLOR, T., GÖKER, A., ISAACS, J. & WARBER, S. 2016. The Online Dissemination of Nature–Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder". *International Journal of Environmental Research and Public Health,* 13.
PENNEBAKER, J. W., MEHL, M. R. & NIEDERHOFFER, K. G. 2003. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology.*
PEW RESEARCH CENTER. 2017. *Social Media Fact Sheet* [Online]. Available: http://www.pewinternet.org/fact-sheet/social-media/ [Accessed 28 November 2017].
PFITZNER, R., GARAS, A. & SCHWEITZER, F. Emotional divergence influences information spreading in Twitter. ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, 2012. 543-546.
PHILLIPS, S. J., ANDERSON, R. P., DUDÍK, M., SCHAPIRE, R. E. & BLAIR, M. E. 2017. Opening the black box: an open-source release of Maxent. *Ecography,* 40, 887-893.
REDDIT INC. 2017. *Reddit API Documentation* [Online]. Available: https://www.reddit.com/dev/api/ [Accessed 2 October 2017].
RENNIE, J. D. M., SHIH, L., TEEVAN, J. & KARGER, D. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings, Twentieth International Conference on Machine Learning, 2003. 616-623.
REUTER, C. & SPIELHOFER, T. 2017. Towards social resilience: A quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. *Technological Forecasting and Social Change,* 121, 168-180.
RUHL, J. B., KRAFT, S. E. & LANT, C. L. 2013. *The Law and Policy of Ecosystem Services*, Island Press.
SAMUEL FOSSO, W., SHAHRIAR, A., HYUNJIN, K., MITHU, B. & MOHAMMED, U. 2016. The Primer of Social Media Analytics. *Journal of Organizational and End User Computing (JOEUC),* 28, 1-12.
SCHWARZ, N. 1999. Self-Reports: How the Questions Shape the Answers. *The American psychologist,* 54, 93-105.

SHELDON, P., RAUSCHNABEL, P. A., ANTONY, M. G. & CAR, S. 2017. A cross-cultural comparison of Croatian and American social network sites: Exploring cultural differences in motives for Instagram use. *Computers in Human Behavior,* 75**,** 643-651.

SHERROUSE, B. C., CLEMENT, J. M. & SEMMENS, D. J. 2011. A GIS application for assessing, mapping, and quantifying the social values of ecosystem services. *Applied Geography,* 31**,** 748-760.

SHERROUSE, B. C. & SEMMENS, D. J. 2015. Social Values for Ecosystem Services, version 3.0 (SolVES 3.0): documentation and user manual. *Open-File Report.* Reston, VA.

SINDHWANI, V., KEERTHI, S. S. & CHAPELLE, O. 2006. Deterministic annealing for semi-supervised kernel machines. *Proceedings of the 23rd international conference on Machine learning.* Pittsburgh, Pennsylvania, USA: ACM.

SOBKOWICZ, P., KASCHESKY, M. & BOUCHARD, G. 2012. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly,* 29**,** 470-479.

STONE, P. J., DUNPHY, D. C., SMITH, M. S. & OGILVIE, D. M. 1966. *The general inquirer: A computer approach to content analysis*.

STRAVA. 2017. *Strava's V3 API* [Online]. Available: https://strava.github.io/api/ [Accessed 16 October 2017].

SUN, Y., MOSHFEGHI, Y. & LIU, Z. 2017. Exploiting crowdsourced geographic information and GIS for assessment of air pollution exposure during active travel. *Journal of Transport and Health,* 6**,** 93-104.

SURAN, M. & KILGO, D. K. 2017. Freedom from the Press?: How anonymous gatekeepers on Reddit covered the Boston Marathon bombing. *Journalism Studies,* 18**,** 1035-1051.

TAYLOR, R. B., STONEMAN, R. J., MARSH, S. P. & DERSCH, J. S. 1984. An assessment of the mineral resource potential of the San Isabel National Forest, south-central Colorado. *Bulletin of the U.S. Geological Survey,* 1638.

THELWALL, M. 2013. Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength. *Conference Proceedings.*

THELWALL, M. 2017. *SentiStrength* [Online]. Available: http://sentistrength.wlv.ac.uk/ [Accessed 11 Sept 2017].

THELWALL, M., BUCKLEY, K., PALTOGLOU, G., CAI, D. & KAPPAS, A. 2010. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology,* 61**,** 2544-2558.

THOMPSON, M. P., GILBERTSON-DAY, J. W. & SCOTT, J. H. 2016. Integrating Pixel- and Polygon-Based Approaches to Wildfire Risk Assessment: Application to a High-Value Watershed on the Pike and San Isabel National Forests, Colorado, USA. *Environmental Modeling and Assessment,* 21**,** 1-15.

TIAN, R. Y., WU, L., LIANG, X. H. & ZHANG, X. F. 2018. Opinion data mining based on DNA method and ORA software. *Physica A: Statistical Mechanics and its Applications,* 490**,** 1471-1480.

TOEPKE, S. L. Temporal sampling implications for crowd sourced population estimations from social media. Proceedings of the International ISCRAM Conference, 2017. 564-571.

TRIPATHY, A., AGRAWAL, A. & RATH, S. K. 2016. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications,* 57**,** 117-126.

TWITTER INC. 2017. *Twitter Developer Documentation* [Online]. Available: https://dev.twitter.com/docs [Accessed 3 October 2017].

US GEOLOGICAL SURVEY. 2017. *Social Values for Ecosystem Services (SolVES)* [Online]. Available: https://solves.cr.usgs.gov/ [Accessed 31 October 2017].

VARGA, S., CHERRY, D. & D'ANTONI, J. 2016. *Introducing Microsoft SQl Server 2016,* Redmond, Washington, Microsoft Press.

VEIGA, M. H. & EICKHOFF, C. 2016. A Cross-Platform Collection of Social Network Profiles. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* Pisa, Italy: ACM.

VISSER, H. & DE NIJS, T. 2006. The map comparison kit. *Environmental Modelling and Software,* 21**,** 346-358.

WANG, J. & ZHANG, X. 2017. Impacts of wildfires on interannual trends in land surface phenology: An investigation of the Hayman Fire. *Environmental Research Letters,* 12.

WOO, H., CHO, Y., SHIM, E., LEE, K. & SONG, G. 2015. Public Trauma after the Sewol Ferry Disaster: The Role of Social Media in Understanding the Public Mood. *International Journal of Environmental Research and Public Health,* 12**,** 10974-10983.

WOOD, S. A., GUERRY, A. D., SILVER, J. M. & LACAYO, M. 2013. Using social media to quantify nature-based tourism and recreation. *Scientific Reports,* 3**,** 2976.

WU, Y., CAO, N., GOTZ, D., TAN, Y. P. & KEIM, D. A. 2016. A Survey on Visual Analytics of Social Media Data. *IEEE Transactions on Multimedia,* 18**,** 2135-2148.

XU, Z., LI, J., LIU, B., BI, J., LI, R. & MAO, R. 2017. Semi-supervised learning in large scale text categorization. *Journal of Shanghai Jiaotong University (Science),* 22**,** 291-302.

YANG, S. & RO, Y. M. Two-layered photo classification based on semantic and syntactic features. CEUR Workshop Proceedings, 2006.

YOO, J., CHOI, S., CHOI, M. & RHO, J. 2014. Why people use Twitter: Social conformity and social value perspectives. *Online Information Review,* 38**,** 265-283.

YOSHIMURA, N. & HIURA, T. 2017. Demand and supply of cultural ecosystem services: Use of geotagged photos to map the aesthetic value of landscapes in Hokkaido. *Ecosystem Services,* 24**,** 68-78.

ZHENG, L., HUA, Y. & ZHEN, L. 2010. An efficient graph-based Flickr photo clustering algorithm. *Applied Mechanics and Materials.*

APPENDIX A: LISTS OF USED QUERIES

Table A: List of queries used for query based searches

Pike-San Isabel National Forest	PSICC
Pike National Forest	San-Isabel National Forest
Pike-San Isabel	Holy Cross Wilderness
Mount Evans Wilderness	Mount Massive Wilderness
Rampart Range Recreation Area	Rampart Range
Picket Wire Canyon	Lost Creek Wilderness
Davenport Campground	Buffalo Peaks Wilderness
Collegiate Peaks Wilderness	Greenhorn Mountain Wilderness
Sangre de Cristo Wilderness	Spanish Peaks Wilderness
Pikes Peak	

Table B: List of queries used for query based searches on subreddits

Nature	Outdoor
Forest	Camping
Rocky Mountains	Travel
Hiking	Backpacking
Colorado	Earth
Mountain	Fishing
Pics	Wilderness
CO	Trail

Table C: List of queries used to define the Nature *class*

Pike-San Isabel National Forest	PSICC
Pike National Forest	San-Isabel National Forest
Pike-San Isabel	Holy Cross Wilderness
Mount Evans Wilderness	Mount Massive Wilderness
Rampart Range Recreation Area	Rampart Range
Picket Wire Canyon	Lost Creek Wilderness
Davenport Campground	Buffalo Peaks Wilderness
Collegiate Peaks Wilderness	Frontier Range
Greenhorn Mountain Wilderness	Cripple Creek
Sangre de Cristo Wilderness	Spanish Peaks Wilderness
Pikes Peak	View
Hike	

Table D: List of queries used to define the None *class*

God	Job
Vacancy	Sale
Birthday	Anniversary
Yelp	Party
Beer	Share
Shea Stadium	Wedding
Liquor	Heaven
Makeup	Make up
Music	BOTDF
Kits	

APPENDIX B: CODE SNIPPETS

FLICKR

```
1.  geophotos = Flickr.flickr.photos.search(bbox = '-106.6965, 37.3341,
                                                    -104.9881, 39.6353',
2.                                              per_page = 500,
3.                                              extras = extras,
4.                                              max_upload_date = max_upload_date)
```

Code snippet A: Using the Flickr API to search based on location

```
1.  userphotos = Flickr.flickr.photos.search(text = query,
2.                                           per_page = 500,
3.                                           extras = extras,
4.                                           max_upload_date = max_upload_date)
```

Code snippet B: Using the Flickr API to search based on queries

```
1.  userphotos = Flickr.flickr.people.getPhotos(user_id = user,
2.                                              per_page = 500,
3.                                              extras = extras,
4.                                              max_upload_date = max_upload_date
5.                                              content_type = 7)
```

Code snippet C: Using the Flickr API to search all photos from the selected user

```
1.  def rateLimit(self):
2.      """Check if the set rate limit for the Flickr API
3.      is exceeded. If so, sleep to fill the 15 minutes
4.      Input:
5.          None
6.      Output:
7.          None
8.      """
9.      elapsed_time = time.time() - self.start_time
10.     if elapsed_time <= 900:
11.         sleep_time = 900 - elapsed_time
12.         print 'Sleeping for %s seconds' % (str(sleep_time))
13.         time.sleep(sleep_time)
14.         self.start_time = time.time()
15.         self.total_count = 0
16.     elif elapsed_time > 900:
17.         self.start_time = time.time()
18.         self.total_count = 0
```

Code snippet D: Function to safe the rate limit from being exceeded

FOURSQUARE

```
1.  venues = Foursquare.client.venues.search(params = {
2.           'll': ll,
3.           'radius': 5000,
4.           'limit': 50,
5.           'intent': 'browse'})
```

Code snippet E: Using the Foursquare API to search based on location

```
user_info = Foursquare.client.users.search(params = {
             'name': user_name})
```

Code snippet F: Using the Foursquare API to find more information about a user

GOOGLE+

```
activities = activity_resource.search(query = query,
                                      maxResults = 20,
                                      pageToken = None).execute()
```

Code snippet G: Finding a list of activities using the given query

```
people = people_resource.get(userId = user_id).execute()
```

Code snippet H: Finding more information about the given user

REDDIT

```
subreddits = Reddit.reddit.subreddits.search(query = query,
                                             params = {
                                                      'after': after,
                                                      'limit': 100,
                                                      'show': 'all'})
```

Code snippet I: Finding relevant subreddits using the given query

```
posts = Reddit.reddit.subreddit(title).search(query = query,
                                              sort = 'comments')
```

Code snippet J: Searching for posts inside a subreddit

```
comments = Reddit.reddit.submission(id = id).comments
```

Code snippet K: Retrieving all comments for a given post

TWITTER

```
search_results = Twitter.twitter.search(geocode = '38.6131,-106.2564,5km',
                                        count = 100,
                                        max_id = max_id)
```

Code snippet L: Harvesting Tweets inside the defined area

```
query_results = Twitter.twitter.search(q = query,
                                       count = 100,
                                       max_id = max_id)
```

Code snippet M: Request for Tweets based on the given query

```
user_tweets = Twitter.twitter.get_user_timeline(user_id = user,
                                                count = 200,
                                                max_id = max_id)
```

Code snippet N: Return the 3,200 most recent Tweets from the given user