University of Zurich[UZH]

Landscape Perception: Analyzing Environmental Discourse in Spanish-Speaking Central and South America Using Twitter Data

GEO 511 Master's Thesis

Author
Philipp Sebastian Rohr
14-056-386

Supervised by
Prof. Dr. Ross Purves
Prof. Dr. Carlota de Benito Moreno (carlota.debenito@uam.es)

Faculty representative
Prof. Dr. Ross Purves

27.01.2025
Department of Geography, University of Zurich

Abstract

This paper attempts to fill a research gap by studying landscape perception through the lens of Twitter data. While existing studies on landscape perception are often based on surveys or visual analysis, there is little research that uses social media data to investigate the topic. Social media platforms, which produce a large amount of user-generated data, offer a unique opportunity to study landscape perception over time and across geographical space. This thesis is based on a dataset of almost 1 billion Tweets in the Spanish language sent in Central and South America between 2012 and 2017. The aim is to explore how landscapes are conceptualized and perceived on social media, especially in the context of the linguistic and cultural diversity of the region. A comprehensive framework is presented that combines Geographic Information Retrieval (GIR) and Natural Language Processing (NLP) techniques. The framework includes machine learning for classifying Tweets to check their relevance in terms of landscape content. Following the classification, spatial, temporal, and thematic analyses are carried out to identify patterns in how landscapes are perceived. Central and South America's incredibly diverse landscapes – from towering mountains to vast jungles to beaches and deserts – is an ideal case study. Furthermore, the linguistic variations across the regions provide an opportunity to understand how landscapes are perceived and discussed in different cultures and talked about on social media. For example, specific landscape features, such as beaches, mountains, or lakes, are associated with distinct emotions and leisure activities. Furthermore, the spatial distribution of Tweets with landscape terms corresponds to the actual distribution of the landscapes. Lastly, the temporal distribution of Tweets allows conclusions to be drawn about significant events and seasonal trends. The thesis shows that unstructured text can be used to gain geographical and thematic insights and highlights the potential of GIR and NLP techniques for interdisciplinary research.

Keywords: Natural Language Processing (NLP), Geographic Information Retrieval (GIR), Spanish language, landscape perception, social media, Twitter

Acknowledgements

First, I would like to express my sincere thanks to my supervisor, Prof. Dr. Ross Purves, who has been instrumental in supporting me throughout the process with his knowledge and support. His input on both a professional and personal level has helped me on this journey, and for that, I am very grateful.

I would also like to thank my co-supervisor, Prof. Dr. Carlota de Benito Moreno. This work would not have been possible without her great ideas and the data she provided.

I want to express my heartfelt thanks to my friends and my sisters. Through their support, their valuable ideas and inputs, and their ability to distract me from work from time to time, they have given me new perspectives and motivation. Your presence and encouragement have made an important contribution to this work, and for that I am very grateful to you. A special thank you to Thilo van der Haegen for your work!

Last but not least, I would like to thank my partner Laura Bozzi, who has supported me unwaveringly and was always available to answer my questions. She not only accompanied me privately and gave me the breaks I needed, but also contributed a very valuable part of her knowledge, time and thoughts to this work – grazie mille.

Author's Note

This research is based on historical data from Twitter (2012–2019). It does not reflect the platform's current state, which is now rebranded as X. I explicitly distance myself from recent developments, policies, and management practices associated with the platform. Also, it is important to note that the author carried out all translations from Spanish to English.

Table of Contents

1 Introduction

Uetliberg is one of Zurich's local mountains. Its name wonderfully exemplifies what this thesis looks to discuss in the following. *Berg* means mountain. However, Uetliberg is more of a hill than a mountain for many people. The aim of this work is precisely to examine such different perceptions of landscape. The thesis' focus is not Zurich but Central and South America, whose incredibly diverse landscapes, ranging from mountains to beaches and rain forests, stretch across half a continent. This makes for an ideal case study to explore the topic of landscape perception in social media.

A Twitter data set serves as the basis for this study and was provided by the Zurich Center for Linguistics. The dataset contains almost one billion Tweets sent in Spanish in Central and South America between 2012 and 2019. However, to use and extract the required information in a meaningful way, the data must be processed using various geographic information retrieval (GIR) techniques. GIR is a subfield of geography and information technology that focuses on extracting information that has no geographical components at first glance. In this context, Tweets, have coordinates or refer to a location that can be geographically assigned in the content of the Tweet itself. GIR techniques can be used to extract information that would otherwise not be apparent.

While there is substantial research on methods for extracting information from social media data and studies on how landscapes are perceived, the intersection of these two fields remains unexplored. Most research on landscape perception relies on surveys or visual analysis, while linguistic studies of social media often focus on lexical variation across regions. What is scarce in research, however, is a combination of these two fields: How are landscapes conceptualized and perceived on social media? How is landscape discussed on Twitter? What emotions are associated with volcanoes? Why does one decide to write a post about nature? These and other questions will be answered in the course of this work. This raises the following research question: In what ways are landscapes discussed and perceived in Spanish-speaking Central and South America based on geographically and thematically filtered Twitter data, and what spatial, temporal, or thematic patterns can be identified?

The aim is to show how the Twitter data set was made geographically legible and how information regarding landscape perception was extracted. The methodology combines machine learning with spatial, temporal, and co-occurrence analyses to uncover patterns in how landscapes are discussed and perceived across diverse linguistic and geographical contexts. Spanish is particularly well-suited for this analysis due to its global prevalence and the documented regional linguistic variations that provide a rich basis for studies. Furthermore, Central and South America makes for an interesting case study because of its diverse landscape features.

The thesis is structured as follows: The second chapter reviews currently relevant research and provides an overview of the widespread methods regarding GIR and Natural Language Processing (NLP). In addition, studies that deal with individual themes pertinent to this research are analyzed, i.e., landscape perception, social media and geography, and language research through social media. Chapter 3 outlines the framework that was developed for this research. It details the machine learning algorithm used to classify the Tweets and the subsequent analyses conducted. The Python scripts, which were written for this work and can be found on GitHub, are explained step by step, thus enabling a reproduction of the results. The fourth and fifth chapters present and describe the most significant results of the analyses. Further results and tables with relevant Tweets can be found in the appendix and on GitHub.

Spatial analysis reveals patterns in the distribution of Tweets containing a landscape term. In doing so, it effectively highlights locations that correspond to the features described in the Tweets. For instance, Tweets using the term *playa* (beach) are predominantly associated with actual beach destinations. The temporal distribution of Tweets reveals event-based or seasonal patterns. Furthermore, the thematic analysis describes the emotions and activities related to the respective terms. In doing so, it offers new insights into the cultural and linguistic diversity of Central and South America while contributing a replicable framework for future studies at the intersection of geography, linguistics, and social media analysis.

2 Literature Review

2.1 Theoretical Foundations

2.1.1 Geographical Information Retrieval (GIR)

Jones and Purves define Geographical Information Retrieval (GIR) as follows: "The provision of facilitates to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection based on queries specifying both theme and geographic scope" (Jones & Purves, 2009). But what does that mean? Breaking this definition down makes it simpler to understand: Everything happening on our planet has a geographical component to it, which can be seen as retrievable information. One can gather valuable insights on specific topics by retrieving and ranking the information's relevance.

Geographical Information Retrieval can be seen as a subarea of information gathering and focuses on searching for information with a geographical component. The information that is looked for can come from different sources, including books, travel blogs, social media posts, or other text documents. GIR systems aim to retrieve information and make it usable and accessible for in-depth research. GIR focuses on four different aspects: indexing, searching, retrieving, and browsing. Indexing means organizing geographic information so that it can be retrieved efficiently later on. Searching equals locating the information that meets specific geographic criteria. Retrieving equals accessing relevant geographical information, and browsing signifies exploring information user-friendly (Monteiro et al., 2016; Purves et al., 2018).

GIR aims to improve the quality and/or quantity of geographic information, especially by extracting unstructured documents that can be found on the web. While a traditional GIS (Geographic Information System) can usually handle structured data with a clear geographic component like coordinates, GIR aims to expand access to a much larger body of unstructured content. This is done by identification and geolocation of places found in documents, primarily by using place names (Toponyms) within the text and parsing them with specific locations using a gazetteer (Purves et al., 2018).

This can, for example, be used to improve delivery logistics by analyzing customer delivery addresses for e-commerce companies (Babu et al., 2015). The author's focus lies in developing a system to classify addresses into predefined subregions within a larger region automatically. The goal was to combine domain knowledge with machine learning to achieve high accuracy in address classification when addresses were unordered, had typographical errors, or lacked additional information, all without the need for typical coordinates. The aim was to improve the efficiency of delivery logistics.

Another example highlights how GIR can be used to improve flood disaster management (Pereira et al., 2019). This study shows how georeferenced social media data, in this case mostly images, can be used to improve disaster management, primarily flooding events. Photos of flooded areas can be used as a real-time source to gain information about the flooding situation to help understand the extent of the damage or even to help identify roads that need to be blocked. Compared to remote sensing imagery, photos from the situation are often more detailed.

In sum, GIR techniques have been applied in various recent areas. They are getting more important for extracting spatial knowledge that otherwise would stay hidden in large volumes of text or in

other media such as photos or videos (Monteiro et al., 2016). A substantial part of such data can be found on different social media platforms, such as Twitter, Facebook, or Instagram.

2.1.2 Social Media and Geography

Social media platforms, especially those with an abundance of user-generated content, can be used for geographic information retrieval. As mentioned above, even photos can be used to retrieve information. But despite the large amount of data created worldwide every day, most of it is unusable for geographic extraction due to either license terms, data crawlability, or the unavailability of geotags (Zhu et al., 2023). Crawlability, in this context, is the ability to access the data. Twitter is mainly used for text-based analysis, and Flickr, for example, excels in image-based studies. What all sources have in common is that they offer a geographic component that is crucial for GIR. This section will discuss potentials, difficulties, and some examples of social media's role in geographic information retrieval.

Surprisingly, one of the more considerable challenges when working with social media data is to extract the precise location of the gathered information. Location extraction can be done in different ways. One would think that every post on social media is geotagged and offers a precise location, but that is not the case. Only a small percentage of social media posts are geotagged (in the case of Twitter, it is roughly 1% (Zhu et al., 2023). However, locations can also be identified through other methods, such as geoparsing, which involves matching mentioned names of streets, cities, or points of interest to their precise geographic coordinates (Middleton et al., 2018).

Another difficulty pertains to data management: social media data often comes in high volumes, varieties, and velocity (Lee & Kang, 2015; Zhu et al., 2023). This makes social media data both valuable and challenging for geographical information retrieval. Users generate a vast amount of content daily, such as images, text, and videos. The wide variety of social media data concerns a wide range of formats, topics, and perspectives. From a technical standpoint, there are different formats (Text, video, images) and varying metadata regarding location, time, and/or user information (Zhu et al., 2023). From a user perspective, different demographics, intentions, languages, and much more must be taken into account (Doyle, 2014; Zhu et al., 2023). Furthermore, as the pace at which social media data is generated is exceptionally high, this signifies that social media produces a high data volume (McKitrick et al., 2023). Posts are created and shared constantly, which results in a continuous flow of new data. This requires efficient data management systems when pursuing real-time analysis (Zhu et al., 2023). In the case of this thesis, data extraction from Twitter was facilitated through a free API (Application Programming Interface), enabling access to substantial periods of data with relatively minimal effort. However, concerning X, collecting data nowadays is associated with a costly subscription model (Corp, 2024). Nonetheless, such APIs provide an unbelievable amount of research potential.

Social media and connected online platforms are a young and still rapidly growing phenomenon. This can be explained because social media has become a crucial component of our lives (Tellez et al., 2023). Research on the growth of social media shows that user numbers on most platforms are increasing. According to Forbes Magazine, young people spend an average of over 145 minutes daily on social networks. The extensive data volume of such platforms is therefore not surprising and a good starting point to conduct research on (Wong, 2024).

When done correctly and difficulties are handled well, social media data can bring fascinating insights into different research areas. Findings from Grieve et al. (Grieve et al., 2019) in their paper

about Mapping Lexical Dialect variation in British English underlines this. They compared regional lexical variation in Twitter corpora with traditional surveys. They have statistically proven that Twitter data and data from surveys correlate strongly. Questions that once required costly surveys to answer can now be addressed by analyzing social media data, specifically from platforms like Twitter.

Adding to the potential of GIR, a study done by Mukherjee et al. (Mukherjee et al., 2022) focuses on using Twitter to analyze the increased numbers of immigrants in the EU. The authors examine Tweets that are related to the crisis from 2016 to 2021. They analyzed popular hashtags related to the migration crisis and found that hashtags like *#refugees* or *#migrants* correlate with the known migration routes into Europe. This shows practically how Twitter aligns with real-world events.

Another topic, which also found more interest during the COVID-pandemic, is tracking diseases and public concern using Twitter. Signorini et al. (Signorini et al., 2011) tried this during an H1N1 pandemic. They collected a dataset of roughly one million Tweets containing keywords related to influenza. Key findings are that public concern, despite a rising number of cases, has decreased over time. Also, their findings suggest that Twitter can be a valuable tool for tracking disease activity and giving insights into real-time disease trends and public perception.

A third study was undertaken by Lansley and Longley (Lansley & Longley, 2016). The authors investigated the geographic distribution of Twitter topics within London. Using a dataset containing georeferenced Tweets from Inner London from 2013, they used Latent Dirichlet Allocation (LDA), a statistical tool to identify topics from text. They found valuable insights into temporal and spatial patterns in how people in London use Twitter and what they talk about.

In conclusion, it can be said that social media in geographical information retrieval can be used for research on a wide range of topics. While sheer data size can be overwhelming, applying GIR to social media poses an opportunity. Additionally, the continuous flow of new data enables real-time analysis. This can help researchers react to changes in the environment immediately and help with issues such as disaster management. Lastly, even though the diversity of formats (Text, photos, videos) is complicated to capture, it allows for a variety of methods and enables new ideas. One of these ideas or trends is to analyze how landscape is talked about and perceived.

2.1.3 Landscape Perception

One subject that is increasingly being researched by utilizing GIR is landscape perception. Researchers look to understand how landscape and its elements, such as mountains, hills or glaciers, are conceptualized and how these conceptualizations are expressed in human natural language (Derungs et al., 2013). Fundamentally, what such research enables is to understand how and why those conceptualizations vary across languages and cultures - a topic that was coined by Mark and Turk (Mark et al., 2011). GIR is often used to parse unstructured data, in this case unstructured natural language that can be found on Twitter. However, as mentioned in recent research, large high-quality landscapes with relevant corpora are scarce. To change that, researchers use natural language processing (NLP) to extract useful information out of those corpora (Baer & Purves, 2023). The kind of corpora and, thus, the underlying data source varies depending on the research question. While data can be extracted from very large corpora such as Wikipedia (Ardanuy & Sporleder, 2017) or Reddit (Fox et al., 2021), it can also be a more specific corpora specifically linked to the topic of landscapes, for example Geograph (Chesnokova & Purves, 2018). Also, Twitter is a data source

that is often used by collecting data that was filtered for specific keywords or locations (Zahra et al., 2020).

But why is this research area relevant? Understanding how humans perceive and talk about landscape helps us to understand how people interact with the landscape and what kind of meaning different landscape features have for various cultures. It helps to better understand the relationship between humans and landscape and how humans interact and make sense of their surroundings (Burenhult & Levinson, 2008; Mark et al., 2011). Landscape perception challenges the assumption that landscape terms are the same for everyone. For example, what one culture might refer to as a *hill* is called a *mountain* in another culture. This difference and variability emphasize the need to understand the cultural and/or linguistic differences in how individuals talk about and thus perceive landscape (Mark et al., 2011). In 2008, Burenhult (Burenhult & Levinson, 2008) asked several questions in this regard: How are landscape features selected as nameable objects (*river, mountain, cliff*)? What is the relation between landscape terms and place names (Villette & Purves, 2018)? How translatable are landscape terms across languages, and what ontological categories do they commit to? This last question is something that is particularly interesting and needs some more explanation.

Different languages categorize and perceive landscape features differently, which is also reflected in the variation of cultural priorities and environmental adaptions. For example, the Yindjibarndi language distinguishes between permanent and temporary water features. On the other hand, in English, permanent or temporal is more likely treated as a feature of a water body than a distinct category (Mark et al., 2011). This suggests that languages don't simply categorize different features in the landscape, but rather, they are actively shaped and vary because different cultures perceive the landscape differently.

Also, toponyms can act as cultural indicators and give insights into how language has evolved. For example the German language, when talking about the landscape terms *Spitze* (Peak), *Horn* (Horn), and *Berg* (Mountain): It is interesting because *Spitze* and *Horn* (Are words, that are frequently found in German (or Swissgerman) names for mountains like Zugspitze and the Matterhorn. But, in contemporary standard German, the terms *Spitze* and *Horn* have different meanings and would not be used independently to refer to a mountain per se. While *Horn* is more often used in the Valais and *Spitze* is more often used in Eastern Switzerland. This suggests that the meaning and usage of such toponyms serve as significant cultural markers, reflecting historical perspectives. Investigating these differences can provide deeper insights into how linguistic elements contribute to the cultural identity of a place (Derungs et al., 2013). This brings to the fore that only more studies can completely untangle such complex relations between landscape terms and toponyms, giving meaning to how these linguistic elements create the cultural identity of a place. Further research should be done by putting together the various linguistic analyses along with a geographical and historical point of view to continue discovering the rich cultural stories hidden within the names given to surrounding places. But what is natural language? To answer this, we must go one step back and discuss communication.

2.2 Language Research with Social Media

Pearson et al. understand communication as the process of using messages to generate meaning; this can be done verbally, non-verbally, or through behavior (Pearson et al., 2008). Britannica defines communication as "the exchange between individuals through a common system of symbols" (Britannica, 2024a). Going into too much detail about the exact definition of communication would go beyond the scope of this work. In order to understand the topic of this Thesis, "Spanish Language Diversity," better, it helps to put communication through language into perspective. Communication can be split into different types: nonvocal or vocal communication. Nonvocal communication includes signals, signs and symbols. These components are found in all cultures and do not use the conception of words or language. However, vocal communication is mostly done by using language. Cambridge Dictionary defines language as a system of communication that consists of sounds, words and grammar (Cambridge University Press, 2024). For Pearson (Pearson et al., 2008), "language is a collection of words with arbitrary meanings that are governed by rules and used to communicate."

Britannica also mentions humans in the definition of language: "Language, a system of conventional spoken, manual, or written symbols by means of which human beings, as members of a social group and participants in its culture, express themselves" (Britannica, 2024b). In Britannica's definition, human beings and even social groups, respectively, and their culture are defining factors for language. Since language is an imperfect means of transmission, the thoughts expressed by one person never precisely match what is decoded by another (Pearson et al., 2008). Further, it is suggested that even among speakers of the same language, communicating information can be challenging due to minor language differences (Tellez et al., 2023). Those alterations might occur due to regional variations, evolution in language, or cultural influences. Language is an enormously diverse topic to research and it is not always easy to find the best data source. While written language (e.g., belletristic) can differ from spoken language, language on social media comes closer to natural language.

The vast amount of data is why in recent years and decades, it has been observed that research into language using social media has grown rapidly. Gonçalves even speaks of an avalanche of content naturally and organically generated by millions or tens of millions of geographically distributed individuals (Gonçalves & Sánchez, 2014). However, social media usage patterns are highly diverse and fluctuating. Usage shifts with age and different platforms gain popularity while others are forgotten. In this chapter, I will focus on how social media (mainly focusing on Twitter) can be used as a source in linguistic research, and I will explain why those platforms are commonly used in this area.

The language spoken or written in social media differs from the language that can be analyzed through surveys. In a survey, language is naturally limited and only several features can be observed (Gonçalves & Sánchez, 2014; Grieve et al., 2019). For example, in a survey, people can only answer questions limited to a certain topic. The words used in the question will influence the chosen words for the answer. Other possibilities, such as already given answers, e.g., multiple choice answers, limit natural language usage even more. In contrast, when people write on Twitter, Facebook, or Instagram, they use their own language and use the words they would also use when talking in real life. On social media platforms, a user can write freely and with fewer restrictions. One of the few restrictions on Twitter is the amount of characters one can use in one single post (280 nowadays). This makes researching social media so interesting because individuals use natural language - which is what linguistic research aims to research - linguistic variation in natural language.

There are some interesting examples in linguistic works investigating regional variation with Twitter or Facebook in the English Language (Blodgett et al., 2016; Doyle, 2014; Eisenstein et al., 2014; Huang et al., 2015; Kulkarni et al., 2016), in Spanish (Donoso & Sánchez, 2017; Gonçalves & Sánchez, 2014, 2016) and German (Hovy & Purschke, 2018; Scheffler et al., 2014). There is research about the variation of English in time and space (Gonçalves et al., 2018; Nguyen et al., 2016). Also, research focuses more on the cultural aspects of language and tries to find cultural regions through Twitter data (Louf, Gonçalves, et al., 2023; Louf, Ramasco, et al., 2023). All these studies have in common that the used data comes from social media and the goal is to uncover regional variances in Natural Speech. However, some languages are more suitable for research with a social media-based dataset. For example, English is a language spoken globally, and there are definitely regional differences. The second most spread language globally is Spanish.

2.2.1 Specific to the Spanish Language

Spanish is spoken by over 600 million people worldwide, including as a second language (7.5% of the global population). This widely spoken language extends its reach across 111 countries across the globe, and in 21 of them as a primary language (Fernández Vítores, 2023). There are significant speaker populations around the globe and on multiple continents. This widespread distribution offers a unique possibility to research Spanish and how the Spanish language has evolved and diversified in different cultures (Gonçalves & Sánchez, 2014; Tellez et al., 2023). Thus, a language like Spanish is also potentially fascinating for researching linguistic variation.

Spanish dialects within and across countries add another layer of complexity and interest to research. Different papers proved the existence of so-called superdialects in the Spanish language (Donoso & Sánchez, 2017; Gonçalves & Sánchez, 2014, 2016) that can be divided into an urban speech spoken in metropolitan areas and cities, and a more rural language, that is spread in smaller towns and villages. Cities uniformize language and eliminate specific words and expressions with a regional character. In the study, the urban dialect presents most of the words used in the corpus, and the rural dialect is less represented and more heterogeneous. The authors propose that this generalization is likely related to official media. This phenomenon can be observed in the Iberian Peninsula and the Americas.

Concerning technology and data sources, Twitter and other social media platforms provide a rich and accessible source of data to investigate the Spanish language (Gonçalves & Sánchez, 2014, 2016; Leis et al., 2019; Pruss et al., 2019; Tellez et al., 2023). As mentioned, Twitter offers user-generated content that is close to natural language.

2.3 Techniques in GIR for Social Media Analysis

This chapter will explain the most important methods and techniques that are useful for this thesis.

2.3.1 Natural Language Processing (NLP)

Natural language Processing (NLP) is a subfield between computers and human languages. Usually, the goal of using NLP is to automatically understand, interpret, or even generate human language in a meaningful way by using computers. NLP has grown rapidly over the past few decades and received even more attention in recent years with the increasing importance of Artificial Intelligence (Deekshith, 2024). While some areas of NLP are fairly straightforward, such as email spam detection and translation, others are more complex, such as information extraction or retrieval, summarization, sentiment analysis, or question answering (Khurana et al., 2023). Due to the many nuances and variations of language, the field of NLP has become a highly diversified sector. Initially, NLP was based on rule-based systems. However, the emergence of statistical methods and machine learning has led to significant advancements in the application of NLP in research. For the first time, it was possible that systems learned from data instead of relying on predefined rules (Deekshith, 2024).

2.4 Challenges in Social Media Analysis

This section addresses the challenges associated when working with social media data. This is important to understand the employed methods and the steps conducted to analyze the data. It is important to understand the challenges when working with social media data and the methods employed to analyze the data. It is well known that Twitter bots exist and that they make up a significant proportion of Tweets. Even according to Twitter, about 1 in 20 posts is not human (Rutkin, 2014). There are different methods to find bots; Louf et al., for example, discarded all Tweets that were posted at inhuman rates (Louf, Gonçalves, et al., 2023). Bot-generated Tweets have the potential to distort results. This is why the detection and deletion of non-human Tweets is important, especially when it comes to investigating language differences.

Secondly, Twitter users often represent a demographic that is skewed towards younger, urban individuals who are more likely to be well-educated and male (Crampton et al., 2013; Donoso & Sánchez, 2017; Grieve et al., 2019; Longley et al., 2015; Wartmann et al., 2015). This bias shapes the data, and a more modern and urban perspective on language use is reflected (Smith & Rainie, 2010). Consequently, this may lead to an underrepresentation of older, rural, or less-educated populations whose linguistic practices and cultural nuances might differ significantly. This limitation has frequently been discussed in recent research. Thus emphasizing the importance of critically evaluating the results of studies based on Twitter data.

A third difficulty that can potentially distort results is the distribution of Tweets. Especially regions in which very few Tweets can be geolocalized have the potential to distort results and generate false patterns. Tellez et al. mention that small amounts of data lead to patterns in the results that do not exist and can, therefore, falsify results. For example, in the paper by Tellez et al. on regionalized models for Spanish language varieties based on Twitter, they found, using cosine dissimilarity, similarities between countries that are only due to the limited data available in certain countries (Tellez et al., 2023).

Fourth, in times of Twitter, short messaging, and smartphones, emojis are omnipresent. About 18% of all Tweets contain emojis (Siever, 2023). Emojis can not be defined as a predominant mean of communication, but they help to transmit the sentiment of a Tweet. Emojis could potentially be analyzed in future research to explore the emotions associated with different landscape terms. Additionally, it would be interesting to investigate whether terms like *mounatin* are frequently replaced by the corresponding mountain emoji.

Fifthly, it is essential to consider the challenges posed by polysemy and metaphors. Polysemy refers to a word having multiple meanings, and as Grieve et al. point out in their study on Twitter, dialect research often does not account for this complexity (Grieve et al., 2019). They found that high levels of polysemy can negatively impact the generalization and accuracy of analyzes. Similarly, metaphors pose a related challenge. In both cases, the specific meaning of a word may not always be the focus. For instance, whether the word *montaña* refers to a literal *mountain* or is used metaphorically to mean *a big task* may not matter, as the spelling remains the same. This can make distinguishing between literal and figurative uses when analyzing language patterns complex.

These challenges must all be kept in mind. The best way to verify the data is by hand-coding it, which is time-consuming. This is why, in this work, a part of the data was manually classified to train a machine learning algorithm similar to a study by Austen (Austen, 2017). In addition, the data must be checked for unusual patterns using macro and micro reading. Hypothetical example: A famous person called *montaña* is running for president in Colombia. This will lead to many Tweets with the word *montaña* even though not a landscape is described. In this case, it would falsify the results.

2.5 Research Gap & Research Question

GIR makes it possible to analyze unstructured data by retrieving, indexing and searching it. Social media platforms like Twitter or Instagram provide a constant stream of user-generated content, some containing geographical components. Analyzing such data can help to examine how landscape is perceived and how it differs culturally or linguistically. The Spanish language is particularly useful because many people across the globe speak it and there is evidence of linguistic variation that has already been studied geographically.

Although there is a vast amount of research on the perception of landscape and linguistic variation, there is a knowledge gap regarding the intersection of these two fields. Studies that address landscape perception usually rely on surveys or visual analysis. At the same time, research on linguistic variation typically focuses on lexical differences between individual regions and languages. How different people conceptualize and perceive landscapes through natural language – especially in unstructured and unformatted texts such as social media – has not been sufficiently explored. Tweets from South America, where the diversity of landscapes and dialects offers a unique opportunity, are particularly fruitful for analysis from a linguistic point of view. This thesis, by proposing a framework, addresses this gap.

This framework uses machine learning to classify tweets. After the classification, spatial, temporal, and co-occurrence analyses are conducted to provide deeper insights and to find patterns. The aim of this approach is to advance our understanding of how landscapes are perceived and discussed within specific geographical and linguistic contexts. The thesis thereby mainly focuses on the following question: In what ways are landscapes discussed and perceived in Spanish-speaking Central and South America based on geographically and thematically filtered Twitter data, and what spatial,

temporal or thematic patterns can be identified? The focus is on the methods used to filter data thematically. In addition, the following multidimensional analyses shed light on the geographical and cultural diversity of Central and South America.

3 Methodology

This section describes the approach used to analyze Tweets from South America to understand how landscapes are described and perceived. Prof. Dr. Carlota de Benito Moreno provided the data set. It contains over 995 million Tweets and almost 30 billion words. The Tweets were published between 2012 and 2019. The analytical process, which is explained step by step, includes techniques from GIR and NLP. The data is preprocessed, filtered, and classified using a specially trained machine-learning model. Then, the data is analyzed for spatial, temporal, and co-occurrence relationships. These methods describe a framework for obtaining meaningful insights from an extensive, unstructured data set while addressing issues such as data noise and polysemy (Grieve et al., 2019). All scripts and code used for data preprocessing, filtering, and analysis are available on GitHub under the following link: https://github.com/phiserohr/Masterthesis.

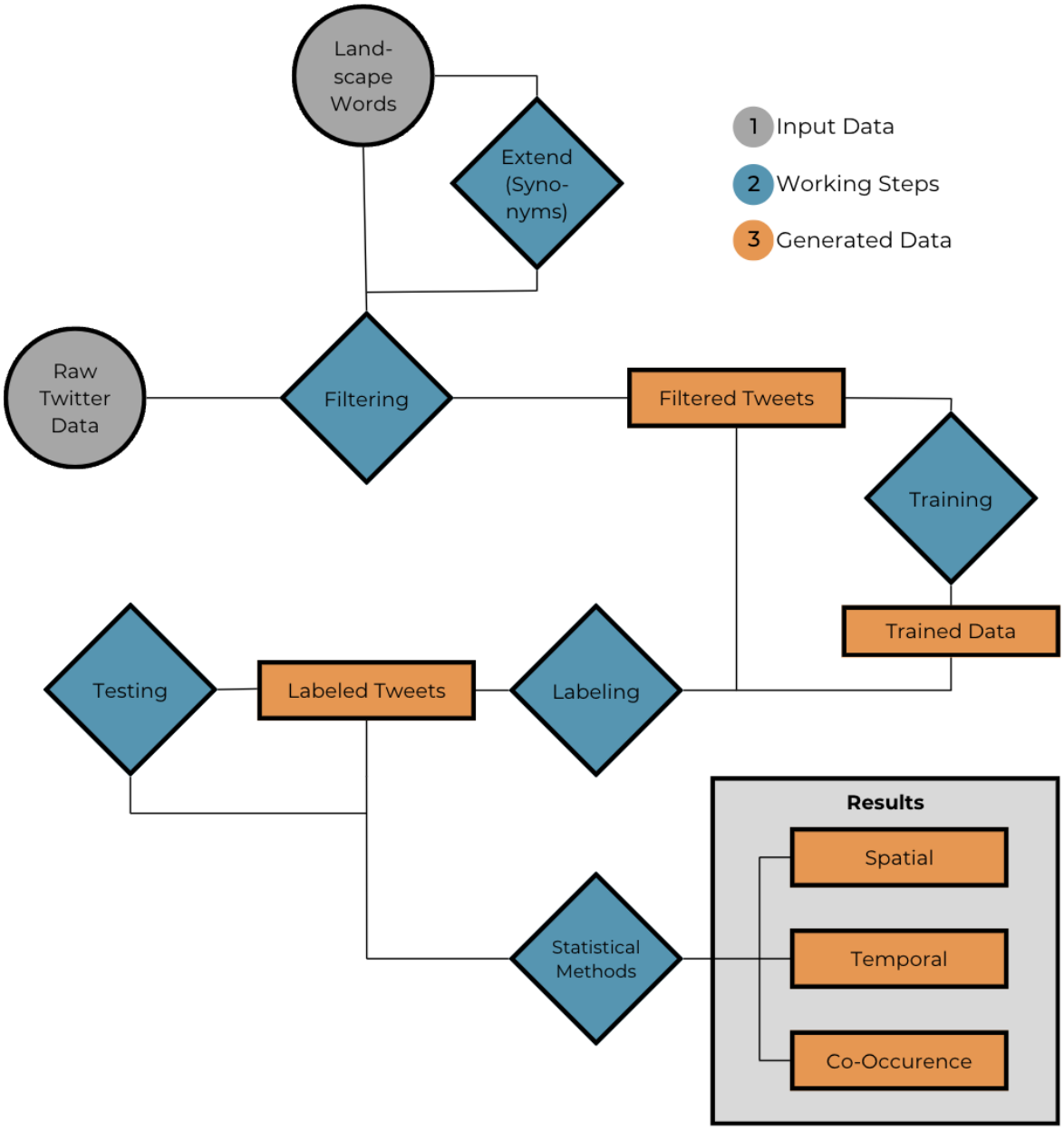


Figure 3.1: Flowchart summarizing the analytical framework (Illustration by the author)

3.1 Building a Landscape Lexicon

As a base to examine Tweets from South America on how landscape is described, the first step is to define a list of landscape terms. A list of words created in a study by van Putten et al. (2020) was used as a baseline for the terms. In their study, the scholars asked participants to freely list exemplars of words to landscape, animals, and body parts. They invited speakers from 7 different languages; one of the languages was Spanish. They found that the participants thought listing landscape terms was the hardest of those three. The Spanish participants were 68 people from Mexico - this is ideal because the research area is in Central and South America and not in the European Spanish-speaking countries. In total, the participants listed 48 terms regarding landscape

in Spanish. Among the most named terms were *montaña* (Mountain), *río* (River), *nubes* (Clouds) and *árbol* (Tree).

In total, 38 terms were saved in the database because van Putten's study also listed words such as *azul* (Blue), *verde* (Green), and *color* (Color). These are obviously colors that are not primarily associated with a landscape term but are used more frequently as a descriptive attribute. Thus, they were not considered in the list. In addition, the word *paisaje* (Landscape) was added because the description of the landscape will be the core content of this thesis.

As one of the goals of this study is to identify whether different cultures or regions talk differently about landscape, synonyms of the 38 terms had to be included in the list. For this, a dictionary called "Diccionario de la lengua española de la Real Academia Española" (Real Academia Española, 2024) was used. The named landscape terms were manually searched for on the website to identify synonyms or different ways of spelling a specific term. The following example using the word *montaña* highlights how that process was conducted. For this term, the dictionary lists several different definitions.

Spanish Definition	**English Translation**
1. Gran elevación natural del terreno	High natural elevation of the terrain.
2. Territorio cubierto y erizado de montes.	Territory covered and bristling with mountains.
3. Gran acumulación de algo.	A large accumulation of something.
4. Dificultad o problema de muy difícil resolución.	Difficulty or problem that is very difficult to solve.
5. Terreno muy poblado de árboles.	Land heavily populated with trees.
6. Monte de árboles o arbustos.	A hill of trees or bushes.

Table 3.1: Definition of *montaña* with English translation (Real Academia Española, 2024).

As illustrated in the translation, definition one is an optimal fit for the description of a landscape feature. In contrast, definition four is an obvious metaphor and the word mountain is used as a descriptive image of a problem that is hard to solve. In this case, the first definition was used to identify the potential synonyms for the term *montaña*. The dictionary lists five synonyms. The difficulty with synonyms is that the respective definitions do not necessarily describe a landscape concept and, in this case, a mountain. The listed synonyms and their respective description was analyzed to identify whether the word is a landscape related term or not.

For example, although the word *pico* can be used to describe a mountain, it also means "Parte saliente de la cabeza de las aves, compuesta de dos piezas córneas, una superior y otra inferior, que terminan generalmente en punta y les sirven para tomar el alimento", which translates to "The protruding part of a bird's head, consisting of two horny parts, one upper and one lower, which generally end in a point and are used to pick up food." In this case, only the 6th definition of *pico* (*Cúspide aguda de una montaña*, which translates to a sharp peak of a mountain.) is relevant as a landscape term.

Spanish Synonym	**English Translation**
Monte	Hill
Pico	Peak
Elevación	Elevation or rise
Prominencia	Prominence
Promontorio	Promontory

Table 3.2: Spanish synonyms for *montaña* and their English translations (Real Academia Española, 2024).

This process was done for all 38 terms to identify the synonyms. In the aforementioned case of *pico*, it was decided that since only the 6th definition has anything to do with a landscape term, the word does not count as a synonym for the term *montaña*. The landscape terms and their relevant synonyms, as well as a translation and the primary definition, were saved in a PostgreSQL database for better understanding. Additionally, typical spelling mistakes and plurals were added. So for the word, for example, *montaña* words like *montana* or *montañas* were added. This is for the simple reason that natural language is analyzed on Twitter, and errors are naturally occurring. Other work, such as from Effrosynidis et al., used a Python library to correct the misspelled words (Effrosynidis et al., 2017). In the appendix is a table (7.1) with the final list of words, including the synonyms of terms that were used for further analysis.

3.2 The Data

This study is based on a large Twitter dataset with Tweets in Spanish written between 2012 and 2019. The Tweets were stored in separate CSV files, of which all with the attribute `user_language` = `es` are recognized as Spanish by the Python library `langdetect` were added to these files. The Tweets were saved with numerous attributes of which the following were used for this work:

- **id**: A unique identifier for each Tweet
- **Tweet_text**: The text content of the Tweet
- **created_at**: The timestamp indicating when the Tweet was created
- **user_location**: The location specified by the user in their profile
- **place_full_name**: The full name of the geographic location associated with the Tweet
- **geo_latitude**: The latitude coordinate of the Tweet's geographic location
- **geo_longitude**: The longitude coordinate of the Tweet's geographic location

Users can leave the `user_location` field empty or input fantasy locations, such as `Narnia` or `The Moon`, making it unsuitable for analysis.

The structure of the Twitter stream is as follows: There is a subfolder for each year. In there, there are folders for every month (e.g., `01`, `02`, up to `12`). Within those, there are, again, subfolders for each day (e.g., `01` to `31`). For every day, there are 24 CSV files named `01_geo.csv`, each containing all the Tweets that were posted within that hour. This corresponds to approximately 9,000 files per year, i.e., around 56,000 files for all years between 2012 and 2019. The number of Tweets per hour can vary greatly. Some files only contain about 5,000 Tweets, while others contain over 100,000. As the time and computational effort required for further analysis is sometimes very high, it was decided to analyze only one year. In 2017, over 100 million Tweets were sent in Spanish. As this is already a large number, it was decided only to run the subsequent steps with these.

3.3 Preprocessing the Data

When working with Twitter data, it is important to perform proper preprocessing. In this case, it was decided to read the data from the respective CSV files, preprocess the Tweets, and then store them in a database. This includes various steps such as filtering out Retweets (Tellez et al., 2023) or deleting messages written by bots. This can be done, for example, by identifying Tweets that are sent at a high frequency. Furthermore, special characters such as hashtags, emojis, and mentions have been removed in many studies (Grieve et al., 2019; Louf, Gonçalves, et al., 2023; Van Putten et al., 2020).

For this analytical step, the words were tokenized and saved in lowercase letters. Tokenizing the input data can lead to a better performance of the NLP techniques (Donoso & Sánchez, 2017; Schmidt et al., 2024). Tokenizing a Tweet means that the actual Tweet, which is a string of words, was split up and broken down into individual tokens. This was done by using a Python function called findall (Foundation, 2024). By breaking up the text into tokens, the Tweets are split into simple, manageable parts that require significantly less computing power. This simplifies the

structure of the text. This step is necessary for the subsequent natural language processing tasks, such as filtering or possible topic modeling. Furthermore, converting to lowercase and removing punctuation makes the text more uniform, thereby reducing variability within a Tweet and in comparison to other Tweets. For example, "Mountains", "mountains!", and "mountains" are all treated as "mountains". This means that a significantly smaller number of words have to be analyzed overall.

> *"The view of the snow-capped Andes mountains is absolutely breathtaking! #nature #mountains"*

Will be tokenized to the following:

> *['the', 'view', 'of', 'the', 'snow', 'capped', 'andes', 'mountains', 'is', 'absolutely', 'breath-taking', 'nature', 'mountains']*

3.4 Filtering the Tweets

The preprocessing described above is part of the script discussed in this section, which filters the Tweets according to landscape terms. The script, available on GitHub as `1_filtering.py`, performs several steps to process the data. It reads the existing CSV files individually, filters the Tweets according to predefined landscape terms, and stores them in a PostgreSQL database. To ensure traceability, the following section summarizes the structure of the script.

In the first step, a database table named `filtered_Tweets` is created to store the filtered Tweets. This table includes attributes for Tweet metadata and an additional attribute, `found_term`, which records the matched landscape term to simplify further processing.

In the second step, landscape terms (including synonyms and common misspellings) stored in the `landscape_terms` database table are loaded for filtering. In the third step, the Tweets are tokenized and converted to lowercase for consistent matching, as described in the *Preprocessing the Data* section.

The most critical step involves loading the CSV files one by one. Each line is analyzed in search of the landscape terms stored in the database. Tweets starting with `rt` (Retweets) are skipped to prevent distortion in the analysis. Processed Tweets are displayed on the command line to provide feedback on progress. Over 100 million Tweets were processed, with nearly 400,000 containing a matched landscape term.

In the script's first version, 38 terms from van Putten's list were used to filter the Twitter dataset, resulting in nearly 3 million matched Tweets. The list included terms such as *car* and *water*, with over 40,000 matches for the term *coche* (Car) and over 250,000 for *agua* (Water). However, during data manipulation and analysis, it became apparent that many terms provided little to no meaningful data for investigating landscapes.

For example, while *coche* was identified as a landscape term in van Putten's study, closer examination revealed that the majority of Tweets containing this term referred to topics unrelated to landscapes, such as jokes, road safety, traffic jams, or theft and crime. Following is an example of a Tweet containing the term *coche*:

"Original: Si bebes o te drogas... no cojas el puto coche. siempre le pasa algo el que menos culpa tiene."
Translated: If you drink or take drugs... don't take the (...) car. Something always happens to the one who is least to blame.

As a result, many terms were excluded, and the updated version of the script narrowed the list to 18 words, as shown in Table 3.3. These selected terms are more directly relevant to the study of landscapes, allowing for a more targeted and meaningful analysis.

Spanish Term	**English Translation**	**Synonyms**
Amanecer	Sunrise	Alba, Albor, Alborada, Madrugada
Atardecer	Sunset	Ocaso, Puesta Del Sol, Anochecer
Bosque	Forest	Floresta, Soto
Cascada	Waterfall	Catarata
Cerro	Hill	Colina, Collado, Loma
Desierto	Desert	
Horizonte	Horizon	
Lago	Lake	Laguna
Mar	Sea	Océano, Ponto
Montaña	Mountain	Monte, Cabezo, Cumbre
Naturaleza	Nature	Natura
Paisaje	Landscape	Paraje
Playa	Beach	Costa
Río	River	
Selva	Jungle	Jungla
Tierra	Earth	
Valle	Valley	Cuenca, Quebrada
Volcán	Volcano	

Table 3.3: Spanish landscape terms with their English translations and synonyms used to filter the Twitter data set.

Furthermore, the second version of the machine learning approach addressed the problem of the low number of relevant Tweets. In the first attempt, only Tweets in which the landscape was written in a literal sense and a certain description of the landscape was provided were classified as relevant. For example, Tweets with the following content were classified as relevant:

> *"The golden sands stretch for miles, meeting the turquoise waves under a cloudless sky. The beach at sunrise is pure magic."*

Tweets with a literal sense but not describing landscape as such were classified as irrelevant. The following is an example:

> *"Just got back from the gym, and now I'm heading to the beach for volleyball practice!"*

However, the strict classification in the training set meant that only a fraction of the manually labeled Tweets were relevant to answering the research questions. As a result, machine learning with the random forest model resulted in poor classification. Therefore, in the second attempt, every Tweet was classified as relevant if it used a literal term for landscape, not only those that also described a landscape. This led to more Tweets being classified as relevant, with the downside of losing accuracy of the random forest model.

3.5 Machine Learning

In the next step, the Tweets that had already been filtered were categorized according to relevance for the study based on landscape topics. To do this, many different methods were tested until the Tweets were ready for analysis. The aim was to use a machine-learning method to classify the Tweets as either relevant or irrelevant for further analysis. Relevant Tweets, in this sense, are all Tweets that use the landscape term literally. Those Tweets that use the Tweets as a name, metaphor, or in a completely different context were classified as irrelevant.

Here is a short, fictitious example with both a relevant and an irrelevant Tweet:

> *"The view from the top of the mountain was breathtaking. The hike was tough, but totally worth it!"*

And an example of a Tweet in which 'mountain' is used metaphorically:

> *"I have a mountain of work to do before the weekend. Guess I'll be climbing my way through paperwork!"*

Why are Tweets that don't literally write about landscape filtered out? The analysis should aim to find out how people talk about the landscape. Tweets in which a landscape term is used in a different sense (e.g., as a name or metaphor) can significantly influence the results and cause a distortion.

One way to filter these Tweets is through machine learning. It was decided to use a random forest model trained in advance and then run over the 400 thousand filtered Tweets. Random forest is a machine learning algorithm that combines multiple decision trees to enhance the performance and is often used for classification tasks (Breiman, 2001). Also, its accessibility and easy-to-understand parameters are additional reasons to use random forest.

3.5.1 Training the Model

To train the random forest model, a script was written, stored on GitHub as `02_training.py`. This script created a new table, which stores the manually classified Tweets and provides the basis for training the model.

The script retrieves a random Tweet from the database and displays it in the terminal for manual labeling. By entering a 0 (irrelevant) or a 1 (relevant), the Tweet is stored in the database and another Tweet is loaded. This process was carried out for approximately 2,600 Tweets to create a sufficiently large training data set. The random forest algorithm requires labeled data to identify the input data's patterns and features.

3.5.2 Classification of Tweets

The training dataset was used to classify the table of filtered Tweets in another script, stored on GitHub as `03_machine_learning.py`. The script utilizes the Python packages `pickle`, `pandas`, and `sklearn`.

The `pickle` package was employed to save the vocabulary and the best random forest model, reducing the computing power required for future calculations. `Pandas` was used to manipulate

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Table 3.4: Confusion Matrix

data in tabular form and to simplify the import and export of data to and from an SQL database. Furthermore, data processing with `pandas` integrates seamlessly with machine learning models such as random forest. The `sklearn` library was used to perform machine learning and to find the optimized model parameters. In addition, the `TfidfVectorizer` function was used to convert strings into numerical vectors, enabling the efficient application of random forest. The words are converted into numbers that can be better processed by the algorithm and evaluated according to their importance. Thus, frequently used words such as *la* (The) or *y* (And) receive a low value, while landscape terms such as *montaña* (Mountain) or *playa* (Beach) receive a higher value.

With these foundations, the manually labeled data was collected from the database, and random forest machine learning was done. The hyperparameters can be adjusted to improve the algorithm's results. A hyperparameter search with `GridSearchCV` was conducted to determine the optimal parameters for the random forest model. This eliminates the need to try out the various options for parameters. The resulting vocabulary and model were saved using `pickle` and stored as `tfidf_vectorizer.pkl` and `best_random_forest_model.pkl` on GitHub.

Following, the filtered but not yet classified Tweets were retrieved from the database and labeled using the trained model. The results — either 1 (relevant) or 0 (irrelevant)— were appended to the dataset and written into a new table in the database called `labeled_Tweets`.

In total, 400,000 Tweets were classified, of which 135,729 were considered relevant by the random forest algorithm.

3.5.3 Performance of the Algorithm

A test dataset was labeled manually to evaluate the machine learning model's performance. This was carried out using a Python script stored on GitHub as `04_sample.py`.

First, 1,000 Tweets were manually classified. The random forest model had already labeled these tweets. Comparing the manual labels to the model's predictions provides a measure of the model's accuracy.

The results could then be compared with each other, and various key figures can be compared. True positives (TP) are Tweets that are correctly predicted by the model. False positives (FP) are incorrectly classified as relevant even though they are actually irrelevant. True negatives (TN) are the ones correctly predicted as irrelevant by the model, and False negatives (FN) are Tweets that are incorrectly predicted as irrelevant. With those numbers, different metrics can be calculated.

For the accuracy, precision & recall of the model, the formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In the attempt with the strict classification mentioned above, the model achieved a precision of 0.51, a recall of 0.33 and an accuracy of 0.95. In the second attempt, with only 18 terms but less strict classification, which was then also used for further analysis, a precision of 0.56, a recall of 0.48, and an accuracy of 0.68 were achieved. The high level of accuracy in the first attempt can be deceptive, but more on this in the discussion part. The F1-Score can also be used for a comparison in the discussion. This is 0.40 for the first attempt and 0.52 for the second attempt. The F1 score can be calculated as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The result of the machine-learning was a table with the filtered Tweets classified by relevance. This table was used for the following analytical steps.

3.6 Geolocation & User Location

In an ideal scenario for GIR techniques, every tweet would be geotagged. However, in reality, this is far from the case. Among the 400,000 filtered Tweets, only 2.4% — approximately 10,000 Tweets — contained geotags. To enhance the amount of Tweets that are geolocated, the idea of using the user location was explored through a geoparsing package in Python, specifically `spaCy`. However, this approach was not as valuable as hoped, as many user location fields were either blank or contained not real locations. Given these limitations, it was ultimately decided to proceed with spatial analyses using only the Tweets that were reliably georeferenced in order to ensure the accuracy and integrity of the spatial analysis. While this is far from ideal, 10,000 geotagged Tweets still provide a sufficient dataset for further spatial analysis.

3.7 Spatial Analysis

The spatial analysis is conducted using a Python script stored on GitHub under the name `chi2_spatial.py`. Spatial analysis was performed to explore the connection between the Tweet's distribution, which contains a landscape term, and physical landscapes. This analysis was done to identify patterns in the way or rather where people talk about landscape. This can help to understand how landscapes are perceived and how they are connected to locations or activities. McKitrick et al. highlight the potential of location-based research of social media data (McKitrick et al., 2023). The analysis utilizes various libraries, including:

- **pandas** and **geopandas**: For data manipulation and handling spatial data
- **scipy**: For statistical analysis, including the Chi-Square test
- **shapely**: For performing spatial operations, such as creating geometries and performing intersections
- **sqlalchemy**: For managing the connection to the PostgreSQL database and executing SQL queries
- **matplotlib**: For data visualization and creating maps
- **numpy**: Used for numerical calculations
- **contextily**: Added for the generation of the basemap
- **os**: Used for directory and file management

The combination of these libraries enabled efficient data handling, spatial computations, and the creation of insightful visual outputs for the analysis.

Defining the key parameters, such as pixel size and the term to be analyzed, made it easier to examine different landscape terms later without requiring changes to the entire script. With the parameters in place, only the term needs to be adjusted, and the resulting CSV files and maps are automatically named according to the landscape term being analyzed.

All Tweets were loaded, but only those with georeferences were considered. A specific bounding box with coordinates around Central and South America (latitude between -56.0 and 33.0 and longitude between -117.0 and -30.0) was applied to ensure that no Tweets from other Spanish-speaking countries were analyzed. These Tweets served as the basis for the Chi-Square analysis. For comparison, only Tweets containing the specified terms and classified as relevant were loaded. Using `geopandas`, Tweets with georeferencing were reprojected to the appropriate projection (in this case, EPSG:3857) and stored in a raster. The raster size was set to 75,000 meters. By aggregating the Tweets into grid cells, which is a common practice in geospatial analysis, further steps are simplified, and results can be better visualized (Donoso & Sánchez, 2017; Funkner et al., 2021; McKitrick et al., 2023). Subsequently, all Tweets in the respective pixels were counted. First, the total number of Tweets in each pixel was counted, followed by the count of Tweets that contain the specific landscape term and that are classified as relevant.

Afterward, pixels with zero values were filtered out, and the residuals were calculated. This was done to reduce noise in the analysis. Residuals represent the difference between the observed and expected numbers, normalized to the square root of the expected number.

After that, the top 10 hotspots, i.e. the 10 pixels with the highest positive deviations, were identified. The residuals were calculated dynamically with adjusted percentiles (95%), to exclude outliers. These were then stored in two separate CSV files. The first contains the ID of the hotspot including the statistical figures and the coordinates. The second is a CSV file containing all Tweets sent at the respective hotspots. This simplifies a later analysis.

Finally, all information was plotted and displayed on a map using `matplotlib`. In addition, the most important statistical key figures from the Chi-Square test were shown in the console and saved. To present the results, the coordinates of the 10 hotspots for each term were then searched for using Google Maps and clearly displayed in a table. This resulted in a map for each landscape term, visualizing the distribution of Tweets, and an accompanying table in which the 10 hotspots are each briefly described with their corresponding coordinates. To find these coordinates, the Tweets were manually searched for keywords to identify the real place.

In summary, this script counted the number of Tweets per term in a predefined grid over South America. These numbers were compared using the Chi-Square test, and deviations from the expected Tweets were shown. This information is displayed on a map, and additional content (Tweets & hotspots) was stored in two separate CSV files, which helped to identify the hotspots and was used as a basis for the results and discussion.

3.8 Temporal Analysis

A similar approach was taken for the temporal analysis. For this, not only geotagged Tweets were used, but all were categorized as relevant. The temporal distribution of tweets is interesting because it provides insights into different patterns, such as seasonal trends or significant events about which many Tweets were sent. This was done, for example, by Ntompras et al. after the COVID-19 pandemic (Ntompras et al., 2022). A Python script, available on GitHub and named `Chi_2_Days.py`, contains the code that analyzes the temporal distribution of Tweets per term and displays them in a matrix. Various packages were used to perform the analysis:

- **pandas**: For data manipulation
- **spicy**: For statistical analysis, including the Chi-Square test
- **numpy**: For numerical operations
- **plotly**: For creating interactive visualizations
- **sqlalchemy**: For managing the connection to the PostgreSQL database and executing SQL queries
- **psycopg2**: Used indirectly via SQLAlchemy for database connections

The relevant Tweets were stored in a pandas data frame. Instead of storing the Tweets in geographically referenced pixels, as in the spatial analysis, the Tweets were assigned to a day of the year (i.e., a value between 1 and 365). Like the spatial analysis, the Chi-square (χ^2) test was used to calculate

the deviations of the actual number of Tweets counted per term and day from the expected number of Tweets. To avoid an error on days when no relevant Tweets with the appropriate landscape term were written, missing data was simply replaced with a 0. Furthermore, the residuals were calculated, and the statistical significance was tested using a p-value.

Additionally, the results were displayed on an interactive heat map using `plotly`. This approach allowed for the identification of unusual activities, such as an unexpectedly high or low number of Tweets per term on specific days. The color bar was divided into five categories from strongly negative to strongly positive, and the X-axis contains the description for the month in each case so that easy readability is guaranteed. The heatmap was then saved as a PNG, which can be seen in the result section 4.9.

A second script, available on GitHub and named `Chi_2_Days_Term.py`, analyzed only one landscape term at a time. This is made possible by dynamic coding, where only the term needs to be changed as a parameter. Additionally, a bar chart was used for the presentation, which will later be included in the results section.

A third script, available on GitHub and named `Hotspot.py`, examined the days with matching terms more closely. This script exports all Tweets for a specified term and day into a CSV file, simplifying the analysis of specific tags in greater detail in the results section of the thesis.

3.9 Co-Occurence

A third part of the analysis dealt with the co-occurrence of words with landscape terms. As in the temporal analysis, all relevant Tweets were used to find re-appearing topics connected to the landscape terms. This analysis can help to understand what emotions and activities are linked to certain landscape terms. A Python script, available on GitHub and named `Co_Occurrence.py`, was used to calculate various statistical key figures. The script relied on several Python packages:

- **pandas**: For data manipulation
- **spaCy**: Used for tokenizing and part-of-speech takking
- **matplotlib**: Used for visualization
- **wordcloud**: Used for the generation of a wordcloud
- **collections**: To count word occurrences and co-occurrences
- **sqlalchemy**: For managing the connection to the PostgreSQL database and executing SQL queries
- **math**: For performing mathematical computations

The script worked through a list of terms one after the other to generate a word cloud and a CSV file with the most essential terms and statistics. The database table with the Tweets was filtered according to the term to be examined. Once retrieved, the Tweets were tokenized, and a context window was extracted for analysis. The context window, defined as the three words to the left and right of the selected term, could also be adjusted via parameters ("Semantics and Discourse", 2018). `SpaCy` was used to tokenize the Tweets and remove specific content. First, non-alphabetical tokens were removed. It was decided to count nouns, verbs, and adjectives as they give insight into the emotions or activities that are connected to certain landscape terms. This is common practice in co-occurence analysis (Liu et al., 2010). In addition, tokens beginning with 'http' (i.e. URLs) or '@' (i.e. mentions) were removed. The co-occurrence of the analyzed term was examined using log-likelihood since this statistical value can be meaningfully applied to a large data set ("Semantics and Discourse", 2018).

The outputs were CSV files showing the 25 words with the highest values for log-likelihood and their actual number of co-occurrences. To create a visually appealing representation, word clouds were created using the 'word cloud' library.

4 Results

This chapter discusses the results of the methods applied and discussed above. The same structure is used as in the previous chapter. First, a rough overview of the results is given. Subsequently, the three individual thematic blocks, spatial, temporal, and co-occurrence, are dealt with separately. To conclude the chapter, the main features of the results are presented, followed by a general summary.

4.1 Data Presentation

The study analyzed more than 100 million Tweets written in Spanish in Central and South America in 2017. The Tweets that were filtered using machine learning focused on the literal use of landscape terms. Along with the main terms, synonyms of these terms were also used.

- **Tweets in total:** 102,524,012
- **Tweets with a landscape term:** 382,988
- **Relevant Tweets (literal landscapes):** 135,638
- **Irrelevant Tweets (metaphorical or other contexts):** 260,891

Of the total Tweets, 0.37% or 382,988 contained a landscape term. The machine learning model achieved an accuracy of 68%, with a precision of 56%. Of the 382,988 Tweets, 35.42% were classified as relevant accordingly, potentially discussing literal landscapes.

- **Precision:** 0.56
- **Recall:** 0.48
- **Accuracy:** 0.68

The following table shows how many Tweets were found for each term, and how many of them (number & percent) were classified as relevant.

Table 4.1: Summary of found terms and relevance, sorted by highest count

Landscape Term	**Total Tweets**	**Relevant Tweets**	**Percentage (%)**
tierra	52,574	28,447	54.11
mar	46,555	11,912	25.59
playa	42,676	33,447	78.37
río	33,133	621	1.87
costa	30,688	1,282	4.18
valle	17,773	1,474	8.29
madrugada	16,727	382	2.28
naturaleza	16,363	15,158	92.64
amanecer	10,808	856	7.92
cerro	10,724	1,688	15.74
montaña	9,211	7,411	80.46
alba	8,785	17	0.19
monte	8,444	1,752	20.75
bosque	8,292	2,662	32.10
lago	7,430	6,214	83.63
cuenca	5,903	256	4.34
desierto	5,867	2,060	35.11
cumbre	5,342	87	1.63
soto	5,207	159	3.05
laguna	5,022	2,937	58.48
atardecer	4,314	3,982	92.30
paisaje	4,064	3,399	83.64
volcán	3,633	3,423	94.22
selva	3,317	1,721	51.88
horizonte	3,120	514	16.47
océano	2,902	1,335	46.00
loma	2,378	183	7.70
quebrada	2,330	463	19.87
colina	1,576	236	14.97
jungla	1,468	266	18.12
cascada	1,091	313	28.69
natura	939	199	21.19
ocaso	918	203	22.11
floresta	786	113	14.38
ponto	648	7	1.08
anochecer	476	134	28.15
collado	450	56	12.44
paraje	354	96	27.12
alborada	289	70	24.22
catarata	281	78	27.76
cabezo	76	12	15.79
albor	54	13	24.07
Total	382,988	135,638	35,42

4.2 Spatial Analysis

4.2.1 Overview

The following section presents the results of the spatial analysis. Table 4.2 summarizes the total number of Tweets, relevant Tweets, and geolocated Tweets for each landscape term. This overview is necessary to understand how the spatial distribution of Tweets is distributed across the individual terms. The terms were selected based on their spatial distribution and displayed here. In general, these are terms for which many geotagged Tweets were classified as relevant and for which there is, therefore, a spatial pattern that can be analyzed later. Terms with few geotagged Tweets were not considered for the spatial analysis.

Table 4.2: Summary of found terms and relevance with geolocation, sorted by number of Tweets with geolocation. The percentage shown is the proportion of geotagged Tweets among all Tweets found for the respective term.

Landscape Term	Total Tweets	Relevant Tweets	No. of geolocated Tweets	Percentage
playa	42,676	33,447	1,758	4.12
mar	46,555	11,912	1,567	3.37
costa	30,688	1,282	739	2.41
río	33,133	621	650	1.96
valle	17,773	1,474	586	3.3
cerro	10,724	1,688	533	4.97
tierra	52,574	28,447	419	0.8
bosque	8,292	2,662	393	4.74
atardecer	4,314	3,982	343	7.95
lago	7,430	6,214	326	4.39
laguna	5,022	2,937	285	5.68
naturaleza	16,363	15,158	260	1.59
monte	8,444	1,752	241	2.85
amanecer	10,808	856	184	1.7
horizonte	3,120	514	131	4.2
montaña	9,211	7,411	130	1.41
cuenca	5,903	256	105	1.78
paisaje	4,064	3,399	103	2.53
colina	1,576	236	95	6.03
soto	5,207	159	92	1.77
loma	2,378	183	89	3.74
quebrada	2,330	463	65	2.79
cumbre	5,342	87	63	1.18
cascada	1,091	313	62	5.68
volcán	3,633	3,423	60	1.65
madrugada	16,727	382	60	0.36
desierto	5,867	2,060	56	0.95
alba	8,785	17	52	0.59
floresta	786	113	46	5.85
selva	3,317	1,721	44	1.33
océano	2,902	1,335	24	0.83
natura	939	199	21	2.24
alborada	289	70	19	6.57
ocaso	918	203	18	1.96
anochecer	476	134	14	2.94
jungla	1,468	266	12	0.82
ponto	648	7	11	1.7
catarata	281	78	10	3.56
paraje	354	96	10	2.82
collado	450	56	9	2.0
albor	54	13	0	0.0
cabezo	76	12	0	0.0
TOTAL	382,988	135,638	9,685	2.53

The terms *tierra* (Earth), *mar* (Sea), and *playa* (Beach) have the highest number of Tweets, with over 40,000 mentions each. In contrast, other terms, such as *cabezo* (Peak) or *albor* (Dawn), have no geolocalized Tweets, which is also due to the overall very low number of Tweets found. *Atardecer* (Sunset) has the highest number of geotagged Tweets, at just under 8%. It is followed by terms such as *alborada* (Sunrise) with 6.57% , *colina* (Hill) with 6.03% and *floresta* (Forest) with 5.85% .

In the next section, each term is briefly illustrated with a spatial distribution on a map and the corresponding hotspots are listed in a table with coordinates. The maps show the distribution of Tweets per term. The scale varies depending on the term. For terms such as *playa* (Beach), where there were also a large number of geolocalized Tweets, the range of the expected distribution is significantly more extensive than for terms such as *collado* (Hill), for which only two geolocalized Tweets were found.

The tables list the hotspots that can also be found on the maps. The ID in the table corresponds to the numbers on the map. The entries in the tables are the pixels that show the 10 highest values for the deviation from the expected distribution. In short, the pixels with the highest values should have a large number of Tweets for the respective term.

4.2.2 Playa

The spatial analysis 4.1 of the term *playa* (Beach) shows clusters in well-known coastal regions and tourist destinations in Mexico, Chile, Argentina and Uruguay. The location that shows the highest deviation from the expected Tweets is Cancún in Mexico (Hotspot 1). This is a very popular beach destination. Many Tweets mention Playa del Carmen, which is one of the beaches that can be reached from Cancún. In Uruguay, Faro de José Ignacio or Ciudad de la Costa are mentioned, prominent places that are also known for tourism and beaches. In Chile, coastal regions such as Playa de Santo Domingo and Iquique are mentioned. Of the 10 hotspots found, all are directly linked to a literal term for beach. The Tweets suggest that there is a strong correlation between *playa* and tourist activities. Overall, the spatial distribution of *playa* is very clearly confined to coastal areas. Even hotspot 7, which at first glance may appear to be inland, is actually in Encarnación in Paraguay, where there are several beaches along a large river, the Río Paraná, which marks the border between Paraguay and Argentina. Along the entire riverbank, there are numerous coves and beaches of varying sizes.

4.2.3 Mar

When looking at the map 4.2, it is also immediately apparent that the pixels with a high value for the term *mar* (Sea) are orientated towards the coast of South America. Of the selected hotspots shown in the table, nine out of ten are located near the coast. Only point 5, in San Bartolo in Bolivia, is inland. The other hotspots either have *mar* in their name, like Mal del Plata in Argentina or are located directly by the sea, such as Isla Puná in Ecuador. It is also striking that points 1, 2, 6, and 9 are located near each other, all on the coast of Argentina. In the vicinity of Mar del Plata, a larger city, many other localities have *mar* in their name.

4.2.4 Tierra

Tierra means earth. This word can therefore refer either to 'the earth,' our planet or to the ground. Here, too, one can see on the map 4.3 directly that the hotspots are spread across the entire continent. It is striking that the points are less clearly located along the coast than with the words *playa* or *mar*. Points 1 and 3 are again very close to each other – as seen in the table, they are both in Tierra del Fuego in Argentina. In this case, the table lacks a column indicating whether the term is directly linked to a landscape feature or not – simply because *tierra*, unlike *playa* or *montaña*, for example, cannot be easily associated with a typical landscape. All the hotspots, except point 6, San Antonio in the USA, are relatively rural areas that suggest outdoor activities.

4.2.5 Lago

The term *lago* means lake. Not much can be seen on the map 4.4, as the distribution seems relatively random. However, it is interesting to see a cluster at points 2, 4, 7, and 10, suggesting a high number of lakes. All 10 hotspots found in the table are directly connected to a lake. As already mentioned, the area around points 2, 4, 7, and 10 is an area with many lakes. Point 1, Lago de Maracaibo, is an inland lake in Venezuela that is also navigated by ocean-going vessels. Some Tweets contain the place Vereda del Lago, a park right on Lake Maracaibo in the city of Maracaibo. In addition, various activities around the lake are described. It is also noticeable that a photo with a link is often included in the Tweets.

4.2.6 Laguna

The term *laguna* was used as a synonym for *lago* (Lake) and generally refers to freshwater areas that are slightly smaller than lakes. The spatial distribution of Tweets reveals clusters in Argentina, Ecuador, and Mexico 4.5. Argentina stands out in the table: 5 of the 10 identified hotspots are in Argentina. For example, Laguna de Chascomúsm, Laguna Blanca, and Laguna Alsina are mentioned. The most deviant hotspot is located in Brazil, and the Tweets mention a city called Laguna, which is located in a province called Santa Catarina. However, the city is located directly by the sea, surrounded by various inland seas, also called Laguna.

4.2.7 Cerro

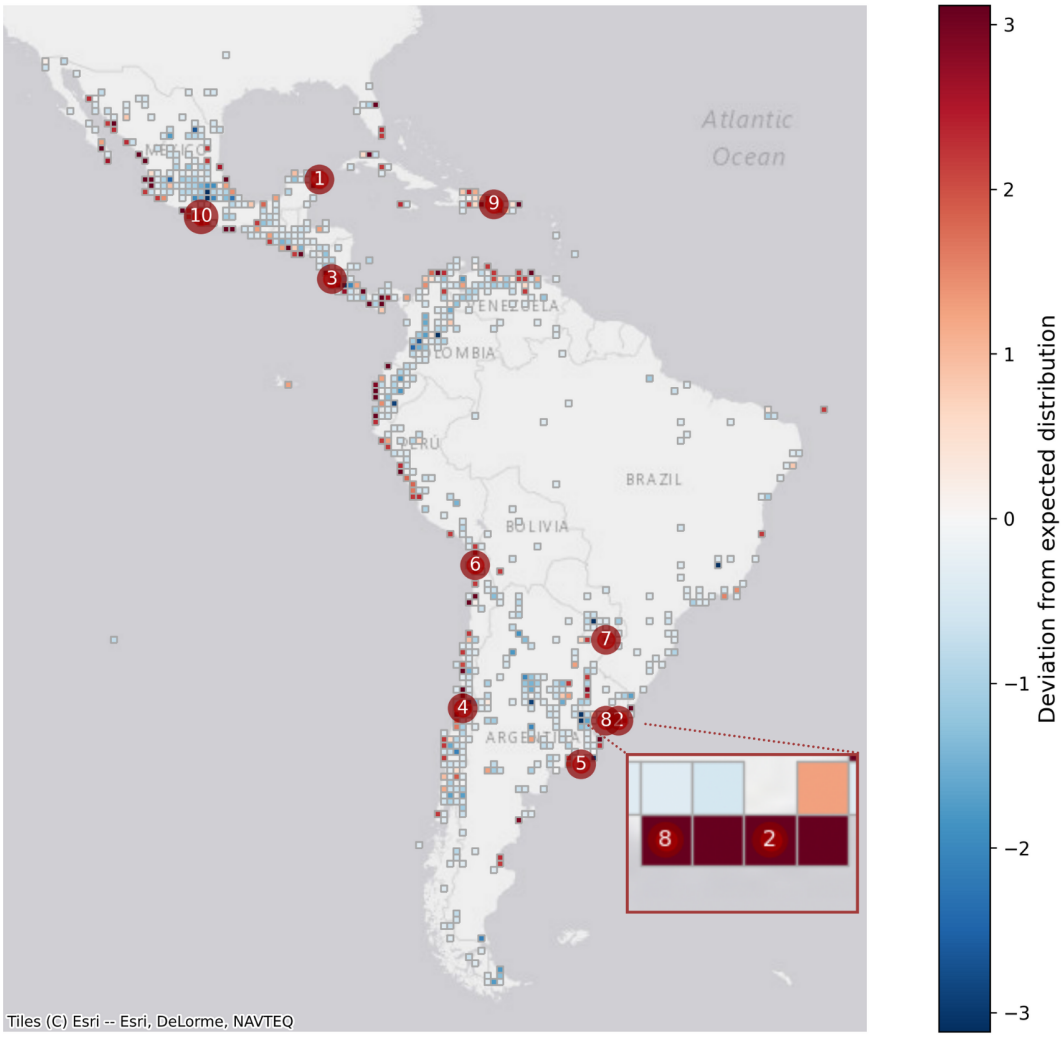
The term *cerro* means hill and is spatially distributed mainly in Chile, Mexico, and Argentina 4.6. Of the top 10 identified hotspots, five are in Chile. Mentions such as Cerro San Cristobál in Valparaíso and Cerro La Cantera are places that are still influenced by religion. Other places, such as Cerro Tres Picos in Argentina and Verro El Quemado in Mexico, indicate an association of *Cerro* with a natural landscape. Furthermore, there are clusters of hotspots in cities like Valparaíso, where hills characterize the cityscape, and some POIs are labeled *Cerro* such as Cerro Cárcel or Cerro Florida. Interestingly, the hotspot identified as number 1 was found in Los Ángeles in Chile: this place is not a literal use of *cerro*, but a reference to a street name. Several Tweets describe a car accident. For example, one Tweet reads '18:53 in Los Angeles vehicle rescue with one injured' (literally translated). In addition, the word *cerro* is also used in a non-literal sense in Paraguay: a football club is mentioned there.

4.2.8 Montaña

Montaña means mountain, and on the map 4.7, you can see that the term is spread across all of South and Central America, with notable clusters associated with primarily mountainous regions or areas with high altitudes. The top 10 identified hotspots include countries such as Chile, Colombia and Peru. Places such as Nevado de Chachi in Chile, Monserrate in Colombia, which is the local mountain of Bogotá, and Huayna Picchu in Peru are mentioned. The latter is one of the two mountains located directly in Macchu Picchu. In addition to the literal mentions of *montaña*, other hotspots are also identified. For example, hotspot 1 in Buenos Aires, Argentina, is a *montaña rusa*, a rollercoaster. Hotspot number 8 is in Mexico City and is also not a mountain but refers to Jardines en la Montaña, a district in the city. In addition, the hashtag *volcán* (Volcano) is used in some Tweets.

4.2.9 Volcán

Volcán means volcano. As the map shows 4.8, clear clusters are recognized in the volcanic regions of Central and South America, particularly in countries such as Chile, El Salvador, and Ecuador. Six of the ten hotspots found are directly related to a volcano. For example, the Volcán de San Salvador in El Salvador (hotspot 1), Volcán Barú in Panama (Hotspot 3), and Volcán Osorno in Chile (Hotspot 4) are identified there. These places are usually associated with tourism, hiking, or beautiful views. In Chile, several volcanoes were identified, including Volcan Osorno, Tolhuaca, and Villarrica. Non-literal usages of the term can be found, for example, in Argentina, where El Volcán is an administrative district. Furthermore, hotspot 2 in Guatemala City is a list of terms. This is a list of a person's highlights since the list contains various things such as volcanoes, football clubs, people, and countries. Nevertheless, the spatial analysis identified various hotspots in Central and South America that are exclusively volcanoes in a literal sense.

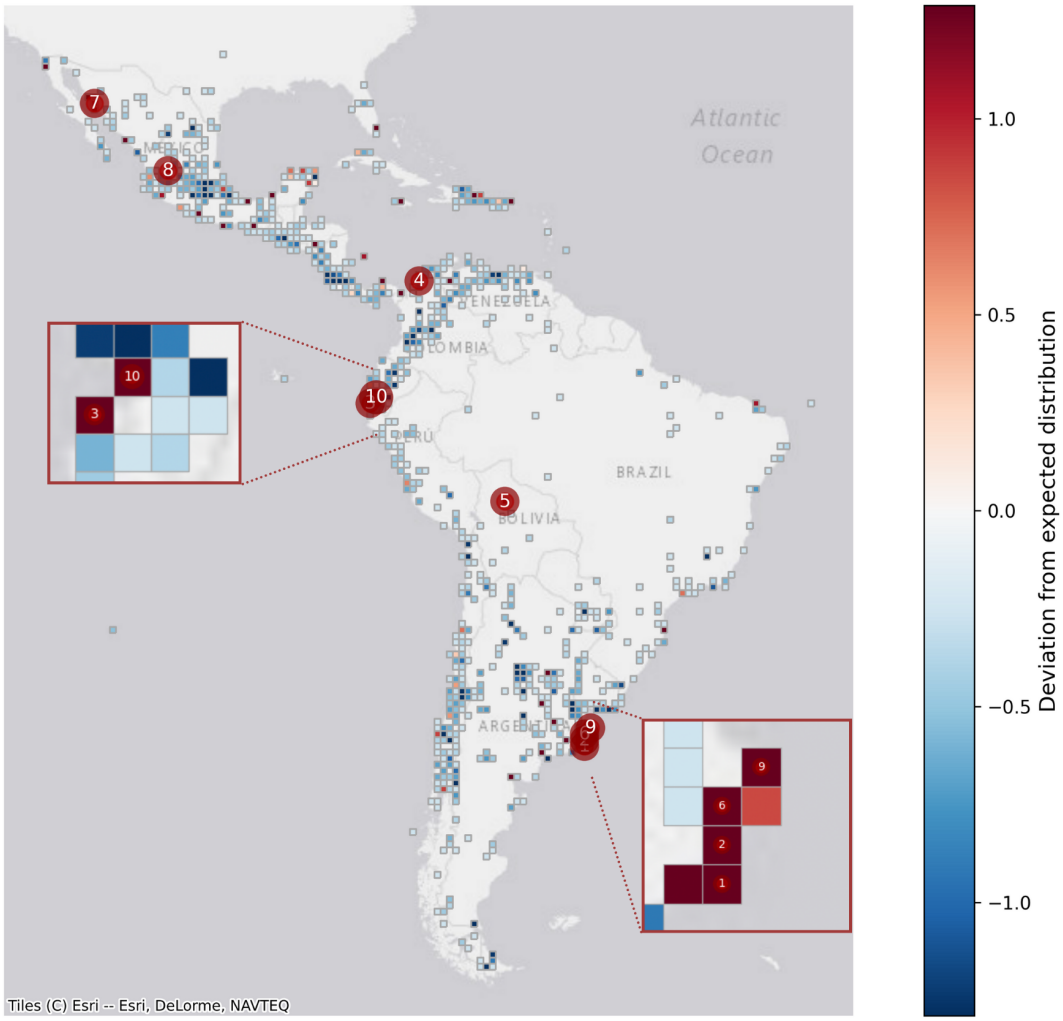


(a) Spatial distribution of Tweets mentioning *playa*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Cancún, Mexico	20.91	-86.95	Yes
2	Faro de José Ignacio, Uruguay	-34.85	-54.62	Yes
3	Playa Hermosa, Costa Rica	10.56	-85.61	Yes
4	Playa de Santo Domingo, Chile	-33.74	-71.46	Yes
5	Necochea, Argentina	-38.63	-58.66	Yes
6	Iqueque, Chile	-19.96	-70.11	Yes
7	Encarnación, Paraguay	-27.36	-55.96	Yes
8	Ciudad de la Costa, Uruguay	-34.85	-55.96	Yes
9	Juanillo, Dominican Republic	18.38	-68.09	Yes
10	Acapulco, Mexico	17.09	-99.76	Yes

(b) Top hotspots for Tweets mentioning *playa*.

Figure 4.1: (a) Spatial distribution of Tweets mentioning *playa*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

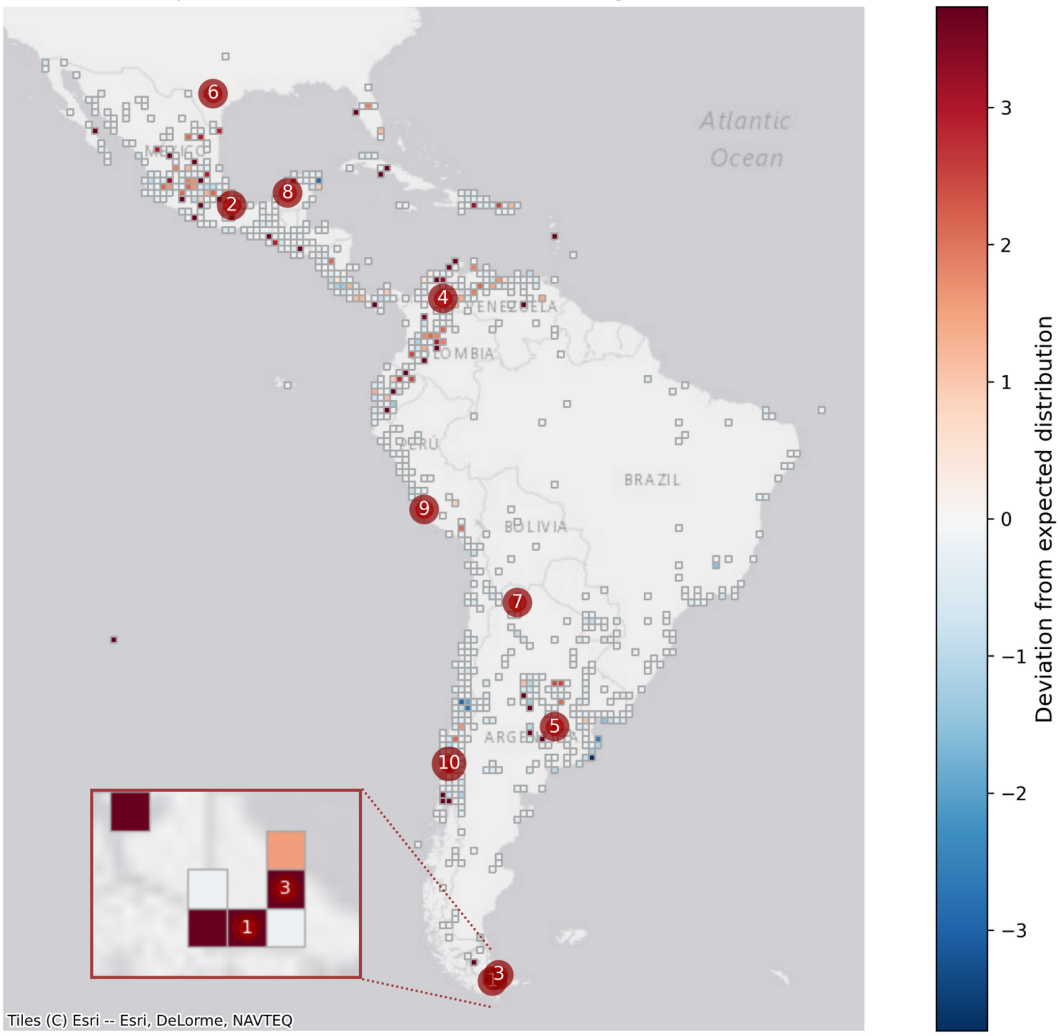


(a) Spatial distribution of Tweets mentioning *mar*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Mar del Plata, Argentina	-38.10	-57.31	Yes
2	Mar de las Pampas, Argentina	-37.57	-57.31	Yes
3	Bocapán, Peru	-3.53	-80.89	Yes
4	El Cerro, Colombia	9.89	-75.50	Yes
5	San Bartolo, Bolivia	-14.17	-66.07	No
6	Valeria del Mar, Argentina	-37.03	-57.31	Yes
7	Cerro Prieto, Mexico	28.26	-111.21	Yes
8	San Miguel, Mexico	21.54	-103.12	Yes
9	Las Toninas, Argentina	-36.49	-56.64	Yes
10	Isla Puná, Ecuador	-2.86	-80.22	Yes

(b) Top hotspots for Tweets mentioning *mar*.

Figure 4.2: (a) Spatial distribution of Tweets mentioning *mar*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

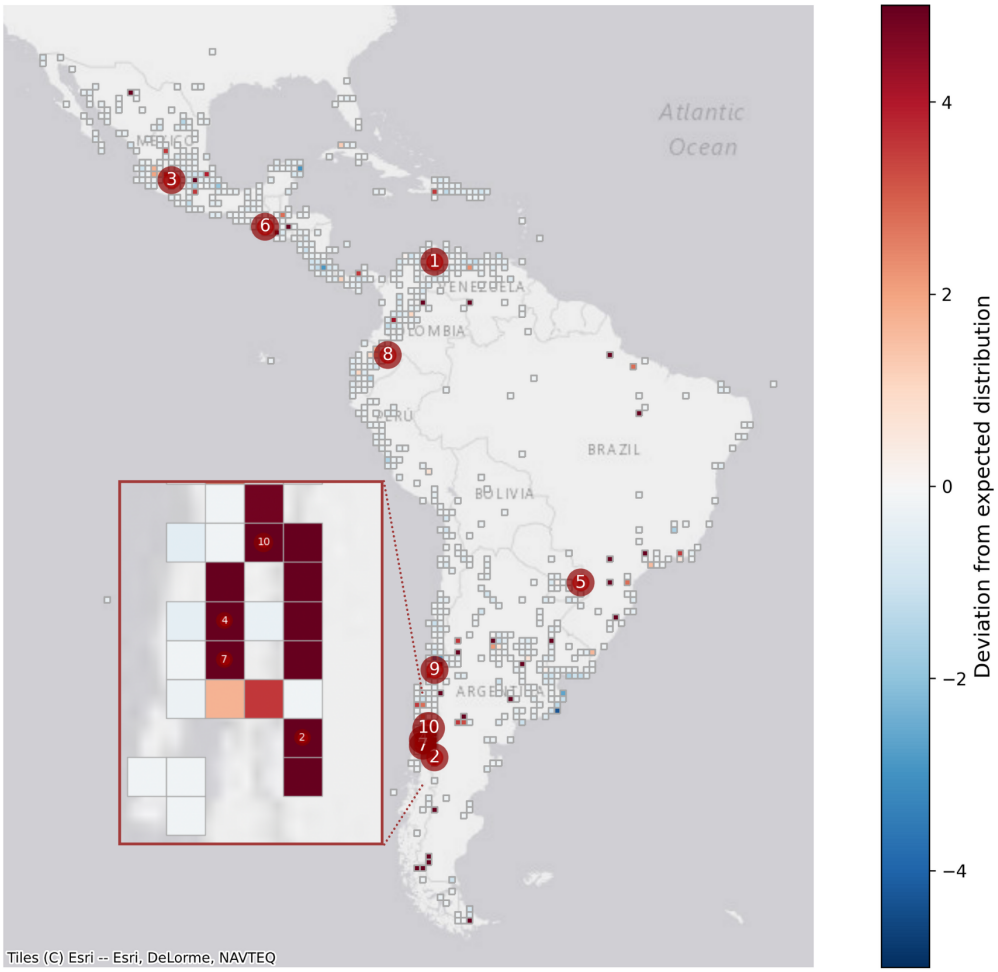


(a) Spatial distribution of Tweets mentioning *tierra*.

Hotspot ID	Place	Latitude	Longitude
1	Tierra del Fuego, Argentina	-54.65	-68.09
2	Tierre Blanca, Mexico	18.38	-96.39
3	Tierra del Fuego, Argentina	-54.25	-67.42
4	Guamalito, Colombia	8.56	-73.48
5	Pehuajó, Argentina	-35.40	-61.35
6	San Antonio, USA	29.44	-98.41
7	Purmamarca, Argentina	-23.71	-65.39
8	Nohakal, Mexico	19.65	-90.32
9	Ica, Peru	-14.17	-75.50
10	Cholchol, Chile	-38.63	-72.81

(b) Top hotspots for Tweets mentioning *tierra*.

Figure 4.3: (a) Spatial distribution of Tweets mentioning *tierra*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

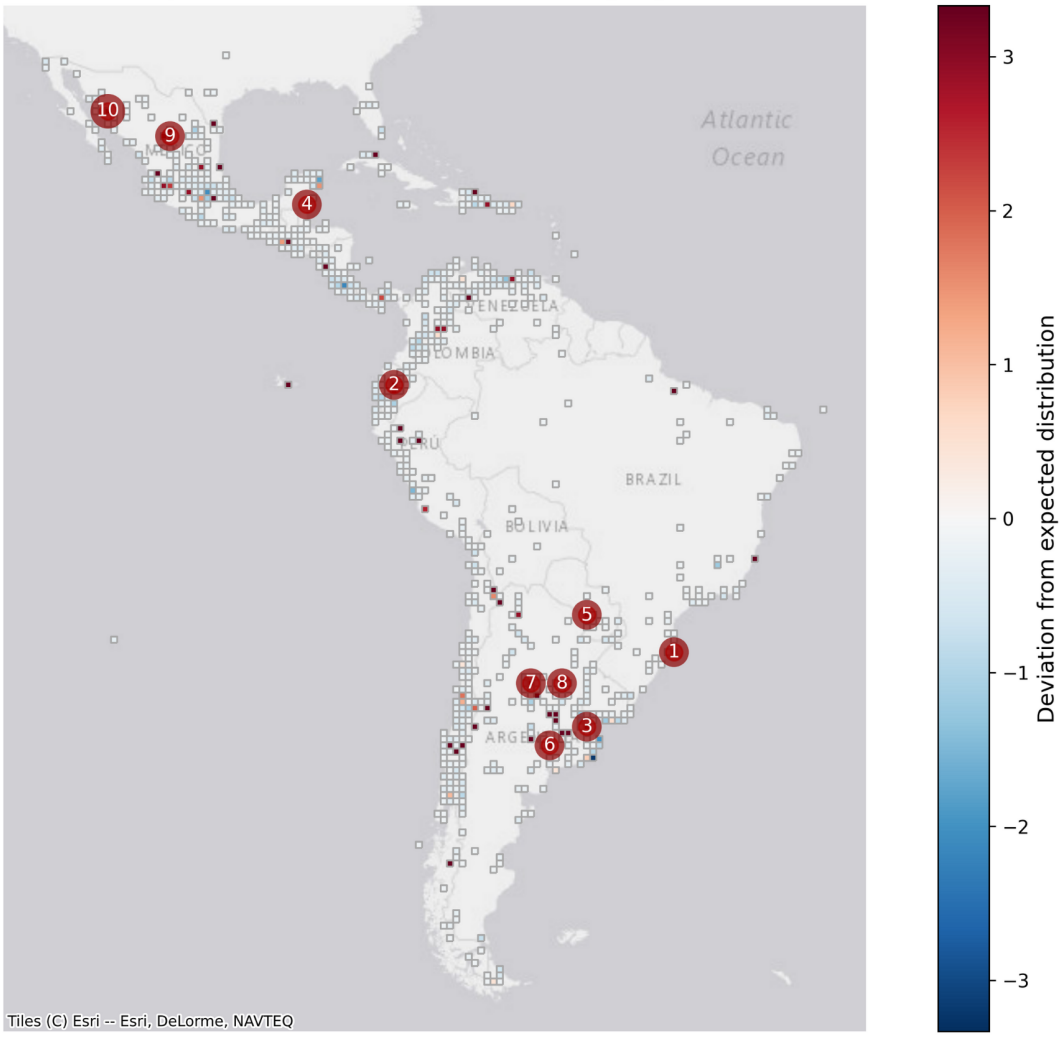


(a) Spatial distribution of Tweets mentioning *lago*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Lago de Maracaibo, Venezuela	10.56	-71.46	Yes
2	Lago Epuyén, Argentina	-42.22	-71.46	Yes
3	Lago de Pátzcuaro, Mexico	19.65	-101.78	Yes
4	Lago Rupanco, Chile	-40.70	-72.81	Yes
5	Lago de Itaipu, Paraguay	-25.55	-54.62	Yes
6	Lago de Atitlán, Guatemala	14.50	-91.00	Yes
7	Lago Llanquihue, Chile	-41.21	-72.81	Yes
8	Jivino Verde, Ecuador	-0.16	-76.85	No
9	Lago Rapel, Chile	-34.29	-71.46	Yes
10	Lago Panguipulli, Chile	-39.67	-72.13	Yes

(b) Top hotspots for Tweets mentioning *lago*.

Figure 4.4: (a) Spatial distribution of Tweets mentioning *lago*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

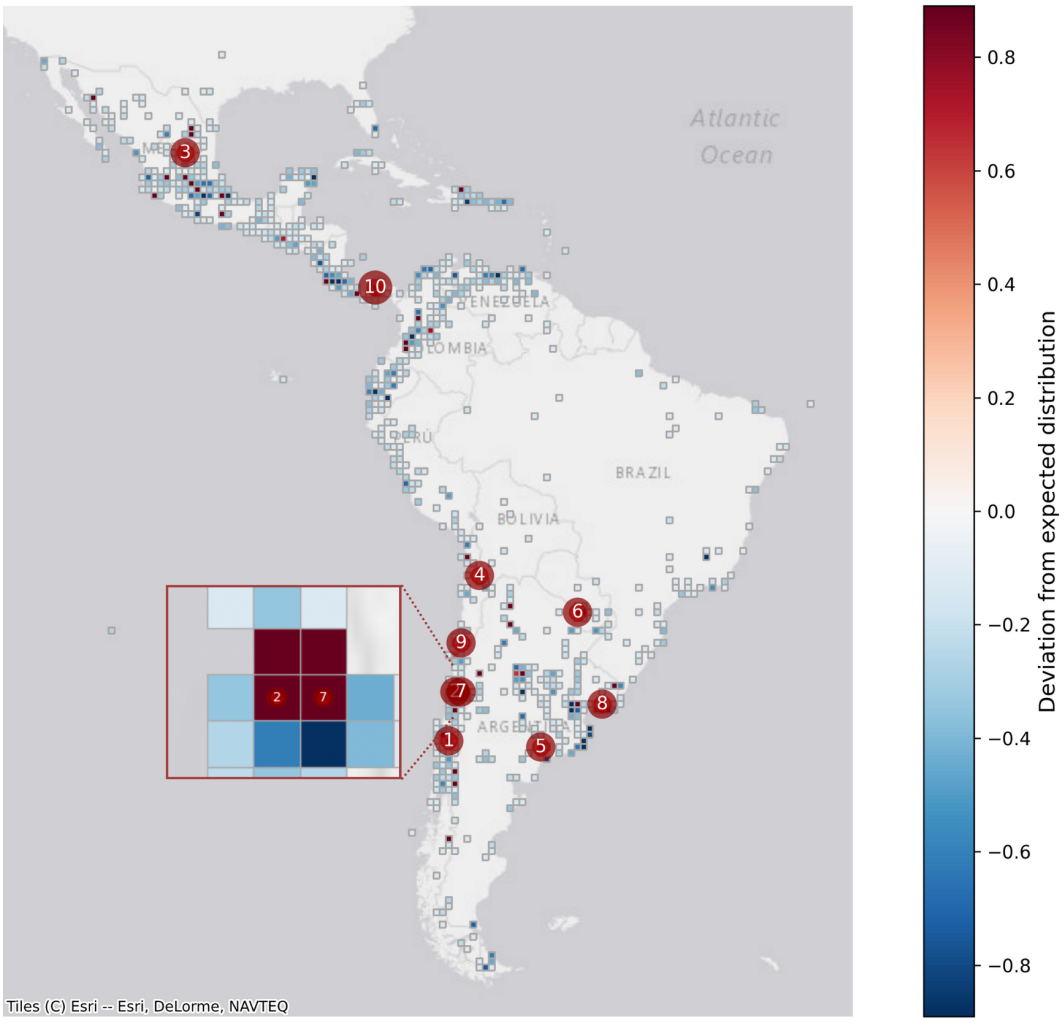


(a) Spatial distribution of Tweets mentioning *laguna*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Laguna, Brasil	-28.55	-48.55	Yes
2	Laguna de Quilotoa, Ecuador	-0.84	-78.87	Yes
3	Laguna de Chascomús, Argentina	-35.40	-57.98	Yes
4	Chetumal, Mexico	18.38	-88.30	Yes
5	Laguna Blanca, Argentina	-24.94	-57.98	Yes
6	Laguna Alsina, Argentina	-37.03	-62.03	Yes
7	Felipa Laguna Azul, Argentina	-31.47	-64.05	Yes
8	Laguna Aguilar, Argentina	-31.47	-60.68	Yes
9	La Laguna, Mexico	25.25	-103.12	No
10	Laguna del Nainari, Mexico	27.66	-109.86	Yes

(b) Top hotspots for Tweets mentioning *laguna*.

Figure 4.5: (a) Spatial distribution of Tweets mentioning *laguna*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

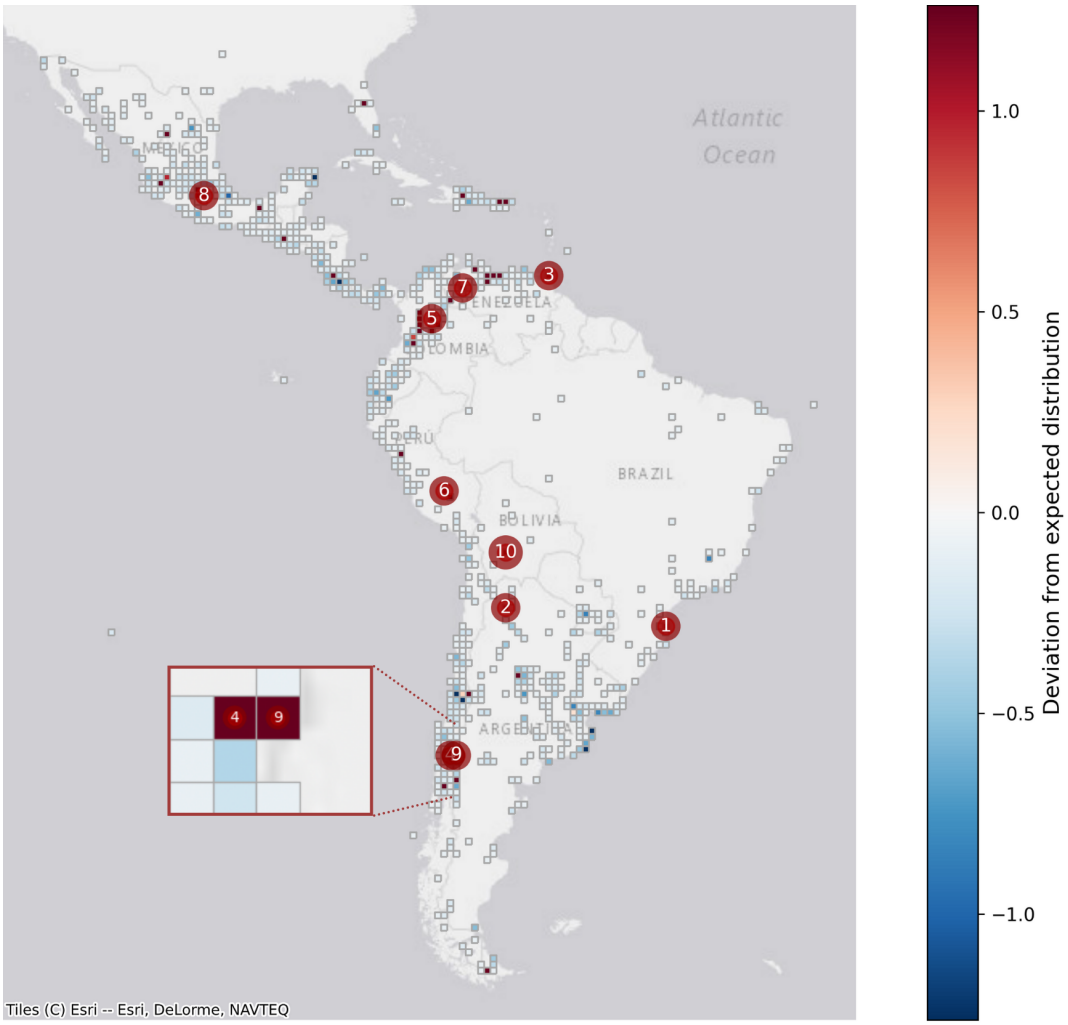


(a) Spatial distribution of Tweets mentioning *cerro*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Los Ángeles, Chile	-37.57	-72.13	No, Streetname
2	Valparaíso, Chile	-33.17	-71.46	Hills in the City
3	Cerro El Quemado, Mexico	23.41	-101.10	Yes
4	Nave Ew El Abra, Chile	-21.85	-68.76	No, Saltmine
5	Cerro Tres Picos, Argentina	-38.10	-62.03	Yes
6	Estadio Defensores del Chaco, Paraguay	-25.55	-57.98	No, Football Club
7	Cerro San Cristóbal, Chile	-33.17	-70.78	Yes
8	Aguas Blancas, Uruguay	-34.29	-55.29	No
9	Cerro La Cantera, Chile	-28.55	-70.78	Yes
10	Cerro Cam, Panama	9.23	-80.22	No, Villagename

(b) Top hotspots for Tweets mentioning *cerro*.

Figure 4.6: (a) Spatial distribution of Tweets mentioning *cerro*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

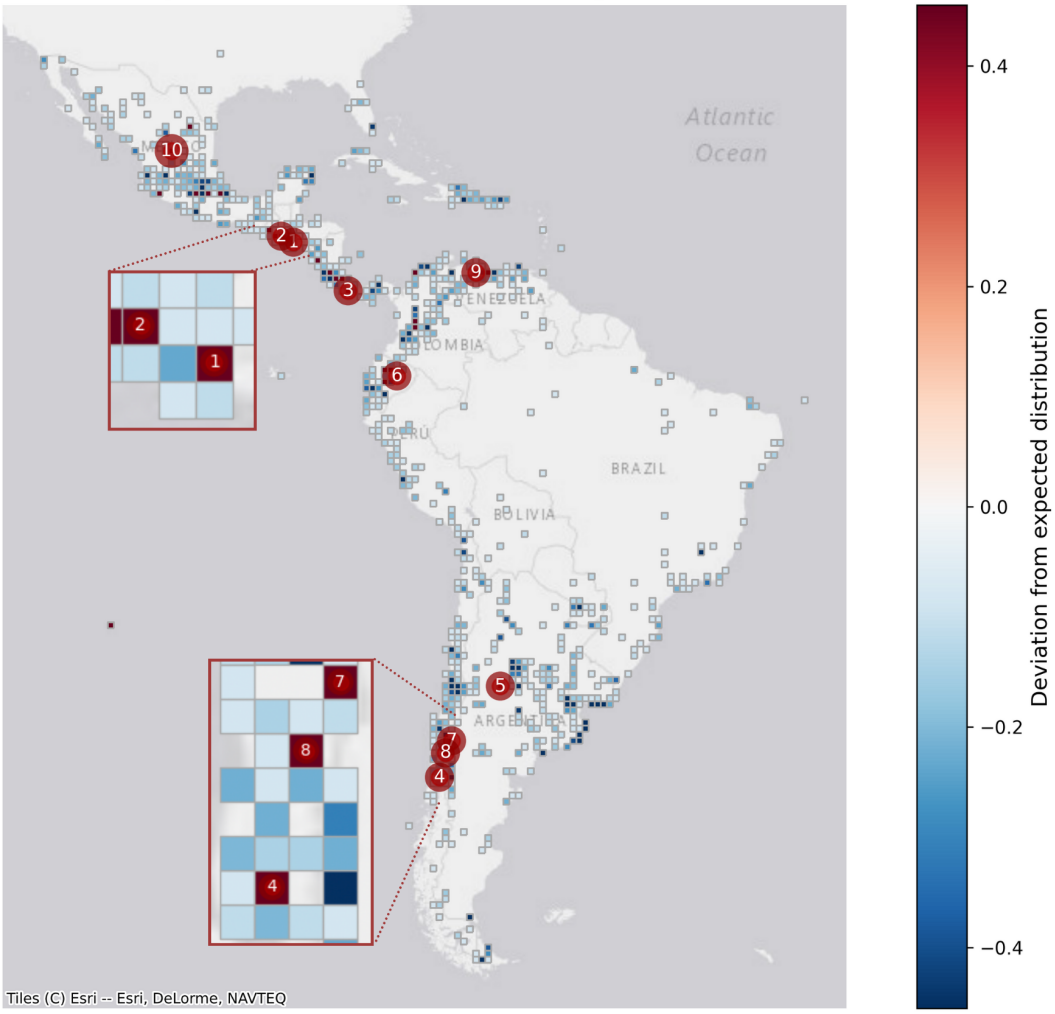


(a) Spatial distribution of Tweets mentioning *montaña*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Buenos Aires, Argentina	-26.76	-48.55	No, Rollercoaster
2	Nevado de Cachi, Chile	-24.94	-66.07	Yes
3	Port of Spain, Trinidad and Tobago	10.56	-61.35	No
4	Llaima, Chile	-38.63	-72.13	Yes
5	Monserrate, Colombia	5.89	-74.15	Yes
6	Huayna Picchu, Peru	-12.86	-72.81	Yes
7	La Puerta, Venezuela	9.23	-70.78	Somehow, Hiking Area
8	Jardines en la Montaña, Mexico	19.01	-99.08	City District, Mexico City
9	Lonquimay, Chile	-38.63	-71.46	Yes
10	Montañas de Colores, Bolivia	-19.33	-66.07	Yes

(b) Top hotspots for Tweets mentioning *montaña*.

Figure 4.7: (a) Spatial distribution of Tweets mentioning *montaña*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.



(a) Spatial distribution of Tweets mentioning *volcán*.

Hotspot ID	Place	Latitude	Longitude	Connection
1	Volcán de San Salvador, El Salvador	13.85	-88.98	Yes
2	Guatemala City, Guatemala	14.50	-90.32	No
3	Volcán Barú, Panama	8.56	-82.91	Yes
4	Volcán Osorno, Chile	-41.21	-72.81	Yes
5	El Volcán, Argentina	-33.17	-66.07	No, City-District
6	Volcán Reventador, Ecuador	-0.84	-77.52	Yes
7	Volcán Tolhuaca, Chile	-38.10	-71.46	Yes
8	Volcán Villarrica, Chile	-39.15	-72.13	Yes
9	San Felipe, Venezuela	10.56	-68.76	No
10	Villa de Cos, Mexico	23.41	-102.45	No

(b) Top hotspots for Tweets mentioning *volcán*.

Figure 4.8: (a) Spatial distribution of Tweets mentioning *volcán*, highlighting significant hotspots based on Chi-Square residuals; (b) Details of top hotspots.

4.2.10 Summary of the Spatial Analysis

It should be noted that not all terms were associated with a large number of Tweets. Overall, only 2.53% of Tweets were sent with geotags, which makes a spatial analysis difficult. Nevertheless, in most cases 8-10 hotspots can be associated with a suitable term. For many terms, 1-2 hotspots were also found where the landscape term was used in a non-literal sense or differently than expected. It should also be noted that the scales for the different maps are different and not uniform. These are adapted to the individual terms and, in principle, it can be said that the more Tweets were found for a landscape term, the greater the range of the respective scale.

Overall, the spatial analysis successfully identifies clusters. Terms such as *playa* (Beach), *mar* (Sea) or *montaña* (Mountain) show strong spatial patterns in relation to each other. However, some hotspots were found in urban centers with high population density. Other places were incorrectly identified as hotspots, because in the Tweets a metaphorical sense of the term is discussed. However, the analysis highlighted the potential of spatial distribution and patterns in relation to understanding how landscapes are discussed and perceived in social media.

4.3 Temporal Analysis

The goal of the temporal analysis was to find differences in the frequency of Tweets. That is, the actual number of Tweets mentioning a landscape term on a particular day was compared with the expected Tweets on that day. The heatmap 4.9 highlights variations in term frequency over time. The rows represent a specific landscape term, and the columns represent a day in the year. Red colors represent a positive anomaly, i.e., a higher number of Tweets than expected, and blue colors represent negative anomalies, i.e., fewer Tweets than expected. This chapter shows the heat maps for the terms with a statistically significant pattern.

Therefore, a high deviation means more or fewer Tweets than expected about this landscape term, ranging from red to blue (shades). For example, an event such as a volcanic eruption can be dated relatively accurately since many Tweets with the term *volcán* are likely to be made immediately before or after an eruption. This graphic serves as an overview, and more detailed graphics of relevant terms will follow in this section.

Table 4.3 shows the significance of the distribution of the individual terms. In this case, *volcán*, *tierra* and *playa* each have a P-value of 0.0, which means that the annual distribution of Tweets, compared with the expected distribution, represents a statistically significant pattern. This suggests that events or seasonal trends influence these terms.

Other terms, such as *bosque* (Forest), *atardecer* (Sunset) or *amanecer* (Sunrise), have a P-value of 1. This means that there are no significant deviations in their frequency distribution. Overall, the statistical tests, together with the visual inspection of the distribution, suggest that there are certain landscape concepts that warrant more detailed analysis. In the following, these terms are presented in more detail and illustrated in individual graphics.

Table 4.3: P-values for landscape terms from daily anomaly analysis using chi-squared tests.

Term	**P-Value**
volcán	0.000000
tierra	0.000000
playa	0.000000
naturaleza	0.000627
montaña	0.264957
lago	0.498004
valle	0.999165
desierto	0.999758
mar	0.999818
selva	0.999995
cerro	0.999998
río	0.999998
cascada	1.000000
paisaje	1.000000
bosque	1.000000
atardecer	1.000000
horizonte	1.000000
amanecer	1.000000

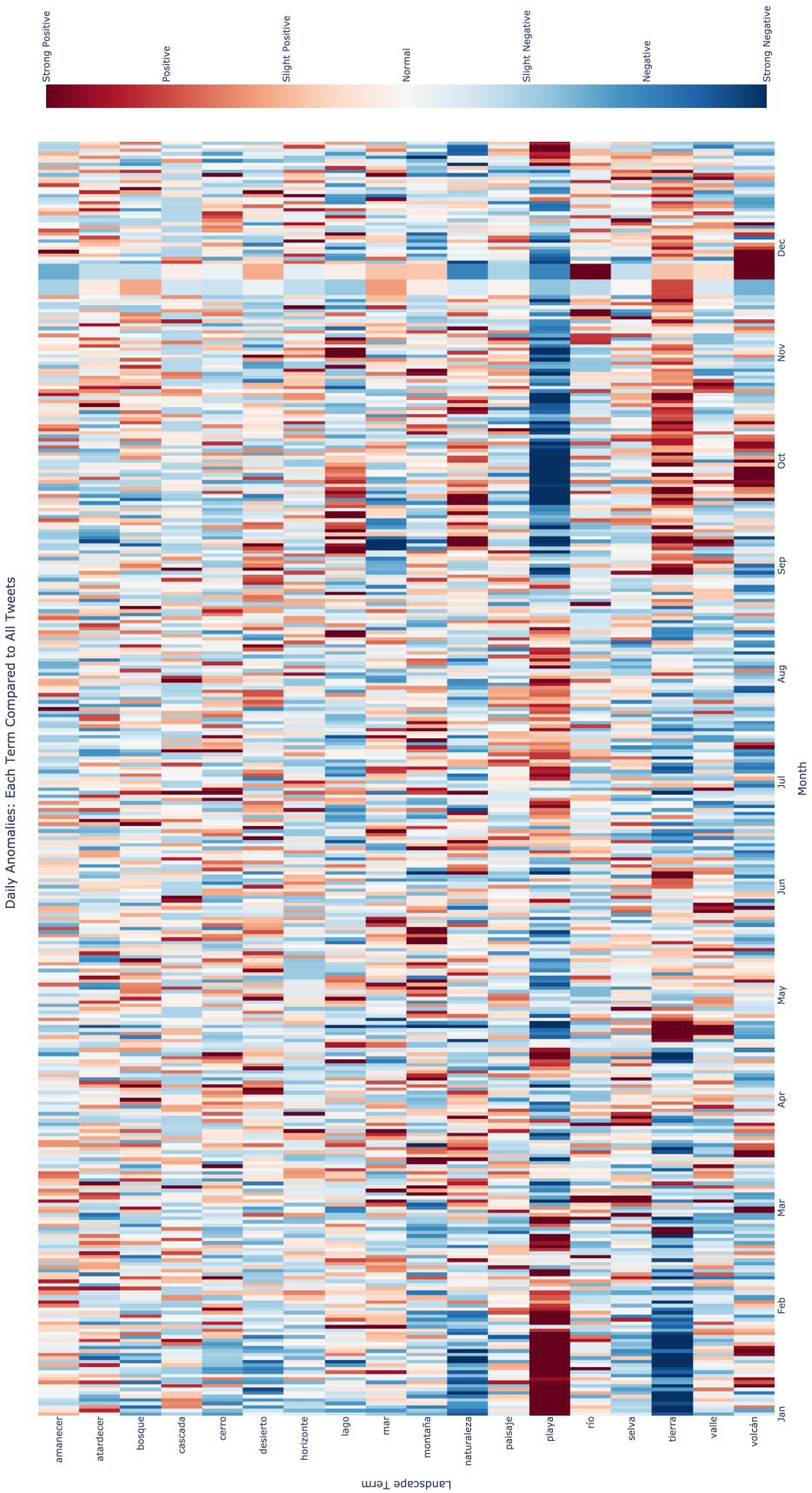


Figure 4.9: Daily anomalies of Tweets including a landscape term compared to all Tweets.

4.3.1 Montaña

The chart shows the daily deviations of Tweets containing the term *montaña* (Mountain). Shown over one year. The abbreviations on the scale were chosen for reasons of space and stand for substantial positive deviation (S+), moderate positive deviation (M+), neutrald deviation (N), moderate negative deviation (N-), and substantial negative deviation (N-). The positive residuals are shown in red tones and indicate days when the frequency of Tweets with *montaña* was unexpectedly high compared to the baseline. In contrast, negative residuals are shown in shades of blue, indicating days when *montaña* was found less often in Tweets. This description also applies to the following four graphs with the words *naturaleza* (Nature), *playa* (Beach), *tierra* (Earth), and *volcán* (Volcano).

In the months of May and June, there are significant spikes in the positive residuals. This indicates that a lot of Tweets about *montaña* were posted during these periods. At the same time, it can also be seen that there is a positive concentration of deviations in spring and summer and at the end of fall.

Negative deviations are constant throughout the year, although it is noticeable that there are hardly any positive deviations and almost only negative deviations in December and January.

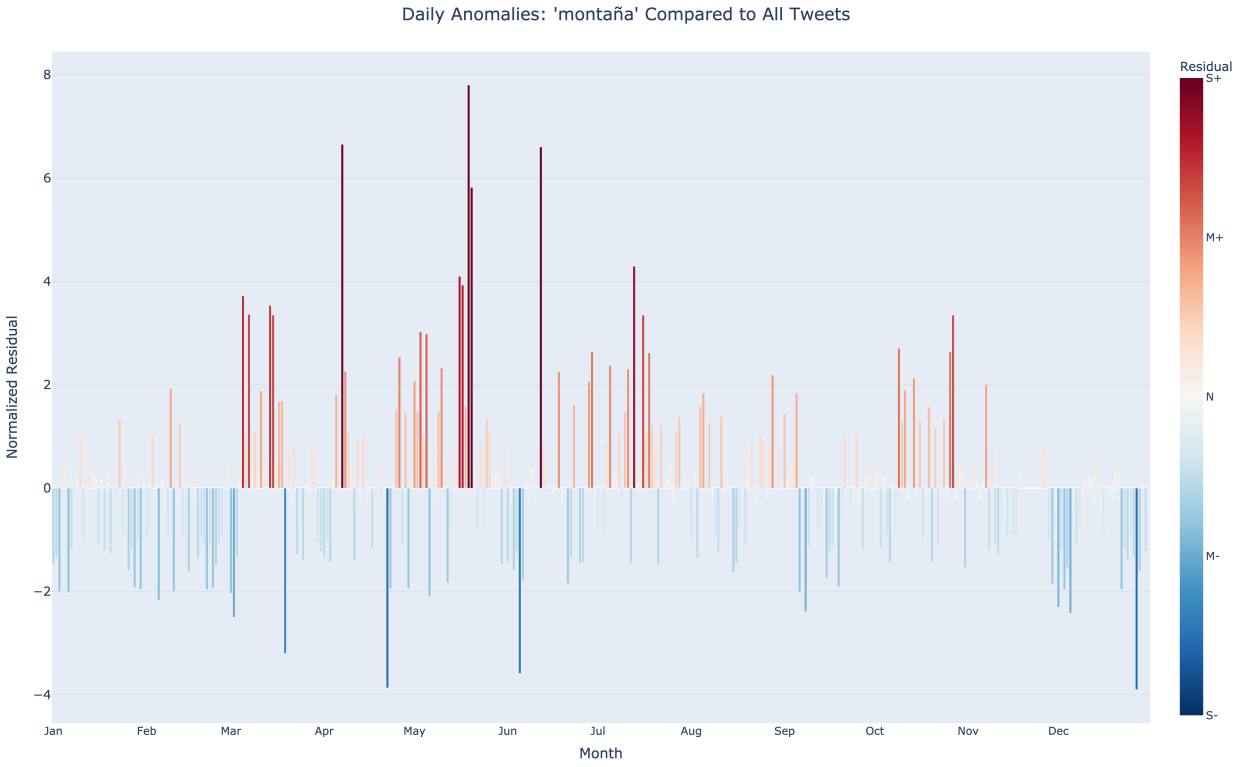


Figure 4.10: Daily anomalies for Tweets containing *montaña*, highlighting significant deviations from the expected distribution based on normalized Chi-Square residuals.

4.3.2 Naturaleza

The term *naturaleza* (Nature) also shows significant peaks in April and September. Especially in September, there are extremely high residuals. These anomalies could be related to specific events. At the same time, a roughly constant positive trend can be seen throughout the year. Only in January and December are the expected Tweets significantly lower than in the rest of the year.

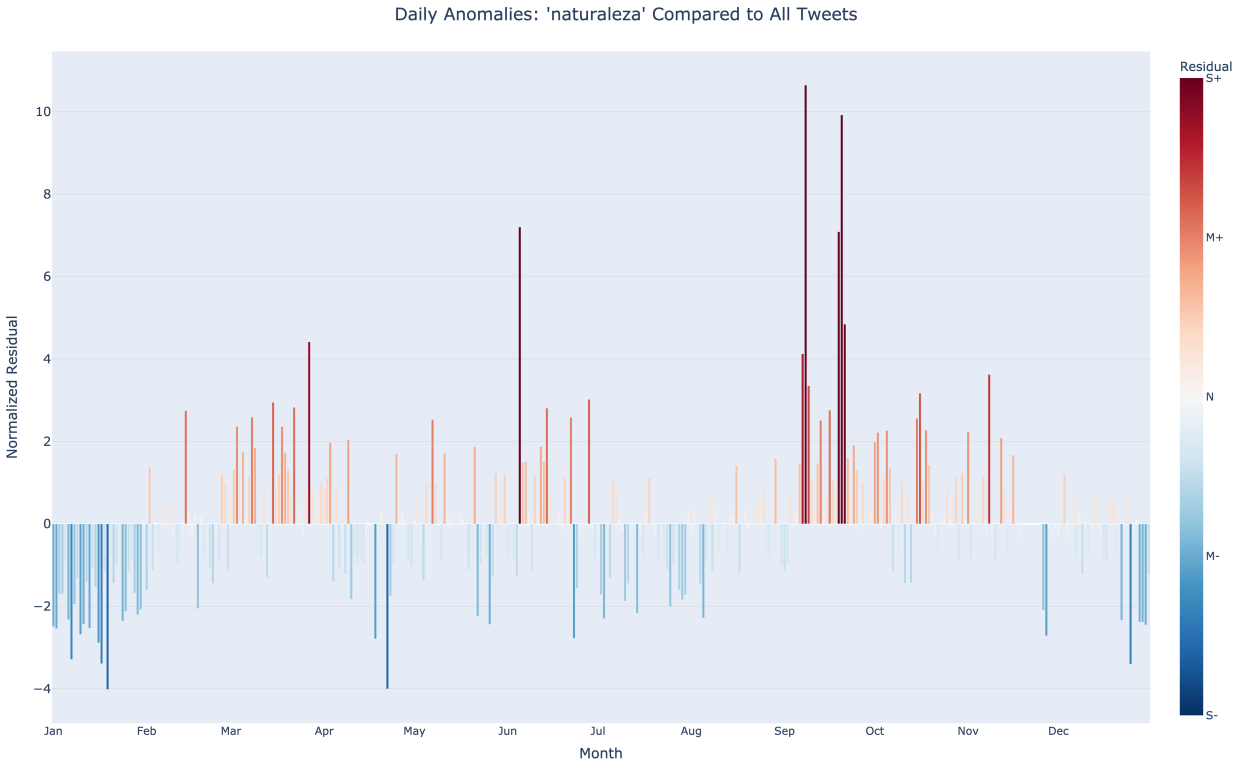


Figure 4.11: Daily anomalies for Tweets containing *naturaleza*, highlighting significant deviations from the expected distribution based on normalized Chi-Square residuals.

4.3.3 Playa

For the term *playa* (Beach), very significant anomalies can be seen at the beginning of the year. In January and February, extreme constantly high, i.e. positive, deviations can be observed. Furthermore, in the middle of the year, two months, July and August, can be seen again, in which positive trends are noticeable. Negative residuals are especially noticeable in fall and spring, in May and later in November. A seasonal trend in the distribution of Tweets containing *playa* can be clearly seen.

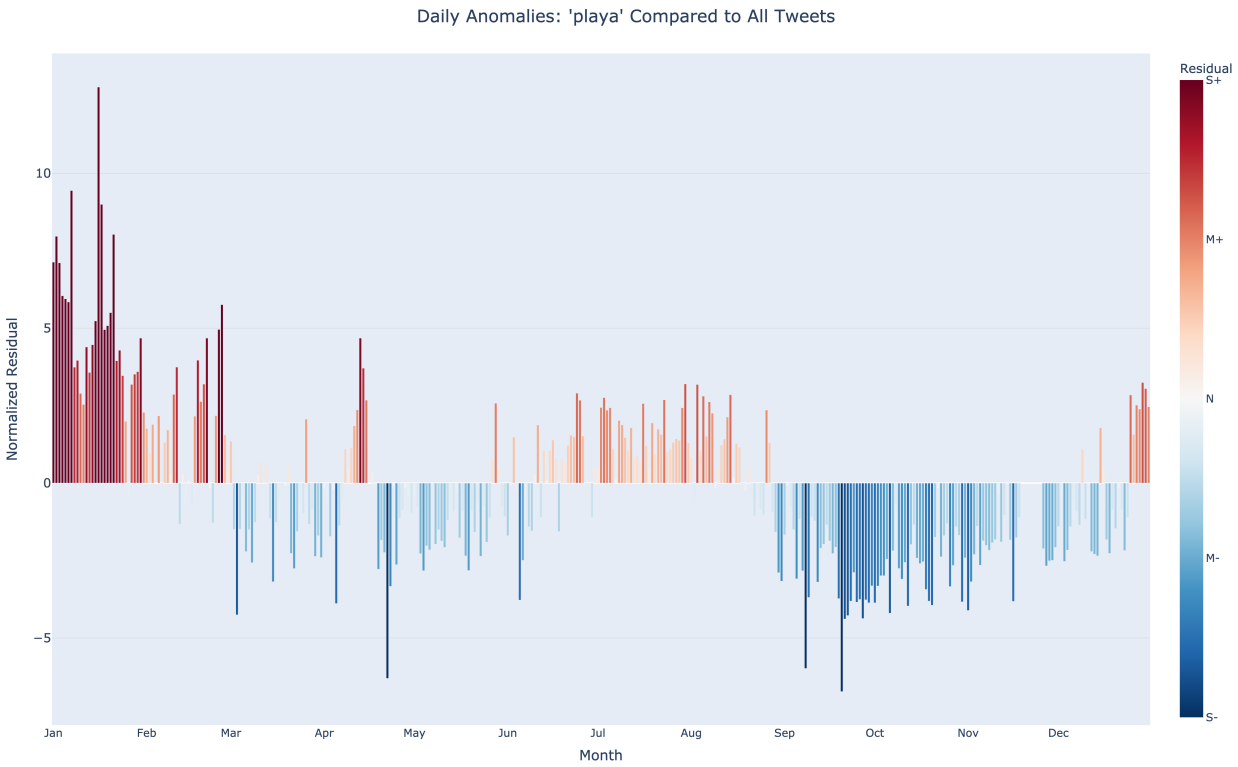


Figure 4.12: Daily anomalies for Tweets containing *playa*, highlighting significant deviations from the expected distribution based on normalized Chi-Square residuals.

4.3.4 Tierra

For the term *tierra* (Earth), there are also noticeable deviations from the expected Tweets. In May in particular, there are a few days – a short event - that show extremely positive deviations. This indicates a major event that triggered discussions around *tierra*. In the rest of the year, the deviations are somewhat more moderate. In the later months of the year, from September to the end of the year, there is a positive trend. Only at the end of February were there two days when a lot was Tweeted about *tierra*, which may also indicate an event.

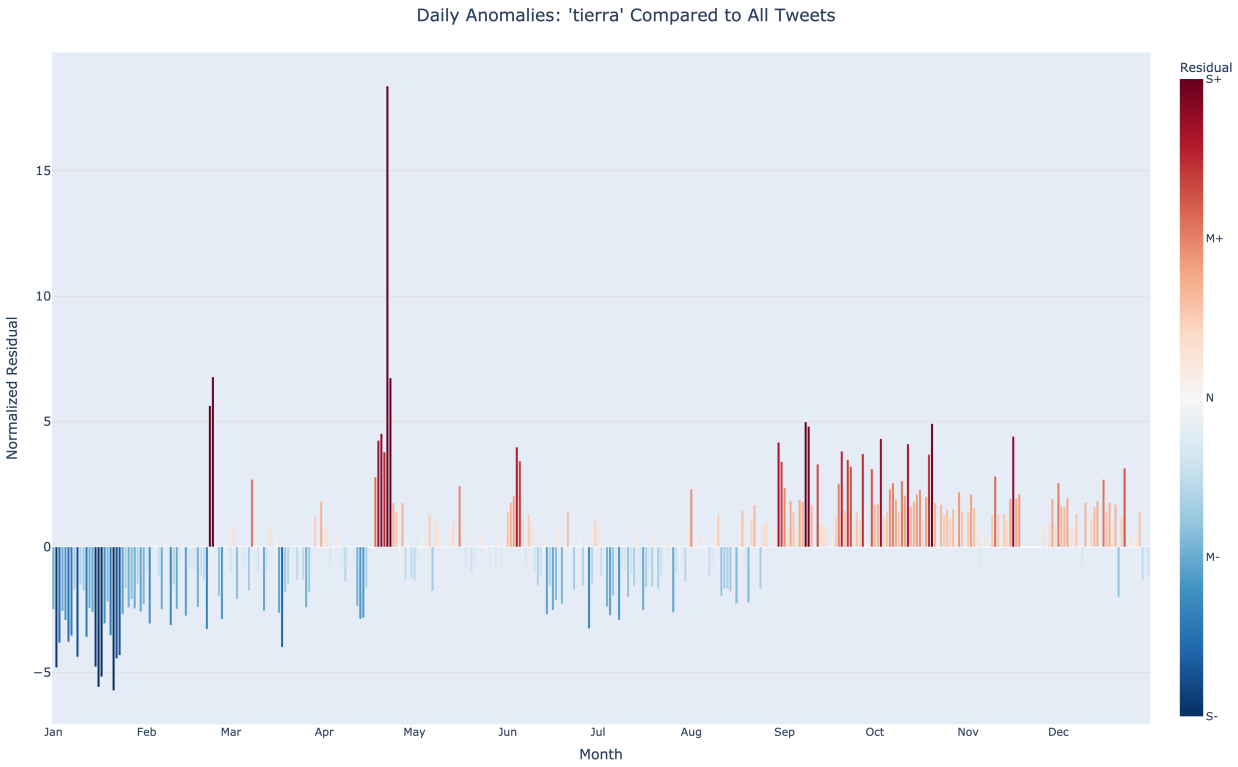


Figure 4.13: Daily anomalies for Tweets containing *tierra*, highlighting significant deviations from the expected distribution based on normalized Chi-Square residuals.

4.3.5 Volcán

For the term *volcán* (Volcano), there are significant deviations from the expected distribution in October and December. Short events, each showing a few days of extremely high positive deviations. In addition, there are always short and smaller peaks throughout the year, occurring in different months, such as in March or even in January. Otherwise, the distribution of the term *volcán* shows a rather negative trend throughout. Especially from May to August, there is a consistent negative deviation from the expected value, which is only interrupted by a few days when a positive deviation can be found.

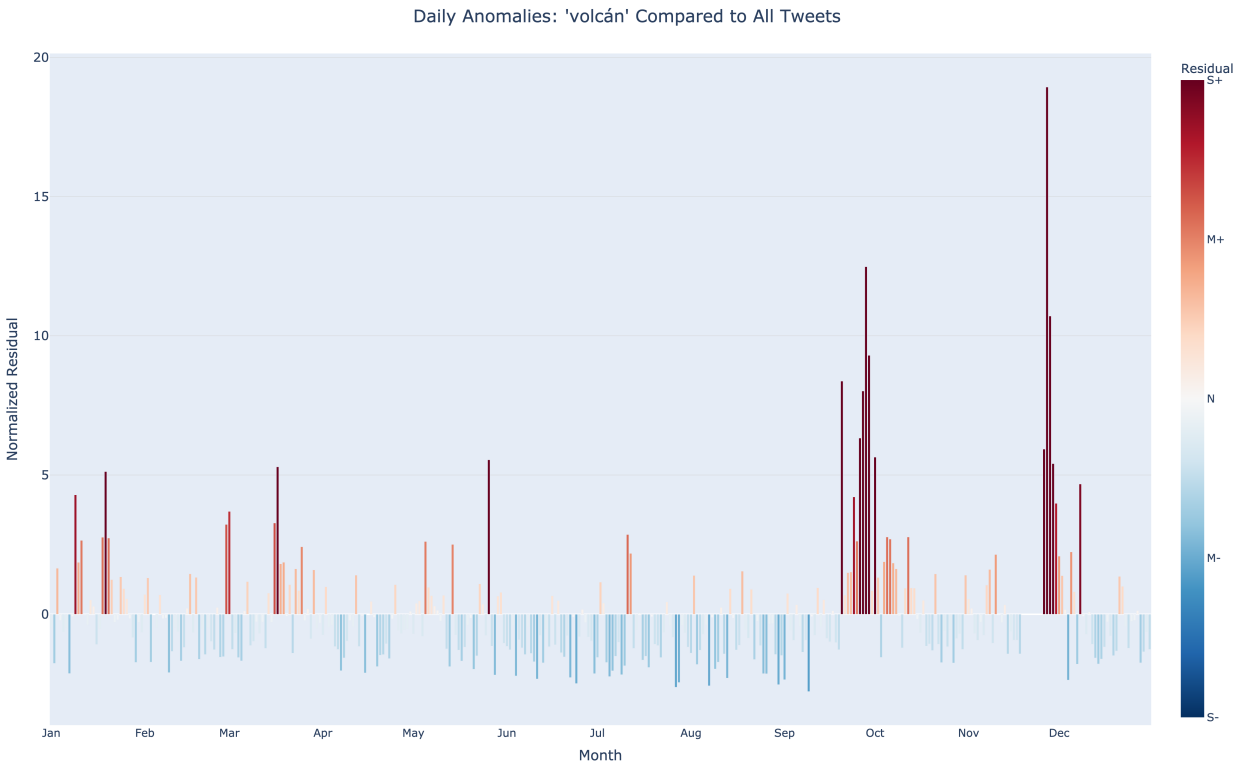


Figure 4.14: Daily anomalies for Tweets containing *volcán*, highlighting significant deviations from the expected distribution based on normalized Chi-Square residuals.

4.4 Co-Occurence

The third analysis looks at the words used in connection with the landscape terms. Below are seven examples of a word cloud, each representing a term on co-occurring Terms. The size of a word indicates the frequency with which this word occurs with the examined word based on log-likelihood. The larger the word, the more often it occurs in the data set. This makes it relatively easy to visually identify significant topics and patterns in the discussion of certain terms on Twitter. For example, for the term *playa* (Beach), words such as *sol* (Sun), *quiero* (I love/like) or Carmen (the name of a beach in Cancún) are dominant. The color of the individual words does not refer to anything in particular; it was randomly generated.

Another special pair of word clouds was created for the terms *amanecer* (Sunrise) and *madrugada* (Dawn or a synonym of *amanecer*). These were selected because it is likely that words that mean the same thing also generate similar word clouds. In other words, for the two words *madruaga* and *amanecer*, a similar distribution of co-occurring words should be observed.

Figure 4.15: Word cloud showing the top associated words for the landscape term *playa*, based on their log-likelihood values.

Figure 4.16: Word cloud showing the top associated words for the landscape term *volcán*, based on their log-likelihood values.

Figure 4.17: Word cloud showing the top associated words for the landscape term *colina*, based on their log-likelihood values.

Figure 4.18: Word cloud showing the top associated words for the landscape term *amanecer*, based on their log-likelihood values.

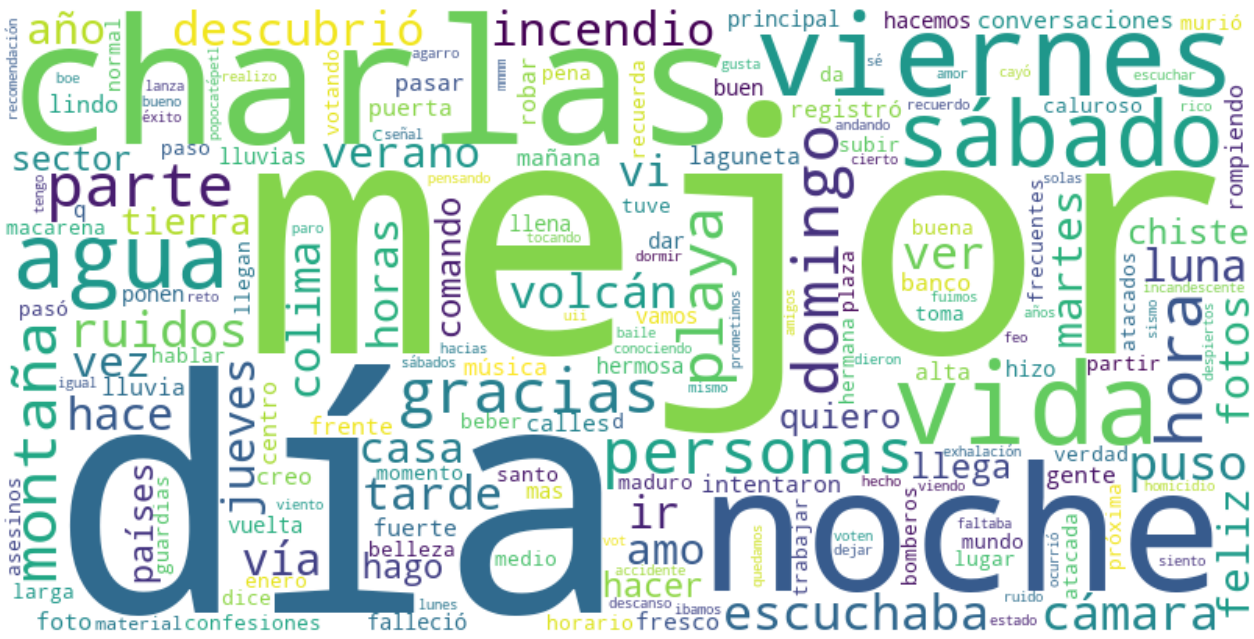


Figure 4.19: Word cloud showing the top associated words for the landscape term *madrugada*, based on their log-likelihood values.

Figure 4.20: Word cloud showing the top associated words for the landscape term *paisaje*, based on their log-likelihood values.

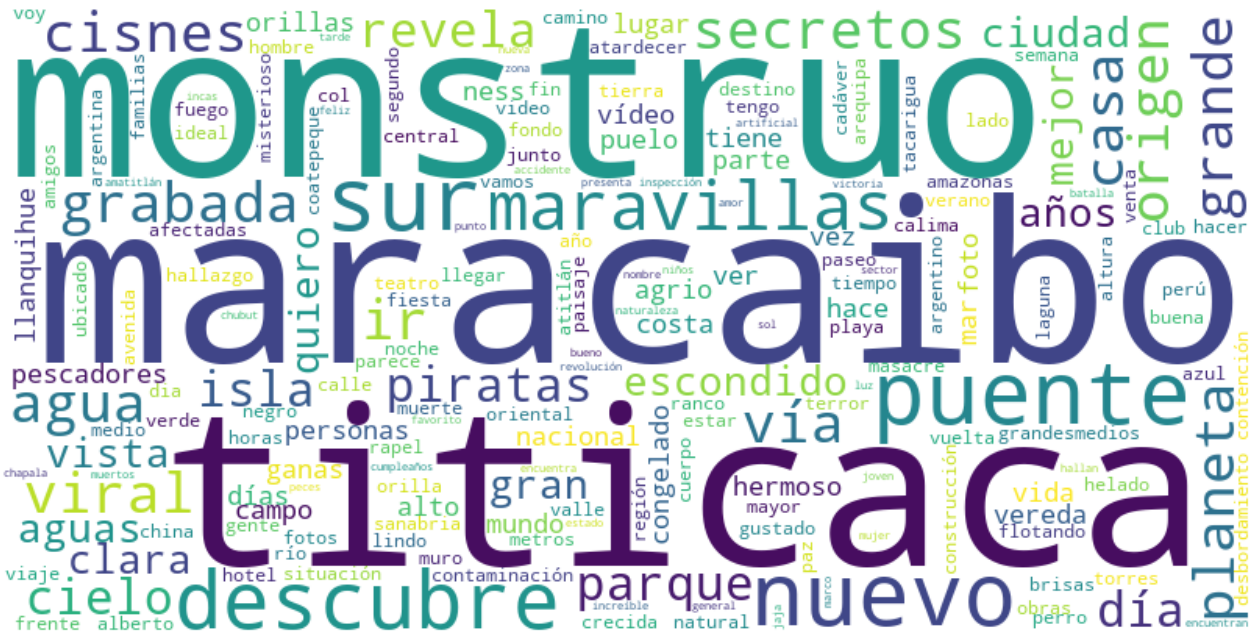


Figure 4.21: Word cloud showing the top associated words for the landscape term *lago*, based on their log-likelihood values.

In general, it can be said that the co-occurrence analysis also generated a useful result. The words that appear in the word cloud are related to the terms examined. A more detailed analysis of the individual outputs is provided in the discussion.

4.5 Summary of Results

The analysis focused on spatial, temporal, and co-occurrence patterns of Tweets containing a landscape term to find out how these terms are discussed and perceived on social media. Each analysis provides different insights into the data, thus enabling a better understanding of the data.

The spatial distribution of Tweets reveals patterns that can be used to draw the following conclusions: Terms such as *playa* (Beach), *mar* (Sea), or *montaña* (Mountain) have strong spatial clusters, often in tourist regions known for their landscapes. Cancún in Mexico is an example of this. Non-literal uses of the terms have also made it onto the lists of hotspots, such as *cerro* (Hill) being the name of a football club. This shows that the terms can also take on different meanings. The difficulty in the spatial analysis was primarily that only 2.53% of Tweets are geotagged.

The temporal analysis shows variations in the frequency of Tweets about landscapes over time. Some terms show statistically significant deviations, indicating specific seasonal trends or events. The term *volcán* (Volcano) displayed individual spikes that are based on events, and the term *playa* (Beach) showed seasonal trends that could be related to vacation times.

The co-occurrence analysis explored the context of landscape terms through frequently occurring words. Word clouds were chosen to highlight the dominant words in relation to the individual terms, allowing for a visual analysis. At first glance, the words that appear correspond to expected associations. Together, the three analyses provide a better understanding of how landscape terms are used on Twitter. This multidimensional approach offers insights into landscapes' cultural, environmental, and social dynamics. The results will be discussed in more detail in the following section.

5 Discussion

Following, the results are interpreted and linked to Tweets. The spatial distribution of Tweets will be examined first, followed by their temporal distribution and finally the topic of co-occurrence. Subsequently, the results will be linked to the literature mentioned at the beginning and to the research questions posed and answered. To round off the chapter, limitations in the process and results will be discussed.

5.1 Spatial Analysis

5.1.1 Discussion of the Results

Overall, it must to be said that the spatial distribution proved more difficult to execute than expected. This is because only 2.53% of all filtered Tweets were geotagged. The spatial distribution of *playa* (Beach) is an example where the spatial analysis worked very well. It can be clearly seen that the hotspots of the Chi-Square analysis are located exclusively in coastal regions. Every single hotspot found is directly related to the beach. In most Tweets, hotspot number 1 in Cancún, Mexico, is associated with Playa del Carmen. This region is a very touristy area associated with various beach activities.

> **Original**: "*El paisaje más hermoso que mis ojos pueden mirar (...) beach in playa del carmen, (...)*
> **Translation**: *The most beautiful landscape that my eyes can see (...) Beach in Playa del Carmen (...).*

There are a lot of Tweets about the beach and the activities undertaken there. Terms such as *hermoso/hermosa* (beautiful) are often used and most Tweets use a number of emojis with hearts, beaches and the sun. A common feature is a link to a photo. Playa del Carmen is by far the most geotagged place with over 220 mentions in the Tweets. No other place has a similar count. The Tweets all point to a positive experience at the beach, which is also reinforced by the emojis used.

Similar patterns can be seen at the other hotspots identified by the spatial analysis. Considering a second example, hotspot 7, Encarnación in Paraguay. At first glance, it is not obvious that this is a beach, because the city is located inland. However, a closer look with the help of Google Maps shows that beaches can be found here too – just beaches on a riverbank.

> **Original**: "*Que empiece una hermosa semana!! @ playa san josé, encarnación paraguay*"
> **Translation**: *How to start a beautiful week! @san josé beach, encarnación, paraguay.*

As can be seen in the quoted Tweet, the beach is also perceived as something very positive. Other users write *mi segunda casa*, which can be translated as *my second home*. Overall, the Tweets here are also all positive and suggest beach activities.

In contrast to *playa*, the analysis of the spatial distribution using other terms also worked, but only to a certain extent. As a second example, *lago* (Lake) is analyzed in more detail. For this term, too, 10 of the 10 hotspots found can be directly linked to a lake. While over 400 Tweets were written which were directly related to the 10 hotspots of *playa*, only 60 were written for the hotspots related to *lago*. The Lago de Maracaibo in Venezuela was identified as hotspot 1. Of the 17 Tweets written

about this place, only seven are related to the lake or Veradea del lago, which is a park area right by the lake in Maracaibo, the city.

> **Original**: "*Mi hermoso lago de maracaibo en maracaibo, venezuela*"
> **Translation**: My beautiful lake of Maracaibo en Maracaibo, Venezuela.

Other Tweets mention, for example, a shopping center (Lago Park) or other residential areas. Certain Tweets are also advertisements for the rental of a beautiful apartment located directly on the lake. Overall, it is striking that although a hotspot was identified at a lake, only a small proportion of Tweets actually use the landscape term *lake*. Compared to Tweets about *playa*, there are also far fewer Tweets, and of those Tweets, a significantly smaller number are directly related.

Nevertheless, some Tweets are also found that describe the landscape as such. For example, a Tweet about hotspot 6, Lago de Atitlán in Guatemala:

> **Original**: "*Todos lo que hemos podido disfrutar de este paisaje coincidimos en que el lago atitlán es el más hermoso (...)*"
> **Translation**: *All of us who have been able to enjoy this landscape agree that Lake Atitlán is the most beauitiful (...).*

Overall, it can be said that the identification of hotspots worked relatively well, and that a lot is actually written about the landscape and its characteristics at these locations. For the various terms, the more Tweets available, the better the analysis worked. For example, for the term *volcán* (Volcano), only 30 Tweets were found for the identified hotspots, and only 6 out of 10 places could be directly linked to a volcano. Furthermore, some of the places were not actually volcanoes, but, for example, a picnic area at the foot of a volcano that has *volcán* in its name. As for hotspot 1, Volcán de San Salvador in El Salvador. In some of the Tweets found there, the volcano itself is not mentioned, but Plaza Volcán, which is a picnic area.

> **Original**: *"felíz cumpleaños mamá @ plaza volcán"*
> **Translation**: *Happy Birthday, Mom @ plaza volcán.*

5.1.2 Limitations & Conclusion regarding Spatial Analysis

The larger the data set, the better the results, since more Tweets will be found for respective locations. The most significant difficulty in spatial analysis is the small number of Tweets that are actually geotagged. When deciding on the topic of this work, it was assumed that a much more significant proportion of Tweets could have been used for spatial analysis and that the user location might be more helpful. But most of the user locations were either empty, filled with a fantasy location or a big city. For such a fine-grained analysis, with the aim of finding hotspots of scenic places, such user locations would only have distorted the result. Suppose a user is at the beach Playa del Carmen and tweets about the beauty of the place but has entered *Mexico City* as the location: the results of a spatial analysis would simply be a map of cities with a high population. The method used was thus able to identify places that serve as important scenic spots for recreation or as excursion destinations in most cases.

5.2 Temporal Analysis

5.2.1 Discussion of the Results

In contrast to spatial analysis, each Tweet has a timestamp and an exact time at which the Tweet was published. This results in a large amount of data that can be used for temporal analysis. Even from a broad perspective, it is clear that there are significant patterns. This was also confirmed by the p-value, which assumes a highly significant value for specific terms (*volcán, tierra, playa*, and *naturaleza*). This means that distribution according to the Chi-Square analysis is not random but that the deviation from the expected distribution of Tweets reveals specific patterns. These patterns will be analyzed and interpreted in more detail in the following section.

The term *playa* (Beach), in particular, shows very strong seasonal trends. This suggests that Tweets containing *playa* are very strongly influenced by the holiday season. For Cancún, for example, the place identified as hotspot 1, the primary holiday season is between December and March, as in the southern hemisphere, the seasons are reversed. Significant positive deviations coincide with vacation periods, with the period from December to March showing an even stronger trend. Nevertheless, the months between mid-June and the end of September also indicate that the word *playa* is used more frequently during vacation periods.

This is also supported by the months that show a negative trend, in particular, the shoulder months from March to June and from September to December. During this time, fewer Tweets are posted about *playa*, suggesting that the beach is mainly associated with vacations. For example, on January 16, more than 300 Tweets were found that were written in connection with *playa*. Again, places like Playa del Carmen in Mexico or Playa de Ayolas in Paraguay were mentioned. In addition, the dominant topics are tourism and leisure. For example, the following Tweet:

> **Original**: *"ir a la playa en vacaciones con mis amigas y quedarnos mucho tiempo"*
> **Translation**: *go to the beach for vacation with my friends and we stay for a long time.*

Often, the desire for beach, sun and swimming is also expressed. As for example in this Tweet:

> **Original**: *"me urge ir a la playa"*
> **Translation**: *I need to go to the beach urgently.*

On the other hand, another topic that is often repeated is a less pleasant event. On January 16, there was a shooting at a music festival that left 5 people dead.

> **Original**: *"cinco muertos y 15 heridos en playa del carmen por un tiroteo durante el festival bpm"*
> **Translation**: *Five dead and 15 wounded in Playa del Carmen because of a shooting during the bpm festival.*

A Google search confirms that this event took place on January 16. This tragic event was tweeted about a lot and is also reflected in the deviation from the expected Tweets. Otherwise, it can be said that the distribution of Tweets about *playa* coincides very closely with vacation periods.

Another landscape concept that shows a noticeable pattern is *volcán* (Volcano). For this term, it is not so much the striking seasonal trends but rather short periods of a few days that represent a

substantial deviation from the expected distribution. For example, we see a relatively large deviation at the end of September, at the end of November, or in mid-March.

On November 27, 2017, 73 Tweets containing the term *volcán* were classified as relevant. However, most of the Tweets are about the Agung volcano in Bali, Indonesia, and the fact that 100,000 people had to be evacuated. A Google search confirms that the volcano erupted on November 25, 2017. At the same time, however, some Tweets are also written about the volcano Popocatépetl in Mexico: ash dispersal and general volcanic activity are mentioned there. It is also written that a second crater has been found.

> **Original**: *"(...) hallan segundo cráter dentro del volcán popocatépetl (Link to a Video)"*
> **Translation**: *(...) second crater found inside the Popocatepetl volcano (Link to a Video)*

This can also be confirmed by an annual report from the "Subdirección de Riesgos Volcánicos" (Volcanic Hazards Subdirectorate), the National Center for Disaster Prevention (Caballero Jiménez et al., 2017). This report summarizes all the data and activities related to the volcano Popocatépetl and confirms the Tweets written.

On September 28, around another peak in the deviation from the normal distribution, a lot is written about the volcano Popocatépetl. The report already mentioned confirms this.

> **Original**: *"volcán popocatépetl lanzó cenizas en el centro de méxico y registró 2 sismos volcanotectónicos"*
> **Translation**: *Vulcano Popocatépetl spewed ash in central Mexico and 2 earthquakes (from volcanic activities) were recorded.*

At the same time, however, the Volcán de Fuego in Guatemala was also often mentioned. The Tweets also contain information about a possible eruption, along with links to videos and photos. This can also be confirmed by a website operated by the National Museum of Natural History (Global Volcanism Program, 2017). It talks about a lava flow that is almost 2 km long, which may not break any records this year, but is one of the more active volcanoes in the *Volcán de Fuego*.

On September 28, other volcanoes are also mentioned, but not on the same scale as the two mentioned. There are also tweets about the *Volcan Agung* in Bali or the *Volcán Reventador* in Ecuador. However, the latter is more about the majestic appearance of the volcano and less about current activities.

The third example of the temporal distribution of Tweets discusses the term *naturaleza* (Nature). Here, too, visually significant differences in annual distribution can be perceived (4.11). In April, there is a moderate peak, and in September and October, there is a huge positive anomaly. At the same time, there are fewer Tweets about *naturaleza* in the holiday months, i.e., January, July, and August. This could indicate that there is generally less online activity during the holiday months.

On September 8th, more than 180 Tweets were written that were classified as relevant and contained the term *naturaleza*. Many natural disasters, such as hurricanes like Irma, Jose, or Katia, were discussed. Earthquakes in Mexico are also mentioned.

> **Original**: *"la naturaleza enseñando a que cuiden el planeta, primero huracán, ahora un temblor en méxico."*

Translation: *Nature teaching us to take care of the planet, first a hurricane, now an earthquake in mexico.*

Original: *"3 huracanes, el temblor, ¿por qué eres así madre naturaleza?"*
Translation: *3 hurricanes, the tremor, why are you like this mother nature?*

However, the word *naturaleza* is often used in connection with *Madre Naturaleza* (Mother Nature). This would actually be a non-literal use of the term and is thus a mistake in the machine learning. Nevertheless, the Tweets are relevant to the topic, even if only negative events in connection with *Madre Naturaleza* are mentioned.

Other Tweets, however, also describe the beauty of nature and link to a photo.

Original: *"la naturaleza es tan maravillosa (Link to Foto)"*
Translation: *Nature is so wonderful (Link to Foto).*

In contrast to the term *volcán*, where one could mainly find current volcanic eruptions through temporal analysis, the term *naturaleza* is less about mentioning direct events and more about nature in general in the context of natural disasters or as a reflection of human activities. In addition, Tweets often state that nature is beautiful and express a specific emotional response to nature. In summary, it can be said that Tweets oscillate between the beauty of nature and the fear of how humanity treats nature. However, a direct event that could have triggered the high positive anomaly could not be found.

For September 20th, another day with a high positive anomaly, the pattern is similar. Various natural disasters are mentioned, including the various hurricanes and earthquakes, and at the same time, general concerns related to nature and climate change are mentioned. Worth mentioning, as tweeted by various people, is also the fear that the nuclear tests from North Korea could have a connection with the natural disasters.

Original: *"no sé sí será paranoia, pero tendrán algo q ver los experimentos nucleares d norcorea con recientes eventos catastróficos d la naturaleza?"*
Translation: *I don't know if it's paranoia, but do the nuclear experiments in North Korea have anything to do with recent catastrophic natural events?*

5.2.2 Limitations & Conclusion regarding Temporal Analysis

In general, temporal analysis provides exciting insights into the events or activities related to landscape terms. Different patterns can be identified for different terms based on positive and negative anomalies. For example, very clear seasonal trends can be identified for *playa* or *lago*, which are probably related to vacations. Especially *playa* makes this trend very clear. For other terms, such as *volcán* or *naturaleza*, it is not so much the seasonal trends that stand out, but rather isolated events, often related to natural disasters. For example, volcanic eruptions can be identified very precisely for the term *volcán*. And for the term *naturaleza*, general natural disasters are mentioned, but less drastic events can also be identified. Furthermore, the analysis is always associated with a manual effort and does not directly provide insights into patterns. These have to be read from the context and analyzed individually for each day. Nevertheless, temporal analysis provides very exciting insights. Also, the analysis works only for a low number of terms. For terms such as *amanecer* (Sunrise) or *río* (River), the distributions are too homogeneous and give no insights on

patterns or events that might have influenced the distribution. However, this is also due to the fact that not so many Tweets were classified as relevant and contained these terms.

It would certainly also be interesting to compare different years with each other. The seasonal trends should appear similar over the years, but special events should be different each time.

5.3 Co-Occurence

5.3.1 Discussion of the Results

For the co-occurrence analysis, a word cloud was created for each term, a number of which can be found in Chapter 4, containing the results. Additionally, the top 25 terms with the highest log-likelihood score were added to the appendix as a table. The co-occurrence analysis had two main goals: Finding various connections and associations between landscape terms and the words that often appear with them and illuminating the context or topics for the respective terms. In addition, the word clouds provide a good visualization for quickly getting an overview of which words are often used in connection with landscape terms.

Here, too, the term *playa* (Beach) is analyzed in more detail. This is also because many Tweets were found and the analysis is therefore meaningful. The most prominent words used in connection with *playa* are as follows:

- **Carmen**: This refers to the beach, Playa del Carmen in Cancún, Mexico. This has already been mentioned in the spatial analysis and the temporal analysis.
- **Querio**: This word means "I want". It indicates a desire or intention to go to the beach, possibly in connection with vacation or leisure activities.
- **Arena** (Sand), **mar** (Sea), **sol** (Sun): These are all words that are strongly associated with the beach anyway.

These words are the most frequently used, while the following words appear less often. The word *photos* is used frequently, not least because a photo was linked in many Tweets. This suggests that people want to share their experiences (positively) on social media. *Disfrutar* means to enjoy and indicates a positive experience at the beach. Overall, most of the words point to vacations, beach activities and leisure. The words that were extracted and that indicate emotions are mostly in a positive context. They are words like *quiero*, *disfrutar* or *necesito* (I want).

For the term *volcán*, a completely different pattern appears when analyzing its word cloud. The following words are dominant here:

- **Erupción** (Eruption): A word that is to be expected in connection with volcanoes.
- **Vista** (View): Also a word that could be expected because there are many volcanoes in South and Central America that are popular hiking destinations and offer a beautiful view of the surrounding area.
- **Vivo** (Live): At first glance, it is not clear why this word was used so often in connection with volcanoes. With the help of an SQL query, it was determined that this was a data error. The actual goal would have been to filter out Retweets, but here the posts are written by a

bot with the same content: a link to a live webcam with a view of a volcano. This bot has written so many Tweets that even the word "bot" appears in the word cloud.

- **Names of volcanoes**: Here, volcanoes such as **Popocatépetl** in Mexico (already mentioned in the temporal analysis), **Turrialba** in Costa Rica, or **Agung** (also mentioned in the temporal analysis) in Indonesia are mentioned. All of them are prominent volcanoes that either had something to do with an eruption in 2017 or are popular destinations for excursions.

In addition, words such as *explosión* (Explosion), *cenizas* (Ashes), *activo* (Active), and *alerta* (Alert) are used in connection with volcanoes, which fits well thematically. Overall, however, one can find words here that describe volcanic activity and its global influence. The focus is on eruptions, ash clouds, fires, and other typical terms for volcanoes. An interesting exception is *chocolate*. Many Tweets with the content *Volcán de Chocolate* were found. Machine learning probably interpreted *chocolate* as a name and, therefore, classified the Tweets as relevant because they have the same pattern. However, this is actually a reference to a chocolate volcano, a kind of muffin with a molten chocolate center.

Thirdly, *paisaje* (Landscape) is discussed in more detail. Many of the most frequent words are found in the context of a song lyric that was also classified as relevant by machine learning: "las nubes grises también forman parte del paisaje" (the gray clouds are also part of the landscape). In this sense, if the song has been interpreted correctly, it is associated with bad experiences in a romance and these belonging to a relationship just as much as gray clouds do to a landscape. What is interesting is that this is not a Retweet, and it is not the same person sending these lines; many different people tweet the same line of the song at different times. This suggests that the song (Ricardo Arjona - Fuiste tú feat. Gaby Moreno) is very popular in Central and South America. On Spotify, the song has almost 500 million views (as of December 24, 2024), and over 1.3 billion on YouTube, indicating widespread popularity.

Other words that often appear in connection with *paisaje* include, for example:

- **Hermoso**, **bonito**, and **lindo** (Beautiful, Pretty, Lovely): These words express positive emotions and refer to the aesthetic aspects of a landscape.

- **Quiero**, **Disfrutar** (I want, Enjoy): Similar to *playa*, these words express a desire or a longing.

- **Landscape terms**: Words such as **naturaleza** (nature) or **montaña** (mountain) are also used here.

- **Urban aspects**: In contrast to natural elements, terms such as **urbano** (urban), **arquitectura** (architecture), or **casa** (house) are also used.

  Original: "amooo el invierno, amo la lluvia, amo el paisaje y por eso me quiero quedar (...)"
  Translation: "I love the winter, I love the rain, I love the landscape and that's why I want to stay (...)"

Overall, the terms connected to *paisaje* are more diverse than, for example, the terms connected to *volcán*. The words are also strongly influenced by song lyrics. However, the words mentioned are also strongly linked to positive emotion and point to the beauty of nature and the appreciation of

nature. In addition, the word *paisaje* is often used in artistic contexts. Be it song lyrics, poems, photographs or paintings. The beauty of nature is always in the foreground, which indicates that landscape has an essential influence on various areas.

For the term *colina* (Hill), the distribution is more as one expects. The following are among the most frequently mentioned terms in connection with *colina*:

- **Cima** (Summit): It makes sense to be at the summit of a hill or at the top of a hill. It's an achievement that you might want to share on social media, perhaps in connection with a photo.
- **Alta** (High): Describes being high on a hill. It suggests elevations.
- **Sendero** (Path): Sendero is probably often used in the context of hiking. It indicates hiking trails that are often found in hilly areas.
- **Parque** (Park): This is associated with recreational areas, which are often linked to hilly settings.

Overall, there is a core theme that recurs: the word cloud revolves around multifaceted aspects of hilly landscapes. Hiking and the associated discovery of the landscape are emphasized by words such as "sendero", "escalar" (to climb) and "disfrutar" (to enjoy). Words like "parque" highlight the importance of hills as recreational areas. Overall, the word cloud shows that there are different topics discussed on Twitter about the term "colina": on the one hand, there are geographical topics and descriptions of the landscape, and on the other hand, there are leisure activities, all in a positive context.

5.3.2 Limitations & Conclusion regarding Co-Occurence Analysis

The co-occurrence analysis worked very well, and the log-likelihood parameter found the most important words that were tweeted in connection with respective landscape terms.

But there are still problems with their evaluation: especially with the term *paisaje*, it was noticeable that the word cloud and the associated words were strongly influenced by a line from a very popular song in 2017. Since the machine learning algorithm had classified this line as relevant, it appears very often in the data set and strongly influences the result. This can also be seen, for example, with the term *montaña* (Mountain): one of the most frequently used words is *radio* because there is a radio station called "La Montaña Rusa" that is often mentioned in Tweets.

Nevertheless, it is interesting to see how the individual analyses connect: for example, the volcanic eruptions of Popocatépetl are found not only in the temporal analysis but also in the word cloud. By contrast, this volcano is missing in the spatial analysis.

5.4 Comparison to Literature

As mentioned in the literature review, many different methods of geographical information retrieval (GIR) were used here. GIR is a method for using restructured data so that geographical analyses can be carried out at the end. In this work, Tweets from 2017 were structured, analyzed, filtered, and processed so that spatial, temporal, and thematic conclusions could be drawn. At the same time, however, difficulties of working with social media data were also made clear: for example, since only 2.53% of Tweets were actually geotagged, spatial analysis is feasible, but it is also challenging to carry out for specific terms because the amount of data is too small. However, this could be remedied by analyzing all years directly rather than just one year. This, however, goes beyond the scope of this work. Despite the relatively small number of Tweets that are tagged with coordinates, spatial analysis was able to consistently find hotspots that are relevant to individual terms. Thus, for terms such as *playa*, *lago*, or even *mar*, places were found that were really related to the searched term, with only a few exceptions.

The literature review also addressed the perception of landscape and how people conceptualize it. This was mainly examined using a co-occurrence analysis. The terms in connection with the various landscape concepts point to the perceptions and the weighting of landscapes for different people. For example, terms such as *playa* or *colina* are very positively connoted and are often used in connection with leisure activities. In addition, adjectives such as *hermoso/-a* or *bonito/-a* are often used. In addition, verbs that express a desire are often used, such as *quiero* or *disfrutar*. These words strongly suggest that certain landscapes, in this case, beaches, sea, and hills, are perceived very positively and also have an emotional value.

The situation is different, for example, with words like *naturaleza* or *volcán*. In the context of these words, natural disasters or even volcanic eruptions are very often mentioned. This suggests that people are worried and afraid of natural hazards. This topic is often discussed on Twitter nowadays. Overall, as mentioned in the literature review, it is possible to analyze user-generated content and thus create emotions and cultural associations with landscapes.

5.5 Limitations

Limitations mentioned at the beginning were repeatedly highlighted while working with the Twitter data set. Data quality is an issue that influences the analysis of social media content. There are many Tweets that consist only of hashtags and short messages and contain no relevant information. Furthermore, the small number of geotagged Tweets makes performing a meaningful geographic analysis more challenging. Initially, it was assumed that with a geoparser, it would be possible to use the user location to increase the number of geotagged Tweets. However, this would only have led to large cities being displayed on a map because a user would add their hometown as a location. Geoparsing this location, e.g., Mexico City would result in thousands of Tweets being mapped to Mexico City.

The influence of bots and Retweets can also be observed. Retweets could be filtered out correctly, as they all start with "@rt". This led to a significantly smaller data volume. Also, the influence of bots was underestimated. In certain studies, Tweets sent at non-human rates were filtered out – this was not done in this work. As a result, results such as the word cloud associated with *volcán* were heavily influenced by a bot that repeatedly tweeted a webcam image of a volcano. This led to several words appearing in the word cloud, even though they had nothing to do with the term *volcán*.

Another limitation is that Tweets containing places other than those found in Central and South America are also included in the data set. For example, the term "volcano" was associated with many Tweets referring to the eruption of an Indonesian volcano. This is only possible because the Tweets were not geotagged but only mentioned another location by name. Nevertheless, such Tweets naturally distort the analysis if one wants to see how people in South and Central America tweet about landscapes.

Also worth mentioning is the high number of links (to videos or photos) in all Tweets, especially in the geotagged ones. Almost all of the Tweets with coordinates classified as relevant also contained a link to an image or video. This may be because people who like to share something want to share the coordinates. Another possibility, which is considered more likely, is that the machine learning algorithm classified Tweets containing a link and a geographical term as highly relevant. It may be a mixture of these two factors.

5.5.1 Machine Learning

Machine learning played a significant role in generating the resulting maps and analyses, though the algorithm has clear room for improvement. The classification of Tweets was challenging, with the model achieving a precision of 0.56, a recall of 0.48, and an accuracy of 0.68. Precision indicates that nearly 60% of the Tweets identified as relevant were correctly classified. However, the recall of 48% highlights that only about half of the relevant Tweets were correctly identified by the algorithm. For further research, which may also be based on this framework, improvements to the algorithm are recommended. This could be achieved, for example, by using a more extensive training data set.

6 Conclusion

The research question posed at the beginning — 'In what ways are landscapes discussed and perceived in Spanish-speaking Central and South America based on geographically and thematically filtered Twitter data, and what spatial, temporal or thematic patterns can be identified?' — is addressed in this section. The three individual areas are discussed separately, according to the structure of this thesis.

The spatial analysis can effectively identify hotspots for various landscape terms. The hotspots are linked to real-world locations, such as Playa del Carmen or Lake Atitlán, and the hotspots are often associated with leisure activities such as hiking or swimming. For most terms it was possible to identify highly discussed areas (in relation to words like *paisaje* (landscape) or *naturaleza* (nature)) or individual locations (for terms like *volcán* (volcano) or *playa* (beach)). Unfortunately, a difficulty was the small number of geotagged Tweets. This limitation significantly restricted the spatial analysis and thus limited the granularity of the spatial analysis.

Clear trends were found for temporal patterns. There are differences between terms such as *playa*, which show a clear seasonal trend, and terms such as *volcán*, which show clear event-based differences. The temporal analysis found patterns related to events or vacation times and thus provided a detailed insight into how landscapes influence human behavior and how humans write about it. The temporal analysis of several years would have gone beyond the scope of this thesis but would be easily achievable with the presented framework. This would allow an analysis of trends over several years.

For thematic patterns obtained by the co-occurrence analysis, associations related to the different landscape concepts were clearly found. To take *playa* as an example again, the beach is often associated with positive emotions and vacations. While other words such as *naturaleza* are often associated with displeasure in dealing with Mother Earth. Thus, landscape terms can be brought into a thematic context and provide information about how they are perceived and with which topics they are associated. It would also be interesting to analyze to what extent the individual terms have linguistic variations.

The methods developed provide an essential basis for further research in this interdisciplinary field. Central and South America are ideal regions for exploring the field of landscape perception and language. This work is a step in connecting the two. It has been shown that different landscape concepts are perceived in various ways and associated with a variety of activities and emotions. Certain landscape types are specifically associated with the holiday season, while other landscape terms show that people are concerned about their environment and worried about the future of the planet. Overall, the study clearly illustrates how different landscape terms trigger different associations in us and how these are communicated in social media. This underlines the potential of social media data for geographical research, especially when investigating the perception of landscape.

Further research could, for example, analyze the entire period from which data was drawn for this study. Furthermore, it would be advisable to narrow the landscape terms even further and focus on specific topics rather than attempting to cover all landscape terms simultaneously. By focusing on fewer terms, the machine learning algorithm could improve and thus allow for a more accurate spatial analysis.

To conclude, this work combining landscape perception and geographical as well as linguistic analysis bridges the gap in research when it comes to the Spanish language. By utilizing social media data and implementing machine learning, multidimensional analyses can be carried out that show the diverse ways in which people conceptualize landscape and express associated emotions and concerns. This not only shows the importance of combining cultural and linguistic perspectives into the GIR world, but it also shows that future studies should delve deeper into the relationship between language, culture, and landscape. This study's findings reflect the themes introduced with the Uetliberg example, where perceptions of landscapes, such as what defines a *mountain* or a *hill*, vary across cultures and contexts. It further shows that the diverse landscape in Central and South America is shaped by their physical features and cultural and emotional perspectives. Specific landscapes are linked to emotions, activities, and even events. By exploring these connections through social media, the complex relationship between people and their environment is highlighted. By doing so, it deepens the understanding of how landscapes are perceived in social media and it lays the groundwork for future interdisciplinary approaches.

References

Ardanuy, M. C., & Sporleder, C. (2017). Toponym disambiguation in historical Documents using semantic and geographic Features. *DATeCH2017: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, 175–180. https://doi.org/10.1145/3078081.3078099

Austen, M. (2017). "Put the Groceries Up". *American Speech*, *92*, 298–320. https://doi.org/10.1215/00031283-4312064

Babu, T. R., Chatterjee, A., Khandeparker, S., Subhash, A. V., & Gupta, S. (2015). Geographical address classification without using geolocation coordinates. *Proceedings of the 9th Workshop on Geographic Information Retrieval*, 1–10. https://doi.org/10.1145/2837689.2837696

Baer, M. F., & Purves, R. S. (2023). Identifying Landscape Relevant Natural Language using Actively Crowdsourced Landscape Descriptions and Sentence-Transformers. *KI - Künstliche Intelligenz*, *37*(1), 55–67. https://doi.org/10.1007/s13218-022-00793-3

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130. https://doi.org/10.18653/v1/D16-1120

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Britannica. (2024a). Communication. Retrieved June 11, 2024, from https://www.britannica.com/topic/communication

Britannica. (2024b). Language. Retrieved June 11, 2024, from https://www.britannica.com/topic/language

Burenhult, N., & Levinson, S. C. (2008). Language and landscape: A cross-linguistic perspective. *Language Sciences*, *30*(2-3), 135–150. https://doi.org/10.1016/j.langsci.2006.12.028

Caballero Jiménez, G. V., Nieto Torres, A., Espinasa Pereña, R., Castañeda Bastida, E., Hernández Oscoy, A., Ramírez Castillo, A., & Calva Rodríguez, L. (2017). *Actividad del Volcán Popocatépetl 2017* (tech. rep.). Secretaría de Gobernación. https://www1.cenapred.unam.mx/DIR_INVESTIGACION/Fraccion_XLVIII/XLVIII_transparencia_proactiva/RV/180212_RV_Informe%20Anual%20monitoreo%20volc%C3%A1n%20Popocat%C3%A9petl%202017.pdf

Cambridge University Press. (2024, March). Cambridge Dictionary. Retrieved April 3, 2024, from https://dictionary.cambridge.org/dictionary/english/language

Chesnokova, O., & Purves, R. S. (2018). From image descriptions to perceived sounds and sources in landscape: Analyzing aural experience through text. *Applied Geography*, *93*, 103–111. https://doi.org/10.1016/j.apgeog.2018.02.014

Corp, X. (2024). X Documentation. Retrieved April 3, 2024, from https://developer.twitter.com/en/docs/twitter-api

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the Geotag: Situating 'Big Data' and leveraging the Potential of the Geoweb. *Cartography and Geographic Information Science*, *40*(2), 130–139. https://doi.org/10.1080/15230406.2013.777137

Deekshith, A. (2024). Advances in Natural Language Processing: A Survey of Techniques. *International Journal of Innovations in Engineering Research and Technology*, *8*, 74–83. https://doi.org/10.26662/ijiert.v8i3.pp74-83

Derungs, C., Wartmann, F., Purves, R., & Mark, D. (2013). The Meanings of the Generic Parts of Toponyms: Use and Limitations of Gazetteers in Studies of Landscape Terms. *Spatial Information Theory*, *8116*, 261–278. https://doi.org/10.5167/uzh-86889

Donoso, G., & Sánchez, D. (2017). Dialectometric Analysis of Language Variation in Twitter. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 16–25. https://doi.org/10.18653/v1/W17-1202

Doyle, G. (2014). Mapping Dialectal Variation by Querying Social Media. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 98–106. https://doi.org/10.3115/v1/E14-1011

Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017). A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis. *Research and Advanced Technology for Digital Libraries*, 394–406. https://doi.org/10.1007/978-3-319-67008-9_31

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLOS ONE*, *9*(11), e113114. https://doi.org/10.1371/journal.pone.0113114

Fernández Vítores, D. (2023). El español: Una lengua viva. In *El español en el mundo 2023: Anuario del Instituto Cervantes* (pp. 23–142). Instituto Cervantes.

Foundation, P. S. (2024). Findall [Publication Title: Codecademy]. Retrieved November 28, 2024, from https://www.codecademy.com/resources/docs/python/regex/findall

Fox, N., Graham, L. J., Eigenbrod, F., Bullock, J. M., & Parks, K. E. (2021). Reddit: A novel data source for cultural ecosystem service studies. *Ecosystem Services*, *50*, 101331. https://doi.org/10.1016/j.ecoser.2021.101331

Funkner, A. A., Elkhovskaya, L. O., Lenivtceva, I. D., Egorov, M. P., Kshenin, A. D., & Khrulkov, A. A. (2021). Geographical Topic Modelling on Spatial Social Network Data. *Procedia Computer Science*, *193*, 22–31. https://doi.org/10.1016/j.procs.2021.10.003

Global Volcanism Program. (2017). Report on Fuego (Guatemala) (E. Venzke & E. Crafford, Eds.). *Bulletin of the Global Volcanism Network*, *42*(10). Retrieved December 26, 2024, from https://volcano.si.edu/ShowReport.cfm?doi=10.5479/si.GVP.BGVN201710-342090

Gonçalves, B., Loureiro-Porto, L., Ramasco, J. J., & Sánchez, D. (2018). Mapping the Americanization of English in space and time. *PLOS ONE*, *13*(5), e0197741. https://doi.org/10.1371/journal.pone.0197741

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing Dialect Characterization through Twitter. *PLOS ONE*, *9*(11), e112074. https://doi.org/10.1371/journal.pone.0112074

Gonçalves, B., & Sánchez, D. (2016). Learning about Spanish dialects through Twitter. *Revista Internacional de Lingüística Iberoamericana*, *14*(28), 65–76. https://doi.org/10.31819/rili-2016-142805

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence*, *2*, 11. https://doi.org/10.3389/frai.2019.00011

Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4383–4394. https://doi.org/10.18653/v1/D18-1469

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2015). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *54*. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Jones, C. B., & Purves, R. S. (2009). Geographical Information Retrieval. In *Encyclopedia of Database Systems* (pp. 1227–1231). Springer US. https://doi.org/10.1007/978-0-387-39940-9_177

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kulkarni, V., Perozzi, B., & Skiena, S. (2016). Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 615–618. https://doi.org/10.1609/icwsm.v10i1.14798

Lansley, G., & Longley, P. A. (2016). The Geography of Twitter Topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., & Sanz, F. (2019). Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *Journal of Medical Internet Research*, *21*(6), e14199. https://doi.org/10.2196/14199

Liu, Z., Yu, W., Deng, Y., Wang, Y., & Bian, Z. (2010). A feature selection method for document clustering based on part-of-speech and word co-occurrence. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, *5*, 2331–2334. https://doi.org/10.1109/FSKD.2010.5569827

Longley, P. A., Adnan, M., & Lansley, G. (2015). The Geotemporal Demographics of Twitter Usage. *Environment and Planning A: Economy and Space*, *47*(2), 465–484. https://doi.org/10.1068/a130122p

Louf, T., Gonçalves, B., Ramasco, J. J., Sánchez, D., & Grieve, J. (2023). American Cultural Regions mapped through the lexical Analysis of Social Media. *Humanities and Social Sciences Communications*, *10*, 133. https://doi.org/10.1057/s41599-023-01611-3

Louf, T., Ramasco, J. J., Sánchez, D., & Karsai, M. (2023). When Dialects Collide: How Socioeconomic Mixing Affects Language Use. http://arxiv.org/abs/2307.10016

Mark, D. M., Turk, A. G., Burenhult, N., & Stea, D. (2011). Landscape in Language: An Introduction. In *Landscape in Language: Transdisciplinary Perspectives* (pp. 1–24). John Benjamins Pub. Co.

McKitrick, M. K., Schuurman, N., & Crooks, V. A. (2023). Collecting, analyzing, and visualizing location-based Social Media Data: Review of Methods in GIS-social Media Analysis. *GeoJournal*, *88*(1), 1035–1057. https://doi.org/10.1007/s10708-022-10584-w

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Transactions on Information Systems*, *36*(4), 1–27. https://doi.org/10.1145/3202662

Monteiro, B., Davis, C., Jr., & Fonseca, F. (2016). A Survey on the geographic Scope of textual Documents. *Computers and Geosciences*, *96*, 23–34. https://doi.org/10.1016/j.cageo.2016.07.017

Mukherjee, S., Hauthal, E., & Burghardt, D. (2022). Analyzing the EU Migration Crisis as Reflected on Twitter. *KN - Journal of Cartography and Geographic Information*, *72*(3), 213–228. https://doi.org/10.1007/s42489-022-00114-6

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational Sociolinguistics: A Survey. https://doi.org/10.48550/arXiv.1508.07544

Ntompras, C., Drosatos, G., & Kaldoudi, E. (2022). A high-resolution temporal and geospatial Content Analysis of Twitter Posts related to the COVID-19 Pandemic. *Journal of Computational Social Science*, *5*(1), 687–729. https://doi.org/10.1007/s42001-021-00150-8

Pearson, J. C., Nelson, P. E., Titsworth, S., & Harter, L. (2008). *Human Communication* (4th). McGraw-Hill Companies, Inc.

Pereira, J., Monteiro, J., Estima, J., & Martins, B. (2019). Assessing flood severity from georeferenced photos. *Proceedings of the 13th Workshop on Geographic Information Retrieval*, 1–10. https://doi.org/10.1145/3371140.3371145

Pruss, D., Fujinuma, Y., Daughton, A. R., Paul, M. J., Arnot, B., Szafir, D. A., & Boyd-Graber, J. (2019). Zika Discourse in the Americas: A multilingual Topic Analysis of Twitter. *PLOS ONE*, *14*(5), e0216922. https://doi.org/10.1371/journal.pone.0216922

Purves, R., Clough, P., Jones, C., Hall, M., & Murdock, V. (2018). Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval*, *12*, 164–318. https://doi.org/10.1561/1500000034

Real Academia Española. (2024). Diccionario de la Lengua Española. Retrieved November 25, 2024, from https://www.rae.es/

Rutkin, A. (2014). Twitter Bots grow up and take on the World. Retrieved March 28, 2024, from https://www.newscientist.com/article/mg22329804-000-twitter-bots-grow-up-and-take-on-the-world/

Scheffler, T., Gontrum, T. S. J., Wegel, M., & Wendler, S. (2014). Mapping German Tweets to Geographic Regions. Retrieved March 23, 2024, from https://tscheffler.github.io/papers/konvens2014-tweegion.pdf

Schmidt, C. W., Reddy, V., Zhang, H., Alameddine, A., Uzan, O., Pinter, Y., & Tanner, C. (2024). Tokenization Is More Than Compression. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 678–702. https://doi.org/10.18653/v1/2024.emnlp-main.40

Semantics and Discourse: Collocations, Keywords and Reliability of Manual Coding. (2018). In V. Brezina (Ed.), *Statistics in Corpus Linguistics: A Practical Guide* (pp. 66–101). Cambridge University Press. https://doi.org/10.1017/9781316410899.003

Siever, C. (2023). Emojis in der digitalen Kommunikation: Trauer um Sternenkinder auf Twitter. Retrieved March 28, 2024, from https://www.uzh.ch/blog/digitalreligions/2023/08/31/emojis-in-der-digitalen-trauerkommunikation-um-sternenkinder-auf-twitter/

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza a H1N1 Pandemic. *PLOS ONE*, *6*(5), e19467. https://doi.org/10.1371/journal.pone.0019467

Smith, A., & Rainie, L. (2010). 8% of online Americans use Twitter. *Pew Internet and American Life Project*. https://doi.org/10.7228/manchester/9780719074462.003.0001

Tellez, E. S., Moctezuma, D., Miranda, S., Graff, M., & Ruiz, G. (2023). Regionalized Models for Spanish Language Variations based on Twitter. *Language Resources and Evaluation*, *57*(4), 1697–1727. https://doi.org/10.1007/s10579-023-09640-9

Van Putten, S., O'Meara, C., Wartmann, F., Yager, J., Villette, J., Mazzuca, C., Bieling, C., Burenhult, N., Purves, R., & Majid, A. (2020). Conceptualisations of landscape differ across European languages. *PLOS ONE*, *15*(10), e0239858. https://doi.org/10.1371/journal.pone.0239858

Villette, J., & Purves, R. S. (2018). Exploring Microtoponyms through linguistic and geographic Perspectives. *Villette, Julia; Purves, Ross S (2018). Exploring microtoponyms through linguistic and geographic perspectives. In: 21th AGILE Conference on Geographic Information Science, Lund, 12 Juni 2018 - 15 Juni 2018, AGILE.* https://doi.org/10.5167/uzh-161909

Wartmann, F. M., Egorova, E., Derungs, C., Mark, D. M., & Purves, R. S. (2015). More Than a List: What Outdoor Free Listings of Landscape Categories Reveal About Commonsense

Geographic Concepts and Memory Search Strategies. *Spatial Information Theory*, 224–243. https://doi.org/10.1007/978-3-319-23374-1_11

Wong, B. (2024, February). Top Social Media Statistics And Trends [Section: Business]. Retrieved April 8, 2024, from https://www.forbes.com/advisor/in/business/social-media-statistics/

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic Identification of eyewitness Messages on Twitter during Disasters. *Information Processing & Management*, *57*(1), 102107. https://doi.org/10.1016/j.ipm.2019.102107

Zhu, X. X., Wang, Y., Kochupillai, M., Werner, M., Haberle, M., Hoffmann, E. J., Taubenbock, H., Tuia, D., Levering, A., Jacobs, N., Kruspe, A., & Abdulahhad, K. (2023). Geo-Information Harvesting from Social Media Data. https://doi.org/10.48550/arXiv.2211.00543

7 Appendix

A: Spanish Terms Glossary

Table 7.1: List of Landscape Terms, Translations, and Synonyms

ID	Spanish Term	English Translation	Synonyms
1	árbol	Tree	palo
2	planta	Plant	
3	montaña	Mountain	monte, cabezo, cumbre
4	valle	Valley	cuenca, quebrada
5	cielo	Sky	firmamento
6	nube	Cloud	bruma, celaje
7	edificio	Building	construcción, edificación, finca, inmueble
8	flora	Flower	
9	sol	Sun	
10	camino	Path	vía, senda, sendero
11	horizonte	Horizon	
12	campo	Field	llanura
13	coche	Car	auto, carro
14	casa	House	vivienda
15	carretera	Road	autopista, autovía
16	piedra	Stone	roca
17	arbusto	Bush	
18	playa	Beach	costa
19	río	River	
20	mar	Sea	océano, ponto
21	volcán	Volcano	
22	arena	Sand	
23	bosque	Forest	floresta, soto
24	desierto	Desert	
25	ciudad	City	urbe
26	selva	Jungle	jungla
27	lago	Lake	laguna
28	cerro	Hill	colina, collado, loma
29	cascada	Waterfall	catarata
30	naturaleza	Nature	natura
31	tierra	Earth	
32	agua	Water	
33	roca	Rock	peñasco, peña, risco
34	luna	Moon	
35	amanecer	Sunrise	alba, albor, alborada, madrugada
36	estrella	Star	astro, lucero
37	atardecer	Sunset	ocaso, puesta del sol, anochecer
38	paisaje	Landscape	paraje

B: Co-Occurence of Words with Landcape Terms

Table 7.2: Top Co-occurring Terms for Amanecer

Word	**Log_LL**	**Frequency**
día	7.17	149
atardecer	6.76	89
playa	6.75	87
ver	6.67	79
noche	6.56	69
oscura	6.30	50
nuevo	6.27	48
quiero	6.20	44
vida	6.16	42
hermoso	5.96	33
mejor	5.96	33
feliz	5.82	28
bello	5.73	25
lindo	5.62	22
luz	5.50	19
días	5.35	16
buen	5.30	15
mundo	5.30	15
regala	5.30	15
colores	5.24	14
dios	5.24	14
gracias	5.24	14
sol	5.24	14
mar	5.24	14
disfrutar	5.17	13

Table 7.3: Top Co-occurring Terms for Anochecer

Word	**Log_LL**	**Frequency**
luna	4.84	13
amanecer	4.84	13
atardecer	4.78	12
vida	4.71	11
luz	4.71	11
desnudos	4.45	8
día	4.45	8
prefiero	4.45	8
ciclismo	4.34	7
cicloide	4.34	7
encontró	4.22	6
lugar	4.07	5
iris	3.88	4
vía	3.88	4
agua	3.88	4
cálido	3.64	3
fuego	3.64	3
noche	3.64	3
tarde	3.64	3
playa	3.64	3
hermoso	3.64	3
relive	3.64	3
duró	3.30	2
verte	3.30	2
landscape	3.30	2

Table 7.4: Top Co-occurring Terms for Atardecer

Word	**Log_LL**	**Frequency**
hermoso	7.86	250
ver	7.74	214
playa	7.47	154
sunset	7.30	126
día	7.20	111
bello	7.19	110
foto	7.04	92
amanecer	7.01	89
lindo	6.99	87
sol	6.96	84
bonito	6.93	81
fotos	6.89	77
colores	6.81	70
ciudad	6.76	66
cielo	6.72	63
mejor	6.56	52
mar	6.51	49
vía	6.51	49
espectacular	6.49	48
luna	6.44	45
vida	6.44	45
precioso	6.44	45
viendo	6.42	44
vista	6.38	42
nubes	6.33	40

Table 7.5: Top Co-occurring Terms for Cerro

Word	**Log_LL**	**Frequency**
incendio	7.16	124
porteño	6.65	66
ir	6.40	49
día	6.38	48
forestal	6.19	38
alegre	6.12	35
vida	6.09	34
silla	5.93	28
camino	5.83	25
quiero	5.83	25
vía	5.80	24
personas	5.76	23
agua	5.72	22
lugar	5.72	22
tiene	5.72	22
colorado	5.68	21
foto	5.64	20
verde	5.64	20
azul	5.60	19
negro	5.60	19
navia	5.55	18
ver	5.55	18
semana	5.55	18
vista	5.55	18
playa	5.55	18

Table 7.6: Top Co-occurring Terms for Colina

Word	**Log_LL**	**Frequency**
cima	5.33	21
alta	5.25	19
sendero	5.21	18
vía	4.64	9
parque	4.55	8
casa	4.43	7
municipio	4.31	6
vida	4.31	6
incendio	4.31	6
alto	4.15	5
fotos	4.15	5
construcción	4.15	5
campestre	3.96	4
agua	3.96	4
ver	3.96	4
escalar	3.96	4
arriendo	3.96	4
bandera	3.96	4
sector	3.96	4
mas	3.96	4
moderna	3.96	4
contaron	3.96	4
regreso	3.96	4
triunfar	3.96	4
chiapas	3.96	4

Table 7.7: Top Co-occurring Terms for Costa

Word	**Log_LL**	**Frequency**
ir	7.60	244
ganas	7.06	121
quiero	6.91	100
playa	6.75	82
parque	6.48	59
día	6.39	53
irme	6.31	48
tengo	6.09	37
lago	6.07	36
sol	6.00	33
oriental	6.00	33
vamos	5.83	27
estar	5.83	27
mejor	5.80	26
vida	5.80	26
vía	5.62	21
noche	5.53	19
tomar	5.53	19
frente	5.49	18
tierra	5.44	17
felicidad	5.39	16
luz	5.33	15
toda	5.33	15
atlántica	5.33	15
verde	5.33	15

Table 7.8: Top Co-occurring Terms for Cumbre

Word	**Log_LL**	**Frequency**
día	4.27	7
volcán	4.15	6
ir	4.15	6
mundo	3.82	4
planeta	3.82	4
minuto	3.82	4
montaña	3.82	4
vida	3.82	4
integracion	3.58	3
ver	3.58	3
mundial	3.58	3
celac	3.58	3
gran	3.58	3
importante	3.58	3
significa	3.58	3
mejor	3.58	3
nacional	3.24	2
playa	3.24	2
tierra	3.24	2
sol	3.24	2
frases	3.24	2
verde	3.24	2
internacional	3.24	2
deja	3.24	2
argentina	3.24	2

Table 7.9: Top Co-occurring Terms for Desierto

Word	**Log_LL**	**Frequency**
planeta	7.94	315
pasó	7.47	174
convertirse	7.45	169
horas	7.32	145
atravesar	7.31	143
continuas	7.30	140
agua	7.27	135
venezuela	7.22	128
narcangelalvarado	6.99	96
dios	6.99	96
pasar	6.96	93
atacama	6.92	88
mundo	6.75	72
mejor	6.70	68
tierra	6.69	67
vida	6.60	60
prometida	6.49	53
corazón	6.48	52
paraiso	6.39	47
grandesmedios	6.38	46
árido	6.36	45
necesario	6.36	45
vía	6.34	44
hable	6.30	42
campo	6.26	40

Table 7.10: Top Co-occurring Terms for Lago

Word	**Log_LL**	**Frequency**
maracaibo	8.33	390
titicaca	7.89	229
monstruo	7.84	215
puente	7.77	196
nuevo	7.68	176
sur	7.65	170
descubre	7.52	146
maravillas	7.48	140
grabada	7.48	139
casa	7.45	134
agua	7.37	122
planeta	7.36	121
cisnes	7.36	121
ir	7.30	113
vía	7.30	113
viral	7.23	104
grande	7.14	93
secretos	7.11	90
origen	7.05	84
piratas	7.01	80
revela	7.00	79
cielo	6.96	76
isla	6.92	72
día	6.88	69
parque	6.86	67

Table 7.11: Top Co-occurring Terms for Madrugada

Word	**Log_LL**	**Frequency**
mejor	6.20	55
día	5.89	37
charlas	5.67	28
noche	5.58	25
viernes	4.82	10
vida	4.73	9
personas	4.64	8
agua	4.64	8
sábado	4.64	8
montaña	4.52	7
hora	4.52	7
parte	4.52	7
domingo	4.52	7
gracias	4.52	7
puso	4.39	6
playa	4.39	6
ver	4.39	6
incendio	4.39	6
volcán	4.39	6
tarde	4.39	6
descubrió	4.39	6
cámara	4.39	6
escuchaba	4.39	6
ruidos	4.39	6
colima	4.24	5

Table 7.12: Top Co-occurring Terms for Montaña

Word	**Log_LL**	**Frequency**
rusa	9.25	1169
alta	8.67	567
radio	8.39	399
oficial	8.38	396
rostro	7.76	187
disfrutaba	7.74	183
impacta	7.74	183
playa	7.72	179
vida	7.70	174
cuartel	7.64	162
accidentes	7.36	116
carrera	7.35	115
bicicleta	7.33	112
emociones	7.28	106
tiene	7.27	105
vía	7.19	95
cima	7.19	95
ir	7.13	89
día	7.12	88
va	7.09	85

Table 7.13: Top Co-occurring Terms for Monte

Word	**Log_LL**	**Frequency**
bello	7.88	319
morgue	6.79	80
ir	6.64	67
hermoso	6.42	51
día	6.33	46
mercedes	6.30	44
vía	6.26	42
quiero	6.10	35
plata	6.08	34
fotos	6.00	31
cabra	6.00	31
grande	5.97	30
playa	5.94	29
ver	5.91	28
colinas	5.88	27
mundo	5.88	27
tira	5.82	25
vida	5.78	24
llevas	5.71	22
culebra	5.63	20
altura	5.63	20
splash	5.63	20
agua	5.54	18
tiene	5.54	18
cadáveres	5.54	18

Table 7.14: Top Co-occurring Terms for Paisaje

Word	**Log_LL**	**Frequency**
hermoso	7.90	266
parte	7.58	179
grises	7.24	118
foto	7.22	116
forman	7.19	111
bonito	7.15	106
lindo	7.15	106
ver	7.13	104
mejor	7.11	101
disfruta	7.09	99
bello	7.02	91
disfrutar	7.02	91
quiero	7.00	89
cafetero	6.74	65
cultural	6.74	65
disfrutando	6.72	64
nieve	6.72	64
nieva	6.64	58
teclado	6.56	53
naturaleza	6.53	51
espectacular	6.51	50
urbano	6.50	49
mirar	6.48	48
belleza	6.46	47
pintas	6.44	46

Table 7.15: Top Co-occurring Terms for Playa

Word	**Log_LL**	**Frequency**
ir	10.24	2795
quiero	9.85	1728
carmen	9.78	1585
sol	9.49	1121
día	9.47	1091
estar	9.13	724
ganas	9.09	696
arena	8.96	598
necesito	8.95	588
vacaciones	8.91	558
mar	8.88	542
mejor	8.87	532
vamos	8.83	512
verano	8.79	487
fotos	8.75	462
voy	8.71	443
semana	8.61	394
días	8.59	386
casa	8.56	373
noche	8.55	367
tengo	8.53	361
ver	8.44	323
foto	8.38	303
hace	8.36	296
casas	8.34	287

Table 7.16: Top Co-occurring Terms for Tierra

Word	**Log_LL**	**Frequency**
planeta	10.25	2911
día	9.27	869
vida	9.27	868
cielo	9.23	829
años	9.04	662
asteroide	8.92	576
vía	8.78	488
agua	8.68	435
fuego	8.64	416
madre	8.62	405
trabajo	8.52	360
planetas	8.46	337
nasa	8.46	336
mojada	8.45	333
gran	8.45	332
pasará	8.45	332
plana	8.44	328
mejor	8.42	320
dios	8.42	320
pies	8.41	318
olor	8.38	307
unión	8.36	299
grandes	8.35	296
tiene	8.33	289
julioleóngobernador	8.29	275

Table 7.17: Top Co-occurring Terms for Volcán

Word	**Log_LL**	**Frequency**
erupción	8.18	393
vista	7.83	251
vivo	7.80	243
turrialba	7.76	230
volcánirazú	7.73	222
popocatépetl	7.44	156
bot	7.42	151
actividad	7.38	144
colima	7.21	118
fuego	7.21	118
alerta	6.80	72
agung	6.77	69
explosión	6.70	64
fallas	6.68	62
activo	6.65	60
entra	6.59	56
hace	6.56	54
chocolate	6.53	52
registra	6.46	48
ceniza	6.42	46
vía	6.41	45
video	6.26	38
años	6.26	38
exhalación	6.22	36
exhalaciones	6.19	35

C: Co-Occurence Wordclouds

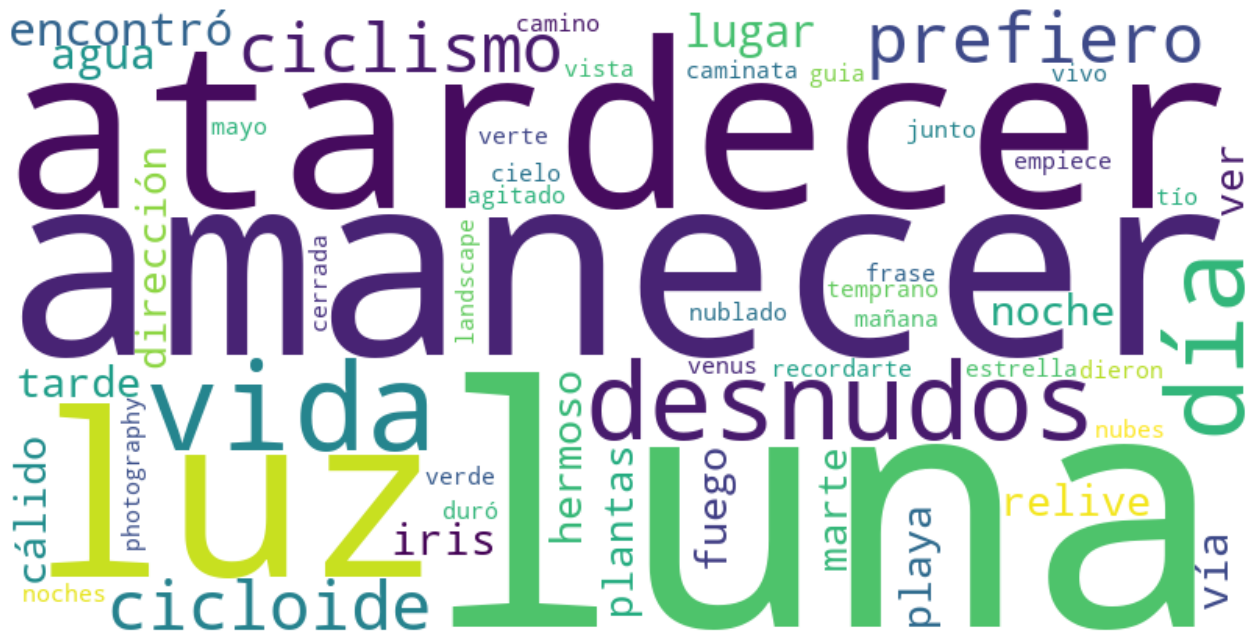


Figure 7.1: Word cloud showing the top associated words for the landscape term 'anochecer'.

Figure 7.2: Word cloud showing the top associated words for the landscape term 'atardecer'.

Figure 7.3: Word cloud showing the top associated words for the landscape term 'cerro', based on their log-likelihood values.

Figure 7.4: Word cloud showing the top associated words for the landscape term 'costa', based on their log-likelihood values.

Figure 7.5: Word cloud showing the top associated words for the landscape term 'cumbre', based on their log-likelihood values.

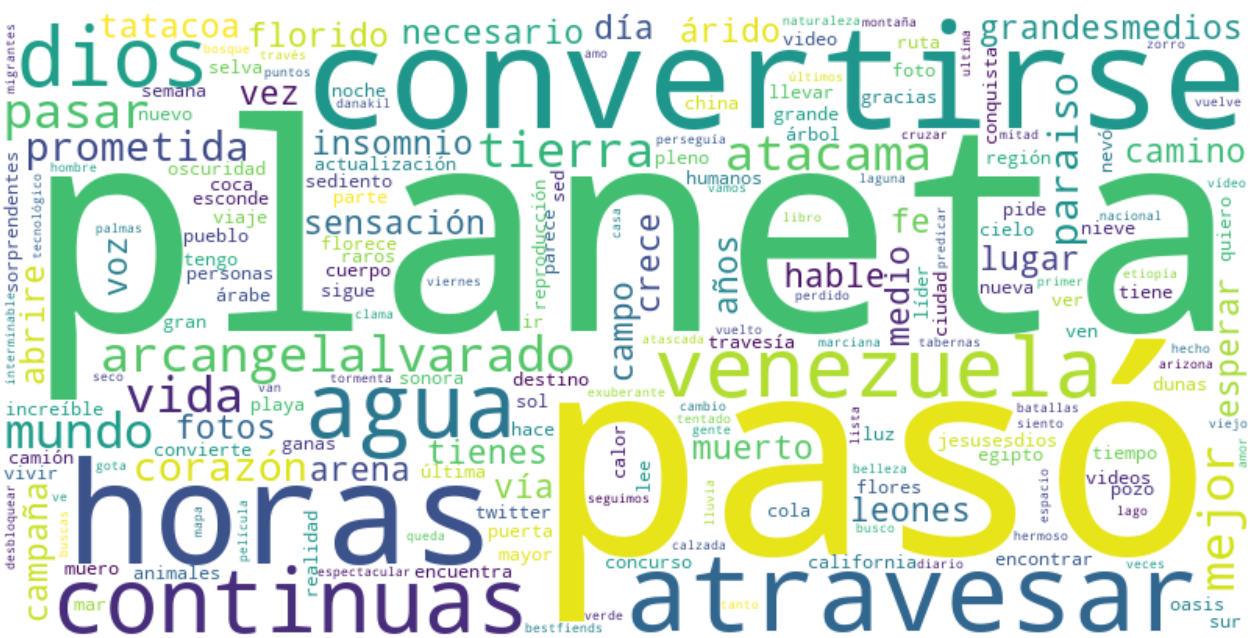


Figure 7.6: Word cloud showing the top associated words for the landscape term 'desierto', based on their log-likelihood values.

Figure 7.7: Word cloud showing the top associated words for the landscape term 'montaña'.

Figure 7.8: Word cloud showing the top associated words for the landscape term 'monte', based on their log-likelihood values.

Figure 7.9: Word cloud showing the top associated words for the landscape term 'tierra', based on their log-likelihood values.

D: Additional Spatial Images

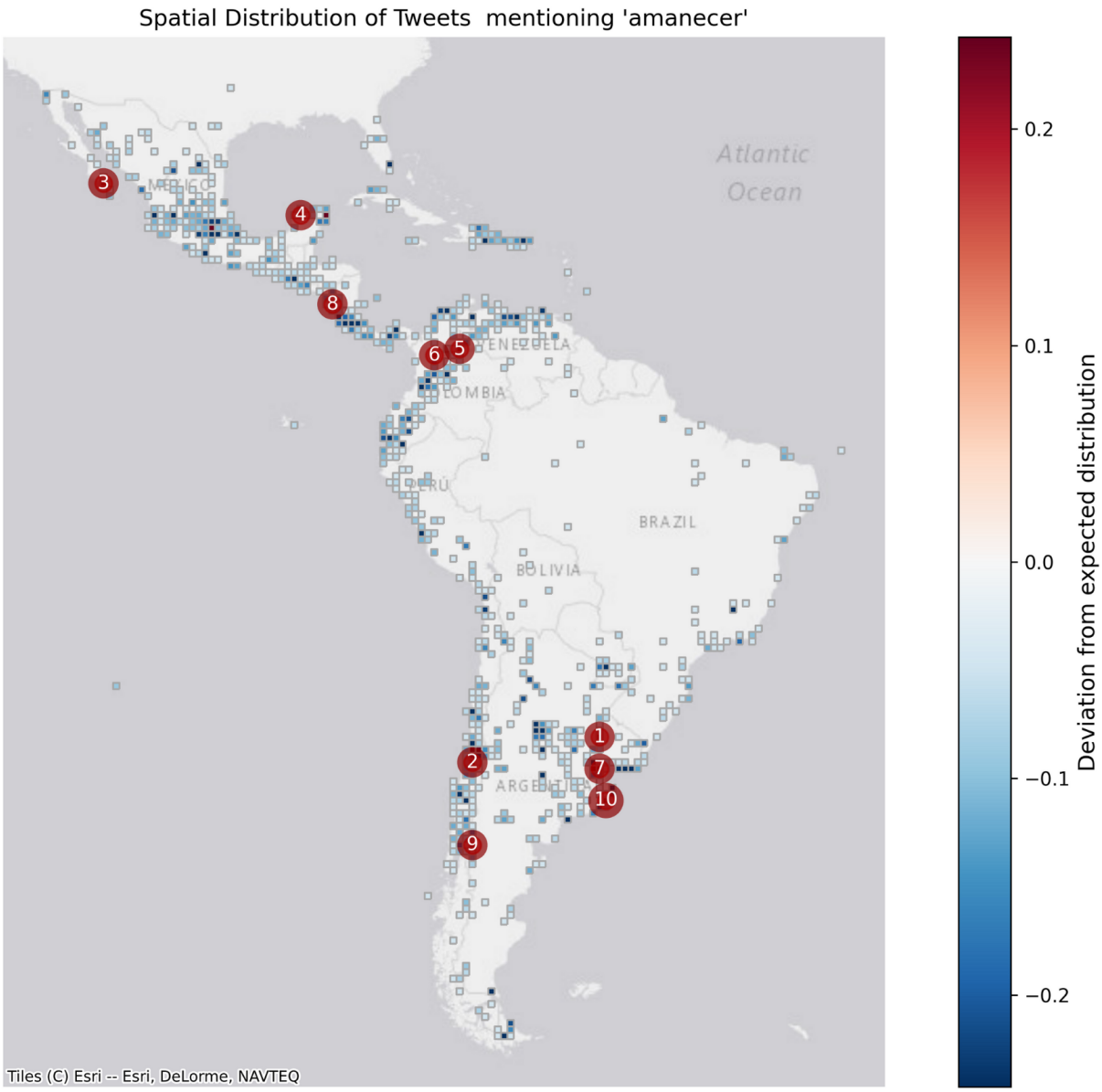


Figure 7.10: Spatial distribution of Tweets mentioning *Amanecer*.

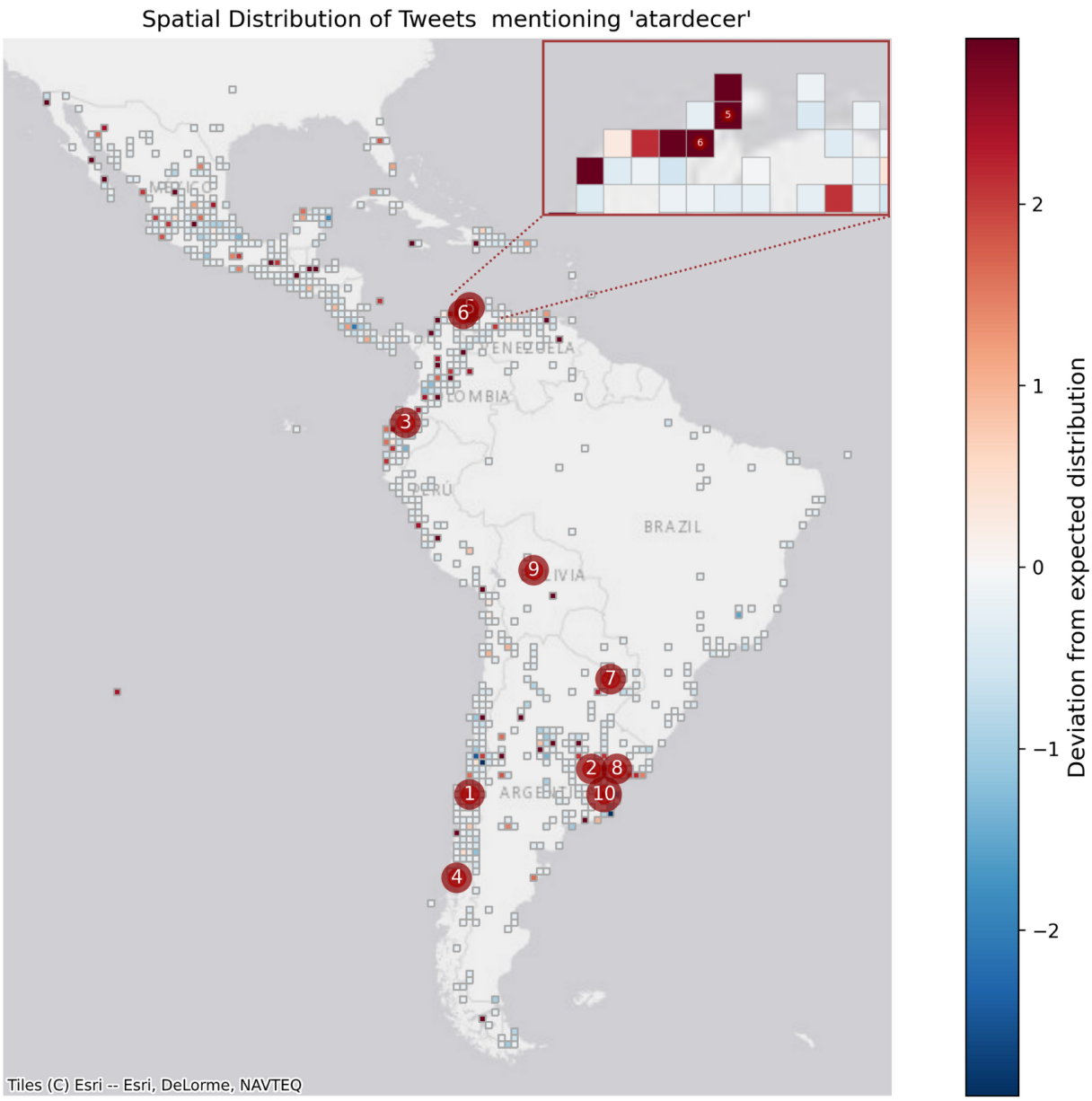


Figure 7.11: Spatial distribution of Tweets mentioning *atardecer*.

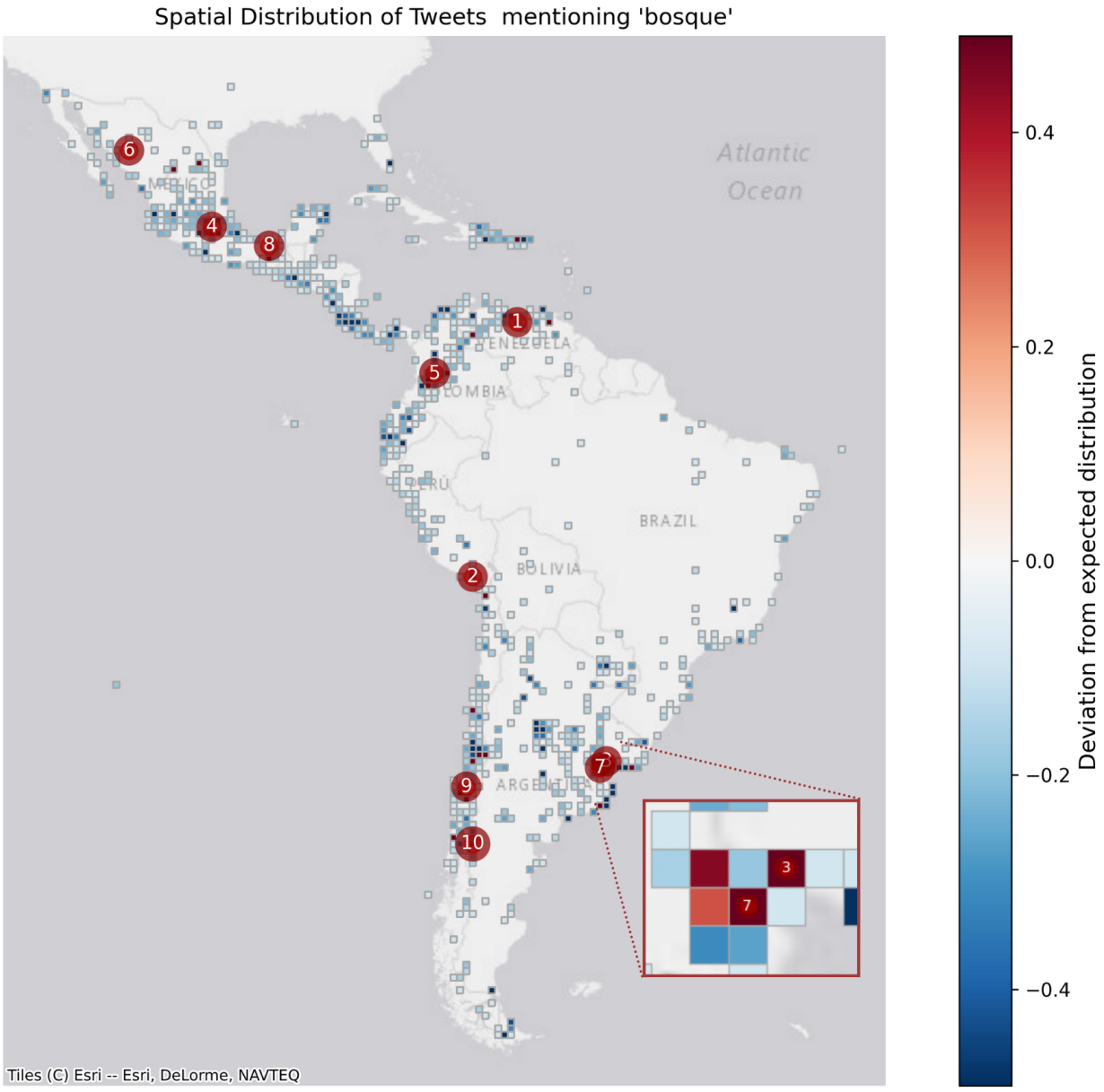


Figure 7.12: Spatial distribution of Tweets mentioning *bosque*.

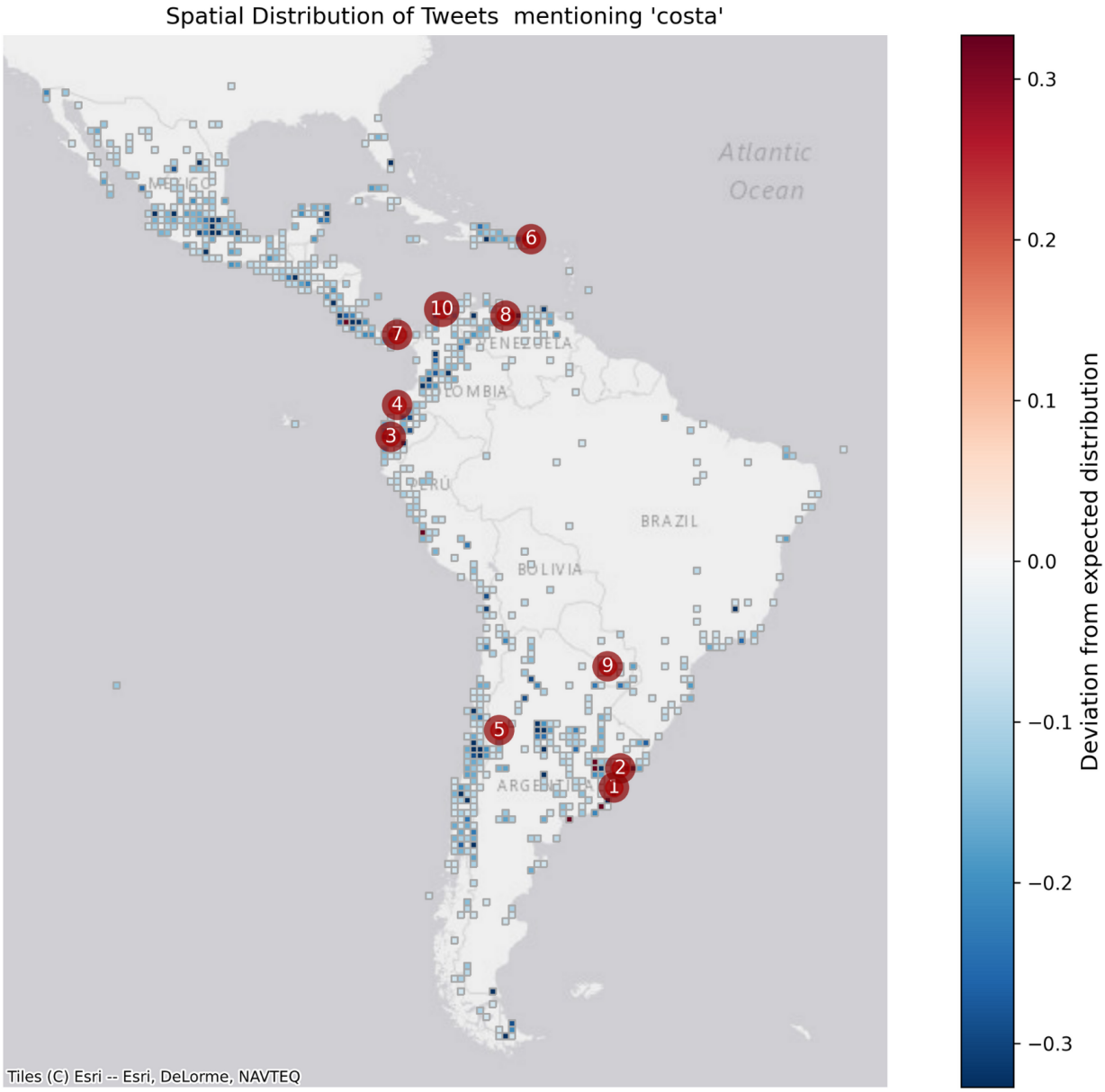


Figure 7.13: Spatial distribution of Tweets mentioning *costa*.

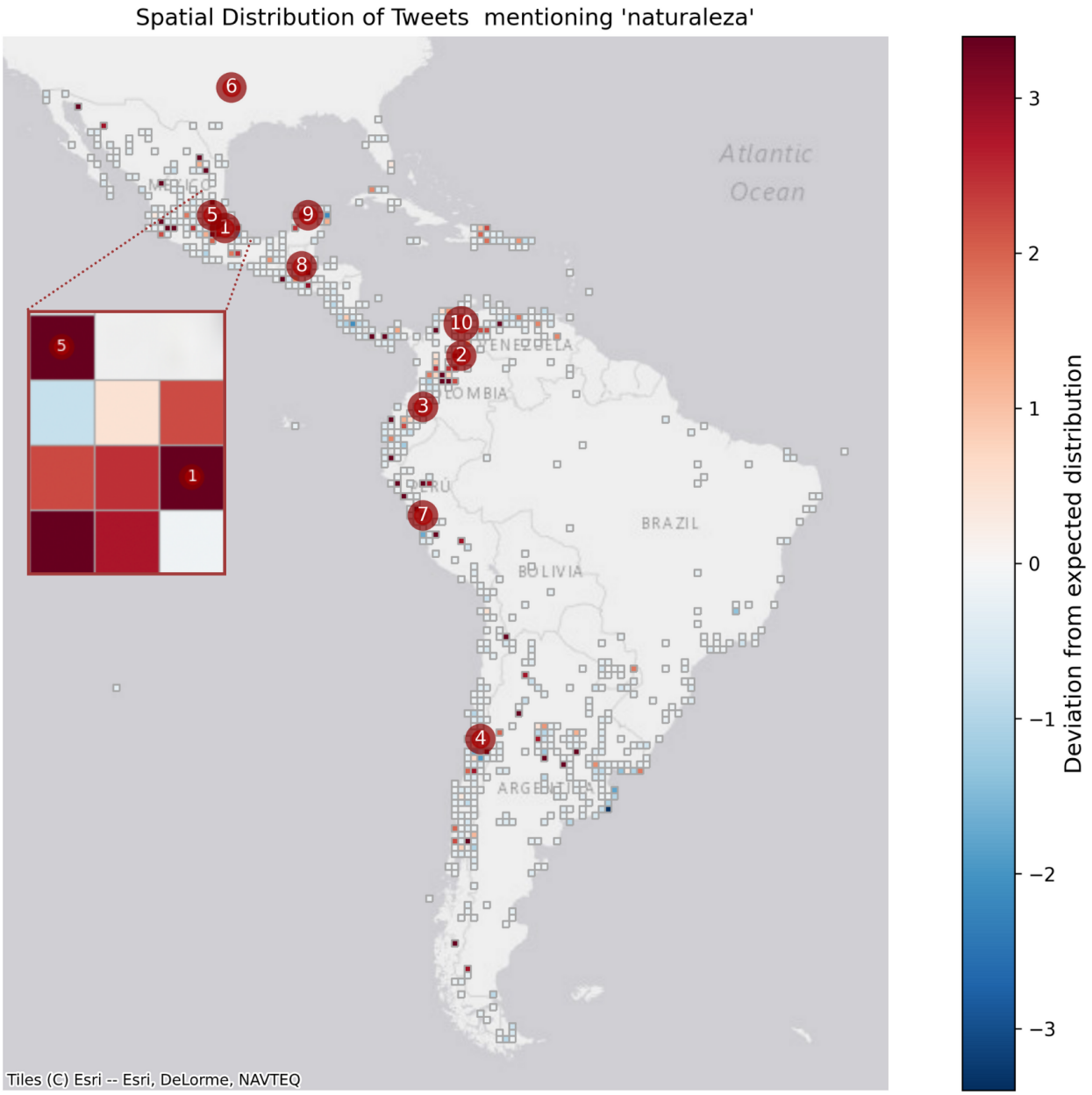


Figure 7.14: Spatial distribution of Tweets mentioning *naturaleza*.

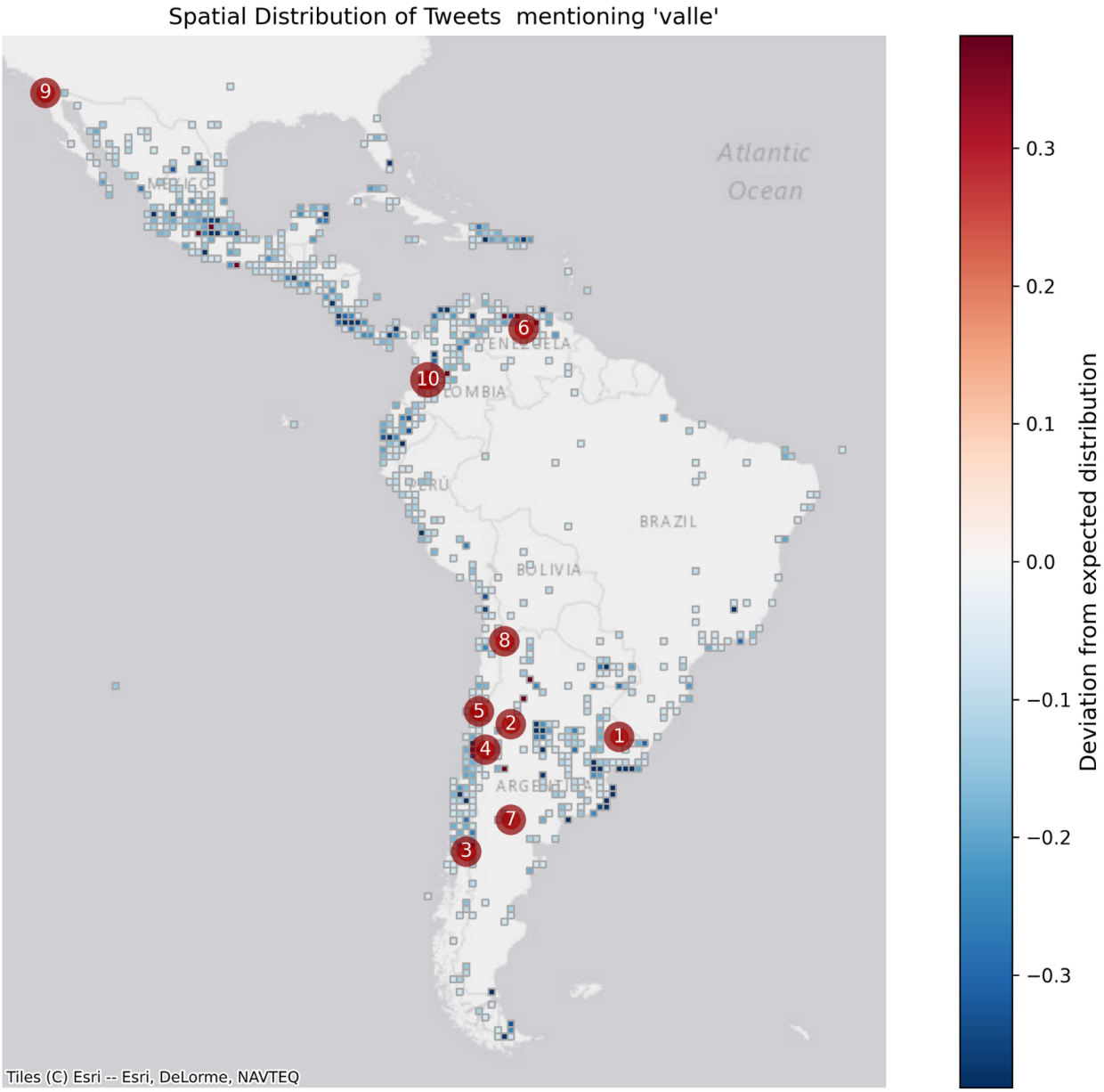


Figure 7.15: Spatial distribution of Tweets mentioning *valle*.

Declaration of Originality

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis. I used artificial intelligence as follows: ChatGPT was used with refining Python code and troubleshooting LaTeX documents. DeepL was used to ensure a fluent reading style and correct translations. The tools mentioned were used only to provide support and did not influence the scientific and substantive core of my work.

Philipp Rohr, Winterthur, 27.01.25