UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF A SYSTEM FOR FEELINGS CHARACTERIZATION IN PARKS USING SOCIAL MEDIA MINING

DIEGO DE VEGA FERNÁNDEZ

JUNIO 2021

TRABAJO DE FIN DE GRADO

Título:	Diseño y desarrollo de un sistema de caracterización de sentimientos en parques mediante Social Media Mining
Título (inglés):	Design and development of a system for feelings characterization in parks using Social Media Mining
Autor:	Diego de Vega Fernández
Tutor:	Carlos Ángel Iglesias Fernández
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF A SYSTEM FOR FEELINGS CHARACTERIZATION IN PARKS USING SOCIAL MEDIA MINING

Diego de Vega Fernández

Junio 2021

Resumen

Actualmente, más de la mitad de la población mundial hace uso de alguna red social. Por lo tanto, las redes sociales son la herramienta idónea para descubrir cómo es la sociedad actual, cuáles son sus hábitos, qué comportamientos tienen en diferentes situaciones o lugares, etc.

Más específicamente, la red social Twitter es usada a diario por cientos de millones de usuarios, que interactúan entre sí publicando todo tipo de información que puede ser recopilada y tratada con el fin de obtener gran cantidad de información útil. Gracias al Aprendizaje Automático, se pueden usar clasificadores que muestren esa información y que determinen cómo influyen en las personas diferentes factores de mundo que nos rodea.

El objetivo será estudiar tweets publicados en parques, tanto urbanos como naturales, para su posterior análisis y clasificación con Aprendizaje Automático y determinar la influencia en una persona de twittear en unos lugares u otros. Las ubicaciones a estudiar serán parques de diferentes ciudades españolas, parques naturales de nuestro país, así como de una población urbana para comparar resultados. El análisis de los tweets será para conocer cómo son los sentimientos de las personas y sus emociones reflejadas en sus tweets. Todo ello gracias al Procesamiento del Lenguaje Natural (PLN). También serán analizados los hashtags que acompañan a esos tweets, así como los temas sobre los que tratan.

Palabras clave: Twitter, Aprendizaje Automático, análisis, sentimientos, emociones, hashtags, temas, Python, parques, PLN.

Abstract

Currently, more than half of the world's population uses a social network. Thus, social networks are the ideal tool to discover what today's society is like, what their habits are, what behaviors they have in different situations or places, etc.

More specifically, the social network Twitter is used daily by hundreds of millions of users, who interact with each other by publishing all kinds of information that can be collected and processed in order to obtain a large amount of useful information. Thanks to Machine Learning, classifiers can be used to show that information and to determine how different factors in the world around us influence people.

The objective will be to study tweets published in parks, both urban and natural, for their subsequent analysis and classification with Machine Learning and to determine the influence on a person of tweeting in some places or others.The locations to be studied will be parks in different Spanish cities, natural parks in our country, as well as an urban population to compare results. The analysis of the tweets will be to know how people's feelings and their emotions are reflected in their tweets. All this thanks to Natural Language Processing (NLP). The hashtags that accompany these tweets, as well as the topics they deal with, will also be analyzed.

Keywords: Twitter, Machine Learning, analysis, sentiments, emotions, hashtags, topics, Python, parks, NLP.

Agradecimientos

Quiero dar las gracias a mi tutor Carlos Ángel Iglesias por el apoyo y la confianza que me ha brindado desde el primer instante para realizar este trabajo, así como al GSI por proporcionarme la ayuda y los recursos que he necesitado en cada momento.

También quiero agradecer a mi familia, en especial a mis padres, por apoyarme durante toda esta etapa universitaria. Por último a mis amigos, los de toda la vida y los que me ha dado la ETSIT, que me han ayudado a superar este grado con éxito.

Contents

List of Figures

List of Tables

Listings

CHAPTER 1

Introduction

1.1 Context

It is very common to hear about advantages and benefits about the existence of natural environments within cities. Multiple studies carried out in many countries say that parks help to promote biodiversity or help to normalize parameters such as temperature, pollution or humidity within a city. In fact, nowadays it is not strange to have an urban green space close to where you live.

The importance given in today's society to natural parks is also well known, this time talking about completely natural spaces far from large cities. They are given great protection, since they are areas considered as the lungs of the planet. In Spain, as in many other countries around the world, it exists the category of National Park for those natural systems that meet certain characteristics and are therefore protected by law. They are defined as natural spaces, of high ecological and cultural value, little transformed by exploitation or human activity. Due to the beauty of their landscapes, the representativeness of their ecosystems or the uniqueness of their flora, fauna, geology, or geomorphological formations, they have outstanding ecological, aesthetic, cultural, educational, and scientific values [1].

Even with all this information about how beneficial is nature to our planet, we don't

have much information about how being in the park directly affects people. The correlation of a person's well-being with green spaces, both urban and natural, is a field in which there is little research. It is true that the existence of urban parks is something that everyone knows of its importance, but we do not know exactly how the fact of being in it affects us or not when we are immersed in the chaos of the cities.

Green spaces are known to help people with things like improving air quality or preventing diseases, but the aim of this study is to learn more about how being in a green area affects an individual. We are interested in how he feels, if he improves his quality of life in a more daily way. We are not simply interested in general, and often unconscious things, like breathing cleaner air. We will focus on investigating the feelings and emotions of people who are in parks.

There may be several methods of knowing how people feel in parks. In our case, we will use something that is known to represent a large part of the population and that serves many people to express themselves. This is the case of social networks, more specifically the social network Twitter [2].

The rise of social networks like this will allow us to obtain a large amount of information, that can later be analyzed in search of those feelings and emotions. Twitter is a platform through which a user can write about whatever he wants, which provides a very large amount of information about people. It is a social network in which most of the accounts are public and no type of relationship is required within it to see what people publish.

For this study, geotagged tweets will be used, thanks to social mining and geolocation tools.

1.2 Project goals

The objetives of the project are the following.

- Locate by geographical coordinates the different locations in which we are going to carry out the study.
- Retrive the tweet datasets necessary for the study according to the established parameters.
- Pre-process the collected information by cleaning it for later analysis.
- Analyze the tweets:

  - Sentiment analysis.
  - Emotion analysis.
  - Hashtag analysis.
  - Topic analysis.

- Draw conclusions from the data obtained in the the analysis.
- Apply the same analyzes to a specific case study.

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1 puts the work in context for the reader and presents the objectives of it. It gives a brief overview of how the information will be handled in the project.

Chapter 2 provides information about the different technologies that have helped to carry out this project. There is information about what they are for, where they come from, and, more briefly, how they will be used.

Chapter 3 deals with how the data was obtained for this work, which filters they had to pass through, and how it is divided. It also shows the distribution of these data in space and time.

Chapter 4 details all the analyzes that have been made to the different datasets, the results obtained, and the comparisons and conclusions drawn from these.

Chapter 5 gives a more concrete vision of the city of Madrid, analyzing it as a separate case study. Here the most important urban parks of this city are analyzed and differences are sought.

Chapter 6 discusses the conclusions drawn from this project, the goals achieved, the problems faced and their respective solutions, and suggestions for future work.

CHAPTER 2

Enabling Technologies

2.1 Technologies for collecting tweets

Twitter is a social networking service on which users post and interact with messages known as tweets. This social network generates a large amount of information, not only the message someone posts, but also the author username, publication date or the location.

Twitter provides a number of APIs with which that information can be collected. But the limit of being able to collect uniquely tweets up to two weeks old, has led us to use technologies independent of Twitter.

2.1.1 Twint

Twint [3] is an advanced Twitter scraping tool written in Python that enables for scraping tweets from Twitter profiles without using Twitter's API. This tool can be configured with several parameters such as user names, dates, or amount of tweets.

Twint requires several packages and libraries to run properly. The most relevant for this project are Pandas, and Geopy, used to locate the coordinates of the analysed parks using third-party geocoders and other data sources.

In our case, we have configured it for collecting tweets depending on specific locations and limiting it in time, in order to obtain datasets with tweets from parks in Spain uploaded in the last three years. Those datasets are collected as CSVs.

2.1.2 Google Maps

Google Maps [4] is a web map application server owned by Alphabet Inc. It offers scrollable map images, as well as satellite pictures of the world among other things like measuring distances, which is the functionality that is going to be used in this project.

2.2 Natural Language Processing (NLP) technologies

This field of artificial intelligence serves computers to understand, interpret, and manipulate human language, thanks to computation and computational linguistics.

2.2.1 MeaningCloud

MeaningCloud [5] is a company specialized in software for semantic analysis. The API provided by them, can be used in order to determine the polarity, the subjectivity, the level of agreement, and many other details from the tweets we have collected.

It also provides you with the information of which words specifically are those that have made the API make the decision about the aforementioned parameters. Finally, it is also capable of detecting the topics on which the analyzed texts are about.

MeaningCloud is a brand by MeaningCloud LLC, a wholly owned subsidiary of MeaningCloud Europe S.L., previously known as Daedalus. Daedalus was founded in 1998 by Jose Carlos González and other colleagues as a spin-off from their Artificial Intelligence research lab at the Universidad Politécnica de Madrid.

2.2.2 RapidMiner Studio

RapidMiner Studio [6] is a comprehensive data science platform with visual workflow design and full automation.

RapidMiner was developed starting in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer at the Artificial Intelligence Unit of the Technical University of Dortmund.

It provides more than 500 operators geared towards data analysis, including those needed to perform input and output operations, data preprocessing, and visualization.

Thanks to this platform, apart from the possibility to use MeaningCloud services with a plugin, we can extract all kinds of graphics that can help us to interpret the obtained information and, also, the datasets before processing them.

2.3 Analysis and Visualization. Pandas

This field covers the technologies that help us to handle the information before and after its analysis.

Pandas [7] is an open source Python library specialized in the management and analysis of data structures. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Developer Wes McKinney started working on pandas in 2008 due to the need for a flexible, high-performance tool for performing quantitative analysis of financial data.

The main use of the tool in this project has been the manipulation of the datasets.

CHAPTER 3

Data acquisition

3.1 Introduction

In this chapter, we cover the data collection phase of this project. We will also detail where all the data comes from, that is, from which places the tweets used in the project come and the exact dates on which they were retrieved. With this, we intend to put the tweets of each dataset into context.

3.2 Datasets

In this section we describe how the datasets has been retrieved. We also detail how the datasets are before processing them, and some graphs are shown to understand where they come from and their characteristics.

The data that is analyzed in this project consists of tweets that have been posted in the last three years, more specifically between January 2018 and February 2021. The capture was done through the tool Twint and following the next characteristics:

- They are geographically bounded to belong to users who have tweeted in parks in

Spain.

- They must be in Spanish.
- Retweets are excluded as they were not actually written in the park.

The total amount of tweets depends on the dataset, since the study is limited in time, not in the amount of information. They are distributed throughout the geography of the country.

Previously, we used Google Maps in order to determine the geographic coordinates of the parks. Firstly, the parks were searched by their name, and then we used the integrated tool for measuring distances. With that tool, the park was bounded by measuring the length of the radius that includes it. Finally, those coordinates were provided to Twint for the search of the tweets.

The reason why we used Twint is because it allows you to capture tweets of any date, contrary to the APIs that Twitter provides.

The capture was carried out with different Jupyter Notebooks. We wrote these notebooks in Python and in the following piece of code we can see an example of them, where can be seen the parameters already mentioned.

Listing 3.1: Example of Jupyter Notebook used to capture tweets

```
#RETIRO 1
# Configure
c = twint.Config()
c.Lang = "es"
c.Geo = "40.41781,-3.68389,0.325km"
c.Since = "2018-1-1"
c.Until = "2021-3-31"
c.Store_csv = True
c.Output = "retiro1.csv"
# Run
twint.run.Search(c)
```

As expected, the amount of data is bigger in the months leading up to the COVID-19 pandemic, as bans since it started preventing people from leaving home a lot. This effect can be seen in Fig. 3.1, in Fig. 3.3, and in Fig. 3.5. The month with least tweets collected is April 2020, as it was the only full month in which home confinement was in effect. In addition, despite the fact that currently going to parks of any kind is not prohibited, the

number of tweets is still lower than before.

Now, we are going to focus on each retrieved dataset separately.

3.2.1 Urban Parks

This dataset consists of tweets posted in the green areas of the main Spanish cities. The total amount is 30620 tweets. In Fig. 3.1 can be seen the distribution of the tweets in time.

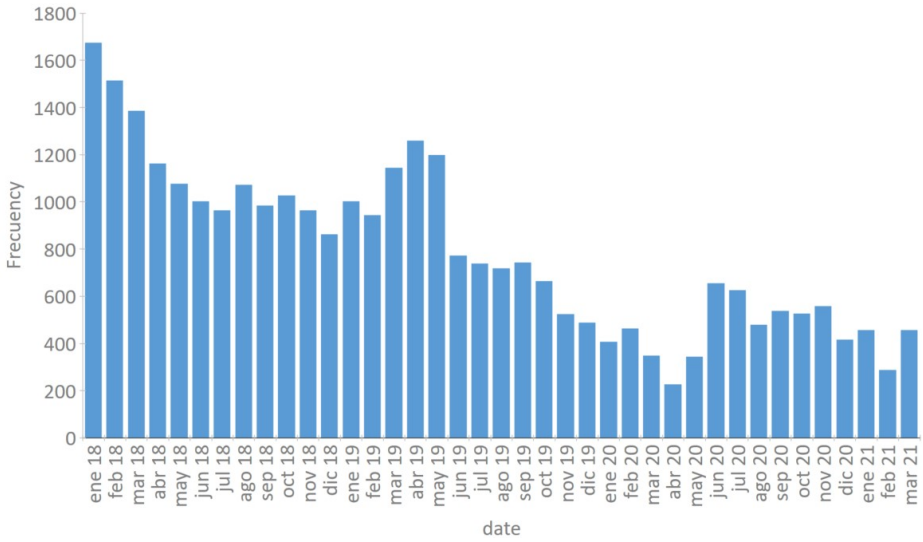

Figure 3.1: Histogram with the frequency of publication of urban parks tweets.

Most of the tweets have been recorded in Madrid, as there is much higher population density and the number of parks is quite large. In the Fig. 3.2 and the Table 3.1 can be seen the quantity of tweets by Autonomous Community.

Figure 3.2: Map with the distribution of the captured urban parks tweets in Spain.

Autonomous Community	**Percentage**	**Tweets**
Comunidad de Madrid	77.07%	23598
Cataluña	11.40%	3600
Andalucía	6.67%	2041

Table 3.1: Autonomous Communities from which the most urban parks tweets come.

In the Fig. 3.2 can be seen that most of the tweets come from the Community of Madrid, as could be expected due to the high population density in this region. Then, we can see more details in the Table 3.1 with the detailed percentages of captured tweets in the most important places. The reason why Cataluña does not get so many tweets, despite its large population, could be that people tweet in Catalan. These tweets have no place in our datasets, since the tools used are to analyze the language only for Spanish, so a filter was previously set to obtain tweets only in this language.

Even having captured tweets from many parks throughout Spain, the great population in big cities such as Madrid or Barcelona, makes the difference between their parks and those in other cities very large. This difference does not mean that we have not captured tweets in other cities with less population, as we can see on the map, but the difference in this aspect is decisive.

The large number of tweets collected in Madrid will allow for a more in-depth analysis in Chapter 5, in which we will focus exclusively on this region, but going park by park instead of putting everyone in a global park dataset.

3.2.2 Natural Parks

The second dataset contains tweets from Spanish natural parks. More specifically, we have retrieved the tweets of the fifteen National Parks of the country. As extensions of land are much larger than in the previous case, the number of tweets increases considerably. The size of this dataset is 232086 tweets, and in Fig. 3.3 can be seen the distribution over time of the publication of them.

We find the peculiarity that in the first months of capture, much fewer tweets were collected than in the rest of the time interval. We have not found any explanation for this phenomenon, but the capture was repeated to rule out possible failures in the capture.

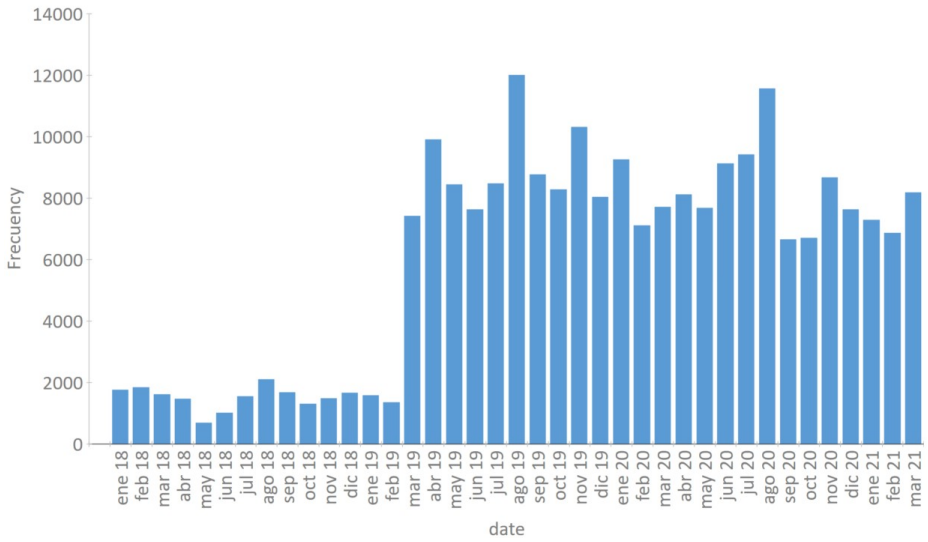

Figure 3.3: Histogram with the frequency of publication of natural parks tweets.

We can appreciate an increase in the summer months, which can be because the people go to these places more in holidays or when the weather is better.

In this dataset we find that there are a large number of tweets from Canarias, where four out of the fifteen National Parks in Spain are located. In the Fig. 3.4 and the Table 3.2 we can see where the tweets come from.

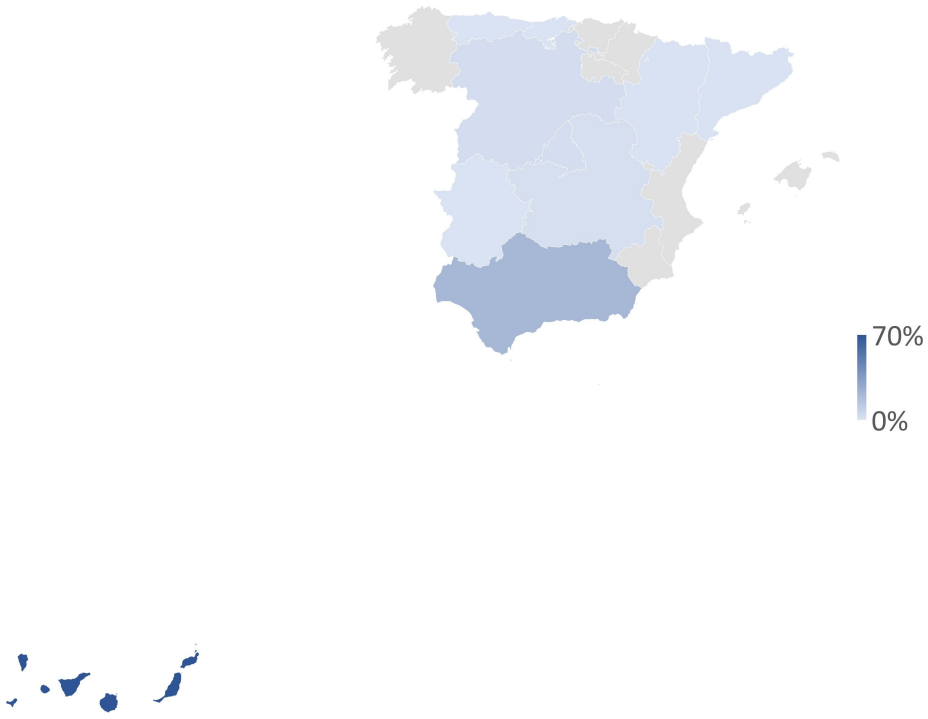

Figure 3.4: Map with the distribution of the captured natural parks tweets in Spain.

National Park	Percentage	Tweets
Teide (Canarias)	58.24%	135175
Doñana (Andalucía)	14.80%	34341
Taburiente (Canarias)	8.67%	20114
Sierra Nevada (Andalucía)	6.22%	14439
Guadarrama (C. Madrid/Castilla y León)	5.20%	12063

Table 3.2: National Parks from which the most natural parks tweets come.

3.2.3 Control Dataset

A control dataset has also been compiled to demonstrate that there are differences between the tweets in parks and the tweets in, for example, a city. The recorded data come from the city of Madrid and belong to the same time period as the main datasets. Its size is 58915 tweets and will be analyzed in the same way as the rest.

The search process is exactly the same as the main dataset and in Fig. 3.5 it is shown the periods of publication.

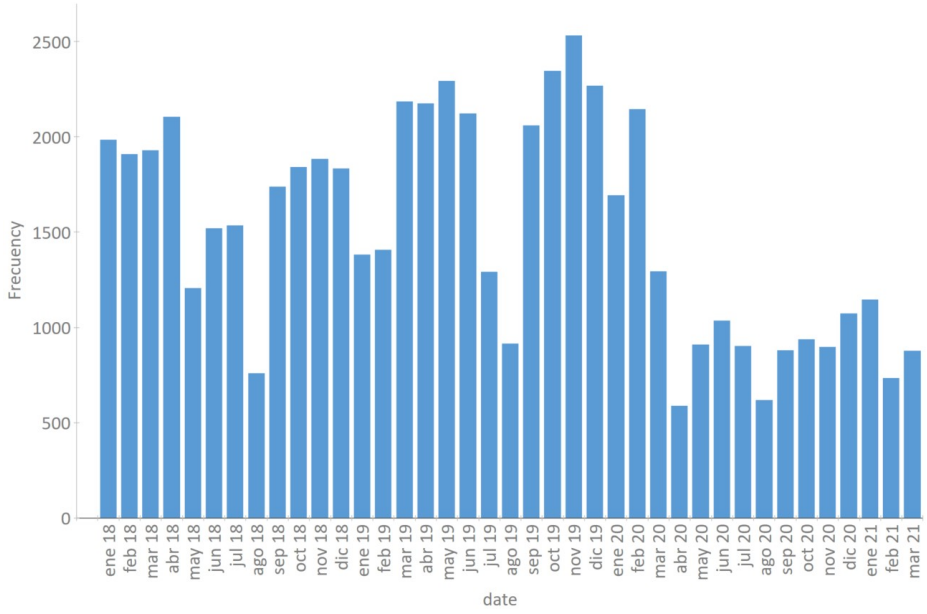

Figure 3.5: Histogram with the frequency of publication of control tweets.

The tweets are from random areas of Madrid, not from the whole city, which is enough for the analysis because the limits of location are less restrictive and many more tweets are retrieved.

This time we can also appreciate a drop in the amount of tweets since the pandemic started. It can also be noted that in the summer months there are fewer tweets, which suggests that during the holidays people tweet less or they are not in the city.

CHAPTER 4

Data Analysis

4.1 Introduction

In this chapter we detail how the datasets have been processed. We can also find some examples of the results, what were the possibilities we could get and what were the ones we finally got.

We will first talk about how the datasets have been prepared before being processed, since the raw data cannot be analyzed directly. Next, there will be explained the four analyzes that will be applied to the three datasets, section by section. Firstly, an analysis of sentiments, then an analysis of emotions and finally we will focus on the hashtags of the tweets, as well as their topics.

4.2 Preprocessing

Firstly, we had to clean our datasets of unwanted information. The parts that could make the analysis process more difficult or simply slower are eliminated. Twint returns the information in CSVs with a lot of information that is not relevant for our study, such as

the number of retweets or the link to the tweet. This has been eliminated manually, simply editing the CSVs and removing the unwanted columns. That is the easy part to eliminate, but for a good analysis the tweets filtering has to be more exhaustive.

We detected that a series of tweets came from bots, for example, bar or restaurant accounts that regularly publish advertisements. The most prominent case was the account of a bar located in the Casa de Campo park, which in the period of tweet collection came to publish more than 7000 tweets. These, being scheduled tweets, have to be deleted from the dataset since they do not express any feeling.

For this, what we have done is ordering the users according to the number of tweets they have posted. Next, we have manually reviewed the ones that generated the highest volume of tweets, and eliminated the tweets of those that were bots. Additionally, we have passed a RapidMiner filter that detects and eliminates tweets that are repeated many times, both by the same account, or if they were at the same time.

We also noticed that many people have their Instagram accounts linked to Twitter, what means that every time the user publishes a photograph on Instagram, his Twitter publishes the link to the photograph and some other information. These tweets have not been discarded, because in this case they will become extra information that is going to be useful later.

Finally, the datasets have been significantly reduced after these filters. We can see the final numbers in Table 4.1. These cleaned datasets still contain the tweets with photographs or with links to Instagram posts.

Dataset	**Tweets before filtering**	**Tweets after filtering**
Urban parks	30620	15540
Natural parks	232086	219147
Control	58915	52292

Table 4.1: Comparison of the number of tweets in the datasets after filtering them.

We are going to use this cleaned datasets for the five analyzes in this project. Even so, depending on it, some will use more or less information since it may be the case that some tweets are not compatible with some analyzes.

4.3 Sentiment Analysis

The first processing of the datasets consists of labeling them according to their sentiment, thanks to the Natural Language Processing (NLP). This is carried out through the MeaningCloud API, making calls from the RapidMiner platform. This sentiment analysis classifies the tweets into five different categories:

- Very positive (P+).
- Positive (P).
- Neutral (NEU).
- Negative (N).
- Very negative (N+).
- None (NONE).

In the Table 4.2 are shown different examples of the tagging.

Tweet Message	**Sent. Label**
Estupenda mañana para pasear por el Parque del Retiro.	P+
Hoy hizo mucho solete y eso me pone muy feliz.	P+
En el cumple de mi nieta.	P
Disfrutando de una temperatura ideal y de la vegetación mediterránea en el Parque del Alamillo, en Sevilla.	P
Hoy toca... por fin. La calor impide salir tanto...	NEU
Que no se nos pase la vida esperando mejores tiempos.	NEU
Ya echo de menos Madrid.	N
Vivo confuso.	N
Pa esto queréis salir a la puta calle? Desgraciados.. Podría estar horas limpiando el parque, de mierda, mierda de gente irresponsable.	N+

Y sigue... creo que la mitad de las plantas muertas por Filomena.	N+
Mi bosque particular.	NONE
Acaba de publicar una foto en Lago De La Casa De Campo.	NONE

Table 4.2: Examples of tweets tagged by sentiment.

Now, we have to compare the results obtained in the analysis of each dataset. There are different hypothesis about what we can find in the results. We can find that tweeting in parks influences people to write more positively or more negatively, or on the other hand we can find that this does not influence their way of tweeting.

In Fig. 4.1 can be found the exact amounts of tagged tweets in each category for each of the three datasets we use in this project. We have used logarithmic scale because there is a big difference between the number of tweets in one dataset and another.

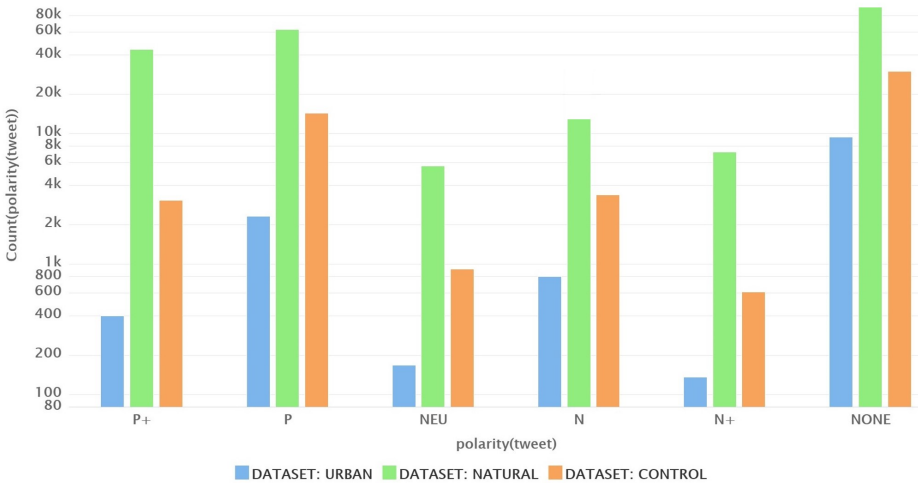

Figure 4.1: Sentiment analysis of the tweets in absolute values.

As the datasets do not have the same number of tweets, the differences between the three are not well appreciated. Therefore, we are going to show the same graph but, this time, in percentages in Fig. 4.2.

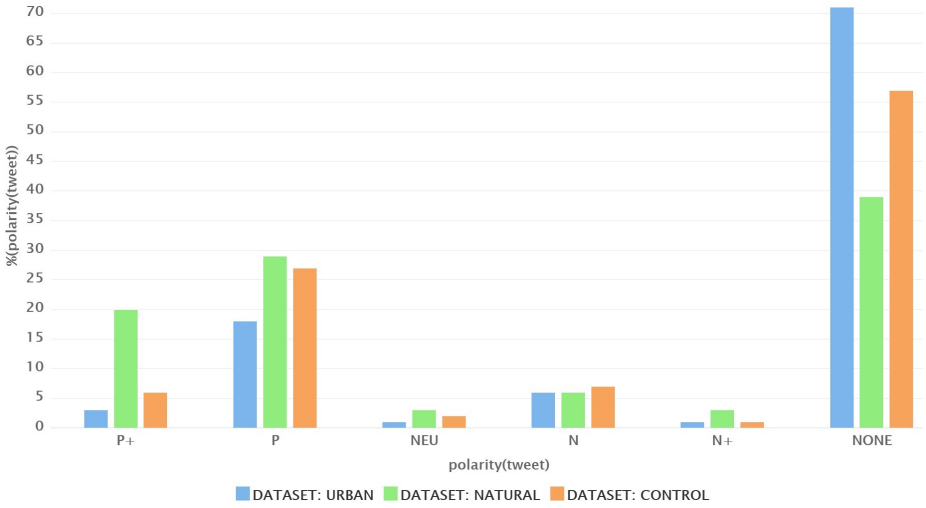

Figure 4.2: Sentiment analysis of the tweets in percentages.

To be even more precise, we put together all the positive tweets and all the negative tweets to get an even better idea. We can compare it in Fig. 4.3 being POSITIVE the sum of P and P+, and NEGATIVE the sum of N and N+.

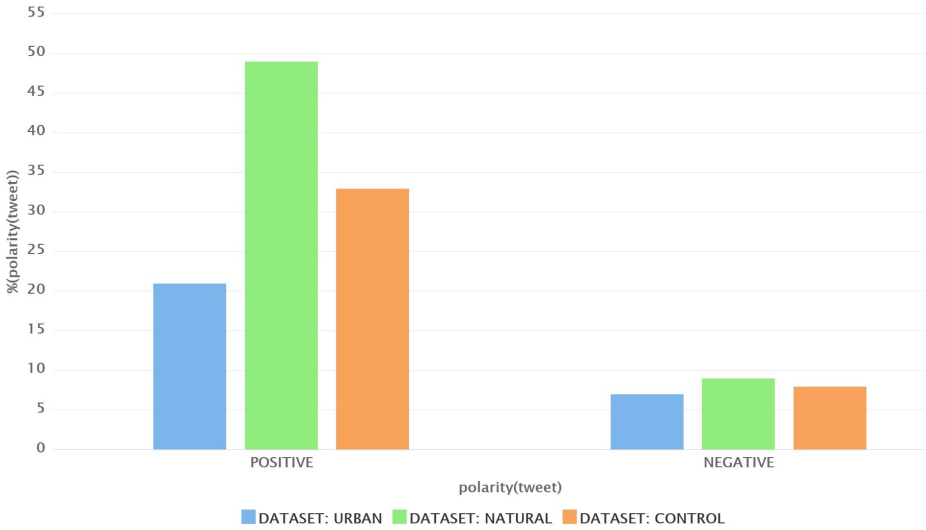

Figure 4.3: Sentiment analysis of the tweets in percentages.

As we can see, there is some differences between the urban parks and the control dataset collected in Madrid, but we notice the biggest difference with the natural parks. It seems that in general, positive sentiments predominate in all datasets, but in natural parks the positive tweets clearly stand out.

We can also see the large number of tweets marked as NONE. If we investigate more

in this category, we find out that most of them are photographs or links to Instagram, as mentioned before, which also suggest that something has been photographed and posted. But this only happens in the case of parks, as we can see in Table 4.3, the percentage of tweets with these characteristics is much higher in the case of parks. This shows that people change their behavior being in places with more nature.

Dataset	$\frac{Tweets\ with\ photo}{Total\ tweets}$	$\frac{Tweets\ with\ photo}{NONE\ sentiment\ tweets}$
Urban parks	44%	97%
Natural parks	31%	83%
Control	21%	42%

Table 4.3: Comparison of the number of images in the datasets.

We can find a big difference between the amount of photographs posted when people is in a park and not. The number of photographs in urban parks is more than double that in the city. We also discover that practically all the tweets without feeling in the parks are because they are photographs. Unlike in the control dataset, where tweets with NONE sentiment are much more varied. It is clear that in the parks many more photographs are posted.

These results suggest that maybe the urban parks should be studied in parts. Since it is evident that in natural parks the tweets are much more positive, it would not be unreasonable to think that a similar situation could have occurred in urban parks. For all this we are going to do an analysis separating urban parks into two subcategories:

- **Type 1:** Parks that people just walk through as one more street. For example, a park in the middle of a city that many people go through every day because it is the shortest way to go to work. An example could be El Retiro park in Madrid, which many people use as a place of passage.

- **Type 2:** Parks where people go to as a leisure activity. For example, large parks that are often used by people to spend the whole day. An example could be Juan Carlos I park, in Madrid, where almost everybody goes as a leisure activity.

If we separate the parks like this, we find out we were right. Tweets that belong to the first subcategory are more similar to those collected in the control dataset. While the

tweets that belong to the second subcategory are very similar to those we retrieved in the natural parks dataset. We can see the results in Fig. 4.4 and Fig. 4.5. We have chosen for the comparison the tweets whose belonging to one type or another is more obvious, while we have left out the parks that were more difficult to discriminate.

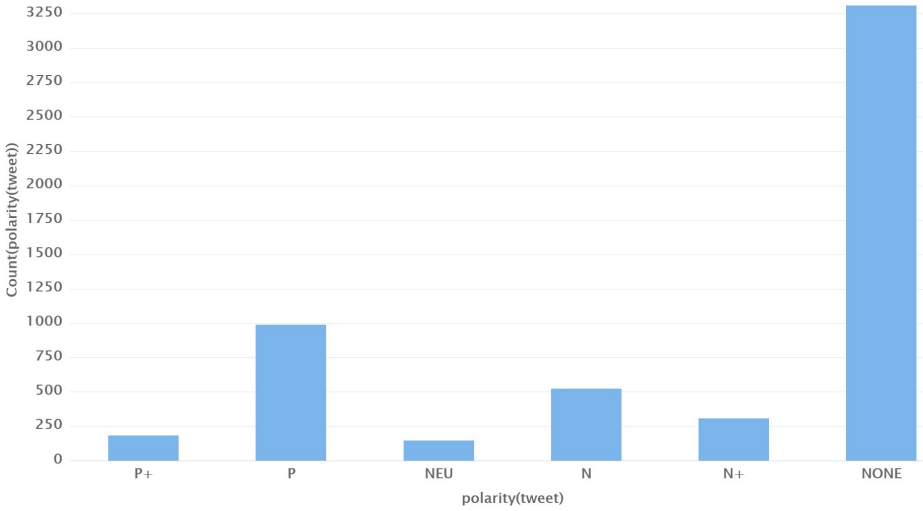

Figure 4.4: Sentiment analysis of type 1 urban parks tweets.

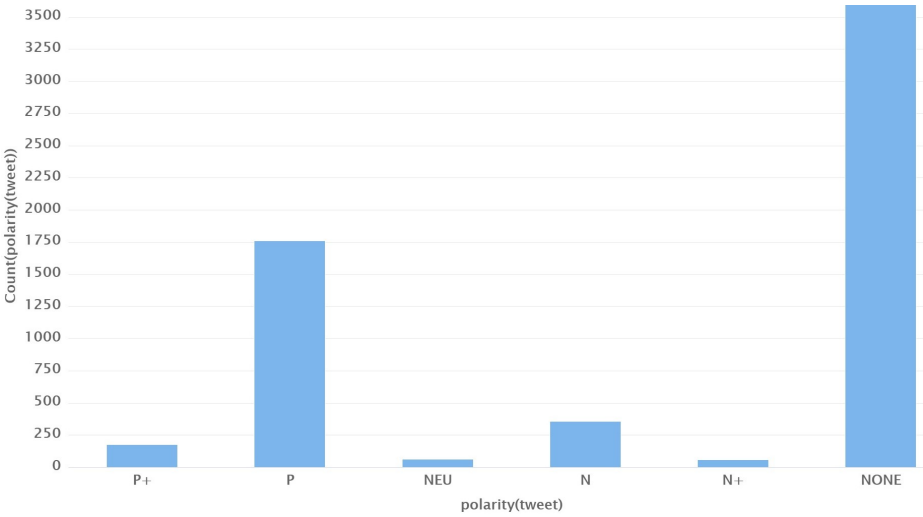

Figure 4.5: Sentiment analysis of type 2 urban parks tweets.

At first glance, the difference between the two graphs might not be appreciated. If we look more closely, it can be seen that the sum of the negative and positive feelings of both is very different. While the sum of the positive and negative sentiment tweets is 1176 and 839 in the first type, in the second type it is 1940 and 416, respectively. The difference is clear between both types.

In tweets of the first type, the number of negative and positive tweets is similar, as in the control dataset. On the other hand, tweets of the second type are more similar to tweets from natural parks, that is many more tweets of positive sentiment than negative.

Now, for the sentiments of each dataset, we will present the descriptive statistics that will allow us to see in more detail the differences between each one. We will do a t-test analysis of the different datasets. This is because the differences in positivity between the different datasets cannot be clearly appreciated. To carry out this analysis, we assign each of the different analysis results an integer as seen in Table 4.4.

Polarity	P+	P	NEU	N	N+
Asigned integer	5	4	3	2	1

Table 4.4: Integer number assigned to each type of polarity.

Now, we proceed to conduct the Two-Sample Mean Comparison T-Test. This depends on the mean and the standard deviation, which we will call M and SD respectively, as well as the size of each dataset. In the Table 4.5 we can see the results of these calculations. It shows the value of t, that informs us about the difference between datasets two by two, each accompanied by its respective mean and standard deviation, which have been calculated with the integers resulting from the previous conversion. The alpha value that we will consider is 5%, that is $\alpha = 0.05$, which gives us a value of t to compare of 1.960. From that number it can be considered that two datasets are different from each other. The larger the value in the table, the more differences we see. If this is positive it means that the feelings are more positive and if it is negative, that zone is more negative.

DATASET	Urban parks	Natural parks	Control
Urban parks (M=3.535. SD=1.044)	-	-37.173	1.540
Natural parks (M=3.934, SD=1,212)	37.173	-	73.509
Control (M=3.52, SD=0.989)	-1.540	-73.509	-

Table 4.5: Results of the calculation of the Two-Sample Mean Comparison T-Test.

We can find in green the comparisons that do pass the test and in red those that do not. Passing it supposes that there are clear differences between them. The difference between

the natural parks and the other two datasets is huge, as we might expect, but between the urban parks and the control dataset we have much less variation.

4.4 Emotion Analysis

In this section we analyze the emotions of the tweets. These have been labelled, in this case, by the emotion that the user is expressing. It has been carried out, like the previous analysis, with MeaningCloud and RapidMiner. In this case, we have used the API of Deep Categorization. The labelling consists in eight primary bipolar emotions:

- Anger.
- Anticipation.
- Disgust.
- Fear.
- Joy.
- Sadness.
- Surprise.
- Trust.

In the Table 4.6 can be seen some examples of the labelling.

Tweet Message	**Emotion Label**
En este mundo traidor.	ANGER
Antonio está enfadado en El Retiro.	ANGER
Acabó el finde. Amenazaba con empezar mal.	ANTICIPATION
Tenía muchas ganas de una tarde así de agustito.	ANTICIPATION
Los colores de la tauromaquia me fascinan pero me repugna 'La Fiesta Nacional'.	DISGUST
Que asco de parques catalanes.	DISGUST
Dios que pesadilla madre mía	FEAR
Sobrecogido en Bulnes.	FEAR
Disfrutando de cada momento.	JOY
Hoy me he levantado feliz.	JOY
Qué triste es verlo todo vacío.	SADNESS

Estoy llorando.	SADNESS
La vida es absolutamente asombrosa.	SURPRISE
Siendo testigo del eclipse lunar. Impresionante verlo.	SURPRISE
Mi estación favorita.	TRUST
De mis lugares favoritos visitados hoy en Cádiz	TRUST

Table 4.6: Examples of tweets tagged by emotion.

Prior to the analysis, we had several hypotheses, but we should expect that most of the tweets in green areas are of positive emotions. This would be what would agree with the predominance of the positive polarity of tweets in green areas. We consider the emotions that can be considered as positive are JOY or TRUST.

In the emotion analysis, tweets whose sentiment is NONE or NEUTRAL are negative in each of the eight emotion tests that are carried out, this is why the number of analyzed tweets obtained in this section is much lower. We can see the numbers in Table 4.7, that shows how many tweets an emotion was found in during the analysis. It should also be noted that a tweet can be positive in several tests. For example, there may be the possibility that a message shows JOY and SURPRISE at the same time.

What is impossible is that a tweet that in sentiment analysis has turned out not to have polarity, now it does have emotion. These tweets are excluded from this analysis since, as mentioned above, they give negative results to all emotion tests. Tweets that contain photographs are included in the discard, which even knowing that they were not going to have any feelings or emotions, were present in the sentiment analysis for the reasons that we already explained. This will also happen in future analyzes in this chapter.

Dataset	**Total tweets**	**Tweets without emotion**	**Tweets with emotion**
Urban parks	15540	11900 (76.6%)	3640 (23.4%)
Natural parks	219147	127375 (58.1%)	91772 (41.9%)
Control	52292	30928 (59.1%)	21364(40.9%)

Table 4.7: Amount of tweets that express emotion in each dataset.

The final results of this analysis will now be displayed for each dataset in Fig. 4.6, Fig. 4.7, and Fig. 4.8.

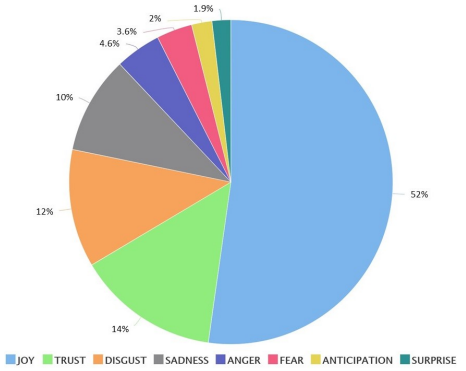

Figure 4.6: Emotion analysis of urban parks tweets.

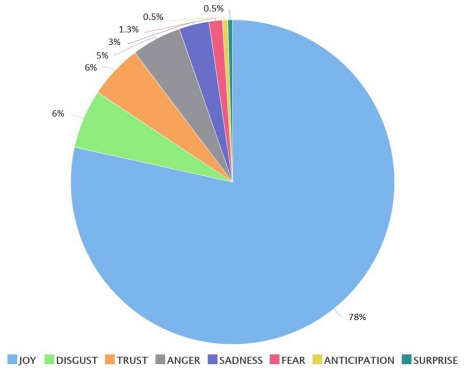

Figure 4.7: Emotion analysis of natural parks tweets.

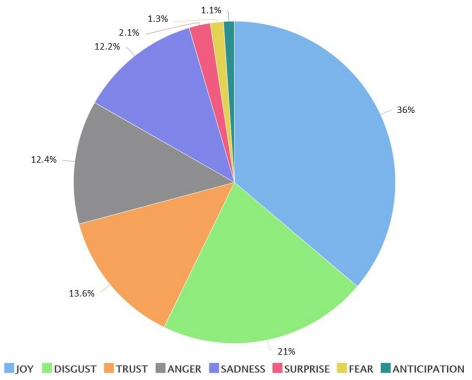

Figure 4.8: Emotion analysis of the Madrid control dataset.

In order to better compare the emotions of each dataset, in Fig. 4.9 we can see a comparison in the same graph. Now we have added the good and bad emotions, considering JOY and TRUST as POSITIVE, and ANGER, DISGUST, FEAR, and SANDESS as NEGATIVE. We have ruled out ANTICIPATION and SURPRISE since they can belong to both categories and, in addition, the amount of tweets they represent is very low and hardly affects the study.

After the study, it is once again shown that in natural environments, people tweet more positively than in urban settings. As in the sentiment analysis, we see that in natural parks the difference is much more noticeable compared to the other two datasets.

This time, in urban parks can also be seen a difference compared to the control dataset. It is clear that, as the environment is 'less urban', the users have more positive emotions.

It begins with tweets with mostly positive emotions in the National Parks, then in urban parks there are many fewer, and, to finish, in the urban control dataset we see a balance between positive and negative emotions.

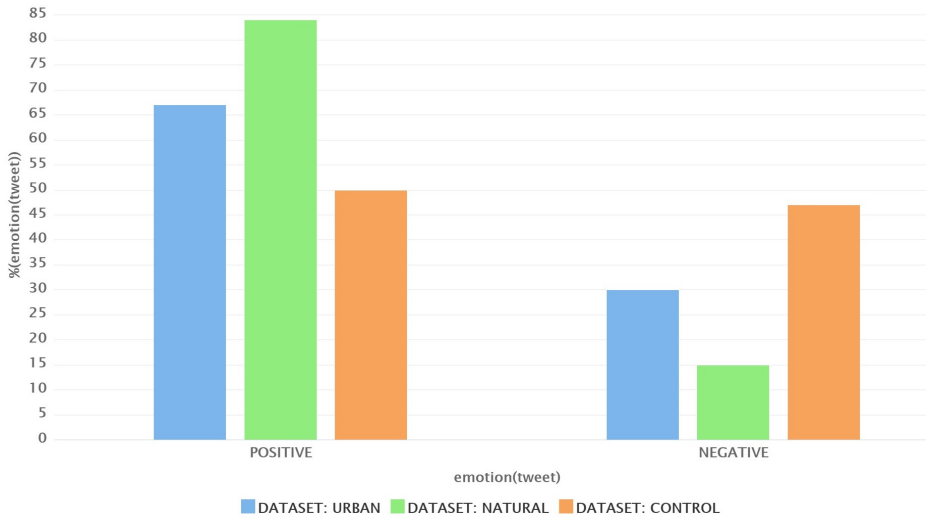

Figure 4.9: Emotion analysis of the tweets in percentages.

As in sentiment analysis, we again separate urban parks into the two types already explained previously. In Fig. 4.10 and Fig. 4.11 we can observe once again the differences between both types of urban parks.

Again, it can be seen how type 1 parks are more similar to the results obtained in the control dataset, while type 2 parks are more similar to natural parks. Thus, we can reaffirm that being in green spaces as a leisure and enjoyment activity has more positive influence than being in a park just passing through as just another city street.

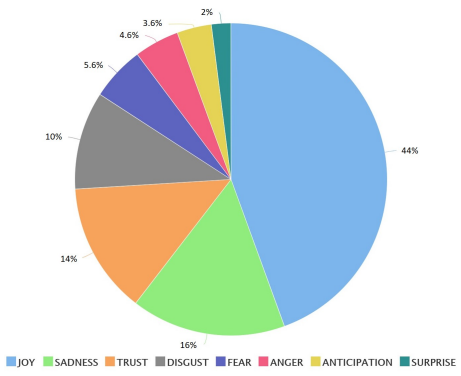

Figure 4.10: Emotion analysis of type 1 urban parks tweets.

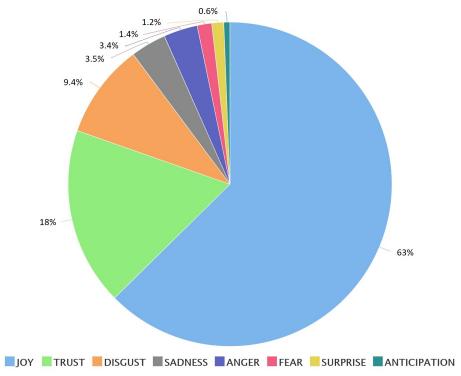

Figure 4.11: Emotion analysis of type 2 urban parks tweets.

4.5 Hashtag analysis

In this section we focus on a very characteristic part of tweets, the hashtags. So far, we have only analyzed the message of the tweet itself, discarding everything else. Now, we will retrieve the hashtags of each dataset to see which are the most common in each place and their frequency. The tool with which we have collected the tweets, Twint, also stored the hashtags of the tweets. Therefore, the same datasets have been used as for the previous analyzes, this time focusing the analysis on another part of them.

Hashtags are generally single words or a small succession of them that are written preceded by the number sign, #. They usually refer to the subject of the tweet, the location where the user is, etc. In the Table 4.8 we can see how many tweets are posted with hashtags.

Dataset	**Total tweets**	**Tweets without hashtags**	**Tweets with hashtags**
Urban parks	15540	8877 (57.1%)	6663 (42.9%)
Natural parks	219147	168878 (77.1%)	50269 (22.9%)
Control	52292	31047 (59.4%)	21254 (40.6%)

Table 4.8: Amount of tweets with hashtags in each dataset.

It seems that in natural parks people are less likely to use hashtags when posting tweets. While in cities, both in parks and not, the percentages of tweets with hashtags are similar. It must be taken into account that the same hashtag can appear in more than one tweet and also that a same tweet can be accompanied by more than one hashtag. To get a more visual idea of the hashtags, they will be displayed in wordclouds. In them, the hashtags that are repeated more frequently will appear larger, while those that are hardly repeated in our datasets will appear smaller.

In Fig 4.12 we can see the wordcloud of our first dataset, urban parks. We can appreciate that all the words are related to things we could expect. There are hashtags referring to the location where the user is, such as #españa, #madrid, #sevilla, #parque, or #retiro. We can also find some related to the photographs that people post when they are in a park, as can be the case of #photo, #instagram, or #landscape. Finally, we can also see hasthtags related to nature or activities that can be carried out in it, as #flores, #nature, #running, or #deporte. This dataset has a total of 29132 hashtags.

In the case of natural parks there is a type of hashtag that clearly predominates over the rest. This is the case of locations, people in these environments use hashtags as information about where they are. Some examples are #tenerife, #canarias, or #picosdeeuropa. Since most of the tweets belong to Canarias, there is a large representation of hashtags referring to that region. We can see the results in Fig. 4.13. The exact number of hashtags in this dataset is 112221.

Finally, the hashtags of the tweets from the control dataset are displayed. These tweets, unsurprisingly, are much more generic than those for the other two datasets. As we can see in the Fig. 4.14, the hashtags are very varied with themes such as photography, music, food, etc. There is also a lot of talk about traffic on the streets, since the control dataset has been collected in Madrid, with hashtags such as #retención or #dgt. This time, the number of hashtags in the dataset is 187521.

Figure 4.12: Wordcloud of urban parks hashtags.

Figure 4.13: Wordcloud of natural parks hashtags.

Figure 4.14: Wordcloud of control dataset hashtags.

4.6 Topic Analysis

In this section we look for the topics that the retrieved tweets deal with. To do this, we make calls to another MeaningCloud API that returns, in binary form, if the text of the tweet matches one of the topics.

The analysis has once again been carried out on the RapidMiner platform. In many of the tweets no topic has been found. The Table 4.9 details the number of tweets that have been sensitive to this analysis and how many have not.

Dataset	Total tweets	Tweets without topic	Tweets with topic
Urban parks	15540	13022 (83.8%)	2518 (16.2%)
Natural parks	219147	175194 (79.9%)	43953 (20.1%)
Control	52292	36498 (69.8%)	15794 (30.2%)

Table 4.9: Amount of tweets with detected topic in each dataset.

It should also be noted that, as there are tweets that do not have a topic, there may be those that have more than one.

Topics that only appear a few times in all datasets have been discarded. In Fig. 4.15 we can look at those that have a significant presence in any of them. Thus, we can compare which topics appear more in some datasets or others.

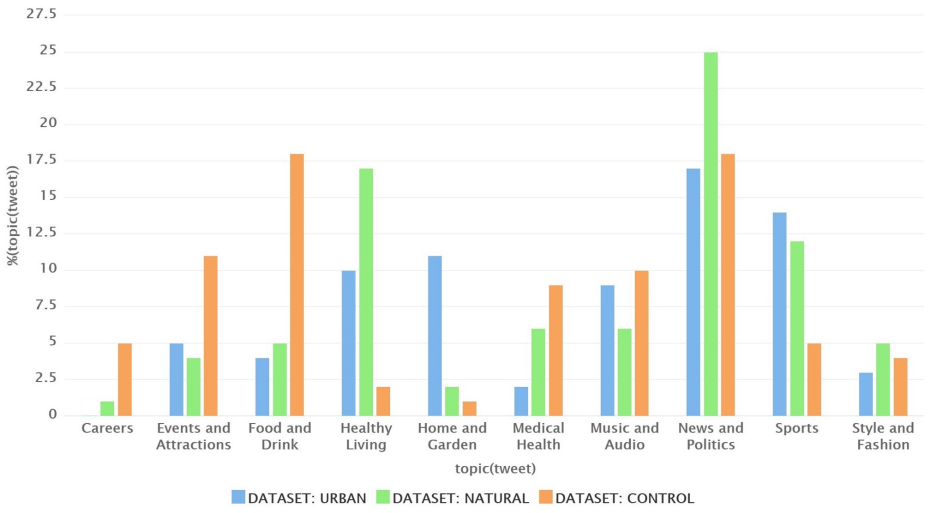

Figure 4.15: Topic analysis of the tweets in percentages

There is a clear difference between the topics of the tweets in parks and those of the control dataset.

To begin with; it can be seen that the first three topics that appear, predominate in the control dataset, while in the parks they have little representation. The first, Careers, represents tweets dealing with job search or career advice, which makes sense that it's not a topic to tweet about in a park. The second, Events and Attractions, includes tweets about sporting events, concerts, and other events, so it also fits. Finally, Food and Drink, is a topic more typical of tweeting in restaurants, bars or houses.

In contrast, we also find topics that clearly predominate in natural environments. The Healthy Living topic is one of the clear examples of a topic that appears more in parks than in the city. It is a topic that includes tweets about fitness and exercise, or wellness. Another topic that stands out in parks is Home and Garden, with tweets about landscaping or gardening, also very typical in parks and little present in the control dataset. Finally, another topic that could be expected to be also much more present in parks than in the urban environment is Sports, since it is very typical to practice any type of sport in a park.

A topic that stands out in urban parks and in the control dataset but not in natural parks is Music and Audio. This could be due to the fact that in the middle of nature, that is, natural parks, people are less likely to listen to music while walking. Unlike in the cities.

What is also clear is that, no matter where you are, one of the favorite topics in Spain and the most talked about is News and Politics. This topic is really present in the three datasets, so much so that it is the one that is most talked about in all of them.

CHAPTER 5

Case study: Madrid

5.1 Introduction

In this chapter we are going to focus more on the city of Madrid. Several green areas of the city will be analyzed to compare and see if there is any difference. The same line will be followed as it has been done on a national scale, putting the retrieved tweets in context and then analyzing them and drawing conclusions.

The compilation of the tweets was carried out individually in each area of the country according to their coordinates, to later be put together for analysis in the Chapter 4. Therefore, this time we only had to take the datasets belonging to Madrid that were part of the urban parks dataset.

5.2 Geography

For this case study, we have chosen the most important and busiest parks in Madrid, since in the others hardly any tweets were collected and they would be insignificant for the study. In the map shown in the Fig. 5.1 we can see which green areas will be studied in this

chapter, while in the Fig. 5.2 is shown the amount of tweets retrieved per zone. Each circle represents a data set with the corresponding area covered.

Figure 5.1: Map of Madrid with the green areas to study.

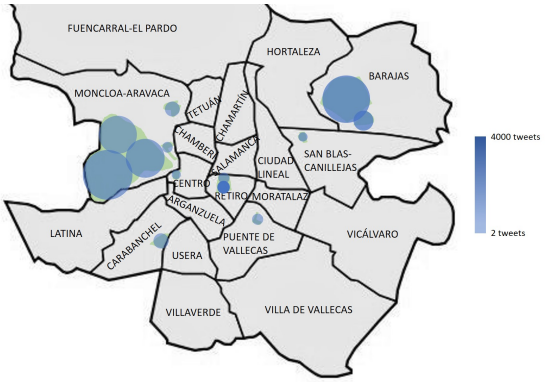

Figure 5.2: Map of Madrid with the amount of tweets per zone.

We can see that there are some parks in which the amount is clearly greater, as is the case of the Retiro park or the Juan Carlos I park. On the other hand, there are others that hardly contribute data to the study. Specifically, the Oeste park has had to be subsequently excluded from the analysis, since it only contributed two tweets and they were not expressed by humans.

Removing that, the rest of the parks in which we have captured tweets have contributed data to the study. In the following sections we will be able to name all the parks that make up the study.

5.3 Sentiment Analysis

In this section we will analyze the sentiments of each green area separately to try to find differences between some places and others.

This time, we will only focus on whether the tweets have a positive or negative polarity. We will discard tweets that lack sentiment. As previously done, tweets with POSITIVE polarity will be the sum of those classified as P and P+, and tweets with NEGATIVE polarity will be the sum of those classified as N and N+.

We will also add to the analysis the control dataset collected in this same city to see how the sentiment of the population varies depending on whether they are in a park or not. The results of this analysis are shown in Fig. 5.3.

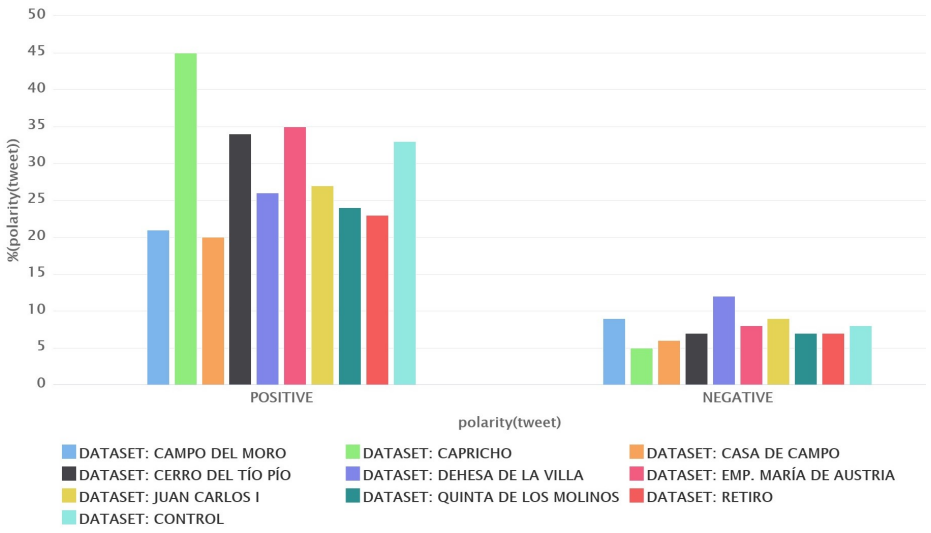

Figure 5.3: Sentiment analysis of the tweets from Madrid parks in percentages.

Looking at the figure, it is obvious that the park that brings the most positivity to the people of Madrid is El Capricho park. It is the park in which more positive tweets have been collected, 45% of them, and in which there are fewer negative ones, only 5%. This park is unique in that it is the only one whose entrance is controlled, and the people who are in it have come solely and exclusively to be in it, not to simply cross it.

Another thing that we can extracted from the graph is that the parks that are located in the southern districts of the city are more positive. We are talking about the Cerro del Tío Pío park and Emperatriz María de Austria park. After El Capricho park, they are the parks where we find the most positive feelings with 34% and 35% respectively. The negative tweets, however, are within the average for all parks.

To continue, the most negative tweets are found in Dehesa de la Villa park. With 12% of the tweets with negative sentiment, it cannot be concluded that this park makes users more pessimistic, since it is still a relatively low percentage in comparison with the positive ones, which in this case is 26%.

Finally, in the graph we can see that people only tweet more positively than in the control dataset in the case of the parks in the south of Madrid and El Capricho park. These results are consistent with those of the analysis at the national level, in which urban parks had fewer positive tweets than the control dataset.

Since we are in the Madrid region, we can also compare these results with the results of the sentiment analysis of the Guadarrama National Park. This is the only natural park from which we have collected data in this study that is located in the Community of Madrid.

People who live in Madrid usually visit this natural park, therefore they are similar population targets. In the Fig. 5.4 we can see the comparison between the sum of all the urban parks that have been analyzed before and the natural park mentioned now.

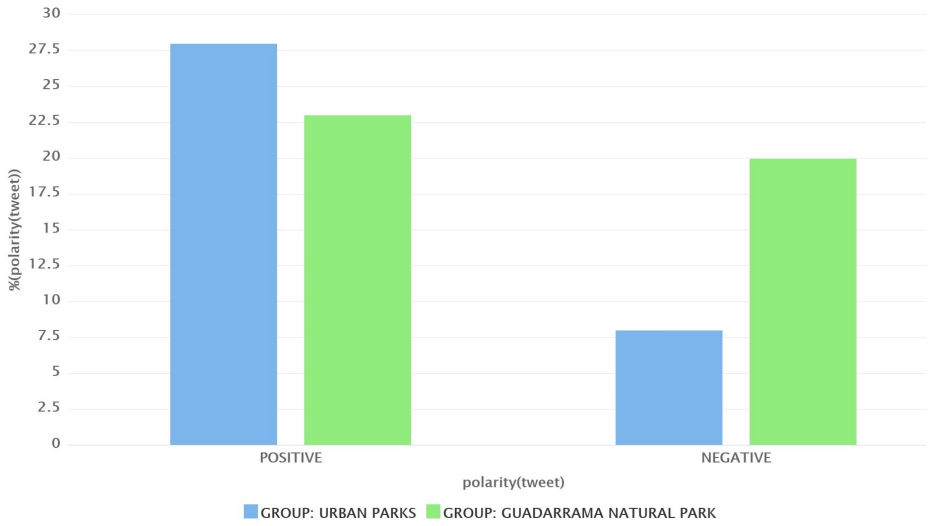

Figure 5.4: Comparison of feelings between urban and natural parks in Madrid.

Surprisingly, this analysis contradicts the one carried out in Chapter 4. Tweets in Madrid parks are more positive within the city than outside.

In Chapter 4 the obvious conclusion was that users who post tweets in a natural park do so with a more positive sentiment on average than those who do so in the city. But in the case of Guadarrama, the negative tweets are almost equal in percentage to the positive ones, unlike in the case of urban parks, in which the positive sentiment tweets are more than triple that those of negative sentiment. In this case study, we can conclude that people tweet more positively in urban parks than in natural parks.

Like we did in the sentiment analysis at the national level, we analyze again the descriptive statistics in Table 5.1. We use $\alpha = 0.05$, which gives us a value of t to compare of 1.960. The values obtained above this number will inform us of clear differences between the datasets, being more positive if the value is positive and more negative if it has a negative sign.

As we mentioned before, we find the tests that have been passed in green and those that have not been in red. Therefore, the comparisons in green mean that clear differences have been found between the analyzed datasets. In the first row of the table, the abbreviations correspond to the names of each analyzed dataset in the same order that they appear in the first column.

DATASET	C. M.	C.	C. C.	C. T. P.	D. V.	E. M. A.	J. C. I	Q. M.	R.	S. U. P.	Ctrl.	G.
Campo del Moro (M=3.537, SD=1.111)	-	-5.258	-0.310	-0.886	0.921	-2.345	0.108	-1.290	0.537	-	0.227	5.450
Capricho (M=4.019, SD=1.12)	5.258	-	10.139	0.697	4.402	1.083	11.287	6.480	13.893	-	15.763	24.315
Casa de Campo (M=3.563, SD=1.032)	0.310	-10.139	-	-0.887	1.346	-2.662	0.899	-1.733	2.091	-	1.553	14.569
Cerro del Tío Pío (M=3.809, SD=1.034)	0.886	-0.697	0.887	-	1.223	0.238	0.935	0.576	1.114	-	1.093	2.442
Dehesa de la Villa (M=3.385, SD=1.203)	-0.921	-4.402	-1.346	-1.223	-	-2.688	-1.081	-1.983	-0.813	-	-1.100	2.298
Emp. María de Austria (M=3.879, SD=1.01)	2.345	-1.083	2.662	0.238	2.688	-	2.953	1.888	3.228	-	3.244	6.348
Juan Carlos I (M=3.528, SD=1.038)	-0.108	-11.287	-0.899	-0.935	1.081	-2.953	-	-2.459	1.109	-	0.312	14.483
Quinta de los Molinos (M=3.652, SD=1.007)	1.290	-6.480	1.733	-0.576	1.983	-1.888	2.459	-	3.379	-	3.174	11.899
Retiro (M=3.493, SD=1.06)	-0.537	-13.893	-2.091	-1.114	0.813	-3.228	-1.109	-3.379	-	-	-1.665	18.843
Sum of Urban parks (M=3.647, SD=1.067)	-	-	-	-	-	-	-	-	-	-	10.977	30.880
Control (M=3.52, SD=0.989)	-0.227	-15.763	-1.553	-1.093	1.100	-3.244	-0.312	-3.174	1.665	-10.977	-	32.038
Guadarrama (M=3.053, SD=1.157)	-5.450	-24.315	-14.569	-2.442	-2.298	-6.348	-14.483	-11.899	-18.843	-30.880	-32.038	-

Table 5.1: Results of the calculation of the Two-Sample Mean Comparison T-Test.

Among most of the urban parks we did not notice great differences, except for the cases in which the differences in feelings were more significant. The areas of Madrid that we had called the most positive continue to be so when compared. El Capricho park stands out again as the most positive and Dehesa de la Villa as the most negative. This statistic also concludes that the parks are more positive than the city, that is, the control dataset, with a t of $10.977 > 1.960$.

Here we can clearly see the great difference between the Guadarrama National Park and the other two datasets, with a t value of ± 30.880 with urban parks and ± 32.038 with the control dataset. This is not a surprise given the previous results. The natural park, with those numbers, is the one that marks the most differences with the rest.

5.4 Emotion Analysis

This section will be similar to the previous one, but this time analyzing the emotions of the tweets.

For this analysis we are not going to group positive and negative emotions as we have previously done. Even so, we continue to consider JOY and TRUST as positive emotions, and ANGER, DISGUST, FEAR and SADNESS as negative emotions. As it happened before, the ANTICIPATION and SURPRISE emotions have hardly any representation in the datasets, they do not even appear in several of them, so we will leave them aside when drawing conclusions.

This time we will also add the control dataset in the study to compare results between green areas and urban environment. In Fig. 5.5 we can see the results with a stacked bar chart, that allows to appreciate the percentages of tweets that each emotion has, as well as if some areas have more of one or the other.

We have placed the emotions named as positive at the bottom of the columns. Thus, the comparison between positive and negative emotions can be better appreciated. Being able to add those that are of the same polarity easily.

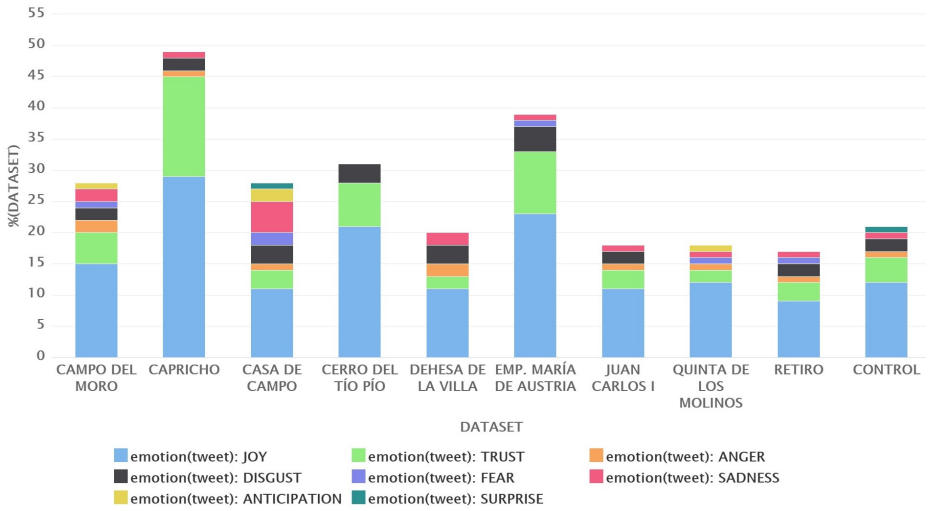

Figure 5.5: Emotion analysis of the tweets from Madrid parks in percentages.

The results obtained are consistent with the sentiment analysis. Once again, El Capricho park is the one where people tweet happiest. The emotions of JOY and TRUST have their maximum there, being also one of the parks with the least negative emotions. The exact numbers are 45% of total tweets of positive emotion versus just 4% of negative, which makes a big difference.

The same differences also occur with the parks in the south of the city. The Cerro del Tío Pío and Emperatriz María de Austria parks are the other two parks with the most positive emotions. It is thus clear that people in this area of Madrid tweet more happily than others. Here we find 28% and 33% respectively of positive emotion tweets, which compared to other parks in the graph is a substantial difference.

This time, where we find the most negative tweets is in the Casa de Campo park. Here, positive tweets have almost the same representation as negative ones, with 14% and 11% respectively. This park is where we find the most tweets with the emotions of ANTICIPATION and SURPRISE. It could be that these are of positive emotion, which would fit

compared to the sentiment analysis.

The results of the rest of the analyzed parks are more or less similar to those of the captured control dataset. It could be said that these parks hardly change the emotions of the users.

Once again, we compare the urban parks of Madrid with the natural park located in this community, the Guadarrama National Park. Thus, we can reaffirm ourselves in the conclusion that was drawn in the sentiment analysis, that in Madrid people do not tweet happier in natural parks. We can see the comparison with the same format as before in the Fig. 5.6.

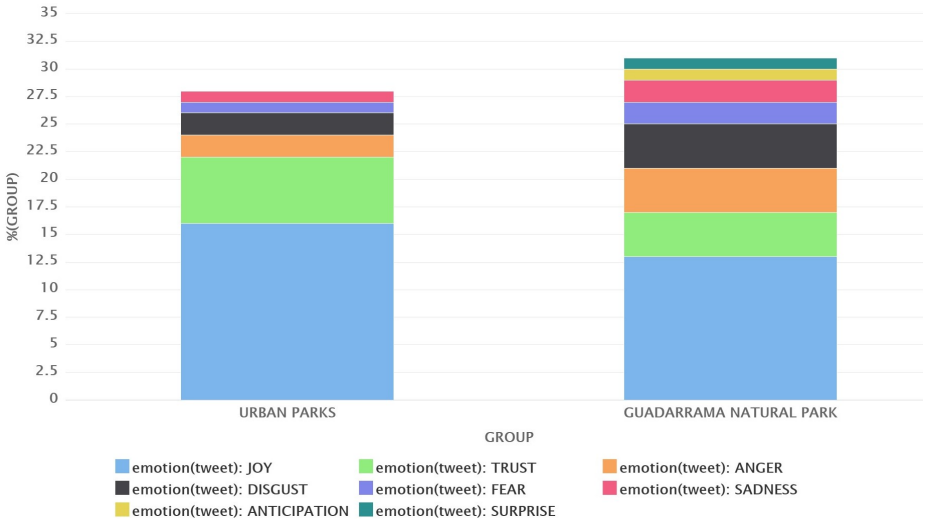

Figure 5.6: Comparison of emotions between urban and natural parks in Madrid.

Indeed, we confirm that in the urban parks of Madrid there are more positive emotions than in the natural park. In urban parks we get 22% of tweets of positive emotions compared to 5% of negative ones. On the other hand, the natural park has 17% of positive emotions, while we obtain 12% of negative emotions. As in sentiment analysis, the difference between positives and negatives is very small compared to urban parks.

Finally, it is once again clear that people tweet differently depending on where they do it and that there are areas that are more positive than others.

5.5 Topic Analysis

In this section, the most repeated topics in the tweets of the different parks in Madrid are analyzed.

This analysis will help us to understand what behaviors people have according to where they are and how each park influences them. As we already did at the national level, we have discarded the topics that had very little representation in all the datasets, since they do not provide relevant information.

In the Fig. 5.7 we can see the results of the analysis. There is a lot of information in it that we will discuss below. Once again, the control dataset accompanies urban parks in order to clearly see the differences among them.

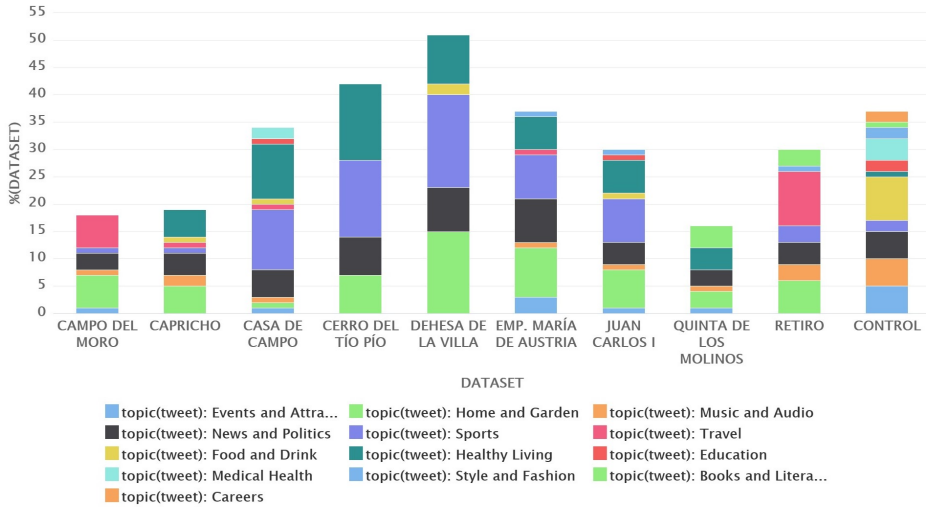

Figure 5.7: Topic analysis of the tweets from Madrid parks in percentages.

At first glance, many things are obvious from each park. There are topics that stand out clearly from others and that represent a large part of the percentage of its dataset.

To begin with, we look at the topic Home and Garden. This topic represents a good percentage in most datasets, however there is no trace of it in the control dataset. Therefore, being in an urban park not only means for the user to be in nature, but it also makes it reflect in their tweets. While on average in the urban parks of Madrid, this topic represents around 7%, in the control dataset it barely has representation.

A similar case occurs with the Healthy Living topic. This topic represents a scant 1% of the tweets in the control dataset, while the representation in almost all urban parks is clearly higher with an average of almost 8%. This makes us think that people associate

parks with a healthy life. There is an exception, which is that there are two parks in which this topic is hardly mentioned, it is the case of the Campo del Moro and Retiro parks.

Speaking precisely of these two parks that we have just mentioned, we could have the key to why this topic is not mentioned there and in others it is. In these two parks a topic predominates over the rest that only does it here. We talk about the topic Travel. It so happens that these two parks are two of the most touristy in the capital of Spain, which could explain this. It could be that we have captured a certain amount of tweets that belong to Spanish-speaking tourists. This topic is hardly represented in the rest of the parks, unlike in the Retiro and Campo del Moro parks, with an average of 8%.

If we continue talking about the Retiro park, we find that it is the dataset in which the Music and Audio topic most appears with the control one. It could be because in this park it is typical to find people playing music or listening to it, for example people who go running in it.

We also find in many of the parks that the Sports topic predominates. Precisely where there is almost no representation of this topic is in the parks that we have previously called touristic parks, as well as in El Capricho park, in which, since it is a restricted-entry park, sports should not be the most common.

Regarding the topic Books and Literature, according to the percentages obtained, we can say that the Quinta de los Molinos and Retiro parks are where people go to read the most, or at least where they talk about reading the most.

Unsurprisingly, the only generalized topic no matter where the user is, is News and Politics. This topic is around similar percentages in all the datasets, and in all of them it has a considerable representation. We have already experienced the same in the analysis at the national level.

Finally, we focus on the control dataset. It looks clearly different from the rest of the datasets on the graph. Apart from having much more variety of topics, those that predominate are very different from those of urban parks. Topics such as Medical Health, Education, Careers, Food and Drink, are some of those that we can find in the control dataset, but that hardly have any representation in the parks.

It is clear that each dataset is different and that a multitude of factors can determine the appearance of some topics or others, but that similarities can be found among them and reasons why some results are obtained or others.

Now we will make the comparison of the average of the urban parks with the Guadarrama

National Park as we have been doing previously. In the Fig. 5.8 we can see this comparison with a similar graph. We introduce in it the topics that are present in a representative number of tweets from at least one of the two groups.

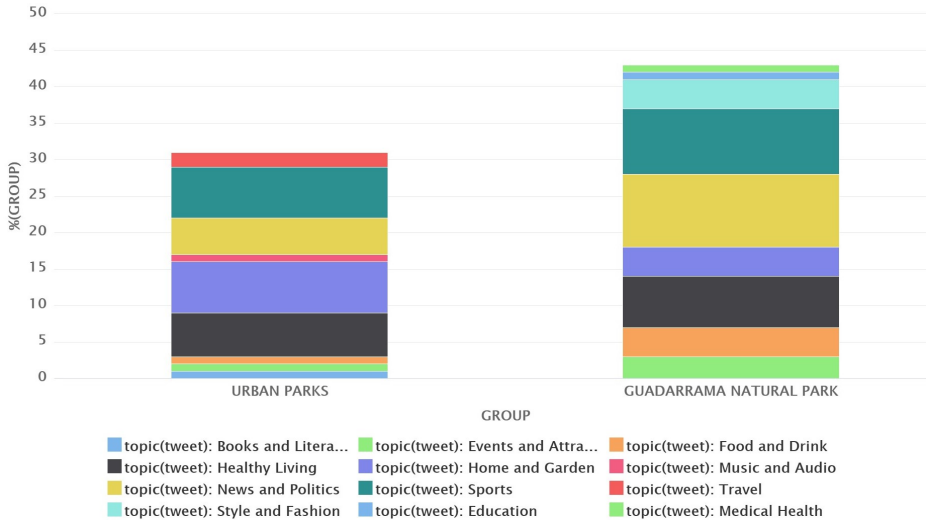

Figure 5.8: Comparison of topics between urban and natural parks in Madrid.

We can see that in natural parks there is a greater variety of topics. The most representative here is News and Politics with a 10%, once again being the most common topic anywhere in Spain.

It is followed closely by the topic Sports, which with 9% here and 7% on average in urban parks, is the one with the greatest presence in general. It is clear that practicing sport is common in parks of any kind and that users also share it on Twitter.

Topics such as Healthy Living or Home and Garden are present in both as expected, since they are very typical of natural environments. The others are less represented and vary depending on the dataset.

CHAPTER 6

Conclusions and future work

6.1 Introduction

In this chapter we will describe the conclusions extracted from this project, the goals we have achieved, as well as some of the problems found with their respective solutions.

Finally, we will expose thoughts and suggestions about future work.

6.2 Conclusions

In this section we present the conclusions that have been drawn throughout all the phases of the project.

Regarding the data acquisition part, located in Chapter 3, we are aware that there are places that contribute many more tweets than others to our datasets. That is why a study of more limited places is carried out later. In the urban parks dataset, the clear contributor is the Community of Madrid, while in the case of natural parks the dominant one is the Canary Islands. In the first case because of the high population density and, in the second, because of the number of natural parks that are located in the region.

The first conclusion from the analysis we draw from all this work is that, in general, people who tweet do so much more frequently in a positive way than in a negative way. In all the studies carried out in this project, regardless of the location or the type of analysis, there has always been a clear difference between positive and negative tweets. If we take all the tweets used for this project, 34.3% are positive and 8% are negative, which is a big difference between both. The number of positive tweets is more than four times greater than the number of negatives.

The other great conclusion that is drawn from Chapter 4 is that in natural parks there are a much higher number of positive tweets compared to the rest of the datasets, while in negative tweets there are hardly any variations in quantities. The relaxation that people experience in nature may be responsible for these statistics. On the other hand, the urban parks and the control dataset captured in Madrid are more similar, the latter even more positive. It seems that users do not achieve in urban parks the disconnection from the city that they do in natural parks.

We also conclude that in every type of park, the number of photographs increases considerably. Most of the tweets that lack polarity in these places are because a photograph is being posted, while in the control dataset there are much more varied reasons. That also gives us insight into how human behavior changes depending on where you are.

In an attempt to explain the notable difference between the results for the two classes of parks, we divided the tweets from urban parks into two subgroups. With this division we conclude that there is a difference between parks that are used as a simple passage area and parks that are used more for leisure. The first is more similar to the control dataset, and the second to the natural parks dataset, so the study determines that when people go to a natural area it is to relax, disconnect, etc. and that this directly influences our social networks, leading us to tweet in a more positive way.

The study of emotions makes us reaffirm our conclusions with feelings. The predominant emotion is JOY everywhere, but this is stronger in green areas. The number of tweets of these characteristics is overwhelming in natural parks, with more than four-fifths classified as positive emotion.

With the analysis of hashtags we draw the conclusion that they are mostly used to report the location where the person is, and that they are used much more in the city than in the countryside. Even so, the results of the previous ones can also be seen reflected in this analysis. In the parks most of the hashtags we see are positive and in the city not so much.

The great conclusion that is drawn from the analysis of topics is that people in Spain love to tweet about News and Politics. It is the topic that has the most representation in all the datasets. We also know that in parks there is much more talk about nature topics than in urban surroundings, as could be expected, while in the city more routine topics predominate.

When it comes to focusing on Madrid, we can conclude that there are areas of the city where tweeting is more positive than others. Here, the south contributes many more positive tweets than other areas of Madrid, being the amount higher than the average and in the control dataset. We also conclude that the happiest park in Madrid is El Capricho.

We also reach the conclusion that, in this case, people do not tweet more positively in Guadarrama, the natural park of this community, but quite the opposite. We found an unusual number of negative tweets in this location, which suggests that Madrilenians cannot find the positivity that users do in other parts of Spain. It is possible that Madrid is different due to the clear differences between this region and the rest of the country, with a very different demography that can cause changes in the way they interact with social media.

Regarding the topics, the conclusions are similar to those at the national level, also seeing that the characteristics of each specific park affect a lot the topics of the tweets that are posted there.

6.3 Achieved goals

This section details the objectives that have been pursued and fulfilled throughout this project.

Capture of three datasets of tweets in Spain by geographic coordinates.

The first objective was to collect three different datasets capturing tweets in many parks in Spain, as well as a control dataset. We had to obtain a dataset of each park that we wanted to study, and then put them together in three global datasets: Urban Parks, Natural Parks and Control Dataset. The tool used was Twint, which in an easy and simple way allows you to capture tweets by entering the search coordinates, as well as other parameters.

Filtering of unwanted data.

Before processing the data, we had the objective of cleaning the dataset of the tweets

that could not be used in the analyzes. We had to eliminate the information that Twint returned and did not serve us, as well as the tweets that were automatically written by bots, etc. For this we have combined the use of RapidMiner as well as manually.

Classification of tweets according to different analyzes.

Our main task was to understand how the way people tweet in Spain changes depending on whether they do it in parks or not.

With this objective, we carry out different studies: feelings, emotions, hashtags and topics. They were carried out with MeaningCloud, which provides different APIs to obtain this information. With all this information, the tweets were classified and the differences and similarities of the tweets posted in one place or another were checked.

Specific study of the behavior of the city of Madrid.

Since Madrid contributes a large number of the tweets used to the project, it was proposed to analyze that city more in depth.

The methods have been similar to those used throughout the country, but this time analyzing park by park, finding the specific characteristics of each area.

6.4 Problems faced

The development of this project has made us run into several problems that we list below, along with the solution that has been given to them.

Use of the technologies needed for the project.

The first problem that arose was learning to use the necessary tools to carry out the work. Knowing nothing about the field, the first months were spent getting acquainted with the proposed technologies.

Data acquisition

When collecting the tweets, a number of problems came up.

At the beginning, the method that was going to be used for the collection was through the Twitter API, but it had time limits that barely allowed us to obtain data. That is why we had to switch to the aforementioned tool, Twint, which has no limitations and thus be able to recover tweets from several years ago.

One of the main reasons why the study cannot cover all the information that Twitter provides, is that we retrieve the tweets that are geotagged. This is something optional for the user, so those who do not have this option activated have not been able to provide data to the analysis. Geolocated tweets represent only about 1% of total tweets, but the huge volume of tweets created daily still generates a sizeable dataset of geotagged tweets thanks to the wide adoption of smartphone devices and the popularity of Twitter worldwide [1].

Then there was the problem that, due to the COVID-19 pandemic, tweets in the last year were lower. This has meant that our study has been influenced by this reason, but we believe that going further back and increasing even more the number of parks studied has been able to correct this.

There was also the setback that MeaningCloud's APIs failed for various reasons. The first that was detected was that the free version of this tool did not offer the quality of service necessary for the work and ended up crashing. This was solved thanks to a license with a higher level of privileges provided by the GSI group at ETSIT.

Later, there were more failures caused by the fact that the CSVs collected by Twint did not contain characters such as the ñ or the vowels with accents of the Spanish language, so they had to be substituted appropriately in each dataset.

6.5 Future work

In this section we discuss the possible improvements that could be made to the project in the future.

Increase datasets.

Increasing the number of tweets in datasets always helps to improve results, as there is more data to analyze. With the tool used for the retrieval of tweets can be collected everything all the way back in time we want.

City level analysis.

Analyze city by city to see in which ones people tweet more positively. For this, it would be necessary to capture tweets from many parks, especially in cities with less population, in order to have an acceptable number of tweets to be studied.

Analysis of the photographs of the tweets.

Study the photographs that users attach to their tweets, thus analyzing what it is that people photograph the most when they are in a park. This could be achieved with an image classifier, which detects what appears in the photography.

Event tracking

Relate peaks of positive tweets in certain places with current events or situations at that time. For example, the victory of a soccer team in a city can generate a pike of positive tweets, or a recent natural disaster can generate a pike of negative tweets.

It can also be given a future approach, to predict when these peaks will occur if we have the knowledge that an event of these characteristics is going to occur.

APPENDIX A

Impact of this project

This appendix reflects, quantitatively or qualitatively, on the possible social, economic and environmental impact, as well as the ethical implications.

A.1 Social impact

This section will discuss the social impact that this project could have.

This project is based on something that today's society uses daily: social networks. The study can help people to know where they can go if they want to improve their mood. Although sometimes it may be obvious that one place can provoke a better mood than another, many times we are not aware of it. This study highlights where people are most positive and that can help all kinds of people.

This type of study can help companies in sectors such as advertising or merchandising to know what to offer in each place. Some examples can be advertising on the social networks themselves or physically in the parks, or also where establishing businesses.

A.2 Economic impact

This section deals with the possible economic impacts that may arise from this work.

Studies of this nature could help companies in the Big Data world that, in a similar way to us, collect information so that those who hire them know what, where and when to offer their services, products, etc. The use of social media mining can reduce time and costs to these studies.

A.3 Environmental impact

This section describes the possible environmental impact that the systems used in this project can cause.

Computers and other devices always leave their mark on nature. This is known as the ecological footprint, which is measured on the ground according to its ability to absorb the CO_2 produced by the industrial process. All systems used for machine learning or big data techniques consume large amounts of energy for data collection and model training.

In particular for this project, a laptop and a cluster have been used for all data collection and analysis, as well as the servers of the various tools that have been used.

A.4 Ethical implications

In this section we will evaluate the ethical problems that our project could involve.

The first ethical discussion that we face is due to the use of Twitter, since we use tweets of users together with their user names. According to Twitter's privacy policy, Twitter is public and tweets can be viewed and searched by anyone around the world [8]. We only use public tweets, so their owners have not put any privacy restrictions on them. Therefore, it is public data and we can use it.

Another ethical problem could come from the possible destruction of jobs that would imply the implementation of these techniques in companies. Although those jobs could evolve to others related to this more technological field.

APPENDIX B

Economic budget

This appendix details an adequate economic budget to bring about the project. We will expose the physical and human resources, the licenses, and the taxes.

B.1 Physical resources

This section details the estimated budget that has been needed for the project in relation to hardware.

The entire budget included in this section is derived from the personal computer on which the entire project has been carried out. Although any equipment with a minimum of processing capacity could have been enough, we detail the characteristics of the one used for this project.

- **CPU:** Intel Core i7-8550U 1.80GHz x4
- **RAM:** 16 GB
- **DISK:** 240 GB SSD

The approximate cost of this laptop is 900€.

B.2 Human resources

This section will be similar to the previous one, this time dealing with budgets for people.

To estimate this budget, we assume that everything is carried out by a single person. The time spent must be the equivalent of 12 ECTs, so multiplying it by the estimated hours to which one of them is equivalent, 30 hours, we obtain a total of 360 hours.

We consider a part-time shift of 4 hours and that in a month there are an average of 21 working days. We also estimate a salary of about 450€/month for a scholarship in the GSI group. Therefore, we calculate a total salary of about 2000€.

B.3 Licenses

All the costs for licenses used for this project will be covered here.

Everything used to carry out this research was open source. The only license that had to be paid was for MeaningCloud, but it was provided free of charge to the GSI group for educational purposes.

Therefore, we conclude that there are no expenses derived from this section.

B.4 Taxes

In the case of marketing something related to this project, different rates and taxes derived from the operation will have to be taken into account depending on the country.

B.5 Conclusion

The final economic budget for the project amounts to 2900€ and the duration is 360 hours.

Bibliography

[1] Parques nacionales: límites y zonas periféricas de protección. https://www.miteco.gob.es/es/red-parques-nacionales/sig/parques-nacionales.aspx.

[2] Richard Plunz, Yijia Zhou, Maria Carrasco Vintimilla, Kathleen McKeown, Tao Yu, Laura Uguccioni, and Maria-Paola Sutto. Twitter sentiment in new york city parks as measure of well-being. *Landscape and Urban Planning*, 189:235–246, 05 2019.

[3] Twint. https://github.com/twintproject/twint.

[4] Google maps. https://www.google.es/maps.

[5] MeaningCloud, Sentiment Analysis API. https://www.meaningcloud.com/products/sentiment-analysis.

[6] RapidMiner Studio. https://rapidminer.com/products/studio/.

[7] Pandas. https://pandas.pydata.org/.

[8] Twitter privacy policy. https://twitter.com/privacy.