Defining the Spanish Innovation Ecosystem, Data-Driven Analysis from 2018-2021

Amat i García, Xavier

Year 2020-2021

Director: Manuel Portela

COMPUTER SCIENCE ENGINEERING

upf. Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Bachelor's Degree Final Project

Acknowledgements

Foremost, I would like to express my gratitude to my primary supervisor Manuel Portela, who accepted the challenge of guiding me throughout this project and provided great advice.

My thanks also go to Ana Freire, David Solans and Miguel Angel Cordobes for their insightful feedback.

I would like to thank Universitat Pompeu Fabra community for their education and accompaniment during the 4 years of my degree. I really appreciate their effort to educate, especially the last two years with remote classes. Likewise, I want to thank the University of Texas at Austin for their learning opportunities during my exchange program. It has been very useful for this thesis.

I wish to extent my thanks to the two companies where I worked as an intern, Cisco and Tokio Marine HCC. The experience there, provided me valuable data and knowledge which I used in this research.

The completion of this thesis and also my bachelor's degree could not have been possible without the deep support from my family and friends.

Abstract

Ecosystems are needed to foster regional innovation. The purpose of this research is to define the composition of the Spanish innovation ecosystem and the relationships between its actors. Taking previous studies as conceptual framework, we define three layers to analyse innovation: Society, Institutions and Investments, Regions. We emphasize the importance of studying social relations in a Society layer . For that, we make a data-driven analysis, using social media data combined with open data. We implement a strategy based on machine learning and spatial analysis that provides updated information on the elements of the ecosystem. The results concluded that this set of actors are attached to the geography and their interactions are indispensable to boost the innovation ecosystem.

Keywords: innovation ecosystem, geographical innovation, data-driven analysis, Science, Technology

Resumen

Los ecosistemas son necesarios para fomentar la innovación regional. El objetivo de esta investigación es definir la composición del ecosistema de innovación español y las relaciones entre sus actores. Tomando como marco conceptual estudios previos, definimos tres capas para analizar la innovación: Sociedad, Instituciones e Inversiones, Regiones. Enfatizamos la importancia de estudiar las relaciones sociales en la capa de Sociedad y realizamos un análisis basado en datos utilizando información de redes sociales complementada con datos abiertos. Implementamos una estrategia basada en el aprendizaje automático y el análisis espacial para brindar información actualizada sobre los elementos del ecosistema. Los resultados concluyen que este conjunto de actores está vinculado a la geografía y sus interacciones son indispensables para impulsar el ecosistema de innovación.

Palabras clave: Ecosistema de innovación, innovación geográfica, análisis basado en datos, ciencia, tecnología

Resum

Els ecosistemes són necessaris per fomentar la innovació regional. L'objectiu d'aquesta investigació és definir la composició de l'ecosistema espanyol d'innovació i les relacions entre els seus actors. Prenent estudis previs com a marc conceptual, definim tres capes per analitzar la innovació: societat, institucions i inversions, regions. Destaquem la importància d'estudiar les relacions socials en la capa de societat i fem una anàlisi basada en dades utilitzant com a font d'informació les xarxes socials combinades amb dades obertes. Implementem una estratègia basada en l'aprenentatge automàtic i l'anàlisi espacial que proporciona informació actualitzada sobre els elements de l'ecosistema. Els resultats conclouen que aquest conjunt d'actors estan units a la geografia i les seves interaccions són indispensables per impulsar l'ecosistema d'innovació.

Paraules clau: Ecosistema d'innovació, innovació geogràfica, anàlisis basada en dades, ciència, tecnologia

TABLE OF CONTENTS

List of Figures

List of Tables

1. INTRODUCTION

The word innovation is present in the news, in board meetings, in politician's speech, or even with the family or friends. However, sometimes it is not very clear what innovation really is and what is the ecosystem that powers it. After one of the biggest global crisis (COVID-19) in recent years, we have heard innovation will be key for the economic recovery. It will certainly be difficult to achieve great innovation performance without understanding the ecosystem that lies underneath. The innovation ecosystems are defined by Ove Granstrand as:

"*the evolving set of actors, activities artifacts, and the institutions and relations, including complementary and substitute relations, that are important for the innovative performance of an actor or a population of actors."* (Granstrand & Holgersson, 2020, p. 2)

In this research, we define the actors and institutions and, how they relate to one another. A good understanding of all these elements can drive better innovation initiatives and enhance the relations between those actors and institutions. Andreea Maria Pece says that improving the ecosystem will drive value, progress and growth of the regions. (Pece, Simona, & Salisteanu, 2015, p. 466)

According to the definition of innovation ecosystem and for the scope of the research we consider and explore three layers: Society, Institutions and Investments, and Regions. We emphasize the novelty in the approach to analyze the social phenomenon. We do not cover the artifacts, which are the innovations *per se*, including product or service innovations.

Taking the definition of innovation ecosystem, in our analysis we follow a data-driven strategy. Specifically, we introduce the use of social media as a data source in the social relations study, as well as a value added when combining it with other external information. Open data public and private platforms offer a great amount of valuable data that can be used to generate insights and fuel innovation. (Obama, 2009, p. 1)

The project covers data from 2018 to 2021. There are two reasons behind this choice. First, census data is not regularly updated. Second, the global pandemic froze the industry in 2020 and although we have seen many innovation initiatives, we could get biased results only focusing on that year.

Globalization is a worldwide phenomenon, but innovation still differs significantly from region to region (Gössling & Rutten, 2007; Schumpeter, 1934). By integrating the foreign knowledge and leveraging the local innovation, a proper ecosystem is created. This research studies the innovation ecosystem from a geographic perspective, focusing on Spain and taking as regional units its autonomous communities.

The research findings indicate that the innovation ecosystem is composed by:
A set of actors, which can be classified depending on their role as policy makers, innovation developers, or innovation promoters. A set of interactions between these actors, which can include different communication types: events, product developments and research among others. Investments that incentivize the appearance of innovation actors and interactions.

All these elements are highly influenced by the regional and urban characteristics. Finally, it is important to comment that the innovation ecosystem cannot achieve great performance if any of the elements are missing.

1.1 Personal Motivation

During my 4 years studying Computer Science, I have attended many data science and innovation courses. In this period of time, I have discovered my passion for combining both, technology and innovation. I had the great opportunity to participate in an exchange program with the University of Texas at Austin, a top university, where I studied and experimented in first-hand innovation. I am lucky to currently work in a global innovation department of one of the greatest technology companies in the world, Cisco. Moreover, I usually participate in initiatives that promote the innovation ecosystem such as worldwide congresses, or volunteer activities related to innovation and technology education.

I always thought that our country could be one of the leaders of tomorrow's digital world, but I believe we should start by defining our innovation ecosystem and what can we do to improve it and drive better innovation policies.

1.2 Structure of the Thesis

In the first part of this research, we look at the state of the art by exploring different innovation definitions and try to find a common understanding. We resume from literature how innovation is created and why it is important for regional development. We present the benefits of taking a data-driven approach in innovation ecosystems. We analyze innovation in regions and the specific case of Spain. Finally, we define the goals of the research.

Next, we define the methodology used in the research. We followed the process of a data science project, so we describe how this research followed that process.

The core of the research is the exploratory analysis. There are several analyses performed, that use natural language processing, networks and geospatial analysis techniques. Then, we look at the results and visualizations. Later, we explain how they help define the innovation ecosystem.

Finally, we state the conclusions and limitations, how we contributed to the state of the art and possible future developments.

2. STATE OF THE ART

Innovation has been a controversial research field since 1934, when Joseph Schumpeter defined it. There are studies on how to create that innovation ecosystem and the factors that influence it, but there are still some gaps to be filled. The most recent studies related to our research are described here and serve as a starting point.

With respect to the use of data for improving innovation performance, a recent study (Loukis, 2016) is an example that uses social media to exploit the extensive knowledge of citizens for the development of innovations in public policies and services. In our study, we take a different point of view and instead of using Social Media Monitoring for promoting the development of new innovation policies, we used it to understand the innovation actors and their relations. Moreover, innovation process consists of three main stages, external knowledge acquisition, assimilation and exploitation. They are focus only on the first one, in our study we cover acquisition and assimilation of knowledge (Roberts, Falluch, Dinger, & Grover, 2012).

Regarding the relationship between innovation ecosystems and the geography, there are four interesting studies centered in Spain. The first one, makes a descriptive and empirical analysis of innovation from a geographical point of view of Barcelona (Pérez, 2007). The second one, builds a regional index for innovation taking into account the Spanish autonomous communities. (Martínez, 2003). The third one, makes a general analysis of the innovation in Spain and compares it with other European regions. (Claromonte, 2019). The last, is a study to evaluate at a metropolitan level the impact of the innovative initiatives in the territory. They combined different data sources to validate the hypothesis that innovation has its own urban form that depends on the characteristics of the site. (300.000 Km /s, 2017).

With this literature review, we can define the theoretical framework for our research. Unlike other studies, this research makes a data driven analysis that uses social media data combined with open data, taking a geographical point of view centered in Spain.

2.1 Defining innovation

Today innovation is used as a buzzword. In addition, there are different definitions of innovation across regions and industries. Many companies and institutions have renamed historical Research & Development departments with innovation departments and incorporated innovation in their mission and vision statements. In order to understand what innovation really is, we will explore the different interpretations and define a common understanding for this research.

Innovation comes from the Latin word *innovationem* and was first used in the 16th Century with the meaning of new idea, device or method. It began to take the root during the Industrial Revolution, associated with science and technology. In 1939 Schumpeter defined innovation as craft inventions into constructive changes differentiating it from invention. (Schumpeter J. , 1939). Today the word although having many different definitions they all have two things in common: the idea of value and nascency (S.P.Taylor, 2017). The Oslo Manual, international reference guide for collecting and using data on innovation, defines it as:

"A new or improved product or process (or combination thereof) that differs significantly from the unit's previous products or processes and that has been made available to potential users (product) or brought into use by the unit (process)" (Eurostat; OECD, 2018, p. 10)

Innovation is creating or capturing a new value by creating new products/methods/ideas or combining existing ones in a nascent way. Unless applied, scaled and capturing the value, innovation is meaningless.

There are different types of innovation. World-renowned author, Safi R. Bahcall in his latest book "Loonshots", encapsulates these types in two groups: Type P and Type S. Type P are the product or service innovations and Type S are strategy innovations (a new way of doing business, organizational innovations, new marketing methods, …). (Bahcall, 2019, pp. 65-94)

2.2 Innovation as a gear of progress

Historically, innovation has contributed to progress and in the recent years we have seen how it has become the engine of successful societies. Just in the last century we have seen innovations (i.e. antibiotics, telephone, or computers) that have created an enormous value for us and generated a bunch of new opportunities. It is important to mention that while innovating, we have to take into account the social and environmental impact and ensure a responsible and sustainable innovation ecosystem.

The pace of innovation in certain domains is growing exponentially similarly to the famous Moore law. A great example of how quickly innovation is being developed is the recent worldwide pandemic. We have transitioned entire workforces to remote, we have created vaccines in less than a year and there have been many innovations to prevent the spread of the virus. It has been a crucial element to deal with the pandemic.

Creating economic value by introducing new products to the market, redesigning production processes, or reconfiguring organizational practices is critical to competitive advantage and growth for firms, industries and countries. (Feldman, 2004). Not only that, but there are also many other positive innovation outcomes such as increasing the well-being, reducing sickness, poverty and hunger, environmental sustainability or communication accessibility. Although there have been some undesired consequences in innovations such as pollution or digital gap, we cannot deny the progress of our society thanks to innovation and technological advancements.

We need to define how to best organize the resources to create, diffuse and sustain the innovation ecosystem that will help increase the standards of living and wealth.

2.3 Factors that influence innovation

There are some factors that promote innovation, and it is important to define and revise them to improve the current system. Even if you create and promote innovation projects, if you do not evaluate them, you will never know how to drive better projects or better systems in the future. This is why having indicators and elements to evaluate is helpful.

These innovation systems are constantly changing and every day we are seeing emerging patterns of how innovation is built and redefined. (Roberts, Falluch, Dinger, & Grover, 2012). It could be the case that a new or completely unrelated element arises and influences the innovation ecosystem in the future.

From an external environment, Oslo Manual defines 5 factors that influence innovation:

- **Spatial and locational factors** such as proximity to product and labor markets, taxation, infrastructure or other factors that vary by location.
- **Markets** are contextual factors including characteristics of the suppliers, structure of demand, competition and digital platforms.
- **Knowledge flows and networks** are one of the most significant elements, because they power knowledge transfer which fuels innovation.
- **Public policy** meaning the regulatory framework as well as the policies to support innovation and the public infrastructure.
- **Society and the natural environment** are the public acceptance, interactions, and the environment's phenomena such as weather.

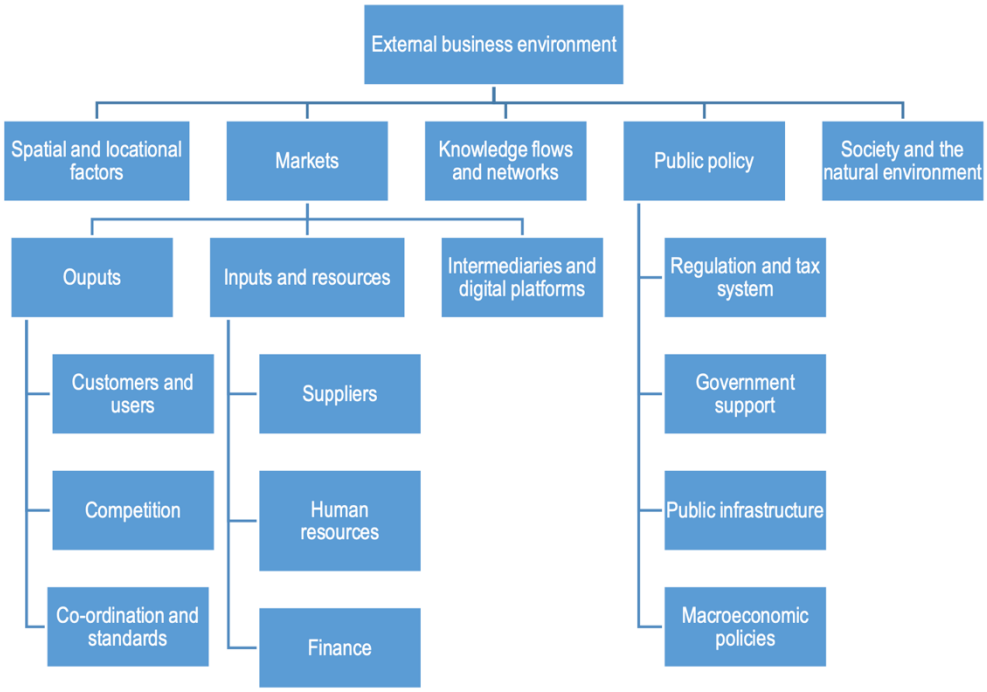

Figure 1: External business environment innovation factors by OCDE & EUROSTAT

In order to measure all these factors, we have to gather information and build indicators. The Spanish Foundation for Science and Technology, every year writes a report with the Spanish innovation indicators. Some of these indicators include companies with innovation expenditure, external commerce or people employed in innovation. (ICONO, 2020)

If we combine all these indicators, we can create indices of innovation performance. There are different regional indices that allow us to see which are the countries with better innovation performances, which mean that probably they have better innovation ecosystems. The most well-known are the European Innovation Scoreboard and the Global Innovation Index.

2.4 Innovation in regions

In 2001, Heijs, J. and Buesa defined regional innovation system as the set of governmental and business institutions that interact between them inside the same geographical framework, with the goal to assign necessary resources to enable all the system to operate together and optimally. (Heijs & Buesa, 2015, pp. 70-74)

There is a relationship between regional specialization (innovation clusters) and innovation performance. The organization of regional innovation systems, also known as clusters, has been associated with greater innovation and growth (Pose, 2020). Regions in Europe without clusters are performing worse than the others such as London, Paris, or Barcelona. Those cities embody an organizational climate enabling and catalyzing innovation and its systems (Concilio, Li, Rausell, & Tosoni, 2018)

Currently, innovation is not just located in developed countries such as the United States (US), there are many other countries investing in it. Actually, Asia is trying to overtake the US position as innovation leader. From its side, Europe is doing its effort to achieve a competitive advantage with the ‘EU New Generation’ funds and the ‘New Green Deal’. These funds are intended to drive innovation opportunities and create nurturing innovation ecosystems in European regions.

a) The Spanish case

In this research we will focus on the country of Spain. A country, which is not specially known for its innovation performance, that has its innovation indicators frozen since 2007. If we look at the before mentioned European Innovation Scoreboard 2020, Spain was at the 14th position of 27 European countries, besides the fact that Spain is the 6th European country with the biggest population and GDP (Gross Domestic Product). We do have talent and research, but we have a big problem in knowledge transfer from research to business. With the proper innovation policies and analysis of the ecosystem, Spain could one day have its spot as leader in the innovation and tech world. (Ferràs, 2021)

The Spanish System for Science and Innovation has 3 main governance instruments and committees: The Council on Scientific and Technological Policy and Innovation; The Advisory Council for Science, Technology and Innovation; The Information System on Science, Technology and Innovation. They assume the responsibility to foster the innovation ecosystem. They incentivize innovation via two national organisms National Research Agency (ACI) and CDTI (Center for Industrial Technological Development).

The Spanish government has produced a "recovery, transformation and resilience plan" that will leverage European funds to modernize Spanish economy, recover the economic growth and employment creation. This plan will probably help improve the innovation performance and ecosystem of the region.

2.5 Data-Driven Innovation Analysis

Technopedia, the digital dictionary for technology terms, define data driven as:

"A process or activity that is spurred on by data, as opposed to being driven by mere intuition or personal experience." (Technopedia, 2021)

Data presents an invaluable opportunity to explore the innovation ecosystem (Wu, 2019). We can have a better definition of the ecosystem by relying on data instead of just intuitions. There are many analytics technologies, especially powered by recent advances in machine learning and the big amount of data that we generate. The systematic data analysis of the situation considers different actors and explores linkages between inter-related elements of the

ecosystem to identify its critical factors. (Talmar, Walrave, S.Podoynitsyna, Holmström, & L.Romme, 2018)

Digital technologies are now indispensable when doing geographic analysis, generating, processing, storing, analyzing and sharing data (the majority of which are born digital); we can create maps, documents, videos and sharing information from all that data. With all those results we can construct the innovation analysis. (Ash, Kitchin, & Leszczynski, 2015)

We used social media as a wisdom of a crowd (collective opinion of a group of individuals) data source. Thanks to the exponential growth of use of the Internet and especially social media content, there are many opportunities to exploit that public knowledge. There have been some previous attempts to use that data for innovation purposes, but mainly in the private sector, to collect customers' opinions for new product development. We want to go further and gather data of the public and private to define the ecosystem. If we want to extract useful information, we have to be sure that the data we are collecting is a fair representation of the society. Social Media Monitoring (SMM) has many advantages such as diverse opinions, real time data, identification of trends and influencers or sentiment analysis. There are also some risks that we should consider, mainly privacy, fair representation, non-biased or manipulated opinions. Knowing those risks and trying to avoid or reduce them is crucial in any similar research. (Euripidis Loukis et al., 2016).

3. OBJECTIVES

Main objective of the research: Analyze the composition of the Spanish innovation ecosystem and the relationships between its actors, using a data-driven approach.

Specific objectives:

a) Society: Understand how innovation leaders and wisdom of the crowd interact within the innovation ecosystem and their influence on it.
b) Institutions and Investments: Understand the role of institutions and their investments in creating and promoting innovation ecosystems.
c) Regions: Understand the regional factors that are related to innovation and the connection between them and the innovation ecosystem.

3.1 Society – Which are the innovation leaders and how do they interact within the innovation ecosystem?

We want to see how innovation leaders build, interact and influence the innovation ecosystem.

Other interesting facts to explore are their role in innovation processes and the importance of talented individuals in the region.

It is important to analyze people as they are the ones that build the ecosystem. In innovation, ideas for their own are meaningless, you need people to shape them and convert them into real impact and value. Understanding how these actors interact between one another inside a geographical framework, it is key to define the innovation ecosystem.

3.2 Institutions & Investments – How institutions and their investment impact the innovation ecosystem?

We want to understand the role of institutions and their investments in innovation and in creating and incentivizing ecosystems.

We cannot analyze innovation without looking at the organizations and investments. Investments to promote innovation are indispensable for the ecosystem. The majority of institutions are now struggling to overcome the challenges of COVID-19, so they need investment incentives for research and innovation.

We can distinguish between two types of institutions: public and private. Public institutions are often the ones that give incentives to innovation, and private institutions are the ones that execute innovation. However, this does not mean that there is no innovation in public institutions. Indeed, universities and research centers are public institutions and are executors of innovation.

Institutions should move away from the lineal innovation model and innovate while interacting with the ecosystem (region, state, community). That context gives a mesh of relationships, that if understood correctly can help thrive and stimulate innovation. (Dagnino, 2001).

According to the recent study of investment in Spanish innovation, although Barcelona and Madrid still are the top hubs of innovation, there has been a boom in other cities, increasing more than 151.5% (Fundación Bankinter, 2020). This shows us how investments are related to the ecosystem, and we want to further explore that relationship in this research.

3.3 Regions – What is the relationship between geographies and the innovation ecosystem?

We want to see how geographic and urban indicators are related to innovation ecosystems and what region characteristics make that ecosystem perform better (innovation clusters).

Innovation is a variable that is discussed as a key factor influencing regional performance and growth (Bottazzi & Peri, 2003). Innovation is different for different regions. This is because the innovation ecosystem changes depending on the regional factors such as density or talent. In the past, economic geographers put their focus on the traditional regional characteristics like infrastructure and number of firms in a region. This approach changed with the importance of inter-organizational relations and networks and the role of social factors. (Porter, 1998). More recently, innovation ecosystems have been affected by the dual innovation phenomena of increasing international linkage and the persisting importance of geographic proximity (OECD, 2013).

Regions were designed as places to overcome time with space, making communications easier. Today, with ICT networks we can overcome space with time by enabling instantaneous transfer of information. (Graham & Marvin, 1996). The spread of knowledge and technology between countries has intensified because of globalization and thanks to internet. The transfer of knowledge has helped boost innovation and has redesigned the interactions of the ecosystem. (International Monetary Fund, 2018)

Regions should work like horizontal networks, where governments, the private sector, the educational institutions, and citizens have an interdisciplinary and holistic vision of the problems and needs of the regions. There are some Spanish regions with organizations that have vertical organisms working in silos that are disconnected, weaking the innovation ecosystem. (Blázquez & Ramón, 2019)

We want to further explore all those factors and their connection with the Spanish innovation ecosystem.

4. METHODOLOGY

In this chapter we introduce the different research strategies at the different phases of the project. We explain the process of the gathering and analyzing data as well as the final dataset that has been used.

4.1 Research Strategy

This research uses Python as the main programming language, it is one of the most widely used in the field of computer science and it has simple syntax but with powerful capabilities. Python has a specialized library (Tweepy) to interact with the Twitter Application Programming Interface that has been used for the Society analysis. It has also libraries for plots such as plotly or seaborn, and it is very easy to deploy Machine Learning (ML) models and perform data analysis.

Moreover, we followed an agile development style when coding the exploratory analysis and visualizations. As studied in many subjects in the degree, agile development provides many advantages in comparison to traditional software development styles such as Waterfall. Some of those advantages are flexibility, quality improvement and predictability.

We divided our research into different phases which correspond to standard data science project.

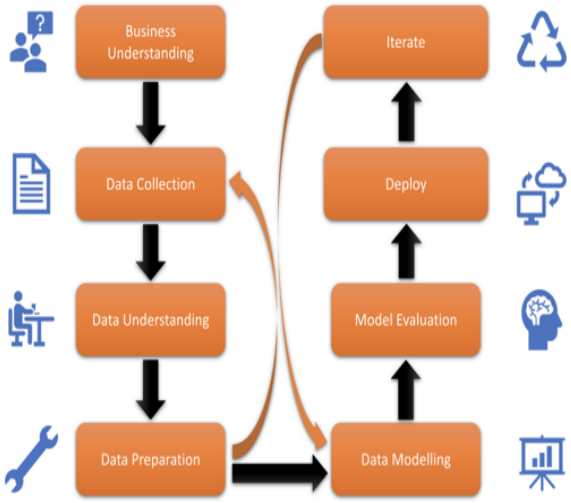

Figure 2: Data Science Project Scheme by Towards Data Science

Following, there is a descriptive analysis of the phases of our research.

Data Preparation:

In the data gathering phase, we collect data related to the research, understand it, clean it and prepare it for data analysis. We use computer science tools such as APIs and scripts to transform raw data into data ready for analysis.

Exploratory Analysis:

We performe different exploratory analysis using data science tools. Those analysis help us achieve the goal of the research. We first explore the data and then go deeper to extract meaningful insights. The analysis starts with the frequency and distribution of the data, continues with the correlations and the structure as a network, content analysis using topic modelling and clustering, spatial analysis combining different data sources and mapping the information, and ends with the innovation classifier.

Visualizations and Results:

Visualizations are the best way to see the result of a data analysis. We use different visualizations according to the different layers; this is because the type of data and analysis in each level is different.

Society:

We visualize the network graph, with the influencers as main nodes and the connections as secondary nodes. We also visualize the text analysis with word clouds and bar plots among others. Furthermore, we combine those visualizations with geographic data, plotting the nodes and interactions on a map.

Institutions and Investments:

We visualize the clusters of companies and research centers in the region. We build a choropleth map with the investments and institutions indicators to see which ones contribute to innovation. Moreover, we combine the information with the Society layer to explore relationships. We have exported that interactive visualization to an html file.

Regions:

We visualize the correlation between the inputs of the classifier to understand the features. Furthermore, we plot the feature importance to see the indicators that are used to classify, which are the ones that help thrive the innovation ecosystem.

Open Innovation:

In order to share the results of this research we have created a GitHub repository with the notebook, the crawler used to gather social media data, the datasets and results. We strongly believe that the results of this research will contribute, in some way, to understanding innovation ecosystems. Making them public will help open innovation for governments, entrepreneurs or institutions. Open innovation as defined by Professor Henry Chesbrough is the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively. (Chesbrough, 2006)

The architecture of the repository is quite simple. We added a README containing the project introduction. You can find the repository here:

4.2Process

In this section we describe the process and methods used in each of the phases and parts of the analysis.

a) Data Gathering

As explained in the introductory part of the project, we have used social media as primary data source for the Society analysis. In order to gather data related to innovation, we have created a crawler (an app that systematically downloads data from a site).

We have chosen Twitter as the main platform to exploit data. The reason is that Twitter offers a very complete and open API, that is often used in social media research projects. It is also one of the social media platforms with more content related to innovation. The other social media platform that we could explore is Linkedin. However, Linkedin does not offer, at the moment, an open API that could fit our purposes. We would have to do scrapping, which is normally seen as an attack to the platform.

Twitter has an API that allows you to download stream data given a query (location, keywords, …) in Python. We have used stream live data because we want to combine current data with existing knowledge. There is a limitation for free accounts which is the API rate, which does not allow you to download more than 50 tweets per 15 min. In future versions, we could buy a business account which would improve that rate.

We have defined the stream filter taking into account three things: keywords, location and language. We have first developed a crawler to choose those keywords. We have built it filtering Spanish people that we knew they are related to innovation such as entrepreneurs, CEOs, journalists or government officials. These Twitter accounts are:

```
Mara Balestrini (@marabales, Innovation Consultant & Faculty IE), Cristina Aranda (@cris_aranda, Technology Consultant Advisor), Xavier Ferràs (@XavierFerras, Innovation Professor ESADE), Pere Condom (@PereCondom, Research Director), Oscar Pierre (@oscarpierremi, Startup Founder), Esteve Almirall (@ealmirall, former CIO and Dtr. Center for Innovation in cities), Elena Gil (@Elena_Gil_Liza, Telefonica Business Operations Director), Hernán Rodríguez (@rodriguezhernan, Innovation & Digital Strategy consultant), Horacio Morell (@HoracioMorellG, Territorial IBM President), David Cierco (@davidcierco, ex-GM Red.es), Nadia Calviño (@NadiaCalvino, VP Economy and Digital Transformation Spanish government), María Luisa Melo (@MLMelo, Director Huawei Spain), Alberto Iglesias (@aiglesiasfraga, Innovation Journalist), Carlos Blanco (@carlosblanco, Entrepreneur & Investor).
```

After deleting stop words, we have looked at which were the words that they use the most and make a list, that was used for the final crawler filter.

```
["innovación", "tecnología", "startup",
"digitalización","transformación digital", "emprend", "digital",
"digit", "innov", "tech", "research", "open innovation", "R&D",
"I+D"]
```

Location is the second filter and there are different approaches to use it. The first one, is using geocoding tweets, this approach has a huge limitation which is that only about 1% of the tweets are geocoded. The second one, is using a bounding box and getting the tweets inside that geographical region. This is an inclusive filter, when combined with keywords it gets tweets either with the keywords or inside the box. Finally, there is the user's location, which is much better because we have many more geolocated tweets but still has its own limitation. People do not set their location or set it in unreal places.

The last approach combined with the bounding box was our choice. After getting the string indicating the location, we added a geocoder to get the coordinates of that location and exclude the ones that are outside Spain. We tried two geocoders, the first one was Google geocoder, but it is a paid service, so we used Folium Python geocoder which is open sourced.

Last filter is the language, and it is necessary because ML algorithms are normally train in one language, mixing could result in worst performance. As a limitation, twitter API does not filter some regional languages, so there could be tweets in Spanish regional languages such as Catalan or Basque.

The sampling period was a month, crawling almost every day from Monday to Sunday. We gather an approximate final amount of 60.000 tweets.

The data gathering process for Institutions and Investments and Regions consisted on doing research to open data platforms such as INE (Instituto Nacional de Estadística) or private reports from Venture Capital associations such as ASCRI (Asociación de Capital, Crecimiento e Inversión) or organizations such as Fundación COTEC. Finally, we have also used a web page crawler that operates with HTML tags, to download information directly from the web.

b) Data Cleaning

We cannot perform Natural Language Processing (NLP) analysis without cleaning first the raw text from the tweets. We decided to do that process by adding an extra column in our tweets dataset which contains the clean terms for each tweet. The cleaning process included lowercase the text; delete all the strange characters such as '@, #' and punctuation; tokenize the sentence, splitting the sentence string into different word strings; delete stop words, most common words such articles or pronouns; perform stemming (Snowball stemmer) to keep the root of the words.

For the other datasets, institutions and regions, we have combined the data from multiple sources into single files, ready to load in the notebook. We have also normalized the data to eliminate redundancy and prepare it for ML techniques.

c) Data Filtering

There are two main filters that we have applied in our data. The first one is in the tweet's dataset, where we got some tweets with the location string only 'Spain', not a specific region in Spain, so we filtered all those tweets. The second filter was in the other two-layer datasets, we deleted Ceuta and Melilla autonomous cities because there was missing data from our sources, so it would not make sense to analyze or compare them.

d) Exploratory Analysis

In the exploratory analysis we have loaded the data and applied different analysis. We have explained in detail all those analysis in the next chapter.

e) Visualizations and Results

After performing the analysis, we have used Python modules to visualize the results. We have looked into all these results in a future chapter.

4.3 Dataset

The final datasets obtained include quantitative data from statistics and the census and qualitative data from Twitter texts. As explained before, we have primary collected some of these data, but we have also used existing studies as a secondary data. We have performed some manipulations in the data to prepare it for analysis. The final datasets obtained are the following ones:

a) Society

Dataset containing different tweet attributes. We have designed this dataset taking into account our research goals and the information needed to achieve them. There are some fields of API response that are deleted to reduce the size of our files.

- Is Retweet: Indicates if tweet is a retweet. 1 means retweet 0 means not retweet.
- RT_UserId: User identification of the original tweet.
- RT_UserName: Username of the original tweet
- Tweet_text: Original tweet text.
- Mentions: Accounts mentioned in the tweet
- Hashtags: Hashtags included in the tweet
- UserLocation: Location of the user profile
- MapsLocation: Geocoded location with coordinates and other geographical information.
- ID: Tweet id.
- Date: Date of the tweet
- Likes: Likes of the tweet. Not original tweet
- URL: Contains the URL of the crawled tweet
- Number_Retweets: Number of retweets
- Terms: Terms of the tweet text after cleaning it.

In our analysis we used two Twitter datasets, one that includes all the retweets and one without the retweets. The reason behind this is that including the retweets could bias some of the results, since it means to duplicate the tweet text and the frequency of the terms.
Other datasets used are: ranking of the 10 cities with the highest number of active users on Twitter in 2020 by Statista.

b) Institutions and Investments

We used different datasets for this part depending on the type of institutions and analysis that we wanted to study.

Private Institutions, including private companies and investments, in each autonomous community.

- Number of companies with innovation expenses, 2019 by INE
- Innovation expenditure in euros in companies, 2019 by INE
- Number of innovative companies, 2017-2019 by INE
- Number of initiatives to assist startups, accelerators, incubators, private programs among others, 2019 by ElReferente
- External Investment in millions of euros, 2019 by Spanish government (Ministry of Industry, Commerce and Tourism)
- Number of Venture Capital investments, 2019 by ASCRI (Asociación Española de Capital, Crecimiento e inversión)
- Percentage of companies that met innovation expectations, 2017-2019 by INE

Public Institutions, including public organizations and investments, in each autonomous community.

- Grants from the state research agency in millions of euros, 2019 by Ministry of Science
- Percentage of GDP used for innovation, 2019 by INE

Other datasets used are: clusters of companies, 2021 by AEI (Agrupaciones Empresariales Innovadoras); clusters of technology centers 2021, by APTE (Asociación de Parques Científicos y Tecnologicos de España)

c) Regions

For the regions dataset we took different geographic characteristics and built indicators for each one.

- Demography
    - Population, 2020 by INE
    - Aging index, 2020 in percentage by INE
    - Foreign population, 2021 by INE
    - Rural Population, 2018 by Spanish government
- Urban
    - Number of houses, 2020 by INE
    - Housing Price for square meter, 2020 by Statista
    - Waste per citizen, 2018 in Kg by INE
    - CO2 Emissions, 2019 in kt by Spanish government
- Health
    - Life expectation 2019
- Politics
    - Election participation in percentage, 2019 by Spanish government
    - Index of transparency, 2021 by Dyntra (Dynamic Transparency Index)
- ICT:
    - Percentage of internet users in the last 3 months, 2020 by Statista
    - Ownership of computers of any type in percentage, 2018 by INE
    - Houses with internet access, 2018 by INE
    - Percentage of houses with mobile phone, 2018 by INE
- Education
    - Achieved level of education (0-2, 3-8, 3-4), 2019 by INE
    - Number of universities, 2019 by Statista

5. EXPLORATORY ANALYSIS

In the following section we describe the different analysis at each layer and the type of insights they provide.

5.1 Society

Frequency analysis is part of descriptive statistics, it consists of counting the number of occurrences of an event. From the information Twitter provides we defined four elements to analyze: mentions, hashtags, terms and emojis. The reason we analyze emojis is because they are an essential communication tool and one of the primary ways of exchanging ideas: "An image is worth a thousand words". We should also bear in mind that they have multiple meanings depending on the context.

We created a function that generates a "bag of words" and given a column name, iterates over all the samples doing two actions: adding new elements to the bag and increasing the frequency of existing elements of the bag. We can convert the result to a Python dictionary, with key 'the element' and value 'the frequency'. Later we sort the results to see the top ones (most relevant words, hashtags, people and emojis). In our research, it is interesting to look at the global frequency and the analysis in each region to compare them.

There are different types of visualization for this analysis that are discussed in the next chapter.

Network analysis [1]is the process of investigating social structures through the use of networks and graph theory. It helps defined the structure of relationships, in our case between users/actors in the social network (Twitter). The users are represented as nodes and the interactions (mentions or retweets) between them as edges. In case we want to consider the direction of those interactions, we use direct graphs otherwise undirect. After defining the network, we describe its attributes such as:

- Degree: number of edges that occur to that node.
- Clustering coefficient: tendency for nodes to cluster together, meaning: users interacting with one particular user they interact between them as well.
- Distance statistic: Average distance between two nodes and diameter (max distance between two nodes)

[1] https://towardsdatascience.com/network-analysis-d734cd7270f8

Two other interesting tasks that we can perform:

- Identifying influencers nodes: with high degree centrality (many connections), closeness centrality (close to other nodes) and betweenness centrality (close and well situated)
- Identifying communities: Subset of users within the graph with dense connections between them. There are different algorithms for that, but probably the most known is the Greedy-Modularity.

As a final step for the network analysis, we have to spatialize our network. In the Visual Analytics subject, we used a software for networks called Gephi and experimented with different algorithms. Forceatlas2 was the one with better results, so we used it for the research.

Correlation analysis [2] is a statistical method to discover if there is a relationship between two variables and how strong is it. It is pretty useful to find patterns, connections and trends between different elements. In our research, we use it with the top hashtags. Note that it could also be interesting to see if there are correlations in mentions for further research.

Co-occurrence analysis [3] is the counting of paired data within a collection unit, for text data we refer to a pair of words in a document. It is a way to understand association between elements and it is really important because it gives us contextual information. It is not the same "innovation failure" than "innovation success". If we only look at the most used word "innovation", we could lose some information. It is helpful to visualize this co-occurrence as a network and understand those associations.

Name Entity Recognition is an information extraction method to locate and classify named entities in text. It can be used to locate names of persons, companies, places, dates or products among others. This can help find actors which are not users but are written in the tweet text. We used well-known library for text processing spaCy and its Spanish model es_core_news_sm. That model is for the Spanish news and has a limited corpus of training, so the results are not as good as with the original English model.

[2] https://blog.flexmr.net/correlation-analysis-definition-exploration
[3] https://www.aabri.com/manuscripts/152265.pdf

Topic modeling is a recurring NLP task, it consists of extracting the topics from a corpus of documents, in our case tweets. There is a peculiarity of our dataset, our documents are really short, and our main topic "innovation" is really specific. That is why the Latent Dirichlet Allocation (LDA) method is a good fit. It is an unsupervised machine-learning model for topic extraction, that also gives as a result the percentage each document talks about each topic. (Revert, 2018)

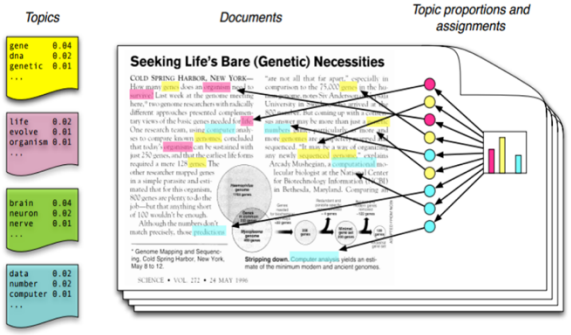

Figure 3: LDA model explanation by Towards Data Science

The model has the number of topics as a parameter, so we tried different values to reduce repeated topics. Gensim Python library has a model for LDA that is really simple to use and comes with an interactive visualization, in Figures 16-29, that helps you explore those topics.

Clustering is an unsupervised learning problem to discover groups of elements based on features they share. In our research, we aim to cluster the different tweets. This can help detect groups of actors by the content and innovation clusters defined before. For that we need a vectorized representation of the tweets. We create a vocabulary model using **Word Embeddings**, specifically Word2Vec, a technique that uses shallow neural networks. With that model, we vectorize all the tweets and get the inputs for our clustering algorithms. We choose two wide used algorithms **K-Means** and **Mini-Batch K-Means**. They work by initially placing k different means (centroids) on the data plane, measuring the distant between each point and the centroids and associating them to the closest (less distance) one. In order to choose that K, there are techniques such as the Elbow method. It runs the algorithm for a range of values and then computes the average score, the inflection on the curve should be the optimal value. Mini-Batch K-Means is an alternative algorithm that reduces the computational cost by only using s subsample of the dataset.

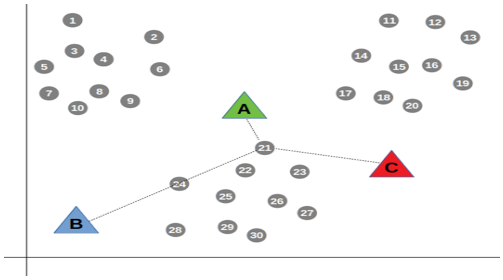

Figure 4: KMeans model scheme by Towards Data Science

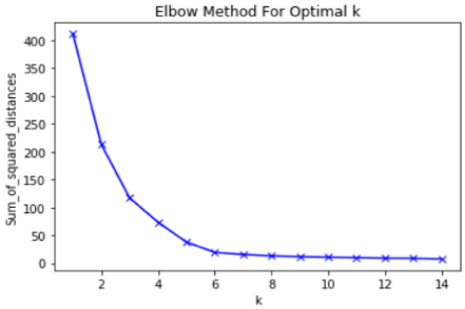

Figure 5: Sum of square distances in the Elbow method

There is already an implementation of those algorithms in Python in the sklearn library. After clustering, we assign the labels to each tweet and see the representative elements in each cluster.

Spatial data analysis is an interdisciplinary field about analytics, visualizations, processes and systems to extract knowledge and business insights from the causal relationship of spatial feature and other data. (Cordobes, 2021). Associating the tweets to a location provides an opportunity to combine all the previous analysis with geographical and location data. It can help us to better understand some of the results by contextualizing information.

In Figure 22, we plot a heatmap to understand the regions where there is more Twitter activity related to innovation. We normalize the data for number of habitants, because it is logical that if a region has more population than another, it will most probably have more users and thus more activity. Moreover, we used an external dataset to check if the top regions talking about innovation are also the regions with the most activity for generic topics.

Finally, in the Figure 23, we overlayed the Twitter network on the geographical map, to understand the relationship from a spatial perspective.

5.2Institutions & Investments

In this layer, we perform descriptive spatial analysis. After building the dataset that combined different investments and institutions information, we create an interactive map that explores the distribution of those investments and innovative institutions across the region.

Folium[4] is a library built on the data wrangling strengths of Python and the mapping strengths of leaflet.js, we use it as a primary tool to build the maps. The visualization that we have done, allows the user to select different attributes related to the topic and see how important they are in different regions.

As a value added, we can see the nodes from the Twitter network interactively clustered with the intensity of their presence in each community. In this way, we can compare intensity from a regional perspective.

We build two other maps that combine Society information and Institutions and investments to see if there is a relationship between them. We recover the twitter heatmap and plot two attributes on top:

- Innovation companies' clusters: association of private companies with innovation activities. Figure 29
- Technology centers: research and innovation public institutions. Figure 30

5.3Regions

Classification is a supervised machine learning problem that learns to assigns labels or categories from previous examples. After the training, we can see which are the weights of the attributes that the algorithm used to classify the samples. Although it may require interpretation of the results, we believe it is an effective and simple way to see which regional factors promote the innovation ecosystem.

The inputs to our algorithm as explained in the Methodology chapter, are different indicators (urban, ICT or education among others) from each of the Spanish regional units.

Our target variable will represent regions with better innovation performance versus regions with worse innovation performance. We build this variable taking three things into account, first the regions with more Twitter activity, second the regions with more investment in innovation and third an innovation regional index. We sort the regions given those attributes. We take the first half for better innovation performance and the second as worse. It is almost impossible to have a precise measurement for innovation, so we experimented with that one.

[4] https://python-visualization.github.io/folium/

Random forest is a classification algorithm that has a large number of individual decision trees which operate as an ensemble, predict a class and the one with the most votes become the prediction. Decision trees are algorithms that determine the class by a set of rules inferred from the data features.

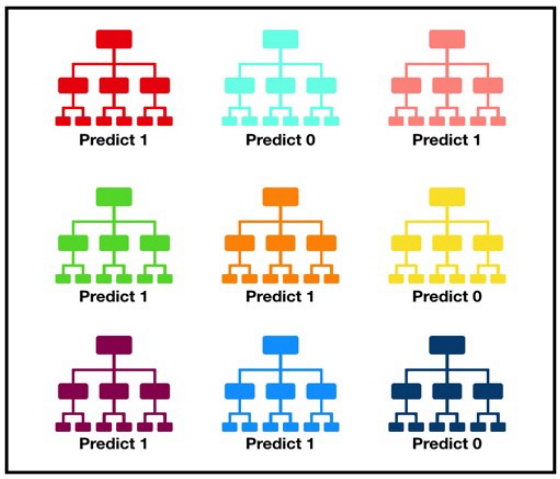

Figure 6: Random forest scheme by Towards Data Science

6. RESULTS AND VISUALIZATIONS

In the following chapter we introduce the results of the research for each layer. We revise how those results help achieve the goals of our research. Some of these results include graphs, maps, tables or plots. In the Annexes section you can find the link with a GitHub containing the complete set of visualizations.

A big part of data science is interpreting the results. Domain knowledge must be incorporated to understand those results. We introduce this domain knowledge from our training, experiences and from external sources such as the internet.

6.1 Society

When defining the most relevant people on our ecosystem network, we do it from two different perspectives: frequency and graph networks. Afterwards, we combine the different results with geo information.

From a frequency point of view, we obtained the following results. We incorporated external knowledge from twitter to check if these users were relevant in the innovation field. We deleted the accounts that were completely unrelated and assigned two colors: orange for low relevancy and green for high relevancy. We can also see the position of the ranking and the type of actor that the user represents.

Username	Mentions	Type of Actor
DES_show	1	Tech Event
PSOE	2	Political party
Yolanda_Diaz_	3	Government representative (Industry)
Sanchezcastejon	4	Government representative
el_pais	6	Generic Newspaper
MADRID	7	Regional Government account
CDTIoficial	10	Innovation Governmental Entity
Telefonica	12	Company
Wwwhatsnew	13	Specific Newspaper
TE_GranEmpresa	15	Company
Vodafone_es	16	Company
Elmundoes	17	Generic Newspaper
AMETIC_es	19	Companies Association
astro_duque	20	Government representative (Innovation)
CienciaGob	21	Government Ministry
AndaluciaJunta	22	Regional Government account
CotecInnova	23	Innovation Foundation
elEconomistaes	25	Specific Newspaper
Expansioncom	27	Specific Newspaper
AngelNinoQ	28	Regional Government Innovation Representative
EconomiaAnd	29	Regional Government Innovation section

Table 1: Most mentioned users in the innovation tweets *dataset*. Green: strong relationship with innovation, orange: weak relationship.

The majority of those users are big entities like government institutions and its representatives, private companies, or newspapers. The first two are the ones that make decisions related to innovation, and the last are the ones that share and spread the information. We also see some foundations or associations that have the role to promote the innovation in the ecosystem. Finally, we see some political parties that given the current context (crisis recovery) believe innovation will make an impact to the citizens. As a future study, it would be interesting focusing on the politicians to check how they define innovation depending on the political parties they belong.

Username	Likes	Retweets	Type of Actors
MarcosSorribas	2	2	Entrepreneur
IdiazAyuso	4	5	Regional Government representative
alejandroTGN	6	8	Regional Government representative
emiliomarquez	7	12	Business Angel
Hugo_saez	15	9	Digital journalist
FLMIRONES	16	26	University Professor
JaSantaolalla	17		Author and Researcher
DFS_MWC	23		Techno Program
carlosferrersot	24	16	Sales Director
ApuntesCiencia	29	25	Science Author
NavisCode	30		Developer
PartidoPACMA		21	Political party
gorka_orive		23	University Professor
PostigoElena		24	University Professor
cuquemar		27	Founder

Table 2: Users with highest mean of retweets and likes in the innovation tweets *dataset.* Green: strong relationship with innovation, orange: weak relationship.

Curiously, when we look at the most retweeted and liked users (taking into account the mean values), the actors we find are individuals related to the innovation. The majority of them are people that share their vision of the innovation ecosystem or build it like founders, developers, entrepreneurs. We also found university professors, who normally do research which is essential for innovation. They also transmit knowledge to other students, who will later contribute on the system.

Non-Textual information is very important specially in social networks. We analyzed which were the most relevant emojis in the communications of the ecosystem. Here is the list of the different emojis:

['👉', '👇', '➡️', '✅', '💻', '👏', '📣', '🚀', '📲', '📌', '🔴', '⬇️', '🗣', '💡', '▶️']

We see hand gestures as indicators, computer and phone representing technology, speaker and person shouting representing announcements, the pin representing events, the bulb historically associated with ideas and the famous rocket that is often associated with the launch of new projects. As we can observe, there are different types of communications: indications, events, celebrations. As a further work, we could filter the tweets depending on the type of communication.

Network analysis provides additional insight to that initial frequency analysis that can help validate the results. We created a network graph for the retweets and exported the file to visualize and analyze it in Gephi. When analyzing big networks, it is common to apply some filters in the topology such as degree of the node. In this way, we reduce the network to the important nodes and reduce the computational cost. The resulting network has the following characteristics:

Average Degree (number of connections): 1.6

Average Path Length (path between two nodes): 4.5

Network Diameter (shortest distance between two most distance nodes): 17

Communities (groups of nodes): 17 with resolution of 3 (more resolution less communities)

Number of nodes: 2493, **Number of Edges**: 3988

We took the nodes with higher degree as the most representative ones. Note that we could also explore betweenness centrality or closeness centrality as a measure of the influence in the network. We found that most of the nodes appeared in one of the previous rankings, because the network is built using the same information. We also found new interesting actors of the ecosystem that are shown in the following table. There is a new type of actor which is startup associations or innovation hubs, they normally incubate the innovation projects of the system.

Username	Degree	Type of Actor
AlbertoEMachado	2	Tech Influencer
begonavillacis	6	Regional government representative
CsMadridCiudad	7	Political party
desdelamoncloa	9	Government account
mincoturgob	10	Government Ministry (Industry)
virginiog	13	Founder
agenciaiisi	14	Regional Government Innovation Entity
SilviaSaavedraI	16	Regional government representative
carmeartigas	20	Government representative (Digitalization)
Startup_VLC	22	Regional Startup ecosystem
franciscopolo	23	Government representative (Entrepreneurship)
gpscongreso	24	Political party
ionebelarra	26	Government representative (Social Rights)
aszapla	27	Board Member Innovation
madridinnova	29	Regional Government Innovation Hub

Table 3: Most relevant people of the network graph generated with the innovation tweets *dataset. Green: strong relationship with innovation, orange: weak relationship.*

We colored the network with the communities identified and visualized the results. These communities represent groups of actors in the innovation ecosystem. Although they are not completely dense and separated, we identified five main communities. Note that there still exists noise in those communities, and you should have domain knowledge to associate them to the different groups.

Community 13: Containing nodes related to the Spanish government and innovation governmental institutions. An example: Moncloa account or the Ministry of Industry.

Community 6: Containing nodes related to the innovation labor force. An example: CCOO.

Community 9: Containing nodes related to the latest most important innovation event: DES_2021.

Community 4: Containing nodes related to each region innovation and their ecosystem of partners. An example: madridinnova or bcntechcity.

Community 3: Containing nodes related to people that share their innovation vision. An example: Esteve Almirall or Pere Condom.

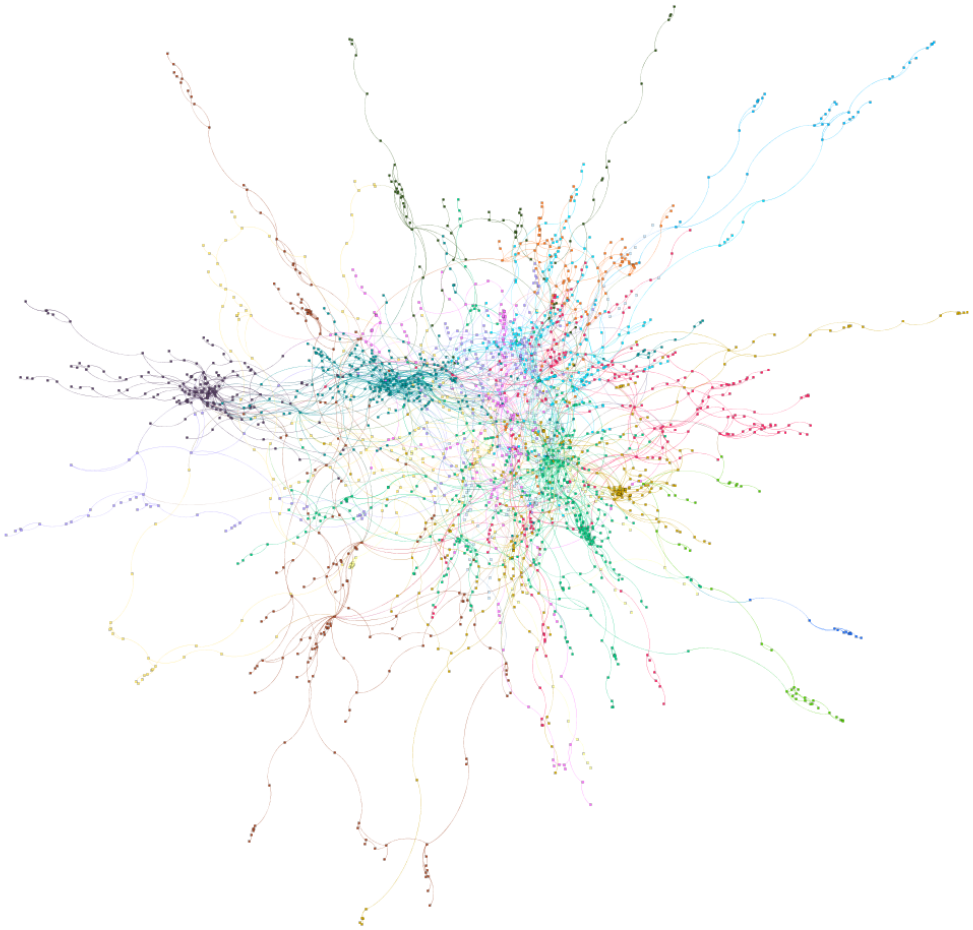

Figure 7: Network graph with identified communities, generated from the innovation tweets dataset.

We also looked at the content of those interactions. We took 3 different approaches: frequency, clustering and topic modelling, entity recognition.

From frequency perspective, looking at the word clouds, we see, as expected, some of the words that we used for the Twitter API search, but we also see other words that did not appear on the search and are related to the innovation ecosystem.

We find words related to technologies such as 5G, artificial intelligence, robotics, big data, blockchain; type of companies such as PYMES (Small and Medium Enterprises), startups, private companies; incentives such as Horizon2020 or the recovery plan; new ways of working such as remote work; events such as the Digital Enterprise Show 2021; the beforementioned COVID-19; finally, we also find sustainability-related words because many companies and regional governments are talking about and associating it to innovation (we need sustainable innovation).

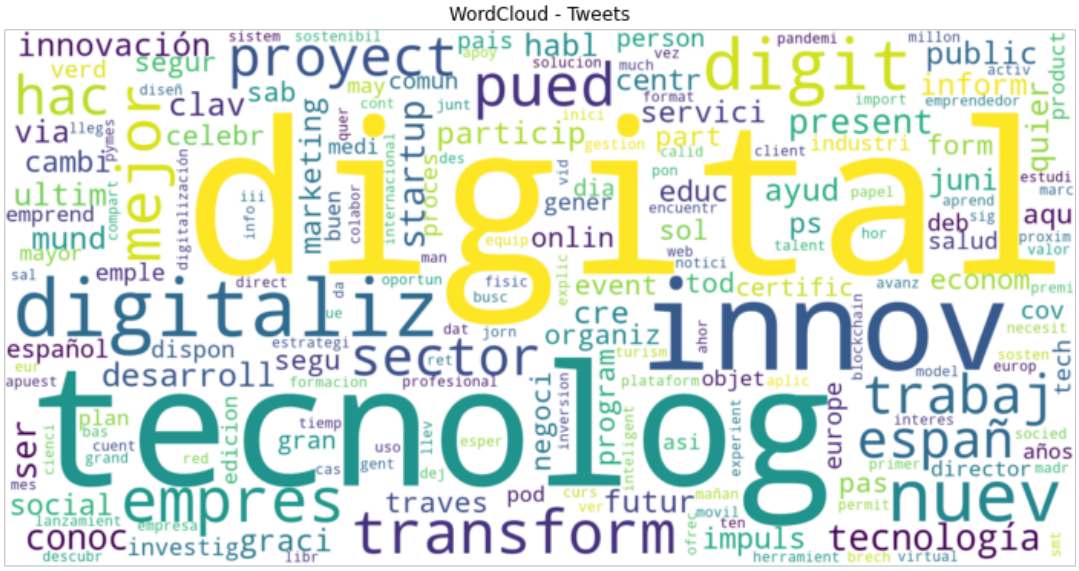

Figure 8: Terms word cloud generated with the innovation tweets dataset

Figure 9: Hashtags word cloud generated with the innovation tweets dataset

We checked if the distribution of words changed depending on the regional location and visualized the results using a mask of the region contour that we edited on Photoshop. Here are two examples of regions and the rest are added on the annexes.

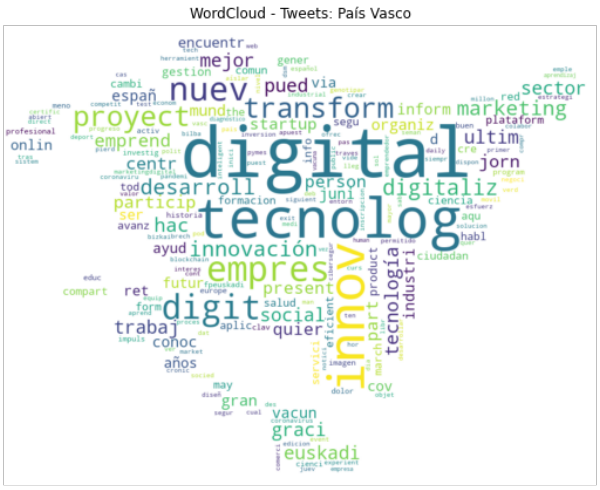

Figure 10: Basque Country word cloud generated from the Twitter innovation dataset

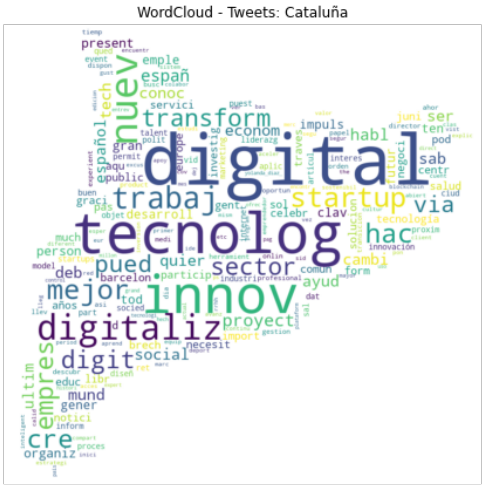

Figure 11: Catalonia word cloud generated with the Twitter innovation dataset

Both countries share a lot of words as expected, but if we look at them, we find two main differences: places which represent the local places that they associate to innovation; local innovation programs. There are also some autonomous communities that have a specific highlighted main topic, but we would need more data to determine if it is a transient phenomenon or if it is a characteristic of the region.

Moreover, we analyzed the correlation between the popular hashtags. This information tells us if there is a relationship between different interactions from the actors of the system. Innovation is strongly correlated with Startup, Technology, Digitalization and negatively correlated with entrepreneurship and sustainability, as we would not deduce it intuitively.

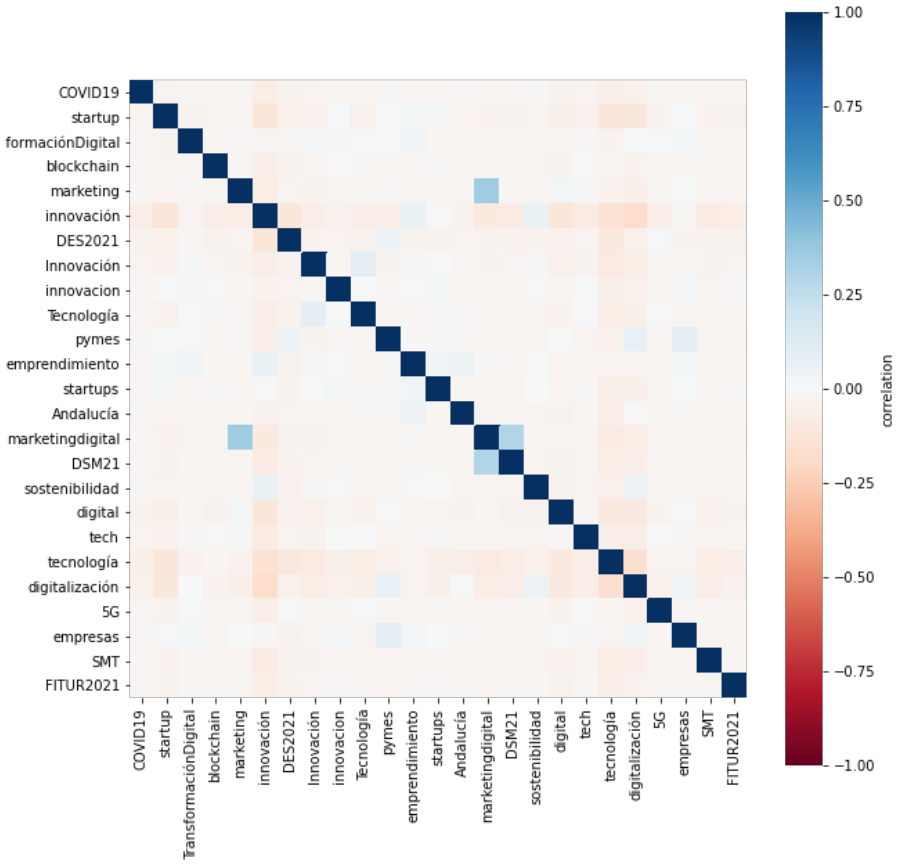

Figure 12: Hashtag correlation matrix generated from the innovation tweets dataset

In order to give a little bit of context to the previous information we visualized the most bigrams, which tell us how the actors associate the concepts of their communications.

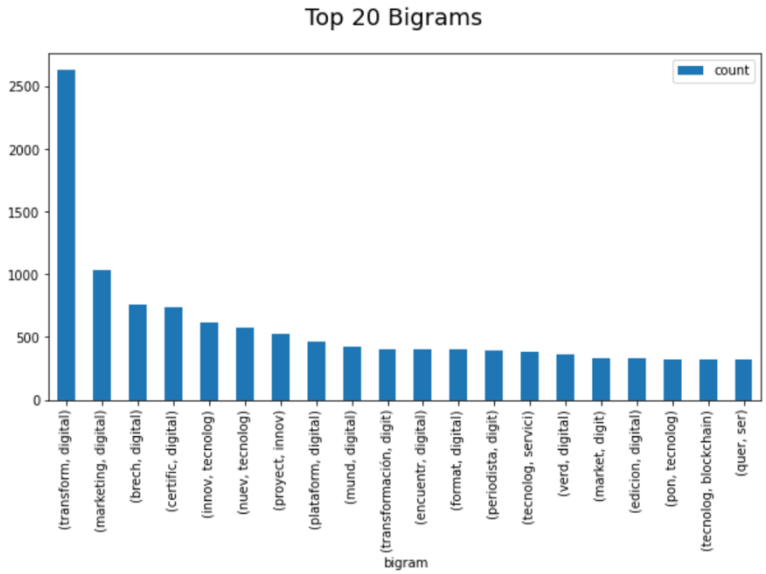

Figure 13: Top 20 bigrams from the innovation tweets dataset

We can see the relationship between digital and transformation, platform, meeting or world which are concepts that separated can have a completely different meaning. There are also two new concepts introduced, which are the digital gap and technology as a service.

Apart from analyzing the frequency of words, we also tried identifying entities with Spacy. The results were not as good as expected but could probably be better if we define a Spanish corpus for innovation. However, we found some locations such as Madrid, Europe, Galicia, Extremadura, Valencia and Barcelona among others and organizations, mainly tech companies such as Google, Apple, IBM and political parties or institutions such as the EU, PSOE or PP.

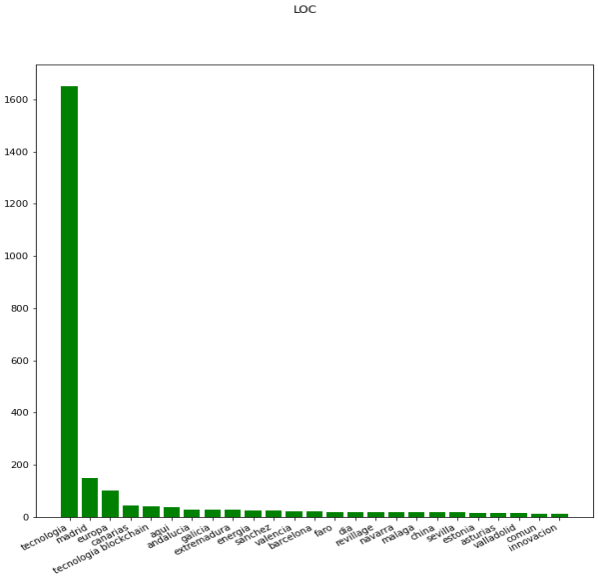

Figure 15: Location entities from the innovation tweets dataset

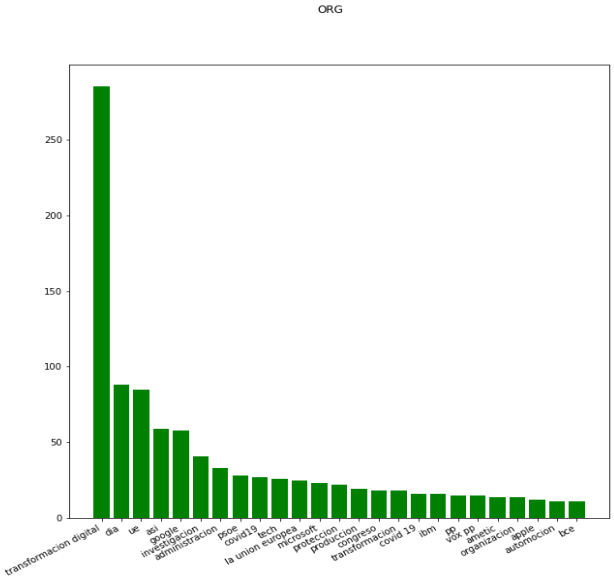

Figure 14: Organization entities from the innovation tweets dataset

To conclude the content analysis part, we looked at topic modeling and clustering. We developed a word2vec model, that could be used as a tool to look at similar words inside the innovation corpus. Here is an example of the output when we search “emprend”. We see the words “ecosystem” and “incubation” as important for those type of interactions.

```
[('extremadur', 0.9626995325088501),
('economiasocial', 0.9527100324630737),
 ('preincub', 0.9484952688217163),
 ('ecosistem', 0.9322354197502136),
 ('solidari', 0.9204318523406982),
('martalmcomp', 0.9148392677307129),
('cartagener', 0.9143445491790771),
 ('startups', 0.9104949235916138),
('emprendedor', 0.9090344905853271),
   ('dirig', 0.908071756362915)]
```

We then, clustered the vectorized tweets resulting in 6 clusters. However, after doing some experiments the silhouette value, an indicator of how dense and separated are the clusters, was very low 0.15 with the first 2 clusters achieving around 0.29, but the others were below 0.2. We looked at the most relevant terms for each cluster and we did not find clear topics. This is because social networks are very noisy, and it is difficult to overcome that.

As explained in the Exploratory Analysis chapter, we used LDA topic modelling also for clustering our tweets, this technique performed better.

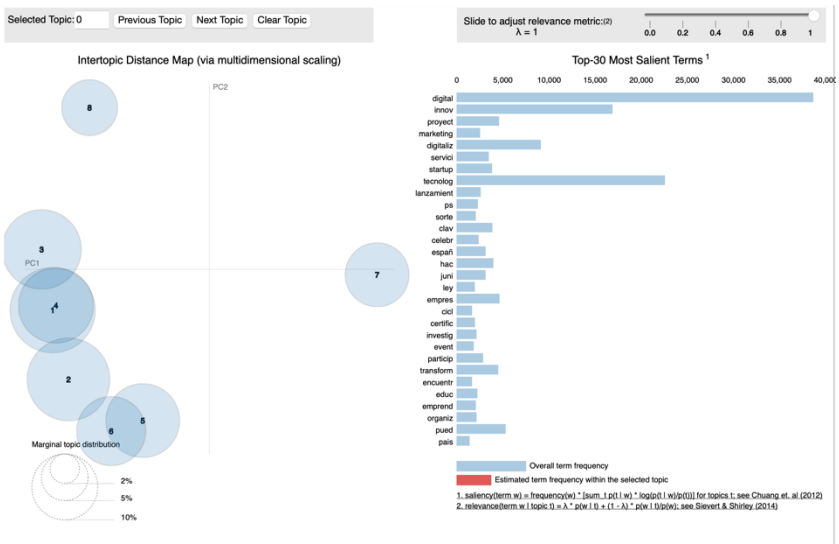

Figure 16: LDA identified topics in the innovation tweets dataset

With the visualization we can see the most relevant terms for each cluster and then infer the topic with external domain knowledge. The obtained topics are entrepreneurship, digital marketing, research and development, digitalization, technology, progress, events and meetings and digital certifications.

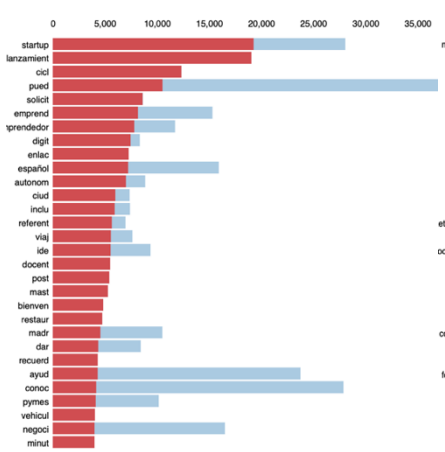

Figure 19: Entrepreneurship and startups topic in the LDA model

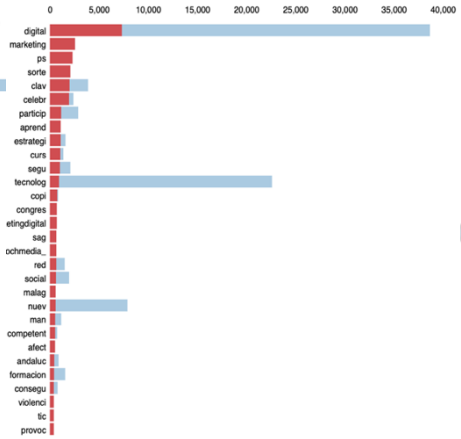

Figure 18: Digital marketing topic in the LDA model

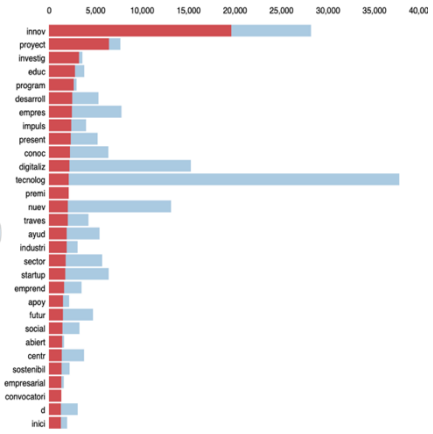

Figure 17: Reserach and Development topic in the LDA model

In the introduction, we talk about the importance of the geographical context in the innovation ecosystem. We analyzed the distribution of the tweets taking into account the location and did some visualizations. The first analysis is precisely the number of tweets in each region. We mapped two versions, the count and the normalized count (tweets for every 100.000 inhabitants).

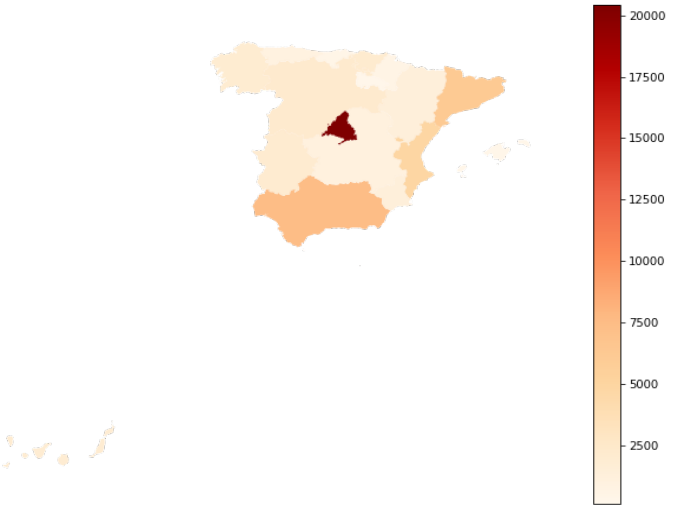

Figure 20: Tweets map with the color intensity representing the number of samples.

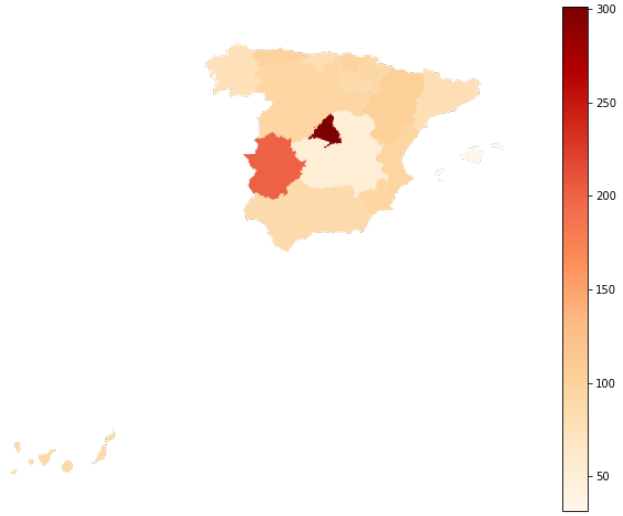

Figure 21: Tweets map with the color intensity representin the normalized number of samples with the population.

If we look at the map, we see that Madrid is a hotspot for our innovation discussions in both maps. In the first map, we also find regions with a lot of content such as Catalonia, Valencia and Andalucía, or Extremadura in the second. We investigated why the difference between Madrid and the other regions was so huge. We found a dataset containing the Spanish cities with more active users on Twitter and mapped those regions. Madrid is also the region with most activity in general.

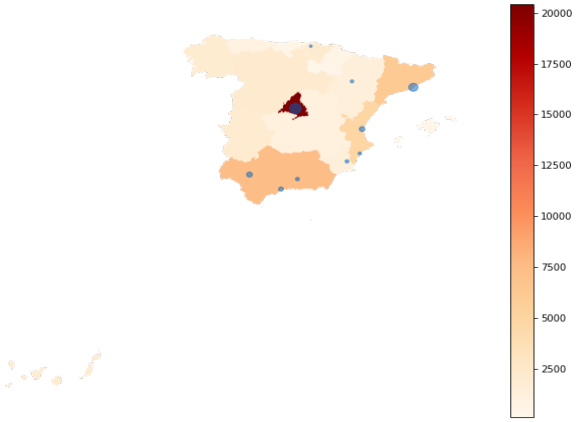

Figure 22: Tweets map with the cities with most Twitter active users.

Our last visualization of the Society analysis is the network on top of the region. We see that Madrid is a radial point because most of the tweets are from there, but we also see how Catalonia (Barcelona) and Valencia produce a lot of communications as explained with the heatmap.

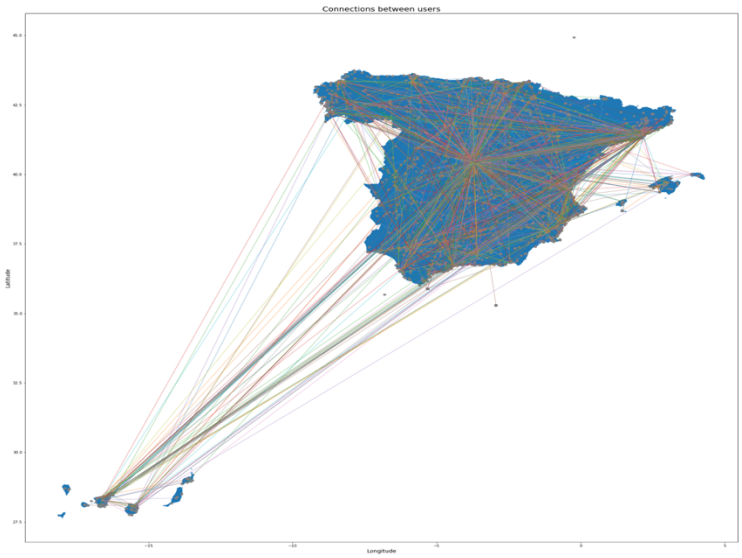

Figure 23: Tweets map with the interactions between a pair of users

5.2 Institutions & Investments

As explained in the introduction of this research, the analysis of this layer is more like an exploration and combination of different types of information. We explore the role of the institutions and investments in the geographical context as well as the relationship with the Society layer.

Our initial visualization is a choropleth map that contains different indicators related to the distribution of innovation investments, the initiatives to leverage those investments, as well as the innovation companies and their performance on the ecosystem. We can further explore the visualization in the interactive html file.

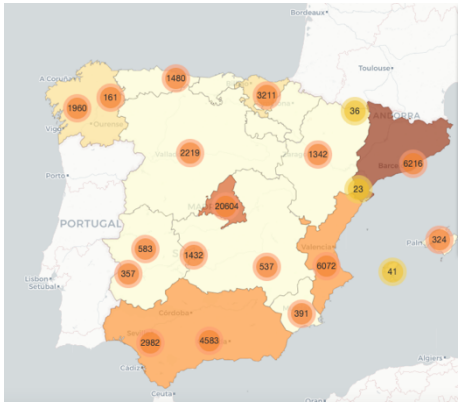

Figure 26: Innovative companyies map

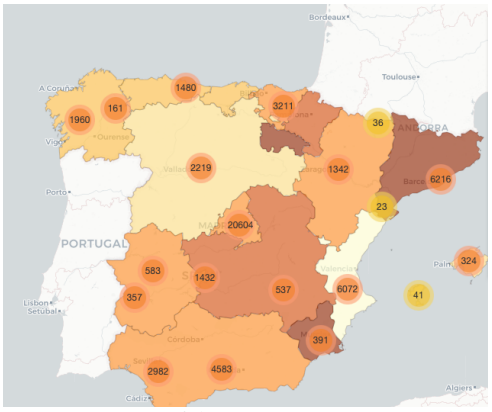

Figure 25: Map of the GDP percentage aimed at innovation

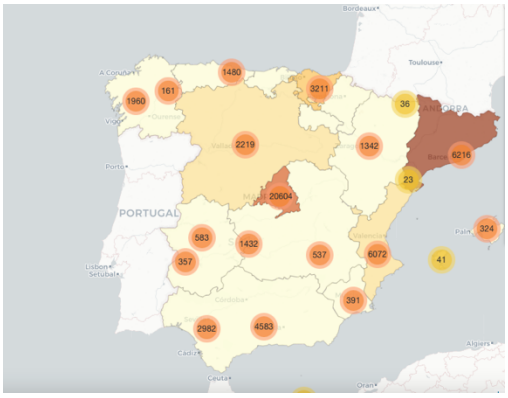

Figure 28: Number of VC investments map

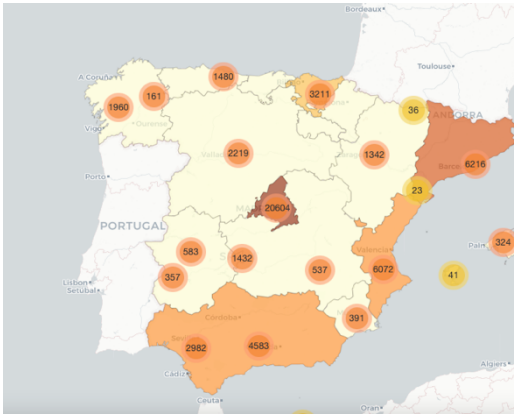

Figure 24: Initiatives to assist startups map

Looking at the different indicators, we can observe that there are some patterns. There are regions that achieve great performance in most of the indicators. An example could be Madrid or Catalonia. We also see how the investments and the initiatives to assist startups are related to the number of innovative companies. In order to achieve great innovation performance, we need both investments and initiatives by research centers and startups that transform the investments into innovation outcomes.

On top of that map, we plotted the tweets clustered by location with a color that indicates the number of tweets on the cluster. As it is an interactive map, we could zoom in and see different clusters inside each autonomous community. However, the regional unit of this research is one layer above.

In the other two visualizations, we link the tweets of a region with the clusters of companies and research centers. It allows us to see both, how those companies and research centers are distributed across the autonomous communities, and check if that distribution resembles to the tweets.

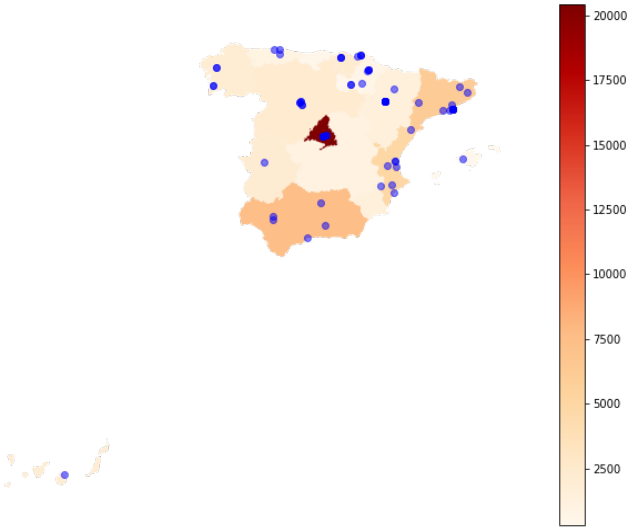

Figure 30: Companies clusters (blue markers) with tweets map

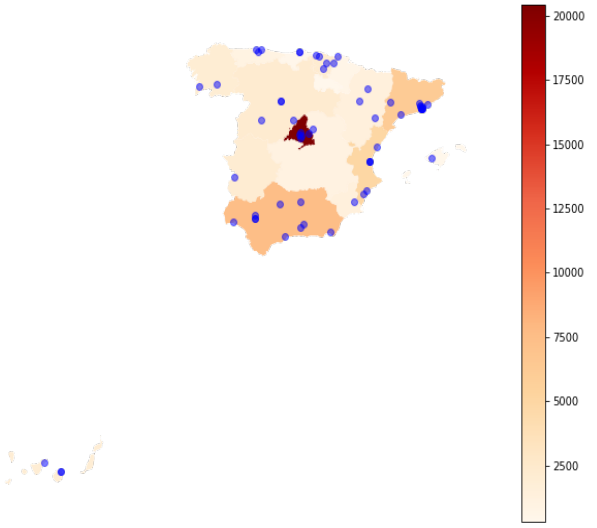

Figure 29: Research clústers (blue markers) with tweets map

If we analyze both private companies and public research centers, we find that more or less they are located in the same places, which are the ones with the most Twitter activity. However, we see some research centers and private companies in places with low activity and low innovation performance. Those are the regions that should adapt their policies to leverage the ecosystem to build innovation.

5.3 Regions

The results in this final layer focus on the regional factors and their relationship with the innovation ecosystem. They correspond to the locational factors, public policy and natural environment of the previously defined external innovation factors. As explained in the Methodology section, we took different indicators that do not have to necessarily be related to

innovation. We developed a classifier using the other two layers and explored which factors the classifier used. As a target value, we used tweets count, startups and an existing regional index. We assigned a weight for each indicator: 0.5 index, 0.25 the startups and 0.25 the tweets count. We tried different values for the number of estimators/trees and found 120 was optimal. It is important to mention that the training sample is too small to learn really meaningful features, because we only have 18 samples corresponding to each autonomous community. We would need historical data to optimize that classifier, which is outside the scope of this research. The analysis shows some preliminary correlations between the factors and the innovation ecosystem.

The first visualization shows the correlation matrix of our inputs. There is a relationship between innovation and number of universities, as mentioned before universities are the actors that produce the research needed for innovation, so they are an important factor of the ecosystem. Interestingly, we also find a correlation with the size of the region (houses and population), historically many citizens from rural regions came to the industrial regions because there were more opportunities. This concentration of people and talent is essential for the ecosystem at this moment, but it can possibly change with remote work. Finally, we see there is a relationship with the CO2 produced, this is because the clusters of factories and industrial zones. We should transition this model to a more sustainable one which has negative correlation between CO2 emissions and innovation.

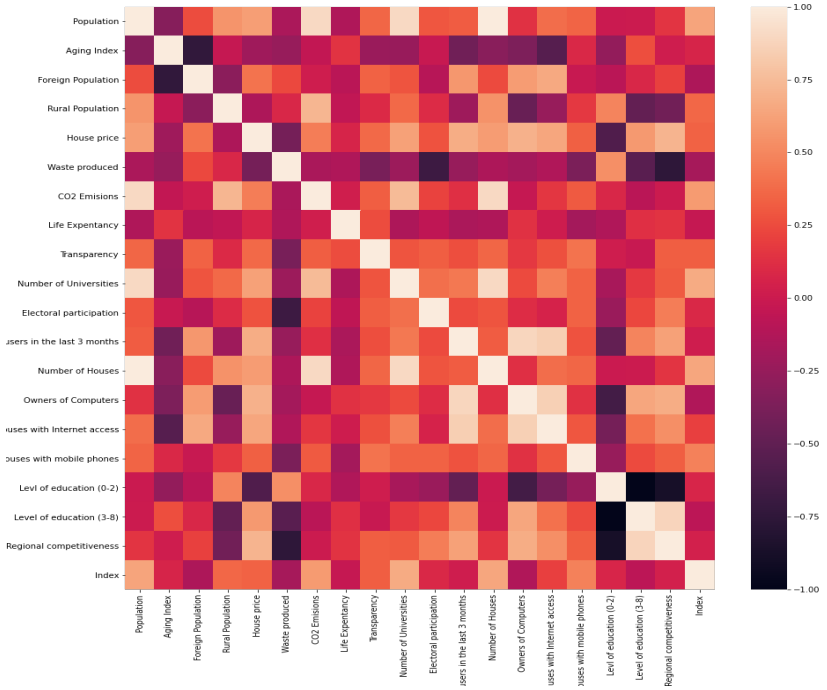

Figure 31: Correlation matrix of the regional characteristics dataset

Finally, we plotted the feature importance of our classifier. As expected, we see very similar features than the ones of the correlation matrix. ML algorithms for classification usually find correlations between the different sample features.

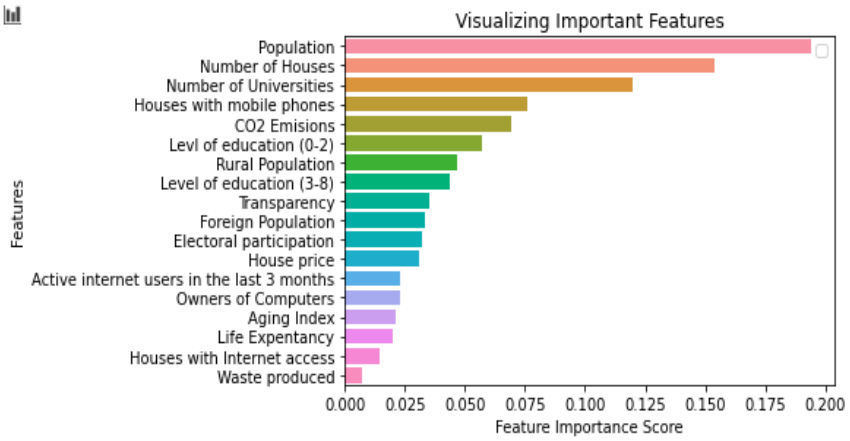

Figure 32: Features importance in the innovation classifier

7. CONCLUSION

In this chapter we define the results and conclusions of our research, the limitations and the possible future developments.

7.1 Research findings

This research provides a complete view of the different actors of the innovation ecosystem, how do they interact between them and the content of those interactions, from a geographical point of view.

By leveraging the wisdom of the crowd and analyzing the results, we have been able to complete the Society layer objective. We group the different identified actors in the exploratory analysis by their role in the ecosystem using our domain knowledge:

Make innovation policies:

- Government entities: Create the regional policies and regulations related to innovation.
- Political parties: Contribute to those policies and regulations.

Develop innovation:

- Universities: Provide the research needed for the development of new innovations.
- Entrepreneurs': Execute innovation and capture the value of it. Normally developers and salesmen.
- Startups: Challenge the status quo and drive innovation.
- Enterprise innovation departments: Generate the next wave of innovations. Two main ways, incremental innovation and disruptive innovation.

Promote the innovation ecosystem:

- Incubation Hubs: Provide the space and resources to generate innovation.
- Tech and Innovation leaders: Provide the vision and advice while promoting the ecosystem.
- Innovation foundations: Promote the innovation ecosystem and the relationship between actors.
- Journalists and digital media: Share the latest news related to innovation. They are needed to spread the knowledge.
- Investors: Incentivize the innovation ecosystem.

We look at the relationships between them and support the idea that normally these actors do not act alone. They need an ecosystem around them, which normally generates clusters or communities.

We describe the main topics of all the interactions between those actors in the ecosystem:

- Research and Development
- Technology and Digitalization
- Innovation Projects
- Innovation Incentives
- Innovation Events
- Ways of working
- Sustainability

We map these actors in the Spanish region. There are regions with the majority of the actors that are needed for a proper innovation ecosystem specially Madrid, but Catalonia or Valencia as well.

We provide an initial definition on how investments contribute to the appearance of the actors on the ecosystem. Particularly, there is a relationship between the VC and public investments with the number of innovation companies or startups in a region. Answering our second layer objective, we have been able to incorporate institutions and investments information and observe the different relationship with innovation.

The results also show that urban indicators such as the size of the region, the number of universities or the CO2 emissions in the environment, are positively corelated with the innovation performance. The gas emissions, as explained before, should transition to a negative correlation in the future, meaning that in regions with better performance there should be less emissions. The last part of the results responds to our initial regions objective. We have been able to understand first and second layers from a geographical point of view. Moreover, we have explained how the urban factors are key to determine that innovation ecosystem.

The research proves that social media provides great insights in the innovation field when performing data-driven analysis and combining it with domain knowledge and contextual information. We have been able to successfully achieve the main goal of the research which was analyzing the composition and relationships in the innovation ecosystem. However, as innovation is a rapidly changing and a growing field, we also state that further research can be performed to complement our study.

We believe in the concept of open innovation. By providing the results of this research we encourage everyone to use them as a framework for further investigations.

7.2 Limitations

Although it provides very insightful information that we would not get otherwise, social media has limitations. In this research, we tried to overcome some of them, but they should be taken into account if further analysis is performed.

Social media do not represent every citizen group for equal, we should not forget about people who may not have access to social media or who do not want to be there. Those people are citizens, and their opinions should be considered as important as the others. We should look for other ways to reach them.

Social media generates a lot of noise and we have seen it in some of the results. We should find methods to filter all that noise. One possibility could be applying topic extraction and after revising all the topics, keep only the ones which are really relevant to innovation. We need domain knowledge to leverage the key insights and remove that noise.

Open data from governments and companies is still difficult to access. It is often outdated and scattered across multiple sources. With all the technology advances and the transition to open governments and innovation, we have to develop the tools and mechanisms to make data accessible.

7.3 Future developments

The analysis of this research could be automatized and leveraged by organizations and governments to better understand the ecosystem where they play. With the recent advances of Big Data and the Cloud, it would be very interesting to see this analysis as a tool hosted on a public cloud, using streaming tools such as Apache Spark Streaming to gather real time information, send it to a database, combine it with historical and external data and visualize the results in a webpage.

It will be key for future developments to overcome the challenges stated in the limitations.

8. Bibliography

Schumpeter. (1934). *The Theory of Economic Development* . Harvard University Press.

Loukis, E. (2016). *Promoting open innovation in the public sector through social media monitoring.* Elsevier.

Pérez, C. (2007). *Factores de localización de las empresas innovadoras: Una aproximación para el caso de la Región Metropolitana de Barcelona.* UPC.

Martínez, M. (2003). *Medida de la capacidad innovadora de las comunidades autónomas españolas: construcción de un índice regional de innovación.* IAIF.

Claromonte, M. L. (2019). *Análisis de la innovación en España y por regiones europeas.* Valencia: Universitat Politècnica de València.

S.P.Taylor. (2017). *What Is Innovation? A study of the Definitions, Academic Models and Applicability of Innovation to an Example of Social Housing in England.* SCIRP Open Journal of Social Sciences.

300.000 Km /s. (2017). *Geografies de la Innovació a l'àrea metropolitana de Barcelona.* AMB.

Feldman, M. P. (2004). *The Significance of Innovation.* University of North Carolina.

ICONO. (2020). *Indicadores del sistema español de ciencia, tecnología e innovación.* FECYT.

Pose, A. R. (2020). *Do clusters generate greater innovation and growth?* London School of Economics.

Wu, W. L. (2019). How Data Analytics Can Drive Innovation. (W. University, Entrevistador)

Chesbrough, H. (2006). *Open Innovation.* McGraw-Hill Education.

Revert, F. (2018). An overview of topics extraction in Python with LDA. *towards data science*.

Cordobes, M. Á. (2021). Geo-visualization and Spatial Data Science - Slides . Visual Analytics 2021 Course UPF .

Ferràs, X. (2021). La España retrasada en innovación. (J. García, Entrevistador)

Technopedia. (2021). *Technopedia.* Recovered from Data Driven : https://www.techopedia.com/definition/18687/data-driven

Blázquez, C., & Ramón, J. (2019). *Impacto de las iniciativas de gestión de ideas internas en el desarrollo de una cultura de innovación en grandes organizacionesUn estudio de casos múltiple en España.* ESIC.

Concilio, G., Li, C., Rausell, P., & Tosoni, I. (2018). *Cities as Enablers of Innovation.*

Ash, J., Kitchin, R., & Leszczynski, A. (2015). *Digital turn, digital geography?* The Programmable City Working Paper.

Talmar, M., Walrave, B., S.Podoynitsyna, K., Holmström, J., & L.Romme, A. G. (2018). *Mapping, analyzing and designing innovation ecosystems: The Ecosystem Pie Model.*

Roberts, Falluch, Dinger, & Grover. (2012). *Absorptive capacity and information systems research: review, synthesis, and directions for future research.* MIS Quarterly.

Graham, S., & Marvin, S. (1996). *Telecommunications and the City.*

Gössling, T., & Rutten, R. (2007). *Innovation in Regions.* European Planning Studies.

Dagnino, R. (2001). *Elementos para una Renovación Explicativa-Normativa de las políticas de innovación Latinoamericanas.*

Fundación Bankinter. (2020). *Tendencias de inversión en España 2020.*

Bottazzi, L., & Peri, G. (2003). *Innovation and spillovers in regions: evidence from European patent data.*

OECD. (2013). *Innovation-driven Growth in Regions: The Role of Smart Specialisation.*

Porter, M. E. (1998). *Clusters and the new economics of competition.* Harvard Business Review.

International Monetary Fund. (2018). *World Economic Outlook 2018.*

Schumpeter, J. (1939). *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process, vol 2.* McGraw-Hill.

Bahcall, S. R. (2019). *Loonshots: How to Nurture the Crazy Ideas That Win Wars, Cure Diseases, and Transform Industries.* St. Martins Griffin Press.

Eurostat; OECD. (2018). *Oslo manual 2018 guidelines for collecting, reporting and using data on innovation.* OECD.

Obama, B. (2009). *Memorandum on Transparency and Open Government.* US Government.

Pece, A. M., Simona, O. E., & Salisteanu, F. (2015). *Innovation and Economic Growth: An Empirical Analysis for CEE Countries.* Procedia Economics and Finance.

Granstrand, O., & Holgersson, M. (2020). *Innovation ecosystems: A conceptual review and a new definition.*

Heijs, J., & Buesa, M. (2015). *Manual de economia de innovación.* Universidad Complutense Madrid.

Mackenzie, A. (2003). *Transduction: invention, innovation and collective life.*

Juma, C. (2014). *Complexity, Innovation, and Development: Schumpeter Revisited.*

Green, E. (2013). Innovation: The History of a Buzzword. *The Atlantic*.

Almirall, E., Garrell, A., Marcer, X., & Ferràs, X. (2021). El mundo no nos espera.

Borkowski, M. (2021). Technology Innovation has changed the world. *The Canadian Business Journal*.

Molina, M. (2021). Por qué las ciudades son la solución? *La Vanguardia*.

University of Cumbria. (sense data). *What is innovation? A study of the definitions, academic models and applicability of innovation to an example of social housing in England.* . 2017.

9. ANNEXES

Definitions

- Hashtag: metadata tag that is prefaced by the hash symbol #. User-generated tagging that enables cross-referencing of content sharing a subject or theme.
- Retweet: Republish or forward a post.
- System vs Ecosystem: In this research we use both system an ecosystem indistinguishably. Although ecosystem has a more nature origin than system alone, it does not have a semantic difference in our research.
- We: Pronoun used for this research, formal language.

Source code, dataset and visualizations results

Link: https://github.com/x4vi99/ES-inno-ecosystem.git

Account: x4vi99

Repository name: ES-inno-ecosystem