# POLITECNICO DI TORINO

**Master's Degree in Data Science and Engineering**


**Master's Degree Thesis**

# Leveraging Wikidata to highlight differences in topics and countries on Instagram Social Network

**Supervisors**

**Prof. Luca VASSIO**

**Prof. Martino TREVISAN** - ***University of Trieste***

**Candidate**

**Carmine DE CRISTOFARO**

**October 2023**

**Abstract**

In today's digital era, Online Social Networks (OSNs) have emerged as powerful platforms that shape the way we connect, communicate, and share information. Among these platforms, Instagram stands out as a prominent player, attracting billions of users worldwide and becoming a significant cultural phenomenon. With its visually appealing content and global reach, Instagram has redefined the landscape of social media, transcending geographical boundaries and bridging diverse communities.

The objective of this thesis is to analyze and study the effects of Instagram usage on the homogenization or preservation of geographical and cultural identities. The proposed approach involves the analysis of Instagram profiles from five different European countries: Italy, France, Germany, the United Kingdom, and Spain, across three distinct social macro-areas, namely politics, entertainment (specifically athletes, models, and actors), and academia (university profiles).

The initial phase of the work extensively utilizes the Wikidata database and its semantic query language to construct a profile database containing all the necessary information for categorizing Instagram users. These users can be either individuals or associations such as universities, belonging to both geographical and social areas of interest. Subsequently, the lists of profiles, categorized by social areas, are loaded into Crowdtangle, a Meta tool, to build the Instagram post database for each user in 2022, along with relevant information (e.g., interactions in the form of likes and/or comments, profile followers, post descriptions, etc.). The obtained dataset comprises 6,939 Instagram profiles and a total of 401,495 posts.

The dataset is then processed to visualize the main statistical characteristics related to profiles and the posts they have published. For profiles, the analyses focus on the number of followers and profile activity over the course of the year, while for posts, types, descriptions, and interaction (passive through likes and active through comments) are analyzed.

Through texts of post descriptions, a topic recognition process is conducted to identify discussion topics addressed throughout the year by the various study categories within the European landscape. Topic recognition is performed using BertTopic, a topic modeling model that extends the extraction of coherent topic representations through a class-based variation of TF-IDF. Specifically, the model generates document embeddings using pre-trained transformer-based language models, clusters these embeddings, and finally generates topic representations using the class-based TF-IDF procedure.

Lastly, this thesis focuses on the top five profiles in each category and country, analyzing their temporal trends in activity, followers, and interaction to derive

empirical insights into real-world events that have influenced Instagram's post-stream.

In summary, this study is significant as it contributes to understanding the role of social media in contemporary society, particularly its impact on politics, entertainment, and academia. By utilizing Instagram data, it provides valuable insights into user digital behaviors and trends. These insights can inform policymakers and businesses in formulating targeted policies and effective marketing strategies. Furthermore, it presents opportunities for individuals and organizations to harness the power of social media for fostering positive engagement, cultural exchange, and knowledge dissemination.

# Acknowledgements

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation and Objectives

In today's digital era, Online Social Networks (OSNs) have emerged as powerful platforms that shape the way we connect, communicate, and share information. Among these platforms, **Instagram** stands out as a prominent player, attracting billions of users worldwide and becoming a significant cultural phenomenon. With its visually appealing content and global reach, Instagram has redefined the landscape of social media, transcending geographical boundaries and bridging diverse communities.

However, this global interconnectedness raises a fundamental question: how does it impact the identity and cultural nuances of individuals with various backgrounds and origins from different European nations?

This thesis aims to explore how the widespread use of Instagram affects the homogenization or preservation of geographical and cultural identities. In order to achieve our goal we focus on analyzing three macro areas: **politics**, **entertainment**, and **academia** across five major European countries: **France**, **Italy**, **Germany**, **the UK**, and **Spain** in **2022**. The decision to focus on these areas stems from their pivotal roles in shaping our contemporary society, each contributing to the cultural fabric and identity of their respective nations. The primary objective of this research is to gain a comprehensive understanding of how Instagram data reflects and influences the dynamics within the chosen macro areas in the context of different European countries by examining the trends and topics that dominate the digital discourse and unraveling the underlying factors that shape these conversations. In this research, our goal is to use **Wikidata**'s vast knowledge base and its **semantic query language** to find entities that have Instagram usernames. As we begin this exploration, we're not only identifying these entities but also storing their associated predicates and objects, which will later be sorted into their respective

study domains. But our research doesn't stop at Wikidata, because we're also turning our attention to Instagram itself. Here, we're examining, thansk to the **official Instagram APIs**, not only user profiles but also specific aspects of individual posts like interactions and descriptions. Through post descriptions, a topic recognition process is conducted to identify discussion topics addressed throughout the year by the various study categories within the European landscape. Topic recognition is performed using BertTopic, a topic modeling model that extends the extraction of coherent topic representations through a class-based variation of TF-IDF. Specifically, the model generates document embeddings using pre-trained transformer-based language models, clusters these embeddings, and finally generates topic representations using the class-based TF-IDF procedure. To conclude we examine the 5 most followed profile for each macro area to provide a detailed view of the most influential profiles. The details of the work's subdivision are revealed in section 2.4.

This comprehensive analysis will shed light on Instagram's global landscape and help us understand how digital discussions relate to real-world events. At the core of this research is the skill of semantic querying, which lets us navigate through extensive datasets and extract valuable information. With each query, we aim to uncover patterns and trends that shape our understanding of the digital landscape, revealing how various entities are interconnected and the role Instagram plays in this complex network.

This study holds significant importance for various stakeholders. First and foremost, it contributes to the academic literature on the role of social media in contemporary society, shedding light on the impact of OSNs on politics, entertainment, and academia. By delving into Instagram data, we can offer insights into the digital behaviors and preferences of users, providing a deeper understanding of online engagement and information dissemination. For policymakers and businesses, this research provides valuable information on public sentiments, interests, and trends within the selected European countries. Understanding the digital landscape can assist in formulating targeted policies, effective marketing strategies, and engaging with the public in a meaningful way. Moreover, this study presents an opportunity for individuals and organizations to harness the power of social media to foster positive engagement, cultural exchange, and knowledge dissemination. Indeed, it can be indicative for the creation of valuable content, meaningful interactions, and collaborations with influential figures. In essence, this study provides a roadmap for harnessing the potential of social media for the benefit of society and organizations.

## 1.2 Online Social Networks

Online Social Networks (OSNs) are digital platforms hosted on the internet that facilitate the creation, sharing, and exchange of user-generated content, as well as the establishment and maintenance of social connections in a virtual environment.

**Figure 1.1:** Leading Online Social Networks logos [1]

These web-based networks empower individuals to create personal profiles, connect with other users, and engage in diverse forms of digital interactions, such as sharing photos, videos, status updates, and messages. As demonstrated in Ellison, Steinfield, and Lampe's seminal study (2007) [2], OSNs, exemplified by Facebook, play a crucial role in shaping digital communication, fostering social relationships, and enabling the dissemination of information on a global scale.

### 1.2.1 Historical Background of Online Social Networks

The roots of Online Social Networks can be traced back to the early days of the internet. In the late 20th century, online platforms like CompuServe and AOL offered rudimentary forms of social interaction, such as chat rooms and instant messaging [3]. However, the true inception of modern OSNs occurred in the early 2000s with the emergence of platforms like Friendster (2002), MySpace (2003), and LinkedIn (2003). These platforms allowed users to create personal profiles, connect with others, and share content with a growing online community [4]. Friendster, founded in 2002, is often credited as the pioneer of social networking sites, providing

users with features to connect with friends and discover new contacts based on mutual connections [5]. Shortly after, MySpace gained immense popularity as a platform for artists and musicians to showcase their work and connect with fans [6]. The rise of Facebook in 2004 marked a turning point in the history of OSNs. Founded by Mark Zuckerberg, Facebook expanded rapidly, becoming the largest social media platform globally. It introduced innovations like the News Feed and personalized profiles, setting a new standard for social networking [7]. LinkedIn, launched in 2003, targeted a professional audience, offering a platform for networking, job hunting, and business-related interactions [8]. Other platforms like Twitter (2006) introduced the concept of microblogging, allowing users to share short messages and links with followers.

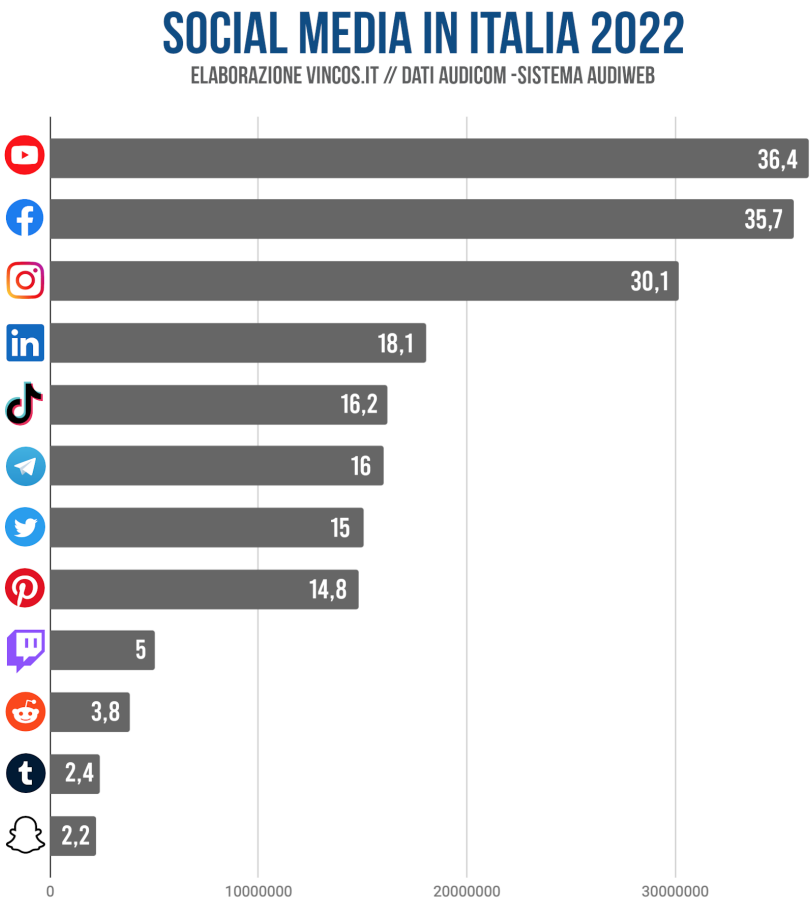



**Figure 1.2:** Estimated number of users of OSNs in Italy in 2022 [9]

### 1.2.2 Current Usage of Online Social Networks

Today, Online Social Networks have become an integral part of modern society, with billions of users actively engaging with these platforms on a daily basis. According to Statista's recent report, as of 2022, the global number of social media users exceeded 4.8 billion, representing more than 60% of the world's population [10]. Among the most popular platforms, Facebook remains the largest social media platform worldwide, with over 2.8 billion active users as of the same year [11]. YouTube follows closely, with more than 2 billion logged-in monthly users [10]. Messaging apps like WhatsApp and Facebook Messenger also have an immense user base, with 2 billion and 1.3 billion monthly active users, respectively [12] [10]. The Italian landscape, [9], is summarize in Figure 1.2).

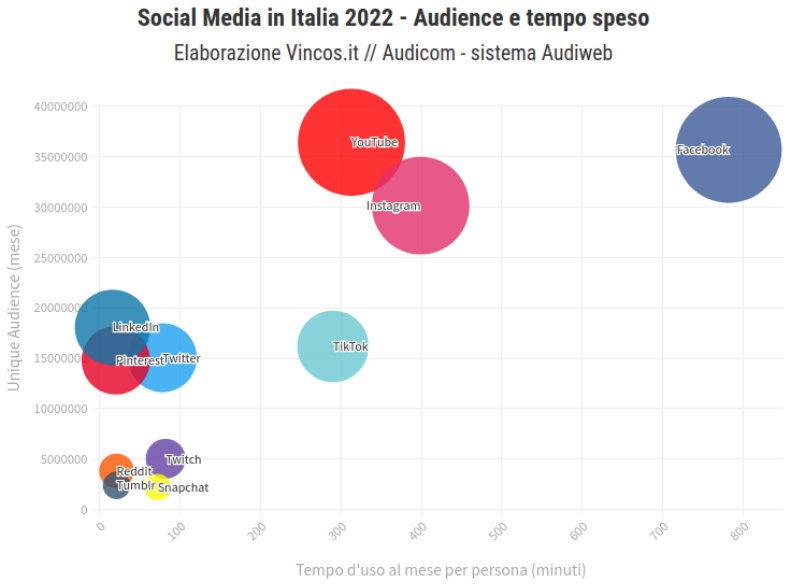



**Figure 1.3:** monthly usage time (minutes) of OSNs in Italy in 2022 [9]

Online Social Networks have revolutionized the way people communicate, share information, and form connections across the globe. According to a Pew Research Center survey, 72% of adults in the United States reported using social media in 2021, up from 5% in 2005 [13]. Moreover, a study by We Are Social and Hootsuite revealed that the average internet user spends approximately 2 hours and 25 minutes per day on social media [14](Italian landascape figure 1.3). Online Social Networks have undergone a remarkable evolution since their inception, fundamentally transforming the dynamics of interpersonal connections and communication. Today, these platforms boast billions of active users and wield an ever-expanding impact on various aspects of our lives, leaving an indelible imprint on our society and shaping

our digital interactions in the modern era.

### 1.2.3 Functionality Framework of Online Social Networks

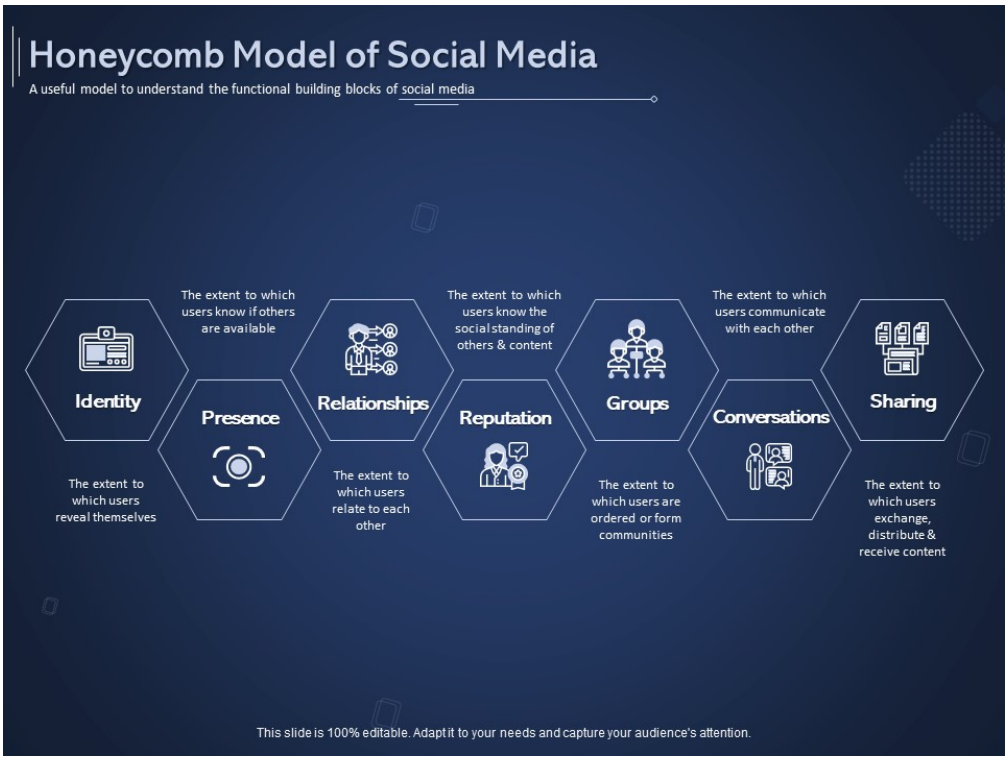



**Figure 1.4:** Online Social Networks Functionality (Honeycomb Model) [15]

OSNs are digital platforms that offer a wide range of functionalities, enabling users to connect, interact, and share content online. According to Peter Kim's Honeycomb model, figure 1.4, social networks are characterized by seven key functionalities: identity, sharing, conversation, participation, emotion, learning, and community [16]. The **identity** functionality allows users to create personalized profiles that reflect their individuality and interests. **Sharing** enables users to post content such as photos, videos, and status updates to share experiences and information with their network [4]. The **conversation** feature facilitates real-time communication through comments, direct messages, and chats, allowing immediate interactions with other users. The **participation** functionality engages users in collaborative activities, including polls, discussions, and project collaborations. The **emotion** dimension pertains to social networks' ability to evoke emotional reactions through likes, hearts, and emojis, enabling users to express feelings and appreciation [17]. The **learning** functionality allows users to acquire knowledge

through educational content, tutorials, and guides. Lastly, the **community** feature creates virtual spaces where users with shared interests can exchange ideas, participate in discussion groups, and build tighter social bonds [18].

### 1.2.4 Our research choice: Instagram Social Network

The selected Online Social Network for our research is Instagram, a photo and video-sharing service founded in 2010 in California. Users can upload media (images and videos) in the form of "posts," which can be edited with photographic filters, organized using hashtags and geotags, and shared publicly or with approved followers (private profile) based on customized privacy settings. Users can explore other users' content through tags and locations, visit their profiles, or swipe through their personal feed. The Instagram Feed is a continuously updating dashboard displaying photos and videos from advertisers and/or followed accounts, visible upon opening the Instagram application. Posts on Instagram can represent a single photo, a video of up to 60 seconds, an album of up to 10 photos/videos, an Instagram Reel (an editable multi-clip video with audio, effects, and other tools), or an Instagram TV (IGTV, allowing up to one-hour videos). IGTV is a standalone application where each user page is referred to as a "channel." Users can interact with posts by "liking," "commenting," and "saving" them privately in the application. They can also send messages including text, posts, Instagram Stories (which are vertical photos or up to 15 seconds videos that expire after 24 hours), photos, or videos taken or uploaded from the photo library to one or more people using the Instagram chat system called "Direct." Additionally, users can share posts in their own Instagram Stories. Unfortunately, due to data unavailability, Instagram Stories could not be studied in this research.

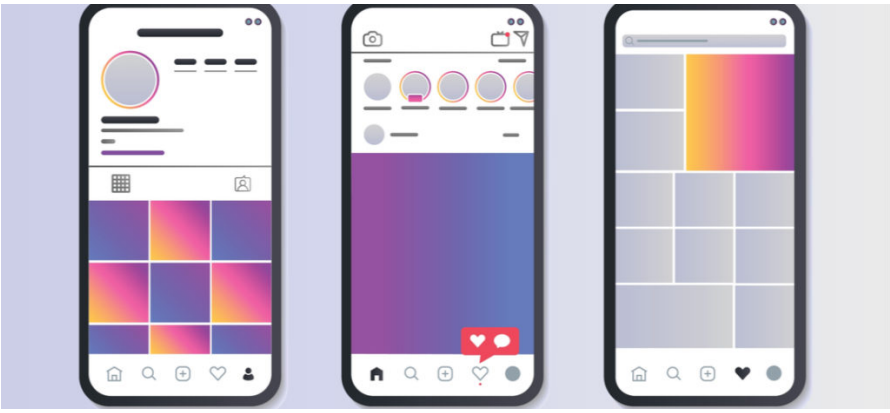


**Figure 1.5:** from right to left, respectively, example of Instagram user's view of profile, home, and feed

### 1.2.5 Instagram influencers

In the context of communication and marketing strategies, including Online Social Networks (OSNs), an "influencer" refers to a popular individual on the Internet with the ability to influence the behaviors, opinions, and choices of a specific group of users, particularly potential consumers. Influencers have the power to create trends on the Internet and impact others' purchasing decisions because of their authority, knowledge, and expertise on a particular topic, position, or relationship with their followers. As a result, brands collaborate with influencers to achieve their marketing objectives, choosing to advertise through the endorsement of influencers rather than traditional advertising methods. Some regulations require influencers to use specific hashtags like #ad or #adv to indicate sponsored content [19]. Instagram influencers regularly publish posts about their chosen topic, such as fashion, food, travel, or technology, and attract a large following of engaged individuals interested in their updates. Influencers can be classified in several ways, such as by the number of followers, the type of content they create (e.g., bloggers, YouTubers, podcasters), the level of influence, or the niche they operate in. The most common classification method is based on the number of followers [19], dividing influencers into five tiers, figure 1.6:

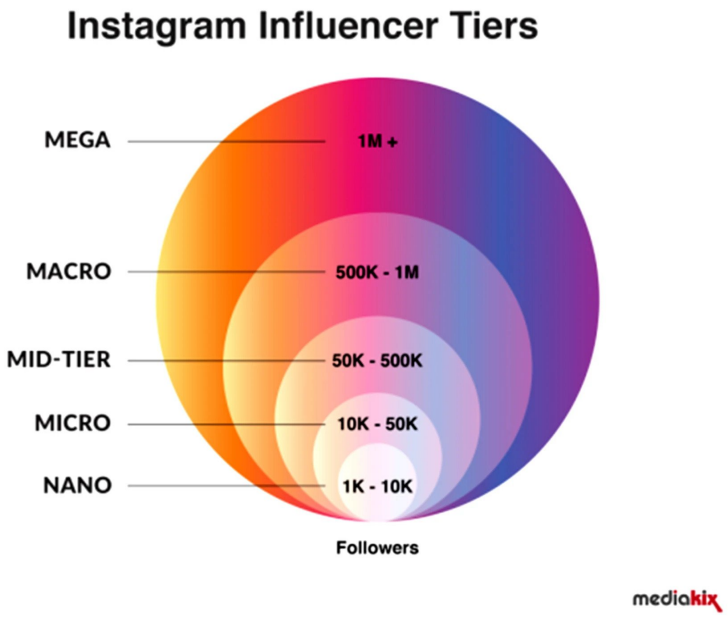



**Figure 1.6:** Tiers of influencers, graphical representation [20]

- Mega influencers: over 1,000,000 followers;

- Macro influencers: between 500,000 and 1,000,000 followers;
- Mid-Tier influencers: between 50,000 and 500,000 followers
- Micro influencers: between 10,000 and 50,000 followers;
- Nano influencers: fewer than 10,000 followers.

According to a study by Zarei et al. [19], most influencers publish unsponsored posts, but they tend to advertise products related to their area of expertise. Generally, mega, macro, and micro influencers prefer to use Instagram Stories to promote sponsored content rather than regular posts, with mega influencers being the most active in posting "advertise-stories." However, the study also indicates that while mega influencers receive the most attention in terms of likes and comments, nano influencers are more effective in sustaining attention within their specific niche.

# Chapter 2

# Related Literature

## 2.1 Online Social Networks General analyses

Online Social Networks (OSNs) have undergone extensive examination in recent times. Zarei and associates [19] directed their attention towards analyzing the categories of products endorsed by Instagram influencers, assessing their potential outreach, and evaluating engagement from their followers. The most interesting part for our study turns out to be the methodology of building the dataset and comparing the various influencer tiers for the final analyses. They gathered Instagram posts and stories from users who appended advertising-related hashtags to their content. This data collection process occurred from September 2018 to April 2019, utilizing the official Instagram APIs. After identifying the relevant accounts in this phase, they closely monitored all posts, stories, and user interactions from July 2019 to August 2019. Subsequently, the data underwent categorization into sponsored/non-sponsored, validated (ensuring influencer identification accuracy), and sampled (via Random Under-Sampling to achieve an equal number of sponsored and non-sponsored labeled posts). Regarding the types of products promoted by influencers, it was observed that 50% of Mega, 58% of Macro, and 70% of Micro influencers either promoted a single product or focused on a specific product category, often aligned with their niche expertise figure 2.1. To enhance data classification, the paper explains the subdivision of content into sponsored, non-sponsored, and hidden sponsored using a Random Forest Classifier, contrasting it with the results of a Contextual LSTM Neural Network architecture. For prediction (manually validated), post captions, profile biographies and post hashtags were identified as the most crucial features. Results revealed a substantial presence of hidden sponsored posts, especially among Micro and Nano influencers, despite guidance from the Advertising Standards Authority (ASA) to use commercial hashtags. This discrepancy is noteworthy, given that influencers' income is subject to taxation.

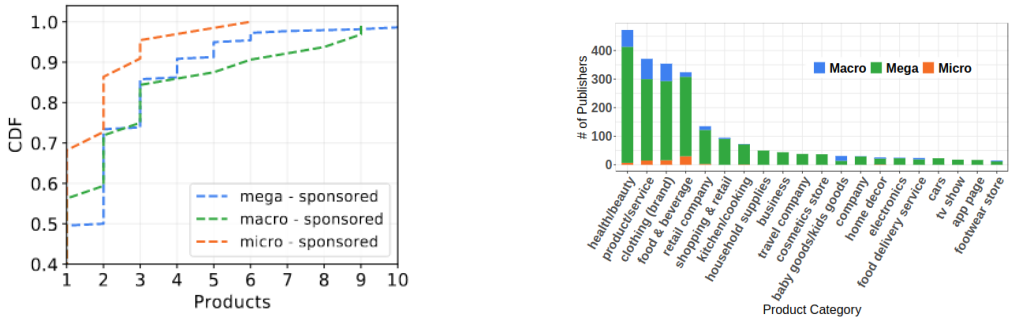



**Figure 2.1:** (a) CDF of number of promoted products. (b) Number of products promoted in stories, taken from [19]

Additional results were discussed in the concluding part of section 1.2.5.

Another study, conducted by Javed et al. [21], centered on analyzing WhatsApp messages and Twitter posts, such as tweets, in Pakistan during the COVID-19 pandemic. This research aimed to scrutinize the nature of COVID-19-related messages, user behavior, and the prevalence of COVID-19 misinformation on Twitter: turning out to be of interest for the idea of correlating and analyzing texts on a given topic to clustering common behaviors among users of social networks. The analysis encompassed 227 public WhatsApp groups, accessible via URL, containing text, images, and videos and using keywords like "Corona" or "2019-cov." Ultimately, 14% of pandemic-related information was found to be incorrect. Due to the complexity of categorizing images automatically, manual tagging was performed by two reviewers, focusing on images containing text related to COVID-19, restrictions, precautionary measures, and social distancing, grouped by similarity using Hamming distance. Messages were then classified into five categories: information, disinformation (unverifiable or false content), joke/satire, religious, and ambiguous (insufficient data for categorization), considering that a single message could belong to multiple categories simultaneously figure 2.2. Most messages were categorized as information, followed by religious content. Analyzing the lifespan of messages, satire messages had the shortest duration, while disinformation persisted for extended periods compared to valid information. Misinformation was further categorized by content, including fake news, false origin theories, bogus treatment methods, claims of vaccine development for economic reasons, seasonal climate-related disappearance of the virus with the vaccine, and comparing COVID-19 symptoms to seasonal flu. Among these, fake news had the shortest duration, while false remedies had the longest. On Twitter, information had a longer "die-time" than incorrect information (three times longer), partially because more users could comment and challenge the accuracy of posts.

Weber et al.'s study [22] examined coordination strategies such as boosting,

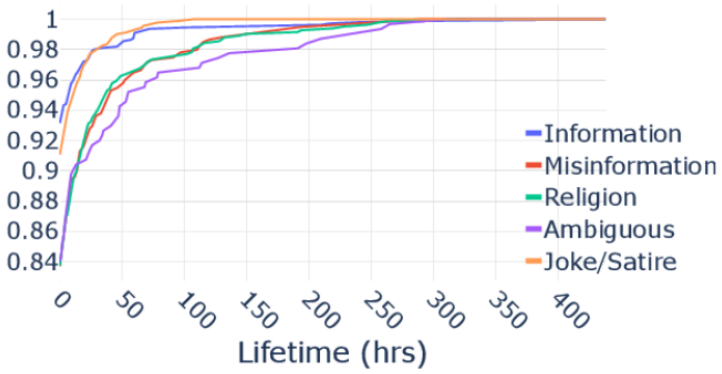



**Figure 2.2:** CDFs of Text of COVID-19 categories, taken from [21]

bullying, pollution, and metadata shuffling to uncover networks of cooperating accounts and groups exhibiting unusual behavior in degree. The research utilized temporal windows of varying sizes, user interactions, and metadata to detect accounts potentially engaged in coordinated, goal-based strategies. The analysis involved comparing a randomized dataset (for validation) with two relevant Twitter datasets: one from the IRA based on general activities in October 2018 and the other containing tweets from the 2018 Regional Australian Elections. The workflow pipeline figure 2.3, spanning from raw data extraction to results, comprised five steps: transforming social media posts into common interactions, filtering relevant interactions based on research criteria, inferring links between accounts by criterion, constructing a Large Coordinating Network (LCN) from inferred pairings, and identifying highly coordinating communities using a community detection algorithm (FSA_V, a variant of the FSA algorithm). This research both affirmed

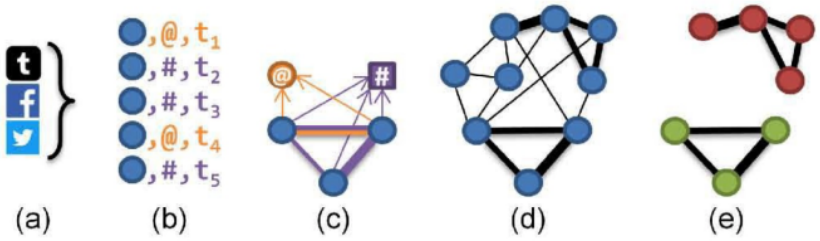



**Figure 2.3:** Conceptual LCN/HCC pipeline (circles are accounts), taken from [22]

the utilization of these strategies in organized operations and opened avenues for real-time applications to detect and potentially counter such schemes. However, the assumption that political accounts would frequently retweet and mention themselves

was not substantiated by the findings.

The utilization of Online Social Networks has real-world implications, which poses challenges in terms of data collection reliability for OSN analyses. Data must not only be accurate but also sufficiently comprehensive to construct meaningful networks and establish meaningful correlations between online and offline events. In this context, Weber et al.'s work [23] adopted a systematic comparative approach, concurrently collecting two parallel Twitter datasets using different methods and tools. They sought to understand how variations in data acquisition influenced the outcomes of social network analyses. The analysts generated two datasets using Twarc (as a baseline) and RAPID. Subsequently, they created three weighted directed social networks based on direct interactions categories: "mention networks," "reply networks," and "retweet networks." Analyses conducted on the datasets encompassed absolute statistics (e.g., tweet and hashtag counts), network statistics (e.g., Louvain cluster count, transitivity, centrality values), and cluster comparisons. The results indicated that while most content attributes remained similar, differences emerged in terms of captured accounts (leading to varying node counts), additional tweets (resulting in more edges), and overall network architecture differences. In recent years, there has been significant interest among scholars in leveraging new Machine Learning techniques and OSN data for predicting users' political orientations. Cardaioli et al.'s work [24] is pertinent to this prospect and of strong interest to us because it does profiling of social profiles based on their offline occupation that the basis of our dataset construction process. They utilized Twitter

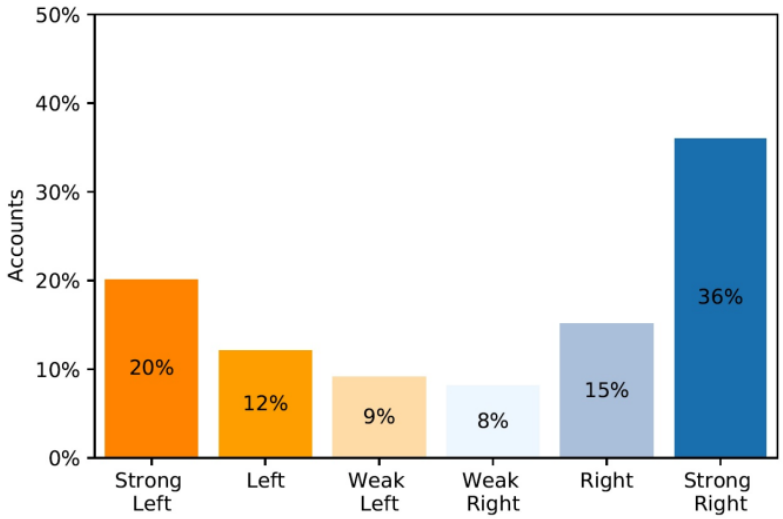



**Figure 2.4:** Agreement of the tested models on M5S supporters, taken from [24]

APIs to download a dataset comprising over 6000 users and nearly 10 million tweets.

The dataset was validated to detect legitimate accounts and eliminate Twitter bot accounts. Pre-processing involved removing URLs, punctuation, and similar elements. Subsequently, accounts were manually labeled by a group of independent human judges, using two different labeling techniques, as supporters of six distinct categories of Italian political parties, spanning from extreme right to extreme left. Profiles identified as supporters of ideologically well-defined parties were employed to train five different classification algorithms (SVM, Linear Regression, SGD, Random Forest, XGB) to assess the political tendencies of individuals labeled as "Movimento 5 Stelle" electorate. To facilitate Machine Learning methodologies, the authors employed TF-IDF to numerically represent tweets and analyze word characteristics and relationships. When predicting left-right associations, the researchers achieved an accuracy of up to 93%. The labeling of "Movimento 5 Stelle" voters revealed a political inclination distribution as depicted in figure 2.4. This analysis indicated that left-wing supporters predominantly tended to critique ideas of opposing parties rather than engaging in discussions on major political topics supported by their party.

## 2.2 Topic Recognition

Topic detection, also known as "topic recognition" or "topic modeling," is a crucial technique in the field of Natural Language Processing (NLP). It aims to identify and categorize the main themes or topics present in a corpus of text. This technique is essential in a wide range of applications, especially in social network data, where understanding the content shared by users is of vital importance. Indeed, it can help, for instance, in advising content to user based on their interests or in finding major events and occurrences around the world, often before they are reported by journals or other forms of classical media outlets. In the field of topic recognition, several approaches have contributed to improving the identification of topics in social network data. Some of these approaches include:

- **Latent Dirichlet Allocation (LDA)**: The LDA approach [25] is one of the most widely used models for discovering topics in textual documents. In the context of social networks, LDA has been applied to extract topics from large volumes of social data, enabling the identification of relevant trends and discussions.

- **Deep Learning**: Deep neural networks, particularly recurrent neural networks (RNNs) and transformer models [26], have demonstrated high effectiveness in topic detection within social network data. These models can capture complex word relationships and identify topics even in intricate conversations.

### 2.2.1 Topic Recognition on Online Social Network

Topic recognition, or topic modeling, aims to identify and categorize the primary subjects present in a corpus of text. Its impact is particularly evident in social networks, where understanding the content shared by users is of vital importance. For instance, on Twitter, this technique is exemplified by the *Trending Topics* feature. Twitter employs advanced algorithms for topic recognition to detect and display the most discussed subjects in real-time. During significant events such as elections or cultural phenomena, Twitter can swiftly identify and showcase the prevailing topics among users' tweets, facilitating global conversations. Facebook, on the other hand, leverages topic recognition to personalize users' news feeds. By analyzing the content of posts and identifying topics of interest based on users' past interactions and preferences, Facebook ensures that users are exposed to content that aligns with their interests and social connections. Similarly, YouTube utilizes topic modeling to recommend videos to its users. The platform's algorithms scrutinize the topics covered in videos that a user has previously watched and seek out similar or related content. This not only enhances user engagement but also broadens users' exposure to a diverse range of videos.

Several approaches have contributed to improving the identification of topics within social network data. Among these, the Latent Dirichlet Allocation (LDA) model stands out, as it is one of the most utilized models for discovering topics in short textual documents such as tweets [27]. Additionally, deep neural networks, particularly recurrent neural networks (RNNs), and transformer models have demonstrated high effectiveness in analyzing complex social content. Detecting topics within social network data holds crucial implications for various applications. For example, it enables monitoring current trends, identifying relevant content for information management, and detecting inappropriate or contentious content within social networks. This aspect is particularly relevant in a context where the dissemination of misinformation, fake news, and harmful content is a growing concern.

Throughout the chapter on the state of the art, we examine some noteworthy academic research addressing topic recognition within social network data. Among these, we find the work of Gregor Heinrich, who introduces an LDA-based approach for extracting topics from short texts like tweets on Twitter [27]. Furthermore, the paper by Liangjie Hong and colleagues proposes an Affinity Propagation-based algorithm for extracting topics and summaries from Twitter tweets [28]. A comprehensive overview of topic modeling models, including those used in social network data analysis, is presented in the article by X. Chen and A. N. Srivastava [29].

In summary, topic recognition within social network data represents a critical discipline for understanding online conversations, and its impact is reflected in

numerous applications, ranging from trend monitoring to information management and combating disinformation. References to significant academic research contribute to a better understanding of the challenges and solutions in this field of study.

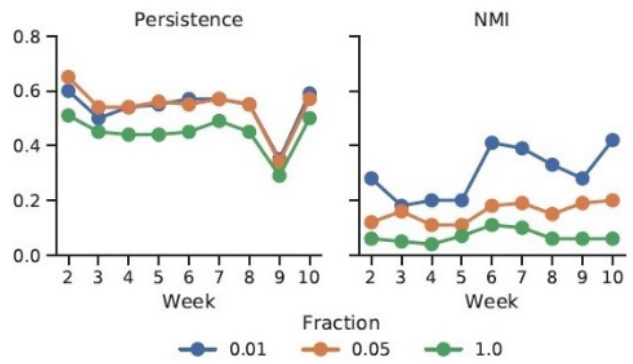
## 2.3 Previous publications by our research group

The foundation of this thesis project originates from extensive research conducted on Online Social Networks (OSNs) by the **SmartData@Polito** research group. This center focuses on Big Data technologies, Data Science (from data management to data modeling, analysis, and engineering), and Machine Learning methodologies applied to various knowledge domains, seeking solutions for both theoretical problems and practical applications in businesses.

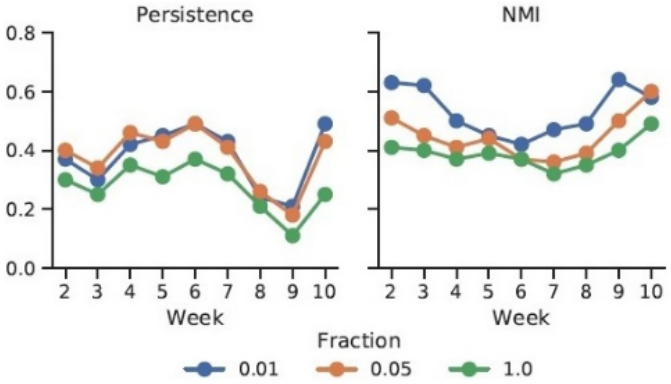
A preliminary work of Data Analytics applied to OSNs is the study conducted by Trevisan et al. [30], which examined how people behaved and interacted with politicians and public figures on Instagram before the European elections in May 2019. For data collection, a custom spider was used to download and archive data and metadata related to the most followed Italian public figures (i.e., influencers), their related activities (their posts), and interactions (likes and comments from users within the first 24 hours of publication) over two months. The study focuses on verifying whether interactions with political figures follow general patterns and whether there are differences compared to interactions with influencer profiles in different categories such as music, sports, and entertainment. The document also provides a characterization of so-called "mentions," analyzed to quantify interactions between commentators and their responses, divided into four categories: mentions with response/without response and requested/unrequested responses. The results suggest that comments on politicians' posts come from a small group of users actively participating in discussions. These comments persist for longer periods, are more numerous and lengthy. Furthermore, users rarely mention other people when commenting on politicians' posts compared to posts from other influencer categories, but these comments attract a large number of responses, most of which are not explicitly requested. This indicates that users are not engaged in the discussion but respond independently after reading previous interactions, possibly with the goal of influencing online political discourse. Differences between profiles of different political parties have also been observed. It is therefore evident that interactions with political and other category posts were significantly different both quantitatively and qualitatively.

Another study related to politics and OSNs was conducted by Ferreira et al. [31], whose aim was to study co-commentator communities to reveal the characteristics and dynamics of interactions on Instagram, highlighting common trends and

peculiarities, the level of engagement, and coordination. For this purpose, a null model was designed to extract interaction network skeletons, obtaining significant interactions among commentators by removing occasional random interactions and, using this model, identifying communities through the Louvain algorithm. The analysis was conducted on a dataset extracted from Instagram, containing the activity of various public political profiles (politicians and political parties) and general influencers (used as a control group) during electoral periods in Italy and Brazil, divided into weekly intervals (a total of ten weeks). Considering all posts from homogeneous groups of influencers published in the same week, each network was composed as follows: nodes represented commentators, edges represented comments between two commentators if both had commented on the same post (co-commentators), with a weight equal to the number of posts on which both had commented. Subsequently, researchers attempted to study profile similarities by grouping politicians based on their community structure and the evolution over time of graph skeletons. This methodology led to interesting observations: the skeletons of commentator networks were divided into fewer but better-defined communities, but these communities were weaker in politics compared to general topics, even though participants were more numerous and active. As expected, political communities peaked in online debate during elections, while in general topics, they were persistent and consistent over time figure 2.5. Most commentators with higher activity levels remained consistently active over time, even as communities changed. The COVID-19 pandemic has profoundly changed the economy, culture, politics, but most importantly, society, and consequently, it was important and interesting to study its impact on OSNs. The study conducted by Trevisan et al. [32] focused on analyzing the effects on social life during the total lockdown imposed during the first six months of 2020, both offline and online since OSNs represented an alternative to physical meetings. In particular, the work analyzed variations in activity trends, interactions, and engagement of Italian influencers on Instagram and Facebook during this historical event. The dataset was created through custom web crawlers and consisted of posts on Instagram and Facebook by 639 popular influencers, previously anonymized irreversibly. Subsequently, a data enrichment step was performed on data derived from each comment, using its psycholinguistic properties (LIWC) and topic extraction (Fig. 2.9). Some interesting variations were observed between the periods before, during, and after the lockdown, and the two social platforms were distinguished: Facebook saw an increase in the number of posts, comments, and likes during the lockdown, while, on the other hand, Instagram was characterized by a flat trend in terms of posts and comments, while the number of likes decreased significantly. Moreover, the ban on social activities in the first weeks of lockdown had effects on hourly patterns: in both social applications, there was an increase in activity in the morning and on weekends and a decrease in the early afternoon and evening. During the




**(a) Italy - Politics**




**(b) Italy - General**

**Figure 2.5:** Temporal evolution in Italy of commenters in communities. Blue: top 1%, Orange: top 5%, Green: all commenters, taken from [31]

same period, as expected, researchers also observed an increase in the popularity of comments related to negative feelings such as anxiety and inhibition, and people began to discuss the health emergency and personal life, such as unemployment. Analyzing online debate, it was observed that people engaged less in discussion, especially regarding politicians, even though they had gained a large number of new subscribers during those weeks. Vassio et al. [33] conducted research on how influencer posts attract interactions (likes or reactions) and how content popularity increases over time, as well as defining the behavior of influencers and followers over time and the progression of interactions over time, from the peak to the conclusion of a post. Researchers examined the activities of Italian influencers and their followers on Facebook and Instagram for more than five years (from 2016 to 2021). Here are some of the key findings: both influencer and user activity

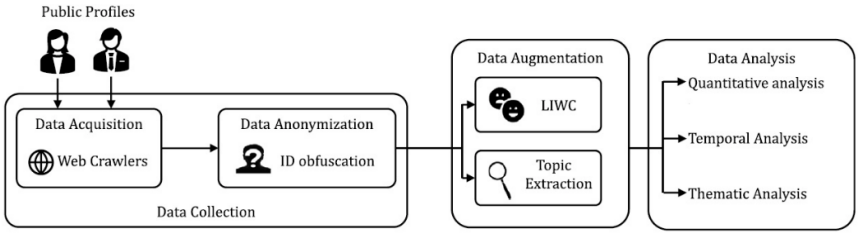



**Figure 2.6:** Data collection and analysis methods, taken from [32]

follow a consistent daily pattern, albeit with a different shape; the post inter-arrival interval follows a long-tailed distribution that is reasonably fitted by a log-normal distribution; on average, 50% of user interactions occur within the first 4 hours of content creation on Facebook and after 2 hours on Instagram, and the number of user interactions grows more rapidly on Instagram than on Facebook (however, after about 30 hours, the two curves converge); influencer posts have a short duration, with an exponential temporal decay of the interaction arrival rate, lasting between 20 and 50 hours (after which they stop attracting interactions), but the decay rate varies significantly from post to post and depending on the OSN in question; the number of interactions obtained during the first hour or even 15 minutes could provide a good indication of content popularity (the freshness of the post has a significant impact on its attractiveness); since generating new posts by the same influencer gradually attenuates the original post's attractiveness, it appears that followers tend to focus their attention at the top of the timeline, and thus the fraction of total interactions obtained in a given time interval is linked to the number of posts just published in the same interval. Furthermore, the findings revealed that, regardless of the online platform, follower and influencer actions follow similar patterns in both OSNs studied: follower and influencer activities decrease overnight, with two peaks during the day, while followers are more engaged in the late afternoon than influencers. Additionally, Facebook posts have a longer interaction duration compared to Instagram posts: the expected interaction time on Facebook is 15 hours, compared to 11 hours on Instagram.

Another study worth mentioning is the one conducted by Bertone et al. [34], which demonstrates that the OSN ecosystem, where prominent influencers strive to recruit new followers to increase their visibility (value), parallels the stock market, where investors choose to buy a particular stock and companies with a large number of investors grow in value. Influencers can be viewed in this context as stocks with a market value that can grow, measured by the number of followers. Regular users act as private investors, following (buying) influencers based on their personal preferences and data collected from other sources. This study will use statistical methods from the financial sector to examine influencer

dynamics on OSNs, contributing to the development of decision support systems to help influencers and advertisers accurately estimate short-term trends. The data was downloaded using the **CrowdTangle** tool (Chapter 3) and covered a period of over three years, from November 2017 to March 2021, and included 60 Italian public figures active on Instagram (obtained through the analysis platform www.hypeauditor.com) divided into three categories: singers/musicians, athletes, and VIPs. The Google Trends API was used to download historical trends for the entire set of influencers and to collect the Google Trends Search Volume Index (SVI), a monthly granularity time series representing normalized search queries (the name/alias of each influencer was used as a search keyword) relative to the highest observed value in the considered period. Bollinger Bands, typically used in decision support system applications to provide stock trading suggestions to operators, could help influencers, advertisers, and social media platforms in the OSN context to model their tactics and dynamically compare completely different time measurements. This approach demonstrated how short-term trends derived from Bollinger Bands for influencer followers on Instagram were similar to those found in external sources like Google Trends, similar to what academics have discovered for the stock market. Researchers also found that influencers with fewer followers had more well-coordinated short-term trends, as they were more likely to gain new followers than those with a large following.

## 2.4 Contribution of This Work

This work distinguishes itself from the others presented in this chapter 2 through its innovative approach to characterizing Instagram profiles, which are subsequently subjected to detailed analysis. In fact, thanks to the use of Wikidata information we are able to do profiling of Instagram profiles even before we get their information related to the platform itself. By doing so we are able to generate lists divided by social status that we will later go on to characterize also by country of origin. In chapter 3 that follows, the process of profiling precisely and the construction of the dataset is presented, which is done by the combination of Wikidata info and information obtained from the official Instagram APIs. Also presented is a description of the main characteristics of published posts such as type, length of descriptions and presence of mentions or hasthags plus a quantitative analysis on passive and active engagement measured in likes and comments respectively. Chapter 4, on the other hand, will present two types of uses of the dataset: the first involves the use of a topic recognition model to analyze the topics of discussion in the posts under analysis, while the second is an analysis of the top5 influencers for each social category and country thus providing us with an overview of the activity and interaction received by the most influential profiles during the year

2022. Finally, Chapter 5 will offer an overview of the results acquired as well as a comparison of them in order to understand how geographic and cultural identity is valued or not within a social platform such as Instagram, which allows for constant interworld connection.

# Chapter 3

# Dataset Structure and Description

## 3.1 Wikidata database

The primary aim of our research is the establishment of a **profile database** for Instagram profiles capable of storing essential information, such as nationality or occupation for populating lists related to our macro study areas. To achieve this, we leverage the open-source database provided by Wikidata.

Wikidata, which was introduced by the Wikimedia Foundation in 2012, has become a crucial participant in the fight for accessibility and knowledge sharing on a global scale. It is fundamentally more than just a database; it is a vibrant repository that supports the global community's search of knowledge. Wikidata thrives on the concepts of collaboration, transparency, and interconnection, in contrast to traditional databases that keep data in separate silos. It acts as a dynamic hub where information is carefully vetted, linked, and made accessible to everyone on a staggeringly varied range of topics. The following are its five key features:

- **Structured Knowledge**: Wikidata's core component is organized knowledge, which includes facts, ideas, and relationships between things. These data are set up in a way that makes them both machine- and human-readable, opening up a wide range of uses, from automated data analysis to scholarly inquiry.

- **Interconnected Data**: Wikidata stands out for its exceptional ability to create links between various types of information. Here, a complex tapestry of interlinked knowledge is created by connecting biographies to literary masterpieces, physical locations to historical events, and much more.

- **Global Collaboration**: Wikidata's strength comes from the vast network of volunteers who work ceaselessly to add to and improve the data. The database will always be current, accurate, and growing thanks to this cooperative approach.
- **Open License**: Every piece of information in Wikidata is distributed under an open license that enables anyone to use it for anything from software development to scientific inquiry.
- **Integration with Wikimedia Projects**: Wikipedia and other Wikimedia projects benefit greatly from Wikidata as a valuable resource. These projects can provide their consumers more thorough and current information by incorporating structured data from Wikidata.

The Wikidata database is structured meticulously according to the **RDF** (Resource Description Framework) data model. Within this model, information is organized into data graphs, composed of RDF **triples** that succinctly represent relationships between subjects, predicates, and objects.

The **subject** signifies the entity or concept to which the triple pertains, representing entities, concepts, or objects in the world. For example, it could represent an author, a geographical location, or an abstract concept.

The **predicate** specifies the nature of the relationship between the subject and the object. Predicates, in essence, denote the nature of the connection, such as *"was born in"* or *"is the author of"*.

The **object** of the triple represents the target entity or value within the relationship. It can assume the form of a literal value, like a date or text, or another subject entity.

Wikidata revolves around the notion of entities, each identified by a unique identifier known as a "QID." These entities encompass a diverse spectrum of objects, concepts, and phenomena, enabling comprehensive knowledge representation. Properties, identified by their own unique identifiers (e.g., "P21" for "sex or gender"), serve as the counterparts to predicates in RDF triples. They define the nature of the relationships and attributes attributed to entities within the Wikidata knowledge graph, figure 3.1. Triples within the Wikidata database can incorporate qualifiers, augmenting the relationship's contextual information. Qualifiers enable the specification of additional details or constraints related to the subject-object association, such as start and end dates for an event. Wikidata extends its scope by encompassing links to external resources, including URLs to web pages, enriching the depth of available information. Each entity within Wikidata is endowed with labels and descriptions, both accessible in multiple languages to enhance global accessibility and comprehension. To foster international accessibility, Wikidata emphasizes interlingual links, enabling cross-reference and translation between the multilingual versions of the same concepts or entities.

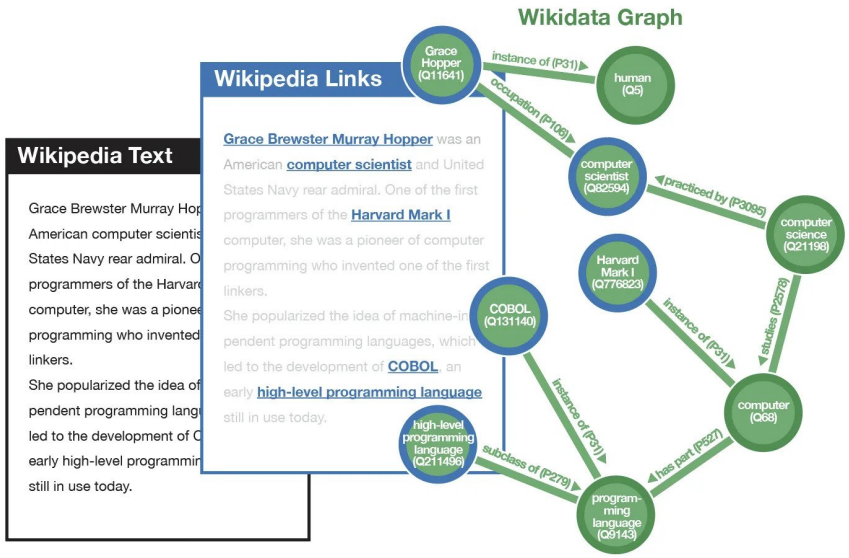



**Figure 3.1:** Wikipedia and Wikidata as three distinct data layers.[35]

### 3.1.1 Interacting with the Wikidata Database: SPARQL Query Language

In the context of this research, the utilization of the Wikidata database necessitates the use of the Semantic Query Language, commonly referred to as **SPARQL** (SPARQL Protocol and RDF Query Language). SPARQL is a meticulously designed semantic query language that serves the specific purpose of extracting data from RDF (Resource Description Framework) graphs, the foundation upon which Wikidata relies.

The foundamental syntax of a SPARQL query consist of the following elements:

- **SELECT:** Specifies which variablese to retrieve from the RDF data;
- **WHERE:** Defines the search pattern to locate the desired data;
- **PREFIX:** Defines prefixes to abbreaviate long URIs in queries;
- **FILTER:** Applies conditions to variables in the query.

Using SPARQL for quering RDF data offers several advantages:

- **Semantic Queries:** SPARQL allows you to perform semantic queries, which means you can query data based on its meaning and relationships rather than just matching text or keywords. This is particularly useful for linked data and the semantic web.

- **Standardization:** SPARQL is an official W3C (World Wide Web Consortium) recommendation, ensuring that it is widely adopted and supported across various RDF data sources and applications.

- **Flexibility:** SPARQL provides flexible querying capabilities. You can retrieve specific data patterns, filter results based on conditions, and even aggregate data using built-in functions.

- **RDF Data Integration:** SPARQL enables the integration of data from multiple RDF datasets, making it a powerful tool for combining information from various sources.

- **Scalability:** SPARQL is designed to handle large datasets efficiently. It can retrieve data from massive RDF graphs without significant performance degradation.

### 3.1.2 Profiling

Considering that our objective is to construct a profile database of individuals who possess an Instagram username on their Wikipedia pages by querying the Wikidata database, our first step is to retrieve all entities that indeed have the predicate P2003, **'Instagram username'**. This retrieval operation using the predicate P2003 provides us with all the usernames of publicly available profiles on Wikidata. Regardless of the nature of the originating entity, such as individuals, associations like universities or companies, we will refer to these profiles as *'influencers'*.

Furthermore, for the purpose of our analysis, which seeks to explore the distinctions and commonalities among Instagram profiles based on their diverse origins and occupations, it is essential to retain the information linked to the **'instance of'** (P31) predicates. This will enable us to determine whether a profile corresponds to an individual human user or to other entities, including universities. This distinction is particularly relevant as we also want to delve into an in-depth examination of social media usage within the academic sphere. Our analysis must also correlate influencers with different countries of origin. Therefore, it is essential to preserve the information provided by the predicates P27 and P37, namely **'country'** and **'country of citizenship'**, depending on whether the profile is human or not. In conclusion, in order to analyze differences and similarities in the social or occupational roles, we save all the information related to the predicate P106, **'occupation'**. This is done specifically for human instance profiles.

This provided us with valuable insights into the social positioning of individuals behind the profiles, a relevant aspect for a better understanding of online interactions and sharing dynamics.

Building the profile database serves two primary purposes. The first is to store relevant information about the profiles and the broader areas of interest. The second purpose is to structure input lists for the official Meta APIs, section 3.2. These lists are constructed to populate profiles in the areas of interest for this thesis. Specifically, an initial screening, on predicate 'country' or 'country of citizenship', is done for the five European countries under analysis: Italy, France, Spain, Germany, and the United Kingdom. Subsequently, the profile lists differ based on the 'occupation' predicate.

For the *'european_politicians'* list, all Instagram usernames from Wikipedia that have the entity **Q82955**, 'politician', as the object of the 'occupation' predicate are considered. Regarding the entertainment macro-area, there is further classification. For social figures, there are three distinct 'occupations' from which we construct the lists: *'european_athletes'*, *'european_models'*, and *'european_actors'*. Excluding models, who are selected if the entity **Q1979154**, 'model,' is present as the object of the 'occupation' predicate, the criteria for athletes and actors are slightly different.

Specifically, for athletes, all profiles that have one of the following keywords in the object of the 'occupation' predicate are selected:*['player', 'sport', 'gymnast', 'athletics', 'swimmer', 'dancer', 'racer', 'athlete']*. For actors, all profiles that contain the word 'actor' in the object of the 'occupation' predicate are considered, thus encompassing differences in actors, such as theatre actors, TV actors, or actors in action movies or romantic movies, etc.

Finally, the *'european_universities'* list for the 'academics/universities' area of study is constructed by selecting profiles that have the entity **Q3918**, 'university,' as the object of the 'instance of' predicate

## 3.2 The Instagram Post Dataset

To fulfill the objectives of this research, a substantial repository of social media data was required. While Twitter's public data is readily accessible through its APIs, gaining access to platforms like Facebook and Instagram can prove considerably more challenging for academic scholars. Consequently, our choice gravitated towards the most accessible platform for managing three of the most prominent social networks: Facebook, Instagram, and Reddit. This platform also extends free access to journalists, academics, and researchers.

Specifically, the data was acquired from **CrowdTangle**[1], a tool owned by Meta, which meticulously monitors interactions on public content originating from Facebook pages and groups, verified profiles, Instagram accounts, and subreddits. It is worth noting that this tool excludes paid advertisements unless these ads

[1] www.crowdtangle.com

originally commenced as organic, non-paid posts and were subsequently promoted using Facebook's advertising tools. Moreover, it does not encompass activity on private accounts or posts that are visible exclusively to select groups of followers.

Notably, CrowdTangle diligently tracks influential public accounts and groups across Facebook, Instagram, and Reddit. This encompassing range includes accounts held by politicians, celebrities, sports teams, journalists, media outlets, publishers, public figures, and other notable entities. The primary objective of this platform is to enhance precision and transparency regarding the dynamics of Online Social Networks (OSNs). It aims to aid researchers in various pursuits, including monitoring specific topics by seeking combinations of keywords and phrases to unveil trends and patterns, tracking the activity of public accounts and communities, analyzing which accounts exhibit the most prolific posting behavior, and identifying who garners the highest level of engagement, comprising reactions, comments, and shares, especially concerning particular issues. Furthermore, it facilitates performance monitoring and the detection of emerging narratives and stories.

### 3.2.1 Selection of the Social Media Platform

Online Social Networks (OSNs) have ushered in a new field known as "social sensing", where user-generated content from platforms like Twitter is leveraged to identify trending events in the offline world. In our research, we have chosen Instagram, a social image sharing platform, over predominantly text-based channels. This choice enables us to explore how social sensing can harness the power of multimodal multimedia content.

Our decision to focus on Instagram is driven by the unique characteristics of this OSN. Instagram is designed to facilitate the creation of public content that is easily accessible to the general public. Consequently, it offers a platform readily available to ordinary individuals. Additionally, Instagram's features for gauging the popularity of posts and identifying anomalies are relatively straightforward. These features primarily include metrics such as the number of reactions, such as likes and comments.

In contrast, platforms like Facebook offer a wider array of reactions, each with distinct meanings (e.g., "love," "sad," "angry"). This diversity in reactions could potentially complicate the analysis process.

## 3.3 Characterization of the Dataset

The dataset was retrieved employing the CrowdTangle public APIs, giving as inputs the five lists presented in section 3.1.2, and reporting the activities of **6,936**

**Instagram profiles** in 2022: the result of this collection was generating a dataset of **401,495 posts** in total.

These data are categorized into various key categories as follows:

## Instagram Profile Categories

### Politicians

There are **1,105** politician profiles in the dataset, divided by country as follows in table 3.1:

| Country | Number of Profiles | Number of Posts |
|---|---|---|
| Germany | 486 | 37,150 |
| Spain | 304 | 25,962 |
| Italy | 129 | 21,366 |
| France | 105 | 10,057 |
| United Kingdom | 81 | 6,823 |

**Table 3.1:** Politician Profiles by Country

### Entertainment Figures

This category includes **5,686** profiles of entertainment figures, further subdivided into subcategories:

- **Athletes**: **2,576** athlete profiles with **88,500** posts, divided by country, table 3.2;

| Country | Number of Profiles | Number of Posts |
|---|---|---|
| Germany | 456 | 11,905 |
| Spain | 538 | 22,627 |
| Italy | 315 | 13,297 |
| France | 773 | 21,650 |
| United Kingdom | 494 | 19,021 |

**Table 3.2:** Athlete Profiles by Country

- **Models**: **688** model profiles with **41,446** posts, divided by country, table 3.3;
- **Actors**: **2,422** actor profiles with **142,594** posts, divided by country, table 3.4;

| Country | Number of Profiles | Number of Posts |
|---|---|---|
| Germany | 81 | 650 |
| Spain | 181 | 8,750 |
| Italy | 104 | 7,699 |
| France | 123 | 6,502 |
| United Kingdom | 199 | 13,845 |

**Table 3.3:** Model Profiles by Country

| Country | Number of Profiles | Number of Posts |
|---|---|---|
| Germany | 224 | 11,360 |
| Spain | 744 | 38,572 |
| Italy | 254 | 20,727 |
| France | 415 | 19,056 |
| United Kingdom | 785 | 52,879 |

**Table 3.4:** Actor Profiles by Country

### Universities

There are **154** university profiles, divided by country, table 3.5:

| Country | Number of Profiles | Number of Posts |
|---|---|---|
| France | 17 | 1,561 |
| Germany | 5 | 939 |
| Italy | 52 | 9,679 |
| Spain | 4 | 773 |
| United Kingdom | 76 | 14,645 |

**Table 3.5:** University Profiles by Country

The dataset comprises data entities representing influencers' posts published during the specified time frame. These data entities are characterized by a wide array of attributes within each record (for each post), which can be categorized as follows:

- **Account Attributes:** These encompass details describing the influencer's account at the time of data collection by CrowdTangle. They encompass the account handle, name, platform-specific IDs, subscriber count, profile URL, account verification status, and various additional metrics.
- **Sponsor Attributes:** If a post includes branded content, these attributes

mirror those of the sponsoring account, akin to the aforementioned account attributes.

- **User-Generated Post Attributes:** These attributes pertain to features influenced by the user's intent during post creation. They include media type (photo, album, video, IGTV), publication date, and the textual content in the post's description.
- **Platform-Generated Post Attributes:** These features are automatically generated by the platform when the post is created. They may include historical data about the influencer's previous posts, the post's URL, and the expected level of engagement (likes and comments) anticipated by the platform.
- **Popularity Attributes:** These statistics gauge the post's popularity based on actual engagement metrics, which are beyond the user's control. This category also encompasses a score, computed by CrowdTangle through post analytics, to assess the post's effectiveness in terms of received interactions.

Figure 3.2 shows an example of this attributes for a single post of an italian politician.

## 3.4 Statistical Analysis

The dataset, subdivided as explained in sec. 3.3, can be therefore characterized from different points of view, by analyzing the statistics of the various features present in each record (post).

### 3.4.1 Follower characterization

One of the first analyses was to verify the distribution of influencers' followers at the time of CrowdTangle's sampling.

#### Politicians

Plotting the Empirical Cumulative Distribution Function (ECDF) of followers,figure 3.3a, it is noticeable that it can be approximated to a log-normal distribution with a peak of **300,000** and the vast majority less than a million: a predictable event since it's easy to imagine that only a handful of politician profiles can reach a large audience. Analyzing the distribution of followers by country is made easy through the use of boxplots[2], figure 3.3b. In this context, it's interesting to note that

[2]every boxplots in this chapter is visualizing values between the 5th and 95th percentiles

```
{'account': {...},
 'date': '2022-01-01 16:46:57',
 'description': 'Il primo gennaio, ma di qualche annetto fa… momento '
                'nostalgia. \n'
                'Auguro una buona serata e un buon inizio a tutti voi.',
 'history': [...],
 'id': '2738656|2741695695847245792',
 'languageCode': 'it',
 'legacyId': 0,
 'likeAndViewCountsDisabled': False,
 'media': [...],
 'platform': 'Instagram',
 'platformId': '2741695695847245792_257728207',
 'postUrl': 'https://www.instagram.com/p/CYMdbWRMUfg/',
 'score': 4.229016331181162,
 'statistics': {...},
 'subscriberCount': 970066,
 'type': 'photo',
 'updated': '2022-08-17 04:21:20'}
```


**(a)** data entity overview


```
{'id': 2738656,
 'name': 'Giorgia Meloni',
 'handle': 'giorgiameloni',
 'profileImage': 'url_profile_image',
 'subscriberCount': 1624870,
 'url': 'https://www.instagram.com/giorgiameloni/',
 'platform': 'Instagram',
 'platformId': '257728207',
 'verified': True}
```


**(b)** account attributes


```
[{'actual': {'favoriteCount': 0, 'commentCount': 0},
  'expected': {'favoriteCount': 930, 'commentCount': 36},
  'timestep': 0,
  'date': '2022-01-01 16:47:05',
  'score': -1932},
 {'actual': {'favoriteCount': 6002, 'commentCount': 153},
  'expected': {'favoriteCount': 2069, 'commentCount': 102},
  'timestep': 1,
  'date': '2022-01-01 17:02:04',
  'score': 2.8350990327038232},
 {'actual': {'favoriteCount': 9823, 'commentCount': 279},
  'expected': {'favoriteCount': 2899, 'commentCount': 150},
  'timestep': 2,
  'date': '2022-01-01 17:17:06',
  'score': 3.3132174483437193},
 {'actual': {'favoriteCount': 12818, 'commentCount': 397},
  'expected': {'favoriteCount': 3592, 'commentCount': 192},
  'timestep': 3,
  'date': '2022-01-01 17:32:05',
  'score': 3.492336152219873},
 {'actual': {'favoriteCount': 15228, 'commentCount': 493},
  'expected': {'favoriteCount': 4194, 'commentCount': 233},
  'timestep': 4,
  'date': '2022-01-01 17:47:06',
  'score': 3.5511633160153604}]
```


**(c)** platform-generated post attributes

**Figure 3.2:** Example of attributes for a single post of a data entity: a) data entity overview and **popularity attributes** reported in 'score' and 'statistics' fields while **user-generated post attributes** are stored in 'media' field; b) **account attributes** such as 'handle' for instagram username and 'subscriberCount' for number of followers; c) **platform-generated post attributes** such as the level of engagement both expected and actual

**Italian** and **British** politicians, on average, command the largest followings, with an average of around 400,000 followers. Additionally, the boxplot helps us identify outliers, which, while traditionally associated with statistical theory, in our specific use case, represent profiles with the highest number of followers. Remarkably, only the **United Kingdom**, **Italy**, and **France** can claim to have at least one **mega influencer with over 1 million followers**.

| | Mean | Std Deviation | Minimum | Maximum |
|---|---|---|---|---|
| General | 38,487.27 | 168,846.28 | 37 | 3,234,827 |
| Grouped by country | | | | |
| France | 72,825.38 | 331,731.63 | 107 | 3,234,827 |
| Germany | 10,923.15 | 45,593.45 | 232 | 578,746 |
| Italy | 115,455.90 | 305,712.55 | 359 | 2,204,098 |
| Spain | 29,574.40 | 86,099.27 | 37 | 845,198 |
| United Kingdom | 71,519.78 | 211,103.72 | 302 | 1,740,667 |

**Table 3.6:** Statistical description of the followers distribution for politicians profiles

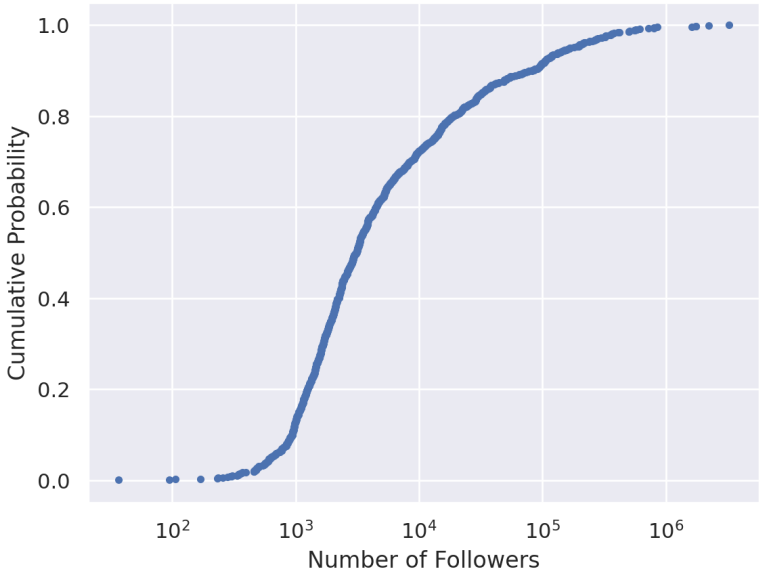
**Entertainment Figures**

- **Athletes:** By means of ECDF, figure 3.4a, we observed that a substantial

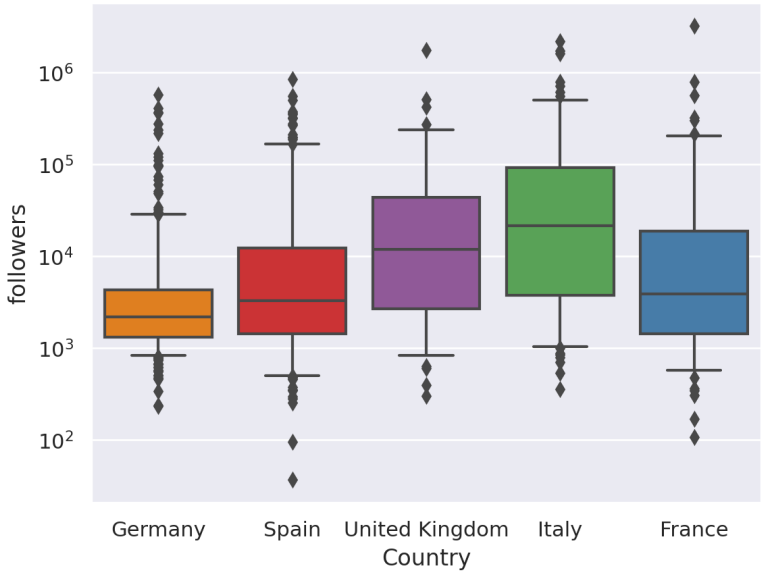
| | Mean | Std Deviation | Minimum | Maximum |
|---|---|---|---|---|
| General | 742,184.09 | 4,397,142.42 | 37 | 104,270,634 |
| Grouped by country | | | | |
| France | 665,059.28 | 5,542,154 | 126 | 104,270,634 |
| Germany | 509,383.28 | 2,793,117 | 52 | 39,839,517 |
| Italy | 548,608.82 | 1,774,690 | 328 | 15,480,143 |
| Spain | 983,369.48 | 4,359,250 | 37 | 58,770,957 |
| United Kingdom | 938,249.15 | 4,755,548 | 299 | 79,676,641 |

**Table 3.7:** Statistical description of the followers distribution for athletes profiles

  majority of athlete profiles have a relatively modest number of followers, indicating that a significant portion of athletes maintains a limited online following. However, what caught our attention is the steep upward slope on the right side of the ECDF chart. This steep rise signifies a small but significant group of athletes with an exceptionally high number of followers. These outliers represent sports superstars who have managed to amass an enormous digital following. Segmenting by country, figure 3.4b, revealed some noteworthy trends. Athletes from Italy and the United Kingdom exhibited the highest average follower counts among all countries. This finding suggests that athletes from these countries have been particularly successful in cultivating strong and engaging online profiles. On the other hand, while Spain's average




**(a)**




**(b)**

**Figure 3.3:** (a) ECDF of number of followers for politicians profiles; (b) boxplot of number of followers for politicians profiles grouped by country

follower count fell below that of Italy and the United Kingdom, it displayed a remarkable variability in follower counts. This variability suggests a more diverse online presence among Spanish athletes, with some garnering significant

followings while others maintain smaller but engaged audiences. France and Germany, though not leading in terms of average follower counts, presented a more uniform distribution of followers among their athletes. While their average counts may be lower, there is less disparity, indicating a more even distribution of digital influence. One noteworthy observation across all countries was the presence of outliers—athletes with exceptionally high follower counts. For instance, France boasted an outlier with over 100 million followers, highlighting the existence of sports icons with massive online followings.

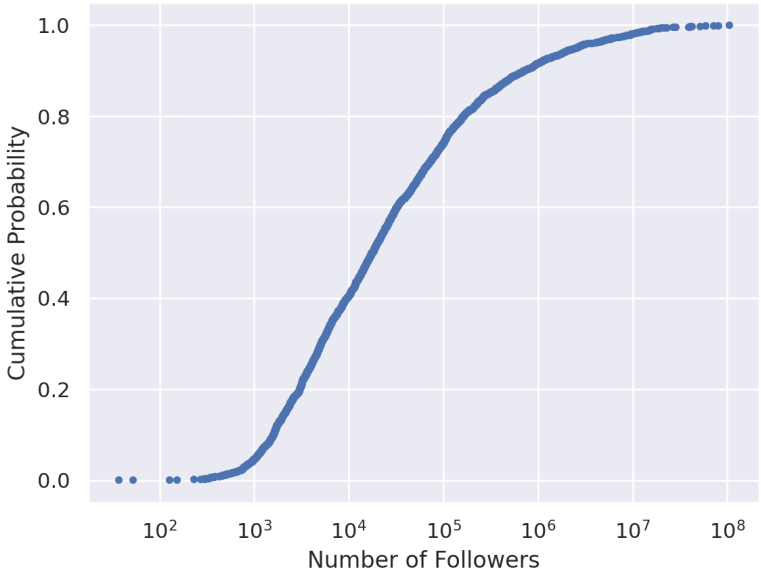
- **Models:** The majority of model profiles exhibit relatively modest follower

| | Mean | Std Deviation | Minimum | Maximum |
|---|---|---|---|---|
| General | 2,100,043.17 | 8,016,274.44 | 144 | 88,211,725 |
| Grouped by Country | | | | |
| France | 752,753.28 | 2,027,427 | 3,599 | 18,720,988 |
| Germany | 666,543.40 | 1,168,914 | 144 | 5,484,503 |
| Italy | 1,003,012.24 | 1,989,171 | 3,832 | 15,954,272 |
| Spain | 1,423,445.29 | 5,122,374 | 373 | 49,502,138 |
| United Kingdom | 4,697,117.52 | 13,573,468 | 1,549 | 88,211,725 |

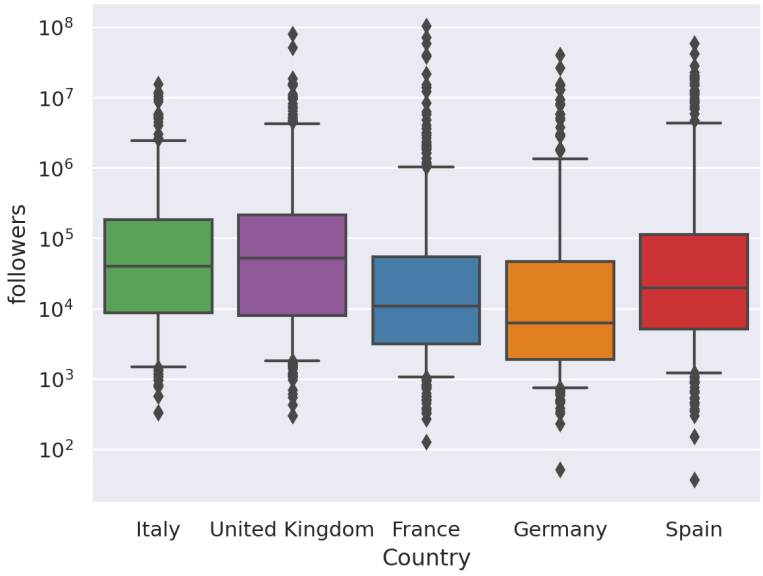
**Table 3.8:** Statistical description of the followers distribution for models profiles

counts, indicating that a significant portion of models maintains a limited online following. However, what particularly stands out is the steep ascent on the right side of the ECDF chart, figure 3.5a. Models from Italy and the United Kingdom showcase the highest average follower counts among all countries. While Spain's average follower count falls below that of Italy and the United Kingdom, it displays remarkable variability in follower counts, indicated by the length of the boxplot, figure 3.5b. This variance suggests a more diverse online presence among Spanish models, with some garnering substantial followings while others maintain smaller yet engaged audiences. France and Germany, while not leading in terms of average follower counts, exhibit a more uniform distribution of followers among themselves.

- **Actors:** The ECDF, figure 3.6a, reveals that a substantial majority of actor profiles have a relatively modest number of followers, indicative of a widespread but moderate online presence among actors. However, the plot's right tail rises sharply, indicating a select group of actors with an exceptionally high number of followers. These outliers represent the digital superstars in the acting world who have amassed millions of followers. Based on their home nations, figure



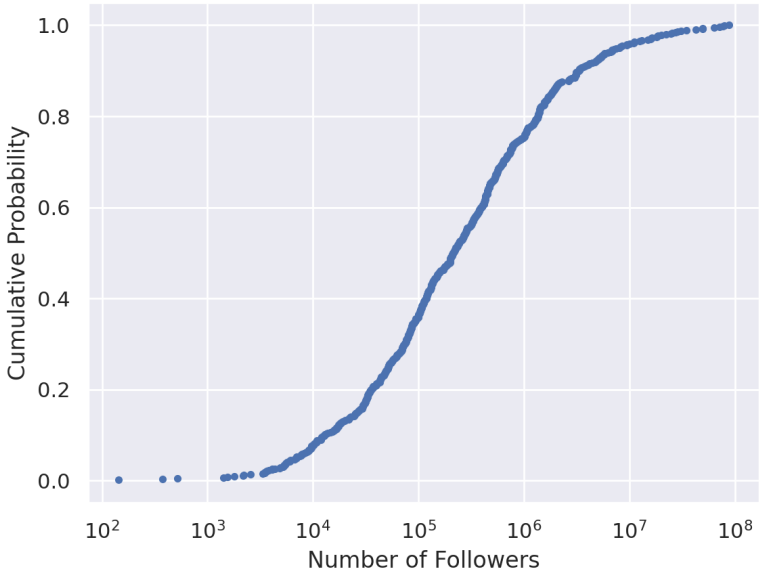

**(a)**



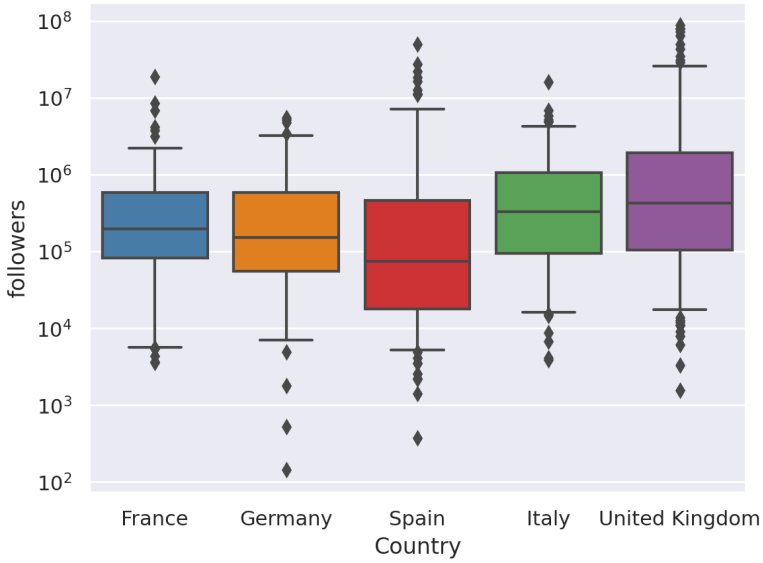

**(b)**

**Figure 3.4:** (a) ECDF of number of followers for athletes profiles; (b) boxplot of number of followers for athletes profiles grouped by country

3.6b, actors from Italy and the United Kingdom exhibit the highest average follower counts among all countries while spanish ones do not lead but display remarkable variability in terms of followers counts. This variance suggests




**(a)**




**(b)**

**Figure 3.5:** (a) ECDF of number of followers for models profiles; (b) boxplot of number of followers for models profiles grouped by country

a more diverse online presence among Spanish actors, with some garnering substantial followings while others maintain smaller yet engaged audiences. France and Germany, instead, present less disparity indicating a more even

| | Mean | Std Deviation | Minimum | Maximum |
|---|---|---|---|---|
| General | 941,139.43 | 4,577,860.53 | 14 | 88,345,088 |
| Grouped by country | | | | |
| France | 403,222.50 | 1,263,845 | 187 | 18,415,929 |
| Germany | 465,702.61 | 1,082,070 | 32 | 7,532,995 |
| Italy | 574,286.71 | 1,428,236 | 93 | 15,930,700 |
| Spain | 559,735.21 | 2,296,798 | 161 | 27,485,428 |
| United Kingdom | 1,841,272.67 | 7,529,902 | 14 | 88,345,088 |

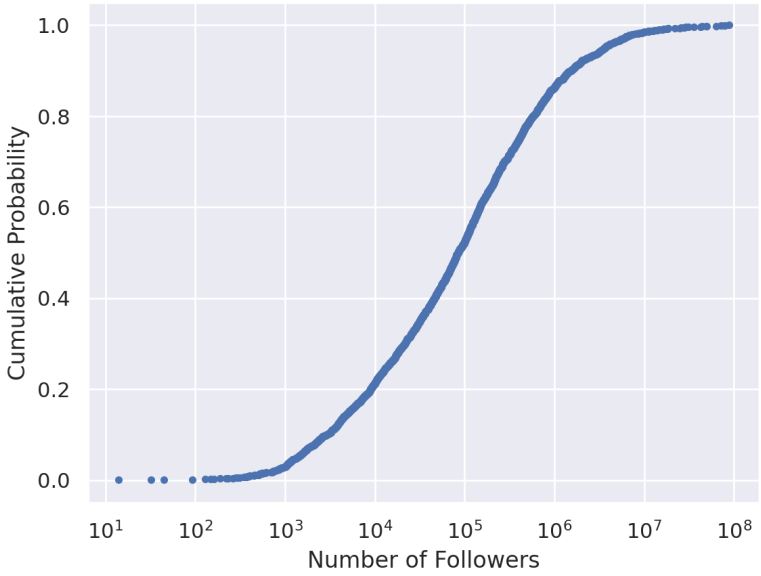
**Table 3.9:** Statistical description of the followers distribution for actors profiles

distribution of digital influence. In addition, all countries boast a substantial number of profiles with over 10 million followers, emphasizing the presence of real social media icons.

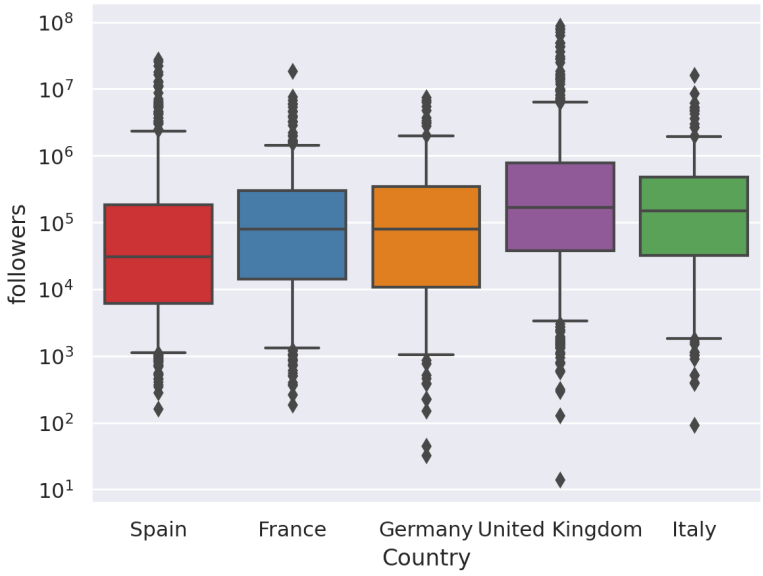
In examining the online presence of politicians and entertainment figures, we uncover an intriguing and diversified landscape where distinctions between these categories manifest in their interactions with the audience and geographic dynamics.

Politicians, despite their engagement in a political context, exhibit significant variability in their average follower counts across different geographical regions. This suggests that their ability to engage the online audience may be strongly influenced by the political landscape and the utilization of social media platforms within their respective countries. For instance, Italy and the United Kingdom feature politicians with an average of approximately 400,000 followers, indicating a robust and engaging online presence. Conversely, Spain and France show lower averages, highlighting potential differences in the online engagement of politicians in these nations. On the other hand, entertainment categories, encompassing athletes, models, and actors, tend to garner higher average follower counts compared to politicians. This trend may be attributed to the intrinsic nature of entertainment, which attracts a broader and more diversified audience. For example, German, Spanish, and British actors boast significantly high follower averages, signifying a strong online following within these sectors.

Geographic variation among politicians and entertainment categories also presents intriguing insights. In Germany, despite having fewer politicians, the average follower count is notably high, suggesting an ability to reach a broader audience. In Spain, the variation is more extensive, indicating a diversified distribution of followers among politicians. Within entertainment categories, such as athletes, models, and actors, we also observe fascinating geographic variations. British actors, for instance, stand out with considerably higher average follower counts




**(a)**




**(b)**

**Figure 3.6:** (a) ECDF of number of followers for actors profiles; (b) boxplot of number of followers for actors profiles grouped by country

compared to actors from other countries. French athletes shine in terms of their online presence, implying a specific resonance of these categories in particular countries.

In summary, these statistical disparities underscore the complexity of online dynamics between politicians and entertainment figures. Politicians can be influenced by the political climate and social media usage in their respective nations, while entertainment categories reflect the specific industries and cultural scenes prevalent in different countries.

**Universities**

| | Mean | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|
| General | 25,506.77 | 44,262.99 | 429 | 462,930 |
| Grouped by country | | | | |
| France | 34,419.75 | 114,361.31 | 1,028 | 462,930 |
| Germany | 3,723.00 | 1,545.15 | 2,403 | 5,531 |
| Italy | 25,137.16 | 27,770.59 | 478 | 135,192 |
| Spain | 16,295.67 | 2,618.66 | 13,359 | 18,388 |
| United Kingdom | 25,386.92 | 26,946.42 | 429 | 177,840 |

**Table 3.10:** Statistical description of the followers distribution for universities profiles

The ECDF for the overall number of followers in European university Instagram profiles, figure 3.7a, displays most European universities have a relatively low number of followers, with an average of approximately 25,506 followers. This indicates that the online presence of many universities on Instagram has not yet reached a broad audience. However, some institutions stand out with a significantly higher number of followers, up to a maximum of 462,930. These universities represent exceptions and have succeeded in engaging a large online audience.

In the boxplot analysis by country, figure 3.7b, variations in the distribution of followers among European universities become apparent.

- Regarding **France**, there is significant variation in followers among universities, with some institutions having reached a substantial audience while others are still growing.
- In **Germany**, the distribution of followers among universities is more compact, characterized by a relatively low standard deviation. This suggests a more uniform distribution of the audience among the institutions.
- **Italian universities** exhibit significant variation in followers, with some institutions having garnered a considerable audience and others less so.

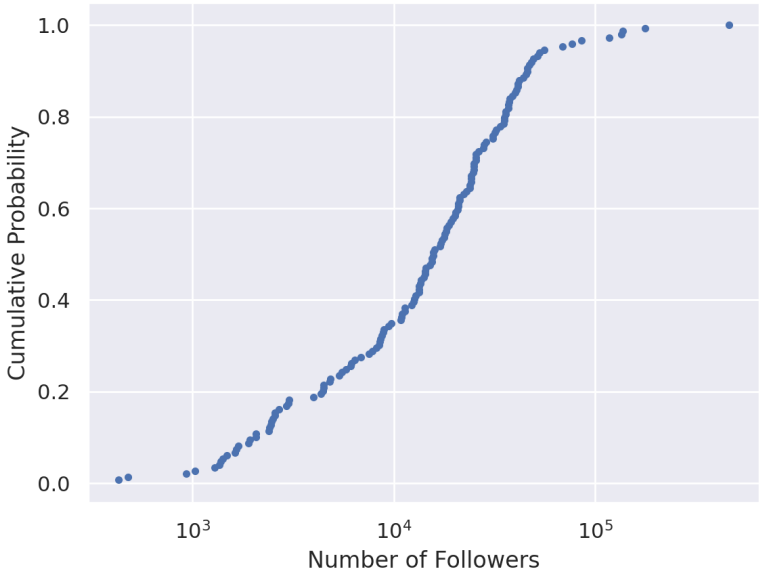
- **Spain** presents relatively low variation in followers, with a narrower range compared to some other countries.
- Within the **United Kingdom**, some universities have reached a very wide audience, as indicated by the high standard deviation, while others have a more evenly distributed following.

### 3.4.2 Posts characterization: Typology and Description

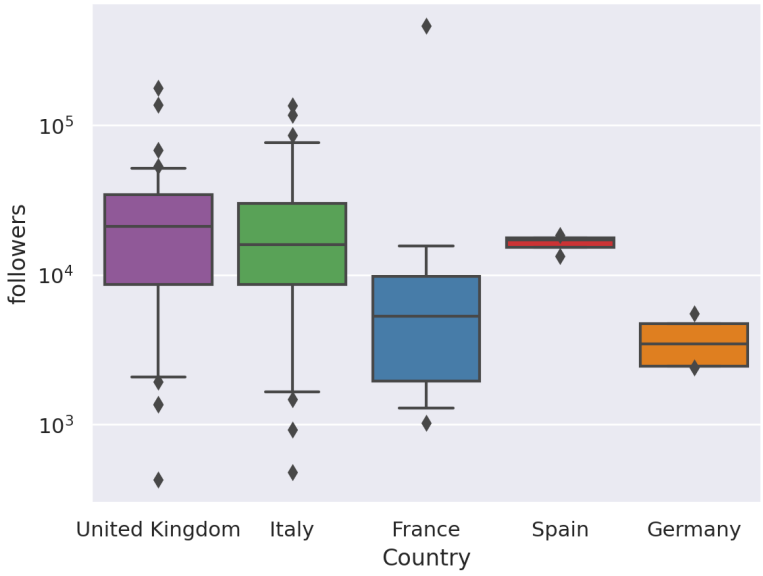
The information provided by CrowdTangle on each individual record (post) allows us to analyze various distinctive aspects of the Instagram platform. Among these, we find the manner in which a post can be published, either as an image, video, or album. By "album", we refer to the concatenation of images and/or videos within a single post, figure 3.8. Additionally, there is the "post description," a brief text of up to 2200 characters used to represent and elaborate on the post's content, tag other users or pages, and include hashtags.

#### Politicians

A unique viewpoint on online political communication practices is provided by the barplot, figure 3.9, that shows the amount of Instagram posts made by politicians in the chosen nations. It demonstrates how politicians in these various countries use slightly different tactics to interact with the public on this social media site. Let's begin with Germany, where there is a definite predilection for "photo-type posts." This implies that German politicians find it effective to convey messages and engage their audience through the use of visuals. While they are less frequent, posts of the "album" and "video" types are also available. In Spain, we observe that the politicians in this nation post more "album"-style content. This decision shows a desire to present collections of photographs that are relevant to political activity or to provide more comprehensive narrative. While "video" type posts are less cPost Analysisommon, they are just as important as "photo" type posts. We see that the total quantity of Instagram posts is lower in the UK than it is in other nations. Politicians here much favor "photo" posts over "video," by a wide margin. Less "Albums" being created, which reflects the popularity of sharing single photographs. The balance between "photo" and "album" post kinds is on the rise in Italy. This shows that Italian politicians are interested in sharing both individual photos and photo collections that are connected to their political endeavors. There are also "video" posts, but they are less common than the other two varieties. Finally, we see that among the countries taken into consideration, France has the most "photo" type posts. This suggests that French politicians strongly prefer to communicate aesthetically appealing photographs.




**(a)**




**(b)**

**Figure 3.7:** (a) ECDF of number of followers for universities profiles; (b) boxplot of number of followers for universities profiles grouped by country

The boxplot in figure 3.10 illustrates the variety in online political communication by highlighting the disparities in the average description lengths between nations. Parallel to this, by analyzing the numerical data from the table 3.11, we can delve

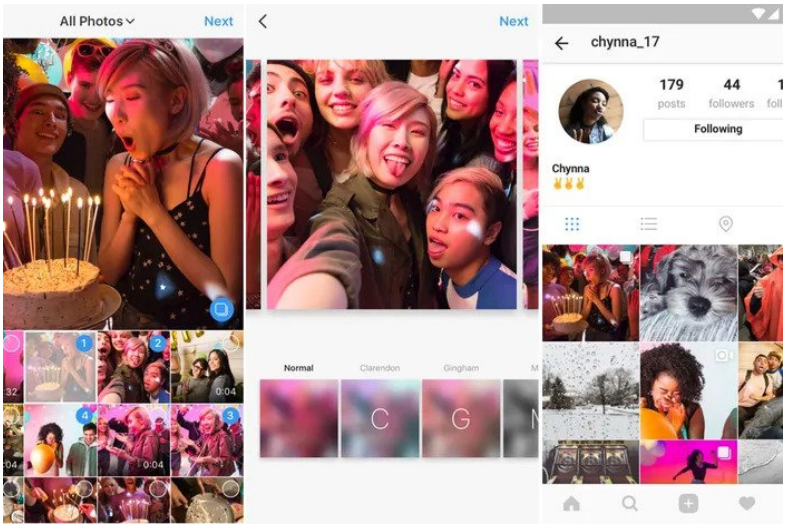



**Figure 3.8:** How to publish an album as post [36]

into the make-up of post descriptions in each nation, concentrating on mentions, hashtags, and simple words as three crucial components. In France, the average length of descriptions is about 322 characters. The average presence of mentions per post is about 0.559, with some descriptions containing up to 50 mentions. For hashtags, the average is about 2,076, with up to 30 hashtags per post. For simple words, the average is about 49,732, with some posts containing up to 407 simple words. For Germany, the average length of descriptions is about 636 characters. The average occurrence of mentions per post is about 0.769, with some descriptions containing up to 45 mentions. For hashtags, the average is about 4,764, with up to 36 hashtags per post. For simple words, the average is about 80,477, with some posts containing up to 352 simple words. In Italy, the average length of descriptions is about 440 characters. The average presence of mentions per post is about 0.454, with some descriptions containing up to 28 mentions. For hashtags, the average is about 1,639, with up to 30 hashtags per post. For simple words, the average is about 65,673, with some posts containing up to 379 simple words. Spain shows an average length of descriptions of about 323 characters. The average presence of mentions per post is about 0.926, with some descriptions containing up to 51 mentions. For hashtags, the average is about 2,042, with up to 30 hashtags per post. For simple words, the average is about 49,717, with some posts containing up to 421 simple words. In the United Kingdom, the average length of descriptions is about 246 characters. The average occurrence of mentions per post is about 0.430, with some descriptions containing up to 39 mentions. For hashtags, the average is about 1,190, with up to 31 hashtags per post. For simple words, the average is about 40,784, with some posts containing up to 392 simple words.

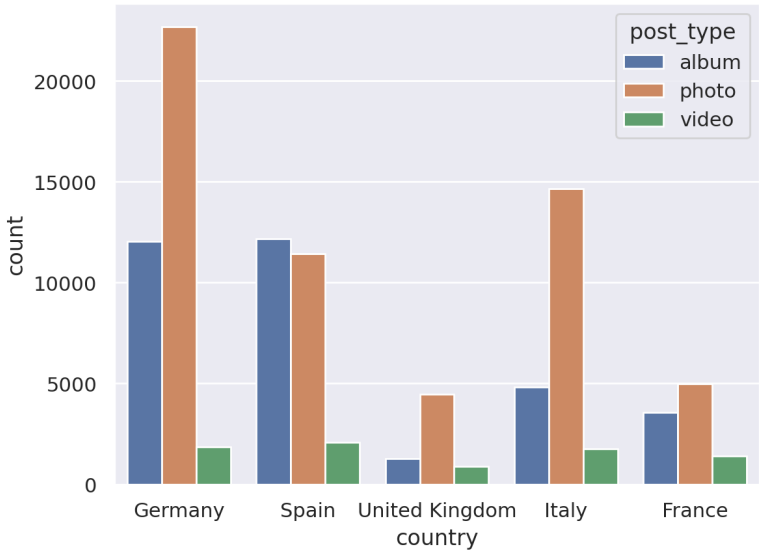



**Figure 3.9:** Distribution of politicians post per typology of post grouped by country

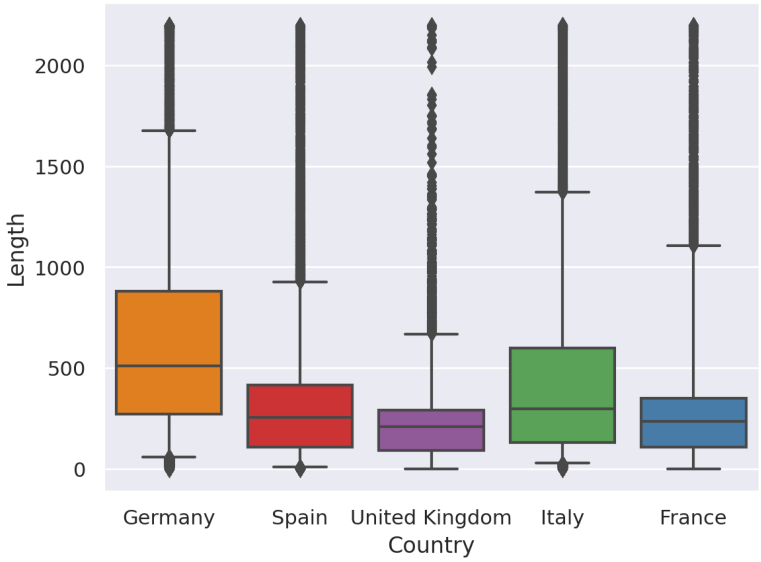



**Figure 3.10:** Distribution of politicians post length grouped by country

### Entertainment Figures

- **Athletes:** To understand athletes' content preferences, we analyzed the type of posts published by each country. The results were represented through a

| | | Mean | Std. Deviation | Min | Max |
|---|---|---|---|---|---|
| Germany | Mention | 0.769 | 1.557 | 0 | 45 |
| | Hashtag | 4.764 | 5.641 | 0 | 36 |
| | Simple Words | 80.477 | 65.534 | 0 | 352 |
| Spain | Mention | 0.926 | 1.902 | 0 | 51 |
| | Hashtag | 2.042 | 3.109 | 0 | 30 |
| | Simple Words | 49.717 | 50.814 | 0 | 421 |
| United Kingdom | Mention | 0.430 | 1.190 | 0 | 39 |
| | Hashtag | 1.190 | 3.194 | 0 | 31 |
| | Simple Words | 40.784 | 39.713 | 0 | 392 |
| Italy | Mention | 0.454 | 1.098 | 0 | 28 |
| | Hashtag | 1.639 | 3.227 | 0 | 30 |
| | Simple Words | 65.673 | 65.541 | 0 | 379 |
| France | Mention | 0.559 | 1.458 | 0 | 50 |
| | Hashtag | 2.076 | 3.337 | 0 | 30 |
| | Simple Words | 49.732 | 57.147 | 0 | 407 |

**Table 3.11:** Statistical description of the words distribution in politicians post, divided by category, for each post, grouped by country.

barplot, figure 3.11, which gives us a snapshot of posting trends in terms of albums, photos and videos.

- **France:** With over 10,000 albums and photos published, France demonstrates a strong visual presence. Video posting is relatively low, with about 770 videos;
- **Germany:** It shows a balanced distribution between albums, photos and videos, with reasonable numbers in each category.
- **Italy:** Italy focuses mainly on albums and photos, with significant numbers of both. Videos are less common, with fewer than 1,000 publications.
- **Spain:** Spain shows a strong preference for albums and photos, with over 10,000 publications in both categories. Videos are also fairly common, with over 1,000 publications.
- **United Kingdom:** The United Kingdom ranks among the countries with a lower social media presence. The distribution of posts between albums, photos and videos is balanced, but with lower numbers than other countries.

We also looked at the boxplot, figure 3.4b that showed the length of post descriptions. France and Germany show similar distributions for the length of post descriptions, with averages around 200 characters. France has a slightly higher standard deviation, indicating greater variability in description length. Italy and Spain also show similar distributions, with slightly lower averages than France and Germany. Spain has a slightly higher standard deviation, suggesting greater variability in description lengths. Finally the United Kingdom shows a similar distribution to Italy and Spain, but with slightly lower averages. The standard deviation is relatively low, indicating greater consistency in description lengths. To assess interaction and content optimization, we further examined the frequency of mentions, hashtags and "simple words" in post descriptions, table 3.12. Germany has a higher average frequency of mentions than the other countries, with an average value of 0.928. The frequency of hashtags is also relevant, with an average value of 3.455. However, the frequency of "simple words" is slightly lower than in France and Spain, with an average of 25,589. Spain shows a significant frequency of mentions, with an average value of 1,556, and a hashtag frequency similar to that of Germany, with an average of 1,809. The frequency of "simple words" is slightly higher than in Germany, with an average of 29,097. In the United Kingdom, the frequency of mentions and hashtags is moderate, with average values of 1,016 and 1,398 respectively. The frequency of "simple words" is similar to Germany, with an average of 26,898. Italy has a moderate frequency of mentions, with an average value of 1,122, and a slightly higher frequency of hashtags than the United Kingdom, with an average of 2,858. The frequency of "simple words" is lower than the other countries, with an average of 24,004. France has a similar frequency of mentions to Italy, with an average value of 1,477, and a higher frequency of hashtags than Italy and the United Kingdom, with an average of 2,399. The frequency of "simple words" is similar to Spain, with an average of 27,824.

- **Models:** By analyzing the types of posts published, figure 3.13, the length of post descriptions, figure 3.14 and the frequency of mentions, hashtags, and "Simple Words" in post descriptions, table 3.13, we come to the following conclusions:

  – Approximately 66 percent of all posts published by European models are images, making them the most common type of post. With 69% of articles using photographs, the United Kingdom stands out among other countries, followed by Spain (64%), France (63%), Germany (62%), and Italy (61%). About 28% of all publications are albums, making them the second most prevalent type of post. With 30% of the albums posted, France is in first place here, followed by Spain (29%), Germany (28%), the


**Figure 3.11:** Distribution of athletes post per typology of post grouped by country

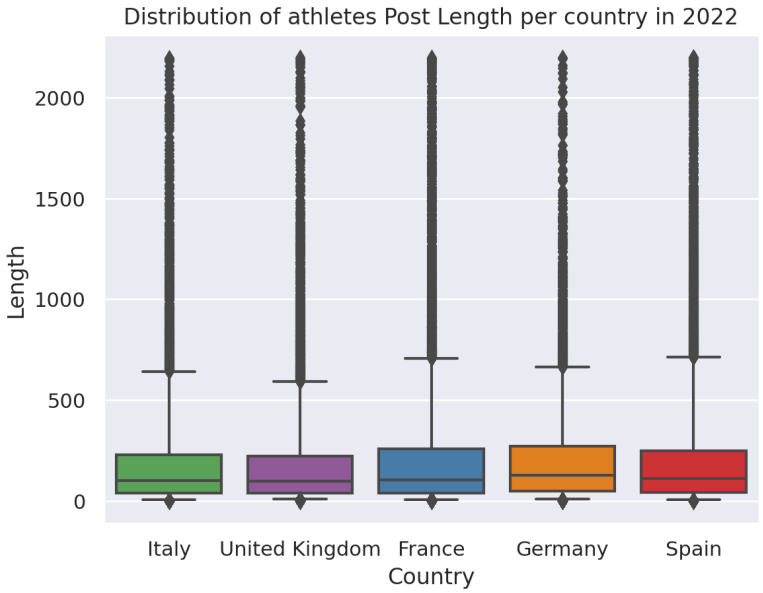



**Figure 3.12:** Distribution of athletes post length grouped by country

United Kingdom (26%), and Italy (26%). About 6% of publications are videos. This is a very tiny amount. The UK exhibits the lowest proportion of videos here, with only 5% of posts in the form of videos, followed by Spain (6%), Italy (6%), France (7%) and Germany (10%).

| | | Mean | Std. Deviation | Min | Max |
|---|---|---|---|---|---|
| Germany | Mention | 0.928 | 1.596 | 0 | 21 |
| | Hashtag | 3.455 | 4.901 | 0 | 33 |
| | Simple Words | 25.589 | 35.600 | 0 | 421 |
| Spain | Mention | 1.556 | 3.671 | 0 | 42 |
| | Hashtag | 1.809 | 3.607 | 0 | 31 |
| | Simple Words | 29.097 | 43.078 | 0 | 417 |
| United Kingdom | Mention | 1.016 | 2.332 | 0 | 34 |
| | Hashtag | 1.398 | 3.364 | 0 | 31 |
| | Simple Words | 26.898 | 37.165 | 0 | 427 |
| Italy | Mention | 1.122 | 2.830 | 0 | 69 |
| | Hashtag | 2.858 | 5.182 | 0 | 48 |
| | Simple Words | 24.004 | 38.412 | 0 | 384 |
| France | Mention | 1.477 | 2.819 | 0 | 41 |
| | Hashtag | 2.399 | 4.389 | 0 | 31 |
| | Simple Words | 27.824 | 43.574 | 0 | 402 |

**Table 3.12:** Statistical description of the words distribution in athletes posts, divided by category, for each post, grouped by country.

  - The length of post descriptions vary depending on the nation. However, in comparison to other nations, French and British models typically write longer descriptions. For instance, post descriptions in France are typically 202 words long, whereas they are 207 words long in the UK. Despite having a substantially greater follower base than most other nations, Italy has descriptions that are, on average, shorter, at around 186 words. Spain and Germany both maintain comparable description length averages, with Spain's coming in at roughly 203 words and Germany's coming in at roughly 213 words.
  - There are huge regional differences in the volume of hashtags and mentions. The majority of "simple words" used in descriptions come from the United Kingdom (approximately 54%), followed by France (53%). In comparison to the other nations, Germany and Spain utilize hashtags more frequently, making up 14% and 13% of all terms, respectively. With 18% of the total words in their descriptions, hashtags and mentions are often employed in moderate amounts in France and Italy.

- **Actors:** Analyses indicate considerable variation among countries in terms of

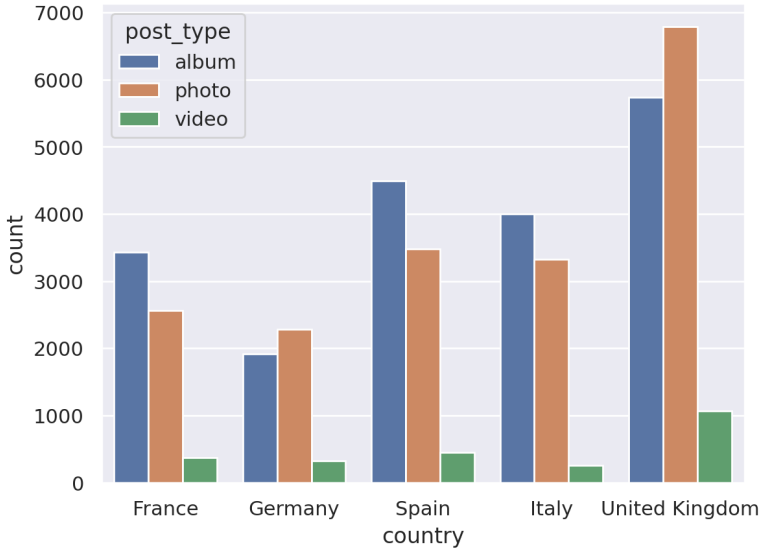



**Figure 3.13:** Distribution of models post per typology of post grouped by country

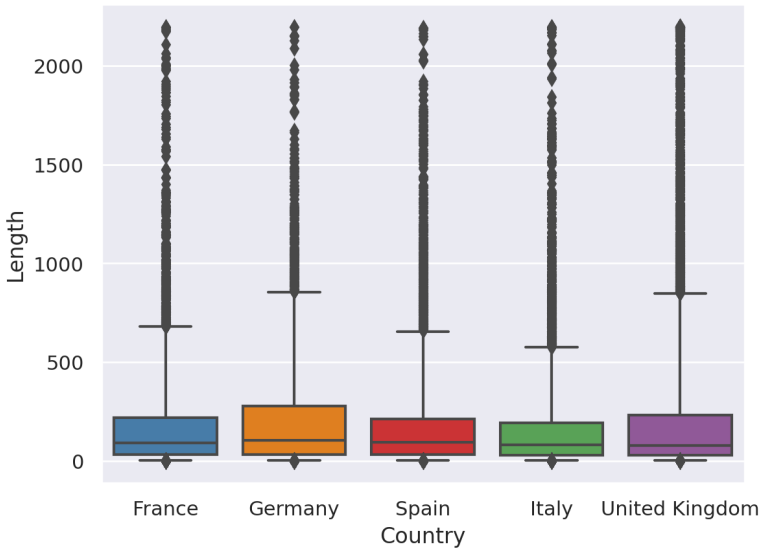



**Figure 3.14:** Distribution of models post length grouped by country

post types, figure 3.15, length of descriptions, figure 3.16, and word frequencies, table 3.14. Spain, for instance, shares primarily photographs and has the longest captions on average, whereas Italy has shorter descriptions. While France features descriptions that are longer than average, the United Kingdom stands out for having a high percentage of photographs and medium-length

| | | Mean | Std. Deviation | Min | Max |
|---|---|---|---|---|---|
| France | Mention | 1.230 | 2.277 | 0 | 37 |
| | Hashtag | 1.724 | 3.762 | 0 | 30 |
| | Simple Words | 27.067 | 44.982 | 0 | 410 |
| Germany | Mention | 1.027 | 2.145 | 0 | 52 |
| | Hashtag | 3.465 | 5.401 | 0 | 34 |
| | Simple Words | 28.081 | 43.357 | 0 | 408 |
| Spain | Mention | 1.512 | 3.208 | 0 | 57 |
| | Hashtag | 2.113 | 4.184 | 0 | 31 |
| | Simple Words | 23.992 | 40.454 | 0 | 409 |
| Italy | Mention | 0.996 | 2.104 | 0 | 25 |
| | Hashtag | 1.865 | 4.122 | 0 | 43 |
| | Simple Words | 21.956 | 37.701 | 0 | 418 |
| United Kingdom | Mention | 1.112 | 2.237 | 0 | 60 |
| | Hashtag | 1.105 | 3.111 | 0 | 30 |
| | Simple Words | 34.074 | 60.020 | 0 | 442 |

**Table 3.13:** Statistical description of the words distribution in models posts, divided by category, for each post, grouped by country.

explanations. In Germany, albums are the most popular media type, and descriptions tend to be lengthy with a lot of hashtags. When it comes to hashtags in descriptions, Germany seems to take special note. In details:

- With only a small number of videos (3,428), Spain produced mostly albums (14,451) and photos (19,943). The most prevalent format in France was photos (10,269), followed by albums (6,128) and videos (2,189). The most prevalent category in Germany were albums (3,987), followed by photos (6,076) and videos (1,010). The most common category in the UK were photos (29,089), followed by albums (16,283) and videos (6,711). A lot more albums (7,897) and photos (10,546) than videos (1,996) were published in Italy.
- All countries present a median of lengths between about 200 and 240 words, and all have a definable wide distribution presenting in fact a standard deviation between 290 and 310 words. Germany and Spain in particular have the largest medians, 257 and 236 words respectively while UK and Italy are the countries with the smallest medians.
- It is evident that the word count, hashtag usage, and frequency of mentions

vary significantly in these European athletes' post descriptions. The highest average number of mentions appears to be in Spain (1.82), whereas the highest average number of hashtags appears to be in Germany (3.11). The average number of simple words is lowest in Italy (27.54), whereas the average number of mentions and hashtags is lowest in the UK (1.16 and 1.85, respectively).

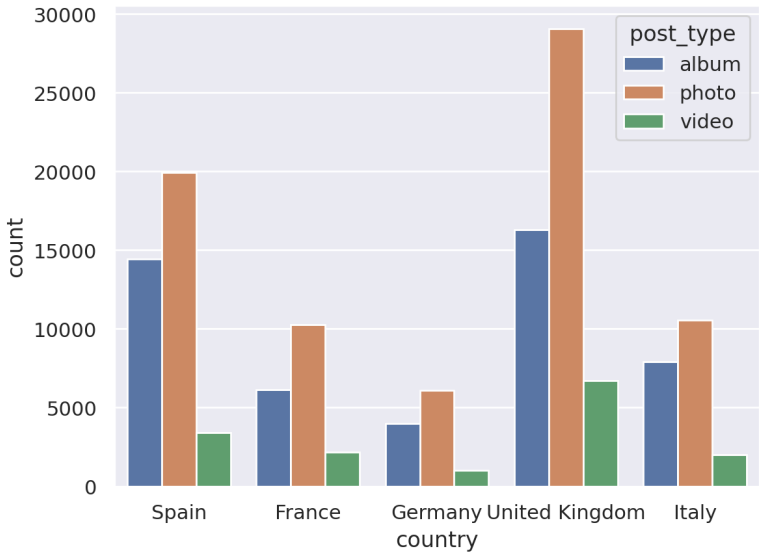



**Figure 3.15:** Distribution of actors post per typology of post grouped by country

### Universities

The postings published by profiles belonging to European universities are represented by the barplot in figure 3.17. With France having the lowest number (545) and the UK having the greatest value (3320), the "album" category is the least prevalent throughout the nations. The UK has the highest value (9690) while Germany has the lowest value (385), making the "photo" typology the most prevalent. The value of the "video" kind is in the middle, with Germany having the lowest value (50) and Italy having the greatest value (703). The distribution of post description lengths broken down by country is seen in the boxplots in figure 3.18. France has the shortest average length (470.32 characters), and the United Kingdom comes in second (394.58 characters). Spain (713.37 characters), Italy (642.50 characters), and Germany (936.80 characters) are the countries with the longest average character counts. Germany has the biggest standard deviation, which suggests that post descriptions vary more widely than other countries. Regarding the frequencies of word types in post descriptions, table 3.15, we can see

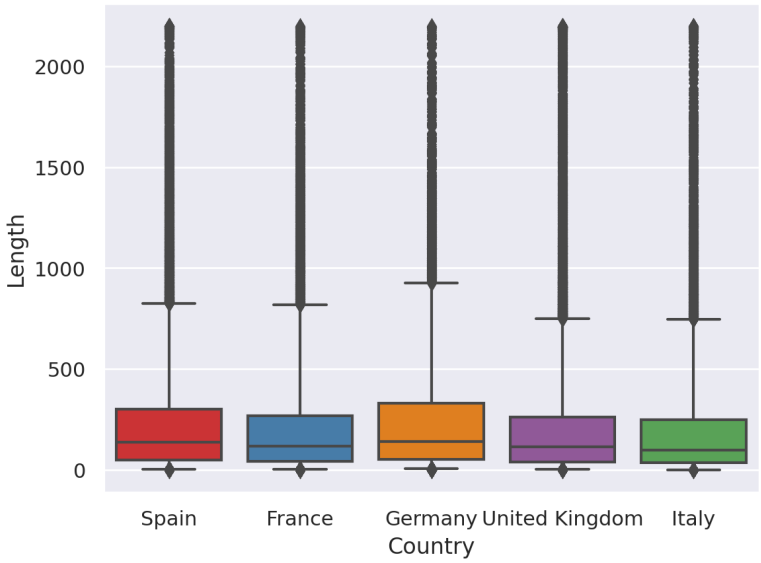



**Figure 3.16:** Distribution of actors post length grouped by country

| | | Mean | Std. Deviation | Min | Max |
|---|---|---|---|---|---|
| Spain | Mention | 1.823 | 3.471 | 0 | 76 |
| | Hashtag | 2.324 | 4.126 | 0 | 35 |
| | Simple Words | 33.446 | 48.426 | 0 | 417 |
| France | Mention | 1.503 | 3.121 | 0 | 70 |
| | Hashtag | 2.378 | 4.667 | 0 | 34 |
| | Simple Words | 31.731 | 49.453 | 0 | 429 |
| Germany | Mention | 1.241 | 2.495 | 0 | 49 |
| | Hashtag | 3.113 | 4.819 | 0 | 34 |
| | Simple Words | 33.576 | 47.179 | 0 | 379 |
| United Kingdom | Mention | 1.163 | 2.341 | 0 | 60 |
| | Hashtag | 1.850 | 4.372 | 0 | 33 |
| | Simple Words | 32.577 | 50.075 | 0 | 439 |
| Italy | Mention | 1.237 | 2.339 | 0 | 36 |
| | Hashtag | 2.539 | 5.042 | 0 | 43 |
| | Simple Words | 27.537 | 45.687 | 0 | 393 |

**Table 3.14:** Statistical description of the words distribution in actors posts, divided by category, for each post, grouped by country.

some interesting trends among nations:

- **Mentions:** Spain has the highest average number of mentions in post descriptions (1,672), followed by Italy (1,044), France (1,196), and the United Kingdom (0,533). Germany has the lowest average number of mentions (0,510).

- **Hashtags:** Germany has the highest average of hashtags in post descriptions (8,046), followed by Spain (2,920), Italy (6,999), France (3,964), and the United Kingdom (3,975). France has the highest standard deviation, indicating greater variation in the frequency of hashtags.

- **Simple Words:** Germany has the highest average of simple words (131,702), followed by Spain (101,952), Italy (80,424), France (66,479) and the United Kingdom (56,106). Spain has the highest standard deviation, suggesting greater variation in the length of post descriptions.

The use of mentions, hashtags, and simple language in university post descriptions is generally noted to vary among European countries. While the usage of basic terms is more prevalent in Germany and Spain than in the other countries, both countries also tend to have higher mention and hashtag frequencies. For these categories, France and the United Kingdom display intermediate values.

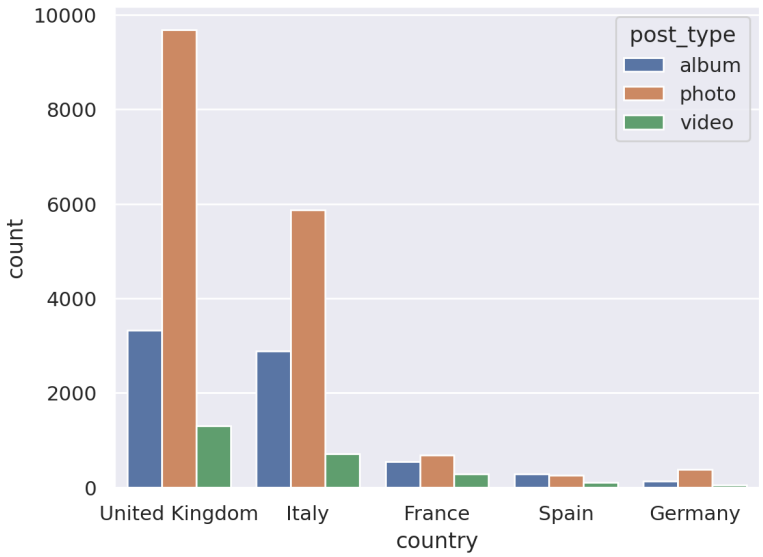



**Figure 3.17:** Distribution of universities posts per typology of post grouped by country

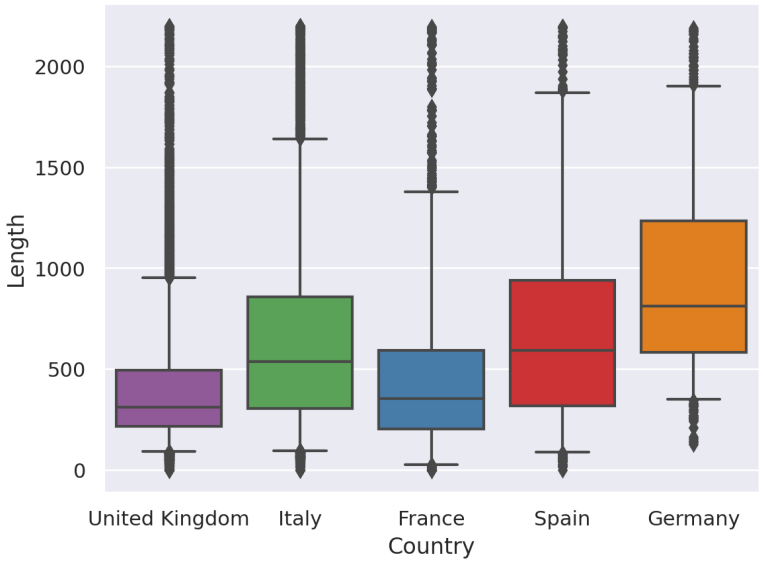



**Figure 3.18:** Distribution of universities posts length grouped by country

| | | Mean | Std. Deviation | Min | Max |
|---|---|---|---|---|---|
| United Kingdom | Mention | 0.533 | 1.396 | 0 | 35 |
| | Hashtag | 3.975 | 6.999 | 0 | 38 |
| | Simple Words | 56.106 | 66.472 | 0 | 408 |
| Italy | Mention | 1.044 | 2.084 | 0 | 30 |
| | Hashtag | 6.999 | 6.749 | 0 | 38 |
| | Simple Words | 80.424 | 66.472 | 0 | 365 |
| France | Mention | 1.196 | 2.420 | 0 | 27 |
| | Hashtag | 3.964 | 4.293 | 0 | 27 |
| | Simple Words | 66.479 | 65.267 | 0 | 360 |
| Spain | Mention | 1.672 | 3.635 | 0 | 30 |
| | Hashtag | 2.920 | 3.326 | 0 | 28 |
| | Simple Words | 101.952 | 80.371 | 1 | 356 |
| Germany | Mention | 0.510 | 1.378 | 0 | 14 |
| | Hashtag | 8.046 | 4.962 | 0 | 28 |
| | Simple Words | 131.702 | 71.267 | 7 | 356 |

**Table 3.15:** Statistical description of the words distribution in universities posts, divided by category, for each post, grouped by country.

### 3.4.3 Reaction characterization

The following essential parameters to analyze were the distribution of the number of reactions, i.e. likes (passive attention) and comments (active attention), to the published posts, to measure "influence" by inspecting engagement levels on a users' posts.[3]

#### Politcians

The boxplots, figure 3.19 show how passive (likes, figure 3.19a) and active (comments, figure 3.19b) interactions on political Instagram pages are distributed throughout the five European nations we study. It is crucial to take into account the background in terms of the follower bases of each nation for a greater understanding, table 3.6.

Despite having a substantially larger network of followers, Italy's distribution of passive interactions has a lower median than in some other countries. This shows that, despite having a large number of followers, the distribution of interactions is more concentrated in certain political profiles, while others may have lesser involvement. Moreover it exhibits a distribution of active contacts that corresponds to its huge fan base. However, the median number of active interactions may be fewer in Italy than in other countries with smaller follower bases, implying that online involvement in Italy may be more concentrated in certain political profiles.

Compared to other countries, the median of passive interactions in Spain is higher. This could indicate more widespread online political involvement or different political involvement strategies in Spain, taking into account the size of the supporter base. With a higher median and a higher number of active interactions, Spain might show more widespread online political involvement and a different political involvement strategy.

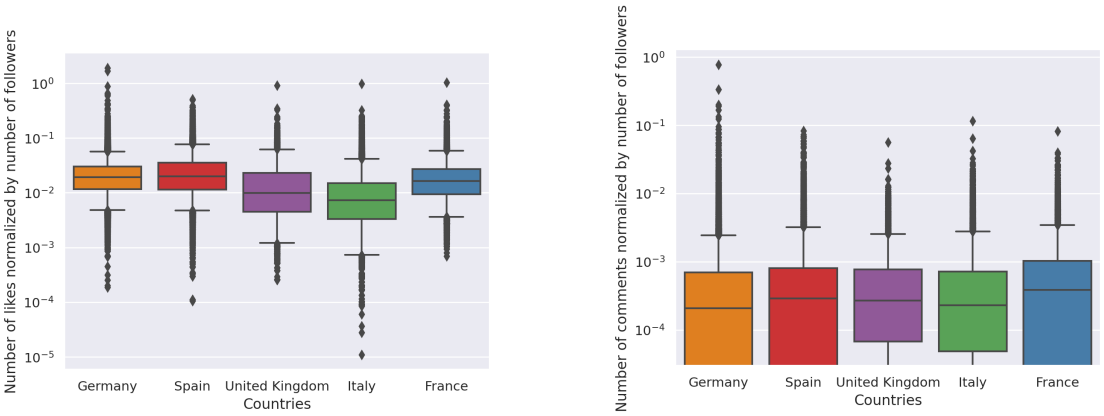
On the other hand, the United Kingdom displays a low median for both passive and active interactions. This finding may be explained by the unique dynamics of British political culture and the communication tactics used by politicians. The size of the follower base may have an impact on engagement dynamics in the United Kingdom, since the median number of active interactions is quite low.

The distribution of passive interactions in France is consistent with the size of the follower base, however the median of active interactions is slightly lower than in other nations, indicating slightly lesser online involvement. Also active interactions is consistent with the size of the follower base, but the median is slightly lower, indicating slightly lower online engagement.

[3]Likes and comments considered in the analysis are collected 24 hours after the post is published

Passive interactions in Germany are distributed similarly to those in the United Kingdom, but active interactions have a wider distribution and a higher median, indicating more widespread online participation and involvement of politicians with their supporters. Both the distribution and median of active interactions are wider in Germany, showing more widespread online political activity and political involvement of politicians with their following.

Focusing then on the types of influencers, Chapter1-figure 1.6, we can deepen the analysis on the interaction involving political profiles. In particular through the **Mann-Whitney U** statistic test we can point out possible statistical differences of passive and active interaction between two nations at a time. In fact, the Mann-Whitney U statistic test is a nonparametric test used to compare differences between two independent groups when the distributions of the data do not follow a normal distribution as in our case where the normalized likes and comments for followers do not follow a normal distribution precisely. The null hypothesis of the test states that, for randomly selected values X and Y from two populations, the probability of X being greater/smaller than Y is equal to the probability of Y being greater than X. The test was run using an upper single-tailed test with an upper **significance threshold of 0.01** and an **alternative hypothesis direction** equal to **"greater"** with the aim of identifying where the amount of passive and active interaction received from influencers with comparable followings is greatest. Spain receives more likes on Instagram than any other country, followed by Germany, France, Italy, and the United Kingdom. This is especially true when we take into account nano influencers. On the other hand, the results for active interaction, as measured by comments, are different: Italy takes first place, followed by Spain, France, the United Kingdom ex equo, and Germany in last place. This finding illustrates how, despite their lack of close appreciation for the posts in this tier, the followers of Italy's nanopolitical influencers generate more discussion than those of politicians from other European countries. For micro influencers, the findings are fairly similar, with Spain consistently placing first, followed by France, Germany, the United Kingdom, and Italy. For this category of influencers in terms of active interaction it is France that excels, followed by Italy, Uk, Germany and Spain. Interesting how Spain, which excels in passive interaction, presents a very small number of comments in relation to the remaining European profiles. On the other hand, mid-tier influencers believe that France, followed by Spain, Germany, the UK, and Italy, will be most well-liked. For this tier both passive and active interaction yield the same results. This begs the question of why likes and comments on a profile tend to remain steady as its following increases. Due to the limited number of profiles we have available by country, the macro and mega influencer categories were not examined.




**(a)** **(b)**

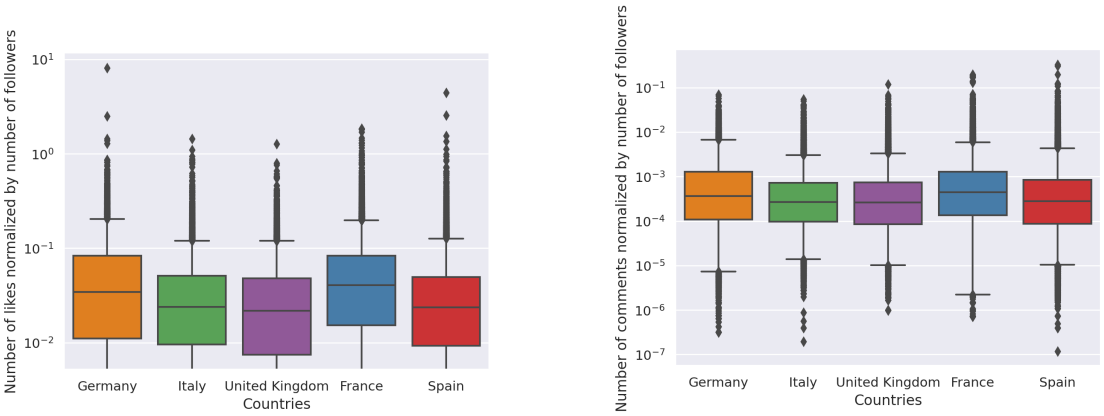
**Figure 3.19:** Distribution of interaction in politicians posts normalized by number of followers: a) likes on number of followers, b) comments on number of followers

### Entertainment Figures

- **Athletes:** The boxplots shown in figure 3.20 allow us to analyze the distribution of passive interaction, figures 3.20a, and active interaction, figures 3.20b, in the Instagram posts of athletes from the European states under consideration. On average, Italian athletes receive about 38.5 likes per 1,000 followers in their Instagram posts. This value is similar to that of Germany and Spain. However, it is worth noting that some Italian posts reached exceptionally high values, with a maximum of 1434 likes per 1000 followers. French athletes have the highest average among the nations considered, with about 62.5 likes per 1000 followers. However, there is considerable variability in interaction rates, with some posts exceeding 1800 likes per 1000 followers. The United Kingdom has a similar passive interaction rate to Italy, Germany and Spain, with an average of about 38.6 likes per 1000 followers. In terms of active interaction, on the other hand, French athletes have the highest average, with about 1.5 comments per 1000 followers. This value is significantly higher than that of other nations, indicating high follower involvement in comments. Spanish athletes follow France with an average of about 1.1 comments per 1000 followers. Although lower than France, this is higher than Italy, Germany and the United Kingdom.

  Dividing the profiles by influencer categories allows us to analyze in detail both active and passive interaction using the Mann-Whitney U test, with the same specifications used for the politician category. Specifically, we obtain the following information:

- **nano influencers:** the German athletes have more interaction than the other nations in terms of both likes and comments while the united kingdom is the state with the least interaction on this scale of followers. Italy and Spain stand out in that the former has a good percentage of passive interaction but a poor one for active interaction while in spain the diametrically opposite scenario occurs.
- **micro influencer:** Instead, for this tier Germany and France lead ex equo the ranking in terms of passive interaction while for passive it is Italy and Spain that have the best percentages in relation to the number of followers.
- **mid-tier influencers:** on these numbers of followers, the results are very similar to the previous tier with the only difference concerning French athletes who surpass German ones in both likes and comments.
- **macro influencers:** here it is the Italian athletes who find the most interaction in both terms followed by the Germans and the French. The British are the profiles that continue to show the least interaction.
- **mega influencers:** for this tier of influencers, however, it is the united kingdom that finds the highest percentages of likes and comments followed by spain and italy.




**(a)** **(b)**

**Figure 3.20:** Distribution of interaction in athletes posts normalized by number of followers: a) likes on number of followers, b) comments on number of followers

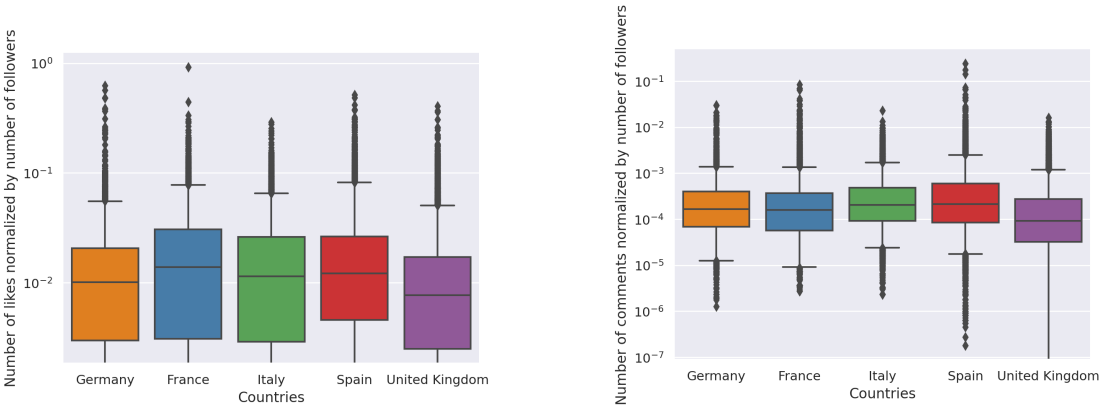
- **Models:** When we look at passive interaction on social media, as measured by the ratio of the number of likes to the number of followers, figure 3.21a, interesting differences emerge among European nations. French models stand

out with an average of about 23 likes per 1,000 followers, meaning that on average, each follower in France generates a significant number of likes. Right behind is the Spanish with an average of about 22.8 likes per 1000 followers, suggesting strong user engagement. Italy ranks immediately after France and Spain presenting very similar values in terms of passive interaction. On the other hand, the United Kingdom stands out with a lower average of about 14.1 likes per 1000 followers, indicating relatively lower engagement than the other countries. If we look at variability, Spain has the widest range of user behavior, with a standard deviation of roughly 31.6 likes per 1000 followers, according to the data. The standard deviation in Germany, on the other hand, is a little bit lower, indicating more consistency in passive contact.This is higher than France, Italy, Germany, and the UK combined. In terms of peak interaction, France dominates with an extraordinary maximum value of about 923 likes per 1000 followers, followed by Spain with about 517.4 likes per 1000 followers. In contrast, the United Kingdom shows a lower maximum value of about 405.2 likes per 1000 followers, suggesting that although average engagement is lower, there are still instances of high levels of interaction. Turning to active interaction, that is, the ratio of the number of comments to the number of followers, figure 3.21b, we notice interesting new trends among European nations. Spain leads the ranking with an average of about 0.741 comments per 1,000 followers, indicating that every Spanish follower is inclined to leave a comment. Following closely, France shows an average of about 0.492 comments per 1,000 followers, demonstrating significant active engagement. Germany, on the other hand, has a lower average of roughly 0.426 comments per 1000 followers, indicating comparatively weaker participation in terms of comments than the other countries. With a standard deviation of almost 4.2 comments per 1000 followers, Spain continues to stand up when we look at variability, indicating a wide range of user behavior in this category. With a spectacular maximum value of approximately 243.6 comments per 1000 followers, Spain once again leads in terms of peak active involvement. France comes in second with approximately 84.2 comments per 1000 followers. The maximum figure for Germany is lower, at roughly 30.3 comments per 1000 followers.

These statistics provide an overview of the dynamics of social media interaction among European nations, highlighting how France and Spain stand out for their significant involvement, immediately following Italy, while Germany and the United Kingdom demonstrate a more substantial involvement, both in terms of passive and active interaction.

The Mann-Whiteny U test for European models supports the boxplot statistics by demonstrating that the models from Germany and the United Kingdom engage with their followers the least, while influencers from Spain and France

receive the most likes and comments on their posts across all influencer tiers. Because there were no model profiles for the nano and micro influencer tiers, these follower scales were not included in the test.



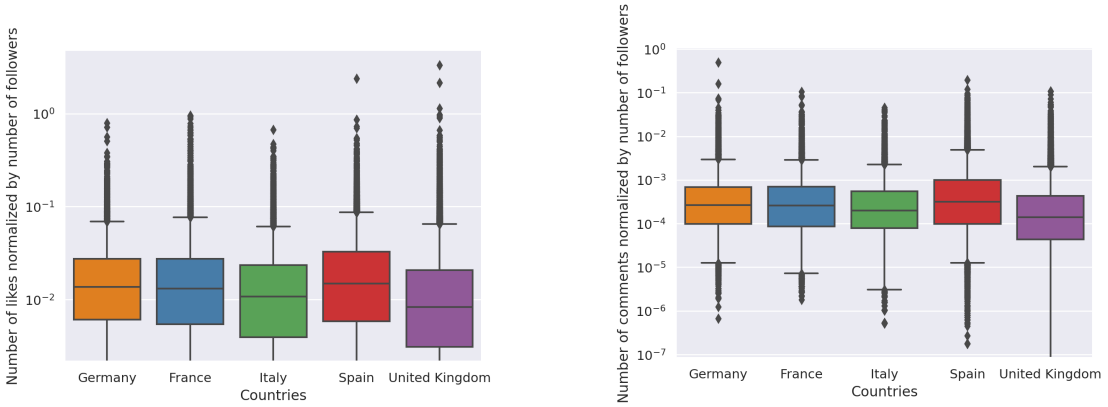

**(a)** **(b)**

**Figure 3.21:** Distribution of interaction in models posts normalized by number of followers: a) likes on number of followers, b) comments on number of followers

- **Actors:** Boxplots in figure 3.22 show the passive, figure 3.22a, and active, figure 3.22b, of european actors profiles. In details, Spain stands out for its high engagement in both categories. Spanish users show appreciable passive interaction, with an average of about 25.7 likes per 1000 followers, meaning that each follower tends to put "likes" on a considerable number of posts. In parallel, active interaction is equally notable, with an average of about 1.2 comments per 1000 followers, indicating that Spanish users are also likely to leave meaningful comments. Data variability is present in both categories, with some exceptionally high figures. France also stands out for substantial involvement in both categories. Passive interaction presents an average of about 23.3 likes per 1,000 followers, confirming that French users are inclined to express their appreciation through "likes." At the same time, active interaction, with an average of about 0.8 likes per 1000 followers, shows that French users actively participate in the discussion in comments. The distribution of the data is varied, with some notable peaks. Both in passive and active interactions, German actors retain a positive level of engagement. Active involvement is roughly 0.9 comments per 1000 followers compared to passive interaction's average of 21.8 likes per 1000 followers. In comparison to other countries, the data distribution is more condensed, which may indicate some consistency in user engagement. The UK is still actively involved, albeit slightly less so

than the other countries. In comparison to active interaction, which averages 0.5 comments per 1000 followers, passive interaction averages roughly 18.0 likes per 1000 followers. It is crucial to remember that the data exhibits variability, with some people displaying incredibly high involvement. Finally, Italy shows similar involvement to the United Kingdom. Passive interaction has an average of about 18.6 likes per 1000 followers, while active interaction is about 0.6 comments per 1000 followers. Variability is present, but in general, Italy shows positive engagement.

The distributions of boxplots are generally confirmed by analysis by influencer tiers using the Mann-Whitney U statistical test, especially for influencer categories at the extremes of the chain of followers. In the micro influencer category, Spain exhibits the highest levels of passive and active contact, whilst France stands out in the mega influencer category. While obtaining impressive results for both types of engagement, Germany and Italy stand out for greater interaction in the categories of mid-tiers and macro influencers, respectively. However, the examination of micro influencers is particularly intriguing, showing that while Spanish actors experience the highest levels of active contact with a high number of comments per follower, Germany and Italy have the highest percentages of likes per follower for passive involvement.




**(a)** **(b)**

**Figure 3.22:** Distribution of interaction in actors posts normalized by number of followers: a) likes on number of followers, b) comments on number of followers

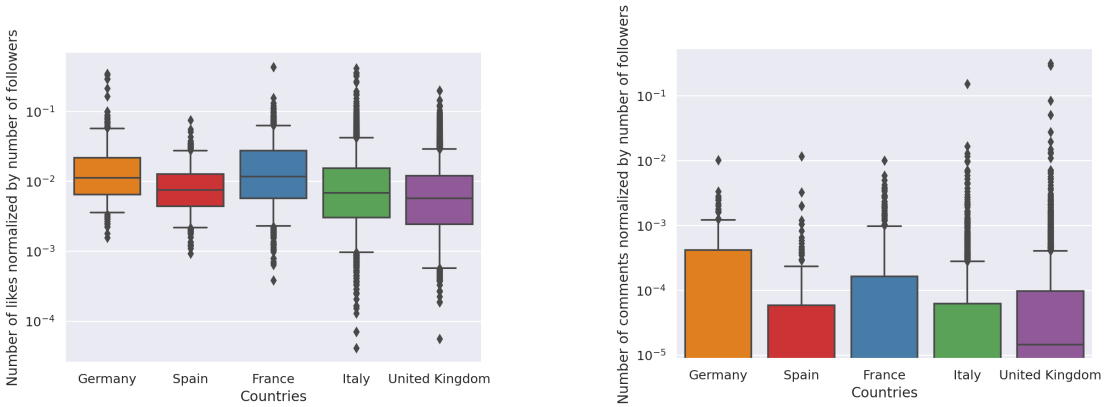
### Universities

Analyzing the passive interaction for European universities, figure 3.23a, we obtain the following information. The country with the highest passive interaction is

France, where there are 19.7 likes on average for every 1000 followers. This implies that fans of French colleges are likely to express a great deal of interest through likes. The distribution of the data shows significant fluctuation, with certain peaks reaching a maximum of 431.8 likes per comment. With 19.6 likes per 1,000 followers on average, Germany follows France with a comparable level of engagement. With a maximum of 346.8 likes per 1000 followers, the data distribution is quite erratic. With averages of 12.3 likes per 1000 followers and 9.1 likes per 1000 followers, respectively, Italy and the United Kingdom exhibit slightly lower levels of passive involvement than the preceding countries. Both scenarios show data variability, with maximum peaks of 415.7 likes per 1000 followers for Italy and 198.1 likes per 1000 followers for the United Kingdom. Spain, with an average of 10.2 likes per 1,000 followers, has the least passive involvement out of all the countries analyzed. The distribution of data varies, with a maximum of 74.9 likes per 1000 followers, even in Spain, it is crucial to note.

With an average of roughly 0.2 comments per 1000 followers, France confirms having the most active involvement when taking into account active interaction, according to figure 3.23b. French users are more likely to participate in the comments' debate. The data distribution varies significantly, with a maximum of 9.9 comments for every 1000 followers. Germany exhibits a comparable level of engagement, with 0.3 comments per 1000 followers on average. The data is sparsely distributed, with a maximum of 10.1 comments for every 1000 followers. With rates of roughly 0.1 and 0.2 comments per 1000 followers, respectively, Italy and the United Kingdom have slightly lower levels of active participation. Again, there is variation in the statistics, with Italy's and the United Kingdom's peaks reaching highs of 152.8 and 316.8 comments per 1000 followers, respectively. Between the countries studied, Spain has the lowest active interaction rate, averaging about 0.1 comments per 1000 followers. The distribution of data is diverse, as can be seen above, with a maximum of 11.6 comments for every 1000 followers.

We concentrated on nano and micro influencers in the examination of influencer tiers for universities because there weren't many profiles with more than 50,000 followers. Interesting interaction dynamics are revealed by the Mann-Whitney U test results. France and Germany show up as the countries with the highest rates of contact, both active and passive, when it comes to nano influencers. This shows that institutions in these countries are successfully utilizing the engagement of this group of influencers to attract a large following. Conversely, Spanish universities stand out as having the most passive connection with micro influencers. This indicates that followers of Spanish universities are particularly inclined to express their appreciation through "likes" and similar interactions. At the same time, British universities top the list in terms of active interaction, suggesting that they are encouraging more active and engaging discussion among their followers.




**(a)** **(b)**

**Figure 3.23:** Distribution of interaction in universities posts normalized by number of followers: a) likes on number of followers, b) comments on number of followers

# Chapter 4

# Uses of the Dataset

In the previous chapter, we examined the structure of the dataset that relates the profile information provided to us by Wikidata and the characterizing information of Instagram profiles provided to us by Crowdtangle. We focused on politicians, entertainment figures (including athletes, models, and actors), and universities while carefully analyzing the distribution of profile followers, the volume and type of activity on the social media platform, the lengths of post descriptions, the presence of mentions and hashtags within them, and the engagement strategies and interaction dynamics within the various categories of profiles on Instagram. With an emphasis on Italy, Germany, Spain, the United Kingdom, and France, we also took into account regional differences. Throughout this chapter, we will explore two key and highly relevant applications of our dataset. We will present **topic recognition** using BertTopic and an in-depth study of the **top 5 profiles of 2022**, providing a detailed analysis of trends and dynamics within digital and social communications in Europe.

## 4.1 Topic Recognition in Instagram Post Description

This section represents a crucial point in our research as it will address the challenge of automatically identifying and categorizing prevalent themes and topics within a large collection of Instagram post descriptions. While Instagram is known for its emphasis on visuals, post descriptions are often a crucial channel through which users communicate messages, share thoughts, and connect their images to larger contexts and stories. A key field of study in text analytics and machine learning is topic recognition, often known as theme recognition. It enables us to instinctively recognize the ideas and important phrases that recur most frequently in texts, improving our comprehension of their substance and organization. Here, we'll use

cutting-edge natural language processing (NLP) techniques with an emphasis on the BertTopic model and machine learning algorithms to examine Instagram post descriptions in particular profile categories and nations.

Our goal is twofold: first, we want to use the BertTopic model to automatically reveal the predominant topics and themes within each profile category and nation, thus providing an in-depth overview of online conversations in these specific contexts. Second, we intend to identify any significant overlaps or differences in the topics discussed across profile categories and across nations. This will allow us to gain a more complete understanding of the social and cultural dynamics at play. In the remainder of this chapter, we will explore the specific methodologies and techniques used to leverage the BertTopic model and data analysis. We will also present the results obtained through the application of this technology to the analysis of Instagram post descriptions. Finally, we will conduct a comparative analysis of emerging themes, allowing us to draw meaningful conclusions about online communication in the profile categories and country contexts examined. This process will pave the way for a deeper understanding of social media dynamics in Europe in light of the latest innovations in topic recognition.

### 4.1.1 Data preparation

The preparation of the data is an important stage in the analysis of Instagram post descriptions since it guarantees that the texts are prepared for additional processing, such as the use of the BertTopic model. Since the data are arranged according to profile type and country, a specialized cleaning procedure is required to handle various languages and eliminate extraneous information from texts. The cleaning of the texts in the various languages in the dataset will be the main emphasis of the detailed description of the data preparation process in this section. The data are first separated by language, which is essentially a representation of the languages used most frequently in the country-specific Instagram post descriptions. Since **punctuation** does not provide useful information for subject analysis, we start the cleaning process by removing it. This step includes removing characters such as periods, commas, parentheses, quotation marks, and special symbols. The following phase is **tokenization**, which separates the text into individual words or "tokens." This subdivision allows us to examine texts in a more granular way, discovering relevant keywords and phrases. The elimination of **stopwords**, which are frequent and ineffective words like "and," "but," "in," etc., is an essential component of text cleansing. These terminology are dropped in favor of others that have deeper meanings because they don't help readers understand the themes in the texts. The last step is the removal of **special characters** and symbols like numerals, emoticons, URLs, and mentions. These components can be viewed as noise in the texts because they don't help readers understand the topics.

### 4.1.2 BertTopic: Neural topic modeling with a class-based TF-IDF procedure

The BertTopic model is an argument model that uses clustering techniques and a variation based on TF-IDF classes to provide coherent representations of the arguments [37]. In particular, the process is broken down into three independent phases. In the first phase, document embedding are developed using a pre-addestrated language model to obtain document-level information. In the second phase, the dimensionality of document embeddings is reduced before semantically similar clusters of documents are created, each of which represents a unique topic. In the third phase, in order to outperform the centroid-based hypothesis, a version based on TF-IDF classes is developed in order to extend the representation of the topic from each cluster, KeyBERTInspered representation model (figure 4.1). These three independent steps result in a flexible topic model that can be used in a variety of use cases, such as dynamic topic modeling. The process of generating topic

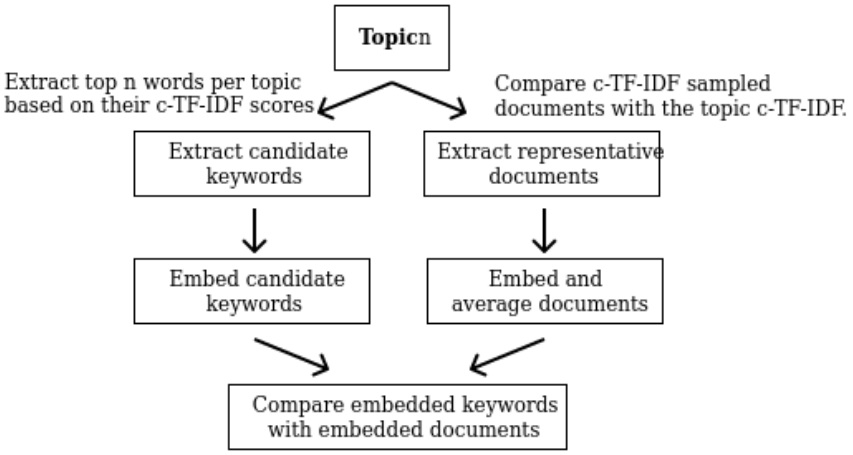



**Figure 4.1:** How KeyBERTInspired representation model works[38]

representations in BertTopic goes through three steps. First, each document is converted into its embedding representation using a pre-trained language model. Then, before clustering these embeddings, the dimensionality of the resulting embeddings is reduced to optimize the clustering process. Finally, topic representations are extracted from the document clusters using a customized class-based variant of TF-IDF. Sentences and paragraphs are transformed into vector dense representations for document embeddings by BertTopic utilizing the Sentence-BERT (SBERT) framework and pre-trained language models. On a number of sentence embedding

challenges, this technique is able to reach state-of-the-art performance. BertTopic employs the UMAP technique to lessen the dimensionality of document embeddings in terms of data dimensionality. In lower projected dimensions, UMAP is renowned for preserving more of the local and global characteristics of high dimensionality data. As a result, BertTopic's clustering is of higher quality. Each cluster is assigned a subject, and topic representations are modeled based on the documents in each cluster. One is interested in learning what distinguishes one topic from another based on the distribution of words in the cluster for each topic. In order to do this, we alter the TF-IDF to permit the depiction of word importance for a topic rather than a document. The usage of TF-IDF-based class representations in BertTopic also enables the representation of topics to be characterized by time. With this method, local topic representations can be created without the requirement to embed and group documents, facilitating quick calculation. In summary, BertTopic is a powerful methodology for topic analysis that leverages a range of techniques, from document embedding generation to the extraction of TF-IDF class-based topic representations, allowing in-depth understanding of topics in documents.

### 4.1.3 Theme Recognition via BertTopic

Theme Recognition allows us to better grasp the topics that each profile group is most interested in. Determining the precise tastes and interests of decision-makers, cultural icons, and academic institutions is crucial for tracking changes through time. Communication methods can be improved by being aware of the recurring topics in each profile category's Instagram posts. We can evaluate overlaps and cultural variances in Instagram usage across various European countries by analyzing topics.

#### Politicians

Figure 4.2 shows the barplots represented the topics of greatest discussion in the post descriptions of European politicians. Specifically, on the abscissae of the barplots are the frequencies with which the topic recognition model finds a post that may belong to the cluster of the specific topic. Going into detail country by country, in France, figure 4.7a, politicians use the Instagram platform partly lightly hence the presence of many holiday greeting posts both as a real propaganda medium since there is much discussion of elections and political agendas in different areas such as nuclear and health care. Peculiar was the fact that among the major topics of discussion did not turn out to be the Russia-Ukraine conflict, which is present in lesser or greater frequency in the other 4 European countries. Particularly for German politicians, figure 4.7b, it was the most discussed topic on Instagram in the year 2022, followed by the issue on energy, which can be considered a child topic of the Russia-Ukraine conflict. In Italy, figure 4.7e, on the other hand, the topic to

prevail is the political scenario between italy and europe, followed by policies for health care having been a covid increase in the past year and the discussion on the war in Ukraine. In Britain, politicians used the Instagram platform mainly for the health care debate in detail about the shortage of staff during the hardest periods of the pandemic. The other topics addressed are policies related to the LGBTQ+ community and the Ukrainian conflict. In Spain, figure 4.7d, on the other hand, the most frequently mentioned topic is raising awareness of women's rights and raising awareness of the issue, which is also mentioned in the posts of Italian and French politicians, but to a lesser level.

### Entertainment figures

From the analysis of topic recognition in posts by entertainment figures, it becomes clear that profiles belonging to this high social status primarily use Instagram as a means to share their personal interests and the activities they are involved in.

For example, actors use the platform not only to promote films, TV series, or other film events they have participated in but also to offer their followers a behind-the-scenes look at their experiences in the world of entertainment.

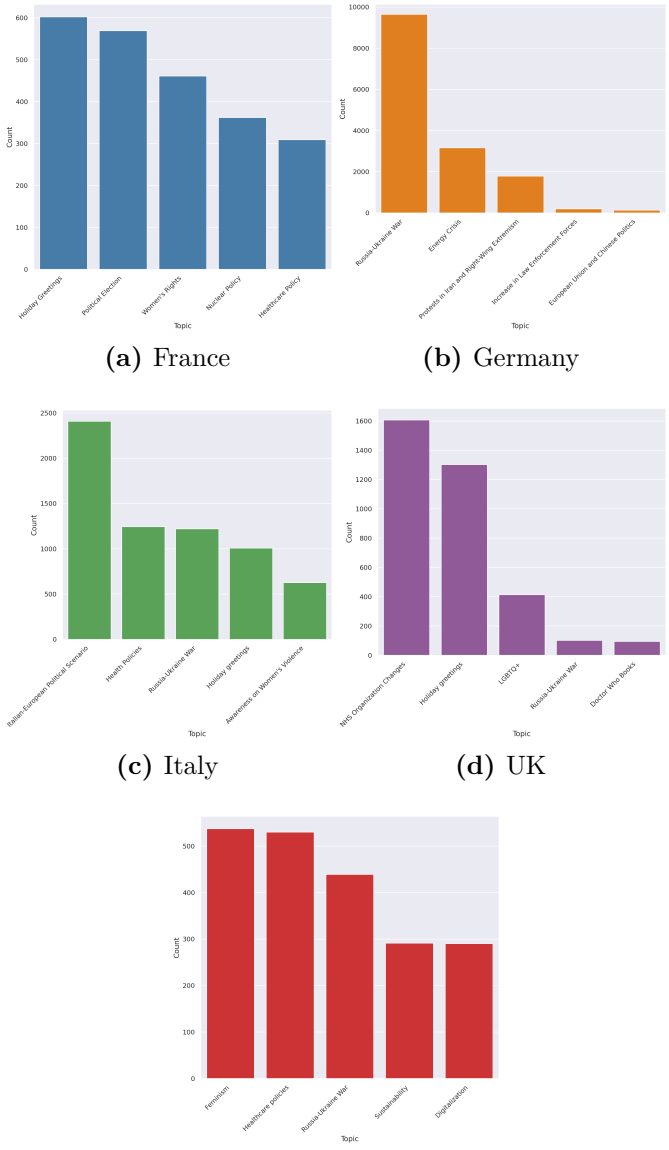
Athletes, on the other hand, share their sporting performances in high-profile international events such as the Olympics or motorsport races, but they also use Instagram to communicate their values, promote a healthy lifestyle, and inspire others through their success stories.

Models, on the other hand, use Instagram not only to document high fashion events like Milan Fashion Week but also to serve as ambassadors for luxury products and brands. The platform thus becomes a tool for promoting fashion and the luxury lifestyle.

What is particularly interesting is how geographical differences are gradually narrowing for these figures. Instagram now represents a crucial work tool for enhancing their image and promoting their professional activities globally. At the same time, these entertainment figures also use Instagram for entertainment and leisure purposes, increasingly aligning with the typical use of the platform by common users. This phenomenon highlights how Instagram's appeal and attraction are extending far beyond geographical boundaries, as these entertainment figures reach out to a vast global community through the platform. In this process, Instagram is becoming a meeting point between their professional and personal spheres.

### Universities

With regard to the analysis of topics of interest university profiles discuss, universities in Italy, figure 4.3a, and Great Britain, figure 4.3b, prefer to promote on their Instagram profiles events that concern the daily life of the universities




**(a)** France **(b)** Germany

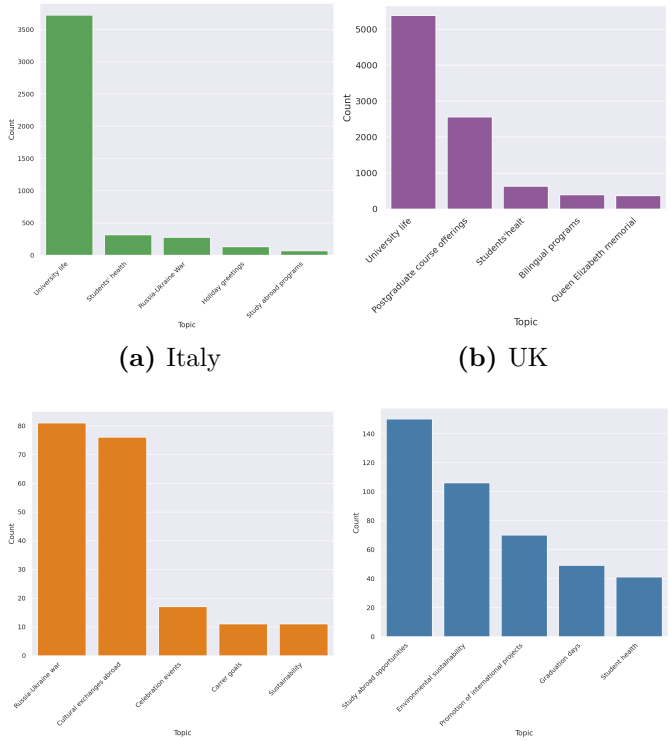
**(c)** Italy **(d)** UK

**(e)** Spain

**Figure 4.2:** Most prominent topics within the corpus of posts of political profile descriptions, divided by country

without excluding dealing with topics of global interest such as the Russia-Ukraine conflict and the death of Queen Elizabeth. Also among the most frequent topics are interest in student health and the proposal of international study programs that introductory paths to the post-academic world. Universities in Germany and France, respectively figures 4.3c and 4.3d, also focus on promoting via Instagram

international study programs and postgraduate pathways but tend to give more attention to issues of global sensitivity such as precisely the war in Ukraine and environmental sustainability. Unfortunately, Spain was excluded from the analysis due to the lack of a substantial body of texts.




**(a)** Italy **(b)** UK

**(c)** Germany **(d)** France

**Figure 4.3:** Most prominent topics within the corpus of posts of university profile descriptions, divided by country

## 4.2 Top 5 Profiles Study in 2022

The second use of the dataset involves a detailed study of the five main profiles within the categories of politicians, entertainment figures and universities. This analysis focuses on the activity of the profiles, the number of followers, and the active and passive engagement of the profiles over the course of 2022.

This study is significant for several reasons:

- We can identify potential trends and significant changes by analyzing the movement of key profile metrics over the course of 2022. It may be useful to understand how real-world events influence the online activity of relief profiles.

- When we compare the performance of top profiles to the median within their respective categories, we may perform benchmarking and assess the effectiveness of communication strategies. This might help you identify areas for improvement.
- The findings of this study could be used to develop future strategies. For example, if a politician has achieved a high level of engagement on specific issues by 2022, he or she may wish to continue focusing on those issues.

### 4.2.1 Politicians

Table 4.1 displays the top five political profiles in the five study countries, as measured by the number of followers. As predicted in the section 3.4, only Italy, the United Kingdom, and France can claim at least one mega influencer, with the others being macro or mid-tier influencers such as senior German and Spanish politicians.

#### Followers temporal trends

In the temporal analysis of followers, shown in Figure 4.4, one can see interesting links to real-world events that may have influenced some radical changes in the number of followers of some of the top5 political profiles in each nation.

In Germany, figure 4.4b, the profiles of Annalena Baerbock and Robert Habeck, co-chairs of the environmentalist Green Party and Germany's foreign minister and vice chancellor, respectively, experienced 31 percent and 55 percent increases in their number of followers between February 2022 and May 2022 and then stabilized on a continuous but much more regular increase until the end of the year. The period of strong change in the terms of the following metric is in conjunction with the Russian invasion of Ukrainian territories. It comes immediately to link this increase to the strong indignation and call for strong condemnation by Germany first and Europe and the world as a whole made by the chairwoman of the Green Party in early March 2022 [39] . Among Germany's top politicians, there is also a strong increase in the number of followers for the chairwoman of the European Commission Ursula von der Leyen during the same time frame mentioned above. It can always be traced back to the Russian-Ukrainian conflict in particular to the speech given at the EUCO meeting [40] at the end of March 2022 on developments in the conflict and punitive measures against Russia.

Other significant shifts in the number of followers happened, most notably in Italy and the United Kingdom, figures 4.4c and 4.4e, for the profiles of Giorgia Meloni, the current Italian Prime Minister, and Rishi Sunak, the current British Prime Minister. Between September and December 2022, the Italian gained little more than 500,000 followers, while the British gained more than a million. The

| Country | Accounts | Number of Followers |
|---|---|---|
| Germany | abaerbock | 578,746 |
| | ursulavonderleyen | 405,861 |
| | christianlindner | 370,479 |
| | robert.habeck | 361,597 |
| | martinhsonneborn | 276,958 |
| Spain | santi_abascal | 845,198 |
| | isabeldiazayuso | 556,122 |
| | luchogarcia14 | 501,399 |
| | cristinacifuentes | 376,500 |
| | maximohuerta | 362,632 |
| United Kingdom | rishisunakmp | 1,740,667 |
| | jeremycorbyn | 509,050 |
| | _katie_hopkins_ | 421,172 |
| | zarahsultanamp | 273,416 |
| | andrewlloydwebber | 239,028 |
| Italy | matteosalviniofficial | 2,204,098 |
| | giuseppeconte_ufficiale | 1,738,736 |
| | giorgiameloni | 1,624,870 |
| | luigi.di.maio | 792,808 |
| | avvocathy | 718,801 |
| France | emmanuelmacron | 3,234,827 |
| | raphaelglucksmann | 794,394 |
| | elodiegossuin | 568,414 |
| | jlmelenchon | 322,777 |
| | marine_lepen | 301,418 |

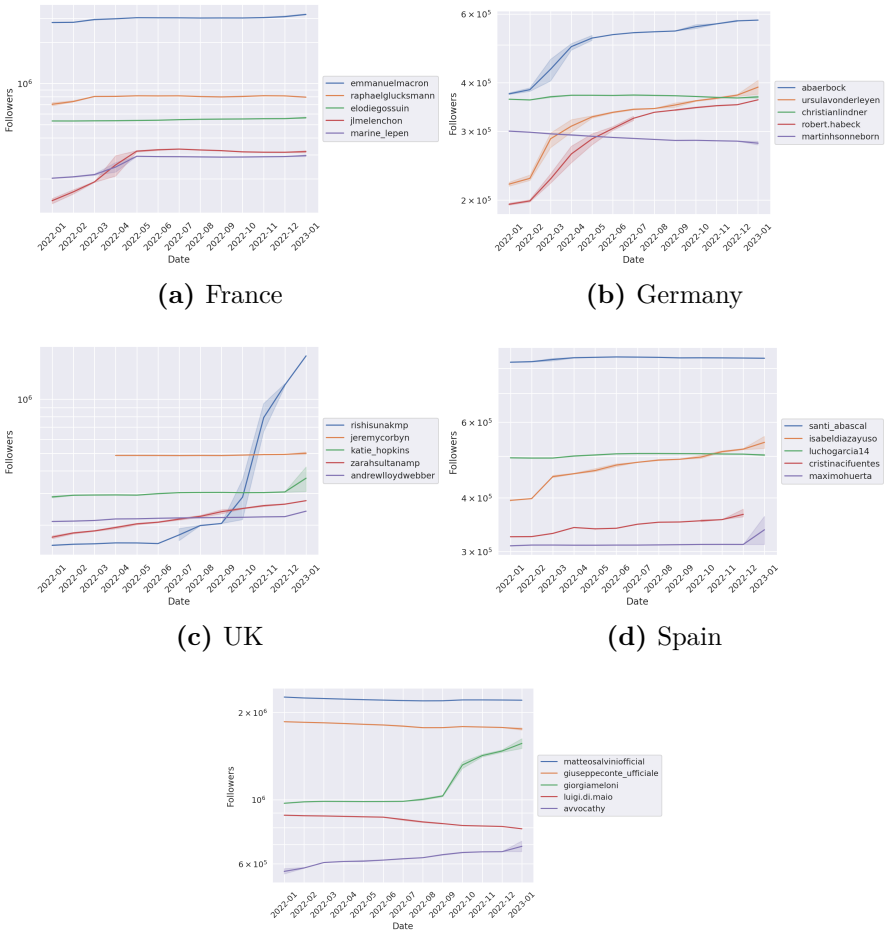
**Table 4.1:** List of top 5 politicians profiles by country with the number of followers.

time frame in which this increase occurred for the Italian prime Minister correlates with his victory in the October 2022 election, and similarly for the British premier in that there was no proper and true election, but he was appointed prime Minister on October 25, 2022.

### Activity temporal trends

Analyzing the temporal trends in the quantity of published posts reveals an intriguing landscape, figure 4.5. The majority of profiles included in the analysis maintained an average posting frequency ranging from 5 to 25 posts per month, adhering to conventional political communication practices.

However, certain political figures exhibited distinct behaviors. For instance,




**(a)** France **(b)** Germany

**(c)** UK **(d)** Spain

**(e)** Italy

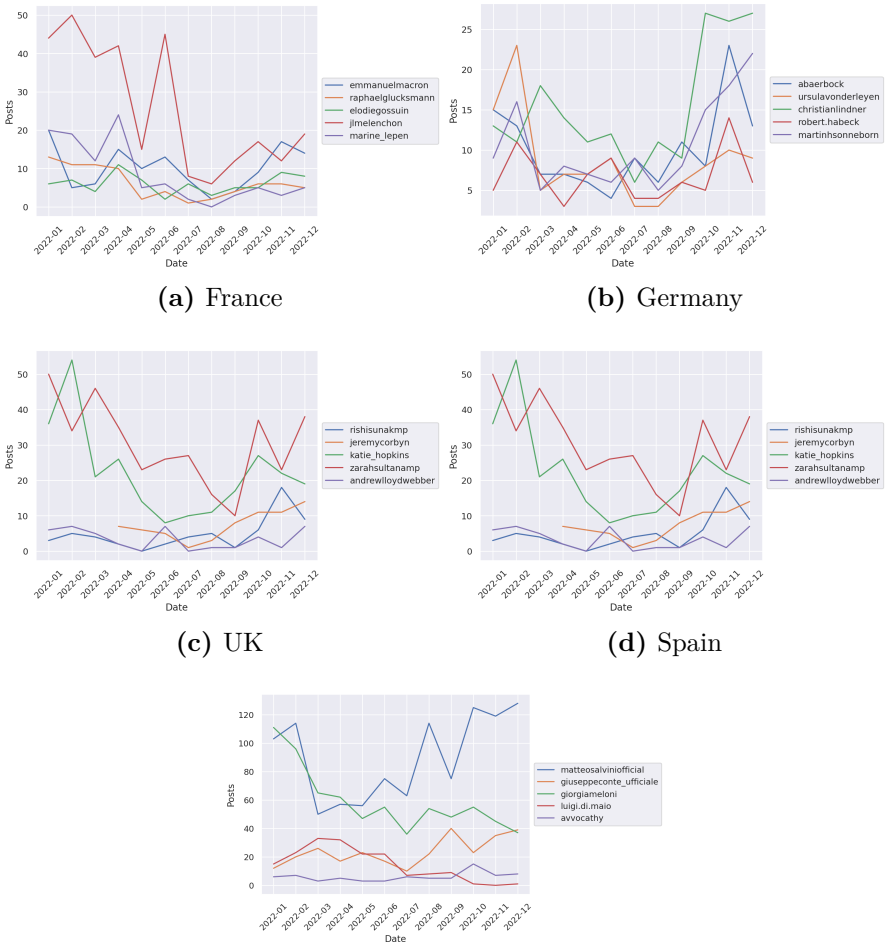
**Figure 4.4:** temporal followers trends for top5 politicians according to the number of followers, diveded by country

Jean-Luc Mélenchon, the leader of La France Insoumise, the prominent left-wing political group in France, displayed a notable increase in postings during the months of January, February, April, and June 2022, with a monthly post count ranging from 40 to 50. These spikes may have been influenced by significant political events or deliberate communication strategies.

Another noteworthy example is Santiago Abascal Conde, the president of the far-right Vox party in Spain, who recorded posting peaks exceeding 40 posts per month, particularly in February and June. These periods may have coincided with pivotal political debates or heightened media attention.

However, among all the European politicians under consideration, two Italian

profiles stood out: Matteo Salvini, the Deputy Prime Minister, and Giorgia Meloni. Their social media activity was notably intense, with a posting frequency ranging between 40 and 60 posts per month. It is worth highlighting that Matteo Salvini reached a peak of 110 posts per month in February and August 2022, indicating a significant commitment to online communication. These Italian politicians appear to be among the most active in utilizing social media platforms to promote their ideas and initiatives. The peaks in their posting activity may reflect their determination to capitalize on strategic opportunities to shape public discourse.




**(a)** France **(b)** Germany

**(c)** UK **(d)** Spain

**(e)** Italy

**Figure 4.5:** temporal activity trends for top5 politicians according to the number of followers, diveded by country

#### Interaction temporal trends

Analyzing the temporal trends of both passive and active interactions, depicted in figures 4.6 and 4.7, provides valuable insights.

In the United Kingdom, Prime Minister Rishi Sunak experienced a substantial increase in normalized likes following his assumption of office, coinciding with a significant rise in the number of followers. Additionally, in terms of active interaction, a considerable uptick in comments is observed around September/October 2022 on Rishi Sunak's profile.

In Spain, the profile of Isabel Díaz Ayuso, President of the Popular Party of the Community of Madrid, stands out with peaks in both passive and active interactions during the months of May and November. These spikes correlate with her reelection as party president and the announcement of new healthcare policies aimed at restructuring the community's healthcare system.

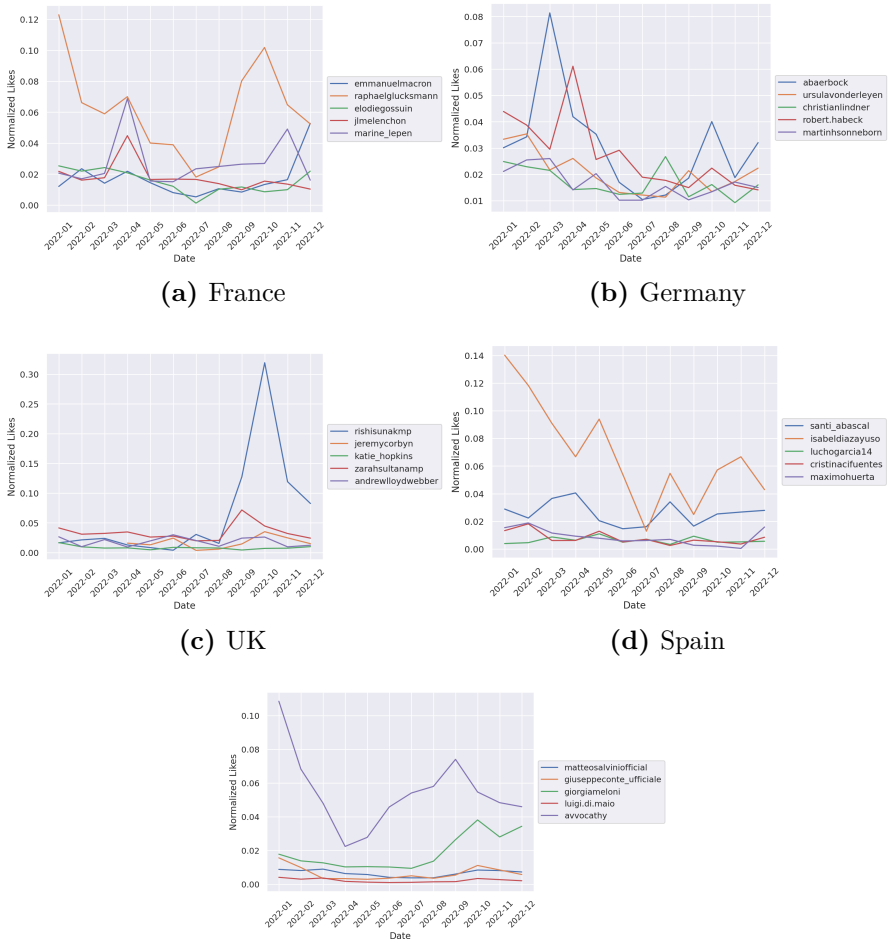
In Germany, the most prominent profiles in terms of interaction are Annalena Baerbock and Robert Habeck, leaders of the Green environmentalist party. These leaders excel in both passive and active interaction, demonstrating the party's effective online communication and adept utilization of Instagram's capabilities.

In France, profiles of Raphaël Glucksmann, a Member of the European Parliament, and Marine Le Pen, President of the Rassemblement National, exhibit positive responses to the analysis metrics. Raphaël Glucksmann has shown significant increases in both types of interaction, both in the months of March and April and in September and October. Marine Le Pen notably stands out in April, both in terms of likes and comments.

In Italy, the curves of interest in terms of interaction, both passive and active, belong to Prime Minister Giorgia Meloni and Chaty la Torre, a former member of the Left Ecology Freedom party, now an attorney and political activist in the field of anti-discrimination law, particularly concerning issues related to sexual orientation and gender identity. Regarding the Italian Prime Minister, peaks of interaction are observed in proximity to taking office in October 2022. The 'avvocathy' profile records significantly higher levels of interaction, particularly in terms of passive interaction, compared to profiles with a much larger following, such as that of Matteo Salvini. This is attributed to the fact that the political activist addresses a more niche audience, highly engaged with the issues addressed on the profile and more closely connected to these topics on the social media platform.

### 4.2.2 Entertainment Figures

The table 4.2 provides a comprehensive overview of the most influential figures on Instagram across five European nations: France, Germany, Italy, Spain, and the United Kingdom. These individuals have been categorized into three distinct entertainment domains that have been explored throughout this research:




**(a)** France **(b)** Germany

**(c)** UK **(d)** Spain

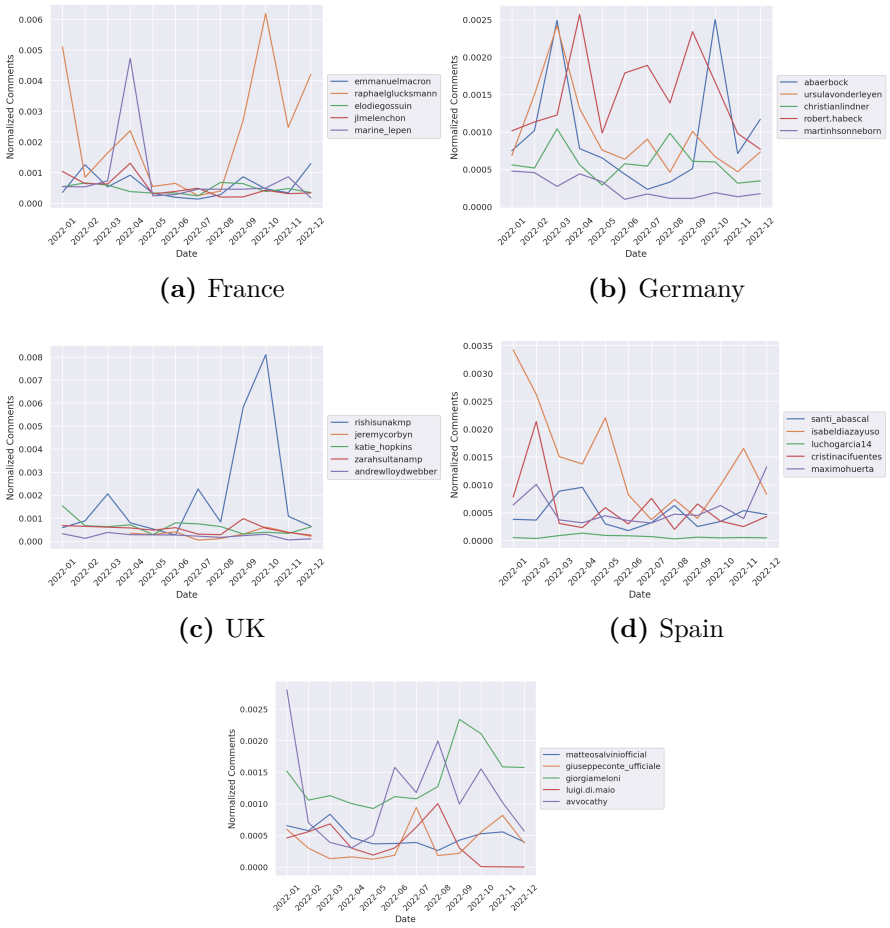
**(e)** Italy

**Figure 4.6:** temporal likes trends for top5 politicians according to the number of followers, diveded by country

athletes, models, and actors. Each of these categories embodies a unique facet of contemporary culture and holds a significant presence in the realm of online influence.

Commencing with the category of athletes, it encompasses some of the most globally recognized names in the realm of sports. To illustrate, in France, Kylian Mbappé, boasting an impressive following of over 104 million, stands as an iconic figure in the world of soccer. Meanwhile, in Spain, Sergio Ramos, with nearly 59 million devoted followers, enjoys international acclaim as a distinguished defender.

Transitioning to the sphere of models, this classification includes personalities who are frequently associated with the fashion industry and the allure of a glamorous




**(a)** France **(b)** Germany

**(c)** UK **(d)** Spain

**(e)** Italy

**Figure 4.7:** temporal comments trends for top5 politicians according to the number of followers, diveded by country

lifestyle. For instance, in Italy, Michele Morrone, commanding an audience of almost 16 million followers, is renowned for his role in "365 Days." Conversely, in the United Kingdom, Dua Lipa, with an astounding following of over 88 million, has carved a successful career as both a singer and a fashion model.

Concluding our exploration with the category of actors, we encounter globally celebrated figures. In Spain, Ester Expósito, boasting an impressive fan base of over 27 million followers, is celebrated for her portrayal in the series "Elite." Similarly, in the United Kingdom, Emma Watson, with an extensive following of over 71 million, is renowned for her iconic role as Hermione Granger in the "Harry Potter" film franchise.

| Country | Athletes | Models | Actors |
|---|---|---|---|
| France | k.mbappe (104,270,634) | tchalamet (18,415,929) | davidmichigan (18,720,988) |
| | karimbenzema (70,919,796) | kevadams (7,557,584) | nabilla (8,533,966) |
| | paulpogba (58,669,948) | thylaneblondeau (6,855,042) | thylaneblondeau (6,815,572) |
| | antogriezmann (40,182,430) | 6pri1 (5,993,267) | lenamahfouf (4,131,543) |
| | zidane (38,892,202) | normanthavaud (5,954,428) | mattpokora (3,781,128) |
| Germany | toni.kr8s (39,839,517) | meryemuzerlimeryem (7,532,995) | lenameyerlandrut (5,484,503) |
| | m10_official (26,567,493) | dagibee (6,781,157) | stefaniegiesinger (5,138,282) |
| | bastianschweinsteiger (15,577,668) | shirindavid (6,284,221) | tonigarrn (4,773,845) |
| | mterstegen1 (14,950,536) | lenameyerlandrut (5,515,913) | evelyn_sharma (3,374,529) |
| | esmuellert (12,976,196) | tonigarrn (4,814,687) | lenagercke (3,226,028) |
| Italy | valeyellow46 (15,480,143) | iammichelemorroneofficial (15,954,272) | iammichelemorroneofficial (15,954,272) |
| | mrancelotti (11,983,410) | itsmarziapie (8,627,932) | marianodivaio (6,886,183) |
| | mb459 (11,458,670) | real$_b$*rown*(6,239,923) | evamenta (5,822,397) |
| | gianluigibuffon (10,535,778) | giuliadelellis103 (5,384,714) | monicabelluccioofficiel (5,191,858) |
| | andreapirlo21 (9,643,518) | monicabelluccioofficiel (5,196,918) | stefanodemartino (4,914,326) |
| Spain | sergioramos (58,770,957) | ester_exposito (27,485,428) | georginagio (49,502,138) |
| | andresiniesta8 (42,161,248) | rosalia.vt (25,529,165) | ester_exposito (27,540,197) |
| | iscoalarcon (28,487,804) | ursulolita (22,106,482) | ursulolita (22,167,628) |
| | 3gerardpique (22,734,497) | enriqueiglesias (18,285,475) | enriqueiglesias (18,236,718) |
| | alvaromorata (20,739,394) | belindapop (16,371,970) | belindapop (16,343,929) |
| United Kingdom | davidbeckham (79,676,641) | dualipa (88,345,088) | dualipa (88,211,725) |
| | garethbale11 (51,532,926) | davidbeckham (79,597,835) | davidbeckham (79,311,151) |
| | louist91 (18,373,451) | aliaabhatt (77,640,553) | aliaabhatt (77,428,084) |
| | waynerooney (16,025,843) | emmawatson (71,797,027) | emmawatson (71,603,453) |
| | marcusrashford (15,715,197) | milliebobbybrown (63,650,737) | milliebobbybrown (63,541,060) |

**Table 4.2:** Lists of top5 athletes, models and actors according to the number of followers divided by country

### Followers temporal trends

In figure 4.8, the temporal trends of the number of followers for the entertainment figures under analysis are presented. These trends reveal a predominantly consistent pattern, characterized by a slight yet steady positive slope. This stability serves as a significant indicator of the established success of the entertainment figures included in our study. They often boast follower counts that are one or two orders of magnitude higher than those of prominent political figures. These entertainment figures have attained a level of fame and engagement on Instagram that has allowed them to build a solid and loyal following base. This privileged position enables them to continuously focus on expanding their online presence, leveraging one of the most effective means of promoting their image. In some cases, this online presence holds an intrinsic value that is nearly incalculable within the context of the cultural world to which these figures belong.

### Activity temporal trends

In figure 4.9, we present the temporal trends regarding the posting activity of the entertainment figures included in our study. In general, these high-profile individuals tend to maintain a relatively limited monthly posting frequency, usually not exceeding 20 posts per month. This behavior can be justified by considering that such influential figures on Instagram often prefer to use other modes of

communication to interact with their audience. For instance, they frequently prioritize the use of stories or reels, although details related to such content are not provided through Instagram's official API.

However, when examining the temporal posting patterns, some noteworthy exceptions emerge. An example is the profile of Dua Lipa, boasting a following of over 88 million followers. In this case, we observe peaks in posting activity that reach approximately 50 posts per month, as seen in the months of March and May.

Other exceptions are found among athletes, such as in the case of Andres Iniesta, a former Barcelona football player in Spain, and Thomas Muller, a footballer from Bayer Munich in Germany. Both exhibit posting activity peaks exceeding the monthly average.

Among models, Léna Mahfouf stands out for the quantity of posts published, a renowned French social personality who gained fame through YouTube.

Lastly, among actors, Emma Marrone from Italy emerges, a pop singer and actress, who has recorded Instagram activity surpassing the threshold of 20 posts per month in February and September.

### Interaction temporal trends

Examining the volume of media interaction garnered by immensely renowned figures on Instagram provides insights of significant interest. Specifically focusing on passive interaction, figure 4.10, it becomes evident that certain individuals stand out notably compared to their international counterparts. Notably, these cases are most prominent in France and Spain.

Among athletes, an example of exceptional passive interaction is exemplified by Kilian Mbappe. In the month of December, coinciding with the Qatar World Cup, he reached remarkable levels of over 100 likes per 1000 followers. This outcome underscores the substantial appeal that football and international sporting events can generate on platforms such as Instagram.

Similarly noteworthy is the former Barcelona defender, Gerard Pique. From March through the end of the year, he recorded an impressive 62.5% increase in passive interaction on his profile. This data suggests significant audience engagement with the content he shares on Instagram.

Transitioning to Spain, we find that among successful models, Georgina Rodríguez emerges, while among actors, Ester Expósito stands out. Both are capable of reaching peaks of interaction exceeding 150 likes per 1000 followers, values significantly higher than the general average of passive interaction, even among influencers of their same tier.

Regarding active interaction, figure 4.11, it is worth noting that the same personalities that have distinguished themselves in passive interaction are also leaders in active engagement. This demonstrates how adept these individuals are

at using Instagram, capable of establishing consistent and interactive connections with a broad audience over time.

### 4.2.3 Universities

The table 4.3 highlights the leading universities in terms of their followers across the five countries under scrutiny. Notably, three universities have garnered a substantial following, surpassing the 100,000-follower mark. Among these, two hail from the United Kingdom: the London College of Fashion, boasting approximately 178,000 followers, and the University of the Arts London, with around 137,000 followers. The third prominent institution is Italy's Politecnico di Milano, which has amassed approximately 118,000 followers. In contrast, the top German universities exhibit the smallest following in comparison to their European counterparts, while those in Spain and France hover between 10,000 and 20,000 followers.

| **Country** | **Accounts** | **Followers** |
|---|---|---|
| France | univdauphine | 15,656 |
| | univ_toulouse | 12,551 |
| | spaceuniversity | 12,251 |
| | psl_univ | 8,914 |
| | cy_univ | 8,504 |
| Germany | ebsuniversity | 5,531 |
| | srhberlin | 4,487 |
| | kuehnelogisticsuniversity | 2,471 |
| | karlshochschule | 2,403 |
| Italy | polimi | 117,589 |
| | unipd | 85,440 |
| | polimodafirenze | 76,509 |
| | unitorino | 55,818 |
| | politecnicoditorino | 49,185 |
| Spain | la_upc | 18,388 |
| | ujiuniversitat | 17,140 |
| | universitaturv | 13,359 |
| United Kingdom | lcflondon_ | 177,840 |
| | unioftheartslondon | 137,029 |
| | newcastleuni | 68,677 |
| | covuni | 53,127 |
| | uniofstandrews | 51,570 |

**Table 4.3:** List of top 5 universities profiles by country with the number of followers

### Followers temporal trends

The temporal trend in the number of followers displays a nearly regular pattern characterized by continuous growth across all the profiles under examination. As evident from the line plots depicted in figure 4.12, none of the curves exhibit sudden changes in their slopes; instead, they maintain a relatively constant and slightly positive trajectory.

### Activity temporal trends

In the context of the temporal analysis of the activity of the top profiles by country, as highlighted in figure 4.13, some noteworthy data emerges. The monthly average of post publications for these profiles generally ranges between 10 and 30 posts, reflecting a consistent and regular political communication.

However, there are some notable exceptions that deviate from this trend. For instance, in France, the International Space University recorded an extraordinary peak with over 80 posts published in the month of July. Similarly, in Spain, the Jaume I University exceeded 50 posts in February and maintained a substantial level of activity with over 40 posts in June.

Italy and the United Kingdom also exhibit periods of heightened activity well above the average. For instance, the Politecnico di Torino published nearly 40 posts in May, demonstrating significant engagement in online communication. Likewise, the University of St Andrews surprised with over 70 posts in March.

### Interaction temporal trends

In terms of interaction, figures 4.14 and 4.15 provide an interesting overview of the universities that have distinguished themselves over the year in engaging with their Instagram followers. Specifically, in France, Paris Dauphine University and CY Cergy Paris University stand out, with peaks of 70 and 60 likes per 1000 followers, respectively, in the month of August and active interaction spikes of over 5 comments per 1000 followers.

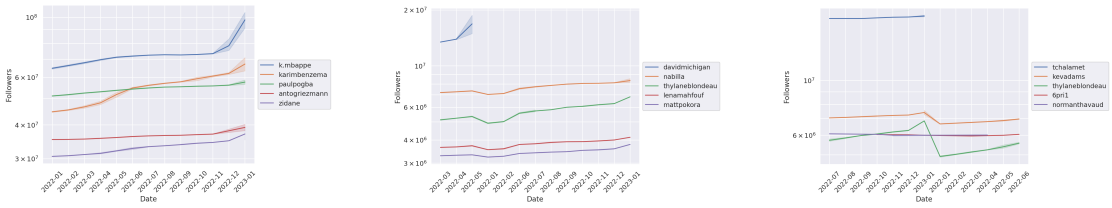
In Germany, it is EBS University and Kühne Logistics University that have the highest interaction with their respective followings. They reached peaks of passive interaction at around 100 likes per 1000 followers, recorded in February and April, respectively. Regarding active interaction, Kühne Logistics University consistently outperforms its national competitors but with lower values compared to other European universities.

In Spain, the public University of Tarragona stands out in terms of interaction, although it lags behind the European scenario in both monthly likes and comments.

In the United Kingdom, the University of St Andrews excels in passive interaction, particularly in September with a peak of over 60 likes per 1000 followers. In terms

of active interaction, London College of Fashion leads with peaks of 0.6 comments per 1000 followers.

In Italy, the scenario is more varied, with the University of Padua slightly leading in terms of passive interaction. However, higher interaction peaks are noticeable for the University of Turin in August, with more than 30 likes per 1000 followers, and for the Politecnico di Torino with a peak of over 50 likes per 1000 followers in December. In terms of active interaction, the scenario is somewhat more consistent, with the University of Padua consistently having the highest interaction on average but with occasional peaks in favor of Polimoda in Florence in the months of March and June.



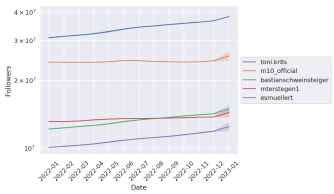

**(a)** France



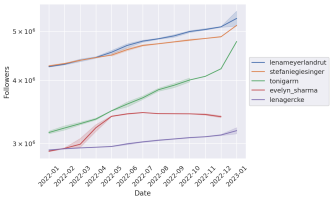

**(b)** France



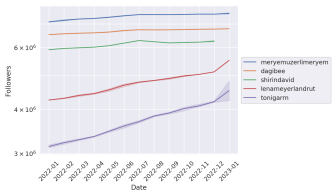

**(c)** France



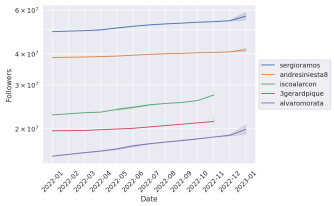

**(d)** Germany



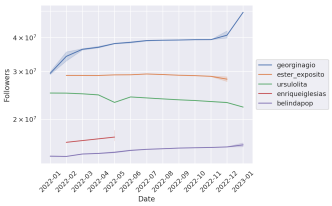

**(e)** Germany



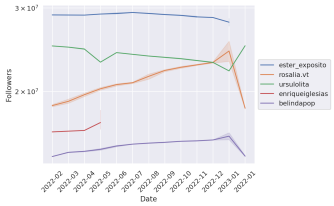

**(f)** Germany



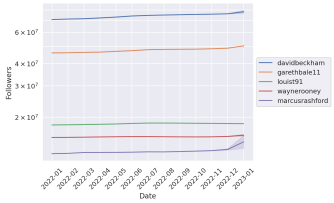

**(g)** Spain



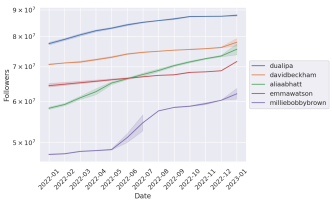

**(h)** Spain



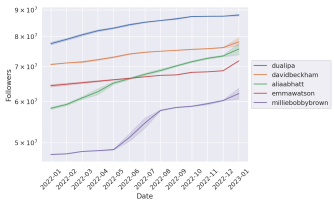

**(i)** Spain



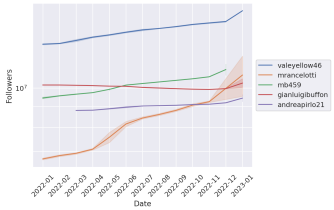

**(j)** UK



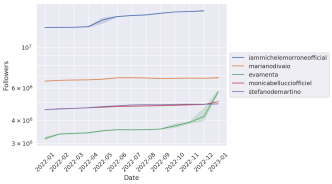

**(k)** UK



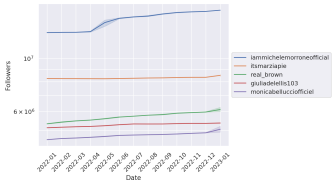

**(l)** UK




**(m)** Italy

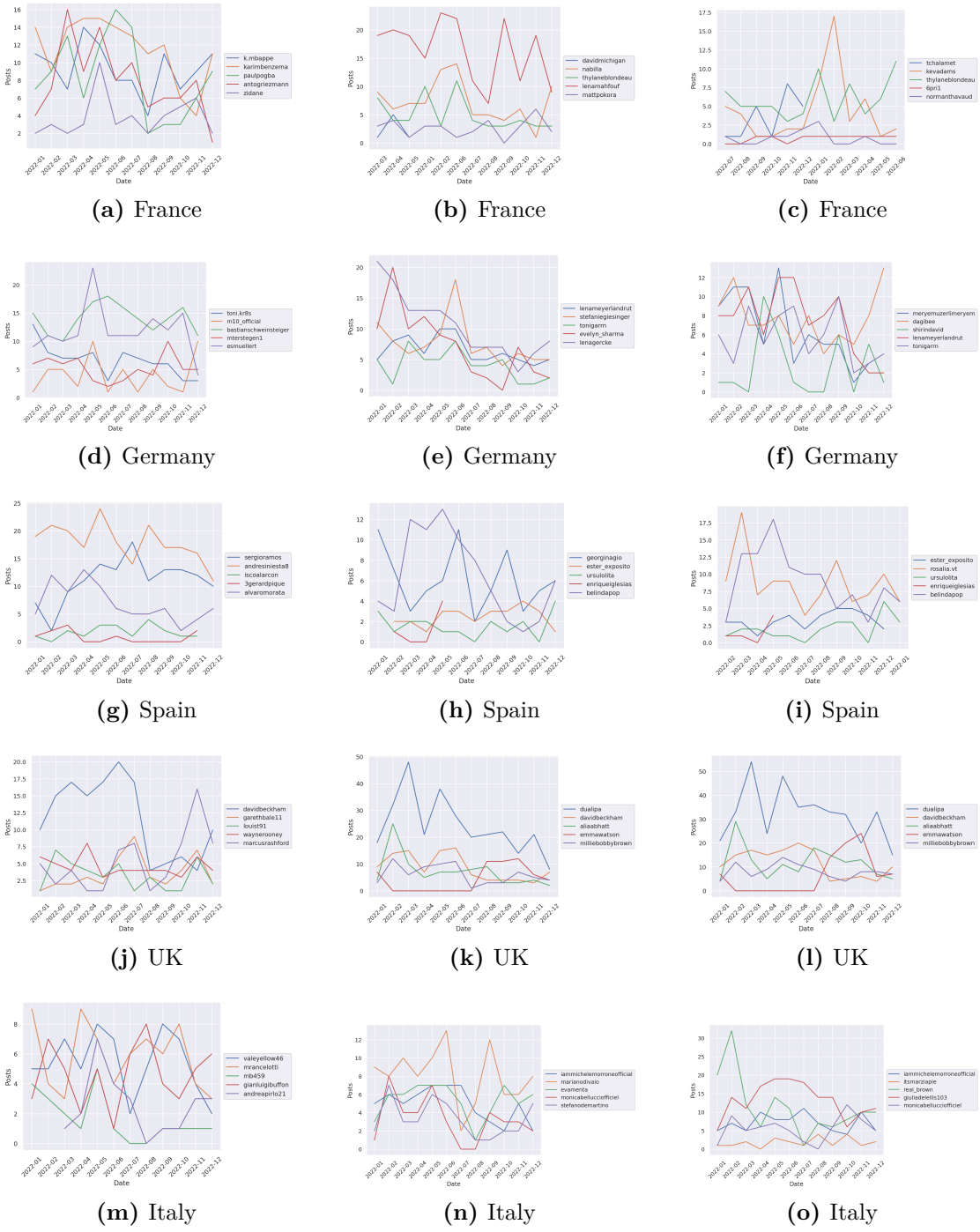



**(n)** Italy




**(o)** Italy

**Figure 4.8:** temporal followers trends for top5 entertainment figures according to the number of followers, diveded by country. In the first column are the graphs of athletes, in the second those of models in the third those of actors




**(a)** France **(b)** France **(c)** France

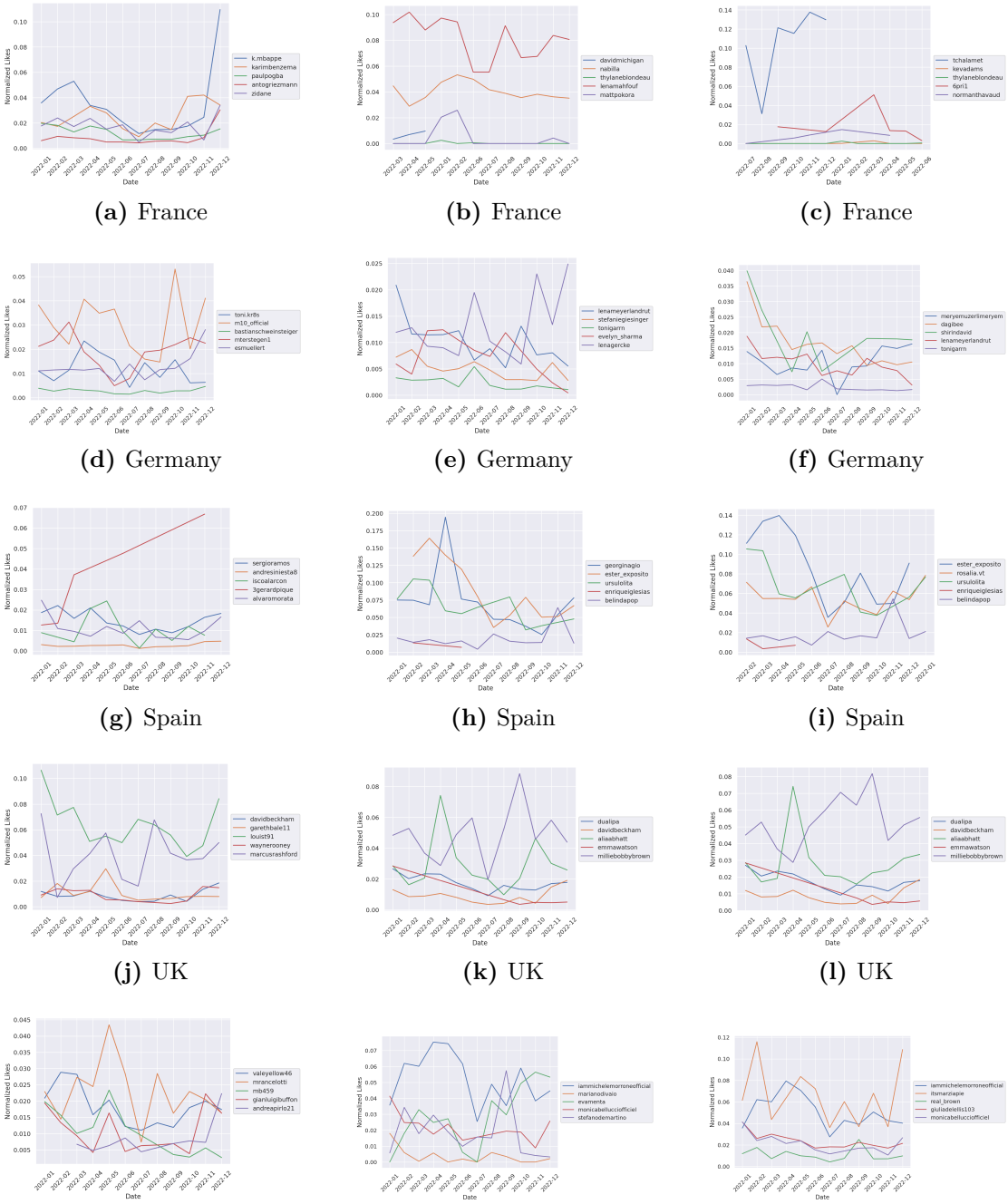
**(d)** Germany **(e)** Germany **(f)** Germany

**(g)** Spain **(h)** Spain **(i)** Spain

**(j)** UK **(k)** UK **(l)** UK

**(m)** Italy **(n)** Italy **(o)** Italy

**Figure 4.9:** temporal posts trends for top5 entertainment figures according to the number of followers, diveded by country. In the first column are the graphs of athletes, in the second those of models in the third those of actors




**(a)** France **(b)** France **(c)** France

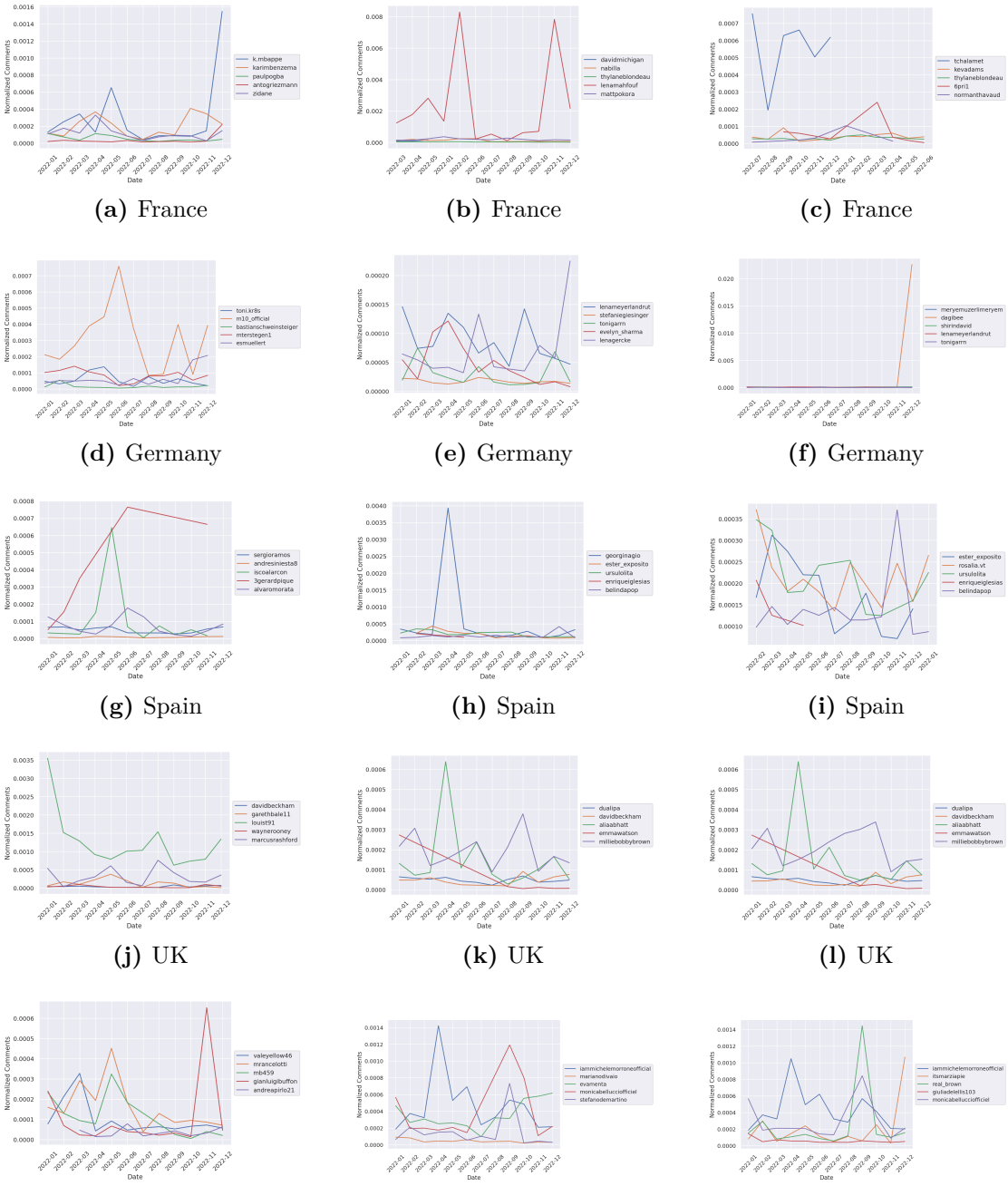
**(d)** Germany **(e)** Germany **(f)** Germany

**(g)** Spain **(h)** Spain **(i)** Spain

**(j)** UK **(k)** UK **(l)** UK

**(m)** Italy **(n)** Italy **(o)** Italy

**Figure 4.10:** temporal likes trends for top5 entertainment figures according to the number of followers, diveded by country. In the first column are the graphs of athletes, in the second those of models in the third those of actors




**(a)** France **(b)** France **(c)** France

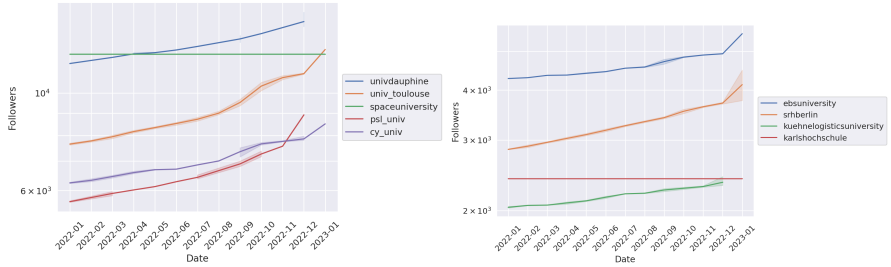
**(d)** Germany **(e)** Germany **(f)** Germany

**(g)** Spain **(h)** Spain **(i)** Spain

**(j)** UK **(k)** UK **(l)** UK

**(m)** Italy **(n)** Italy **(o)** Italy

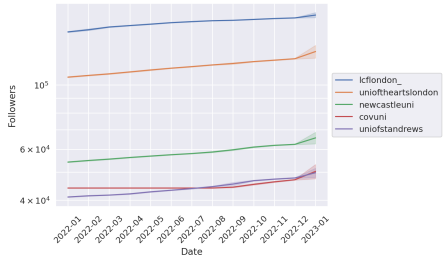
**Figure 4.11:** temporal comments trends for top5 entertainment figures according to the number of followers, diveded by country. In the first column are the graphs of athletes, in the second those of models in the third those of actors




**(a)** France

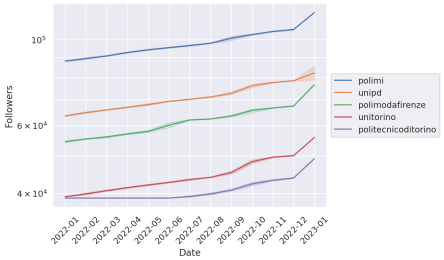



**(b)** Germany



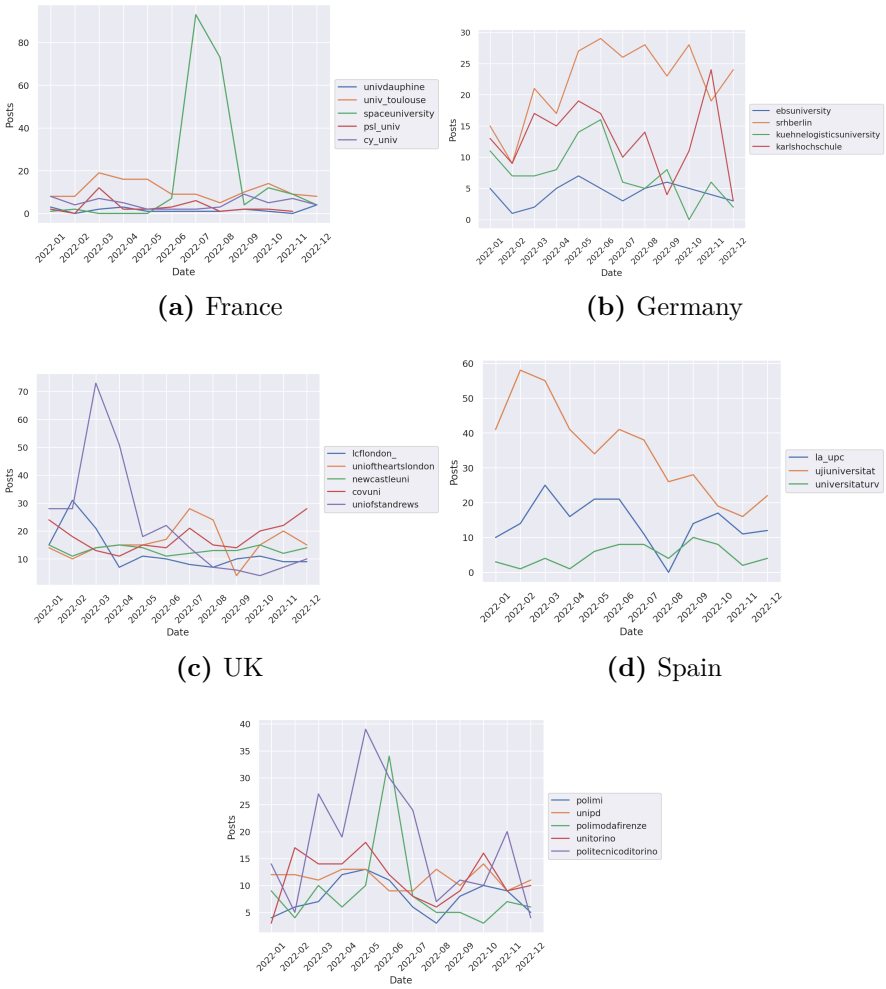

**(c)** UK


**(d)** Spain




**(e)** Italy

**Figure 4.12:** temporal followers trends for top5 universities according to the number of followers, divided by country

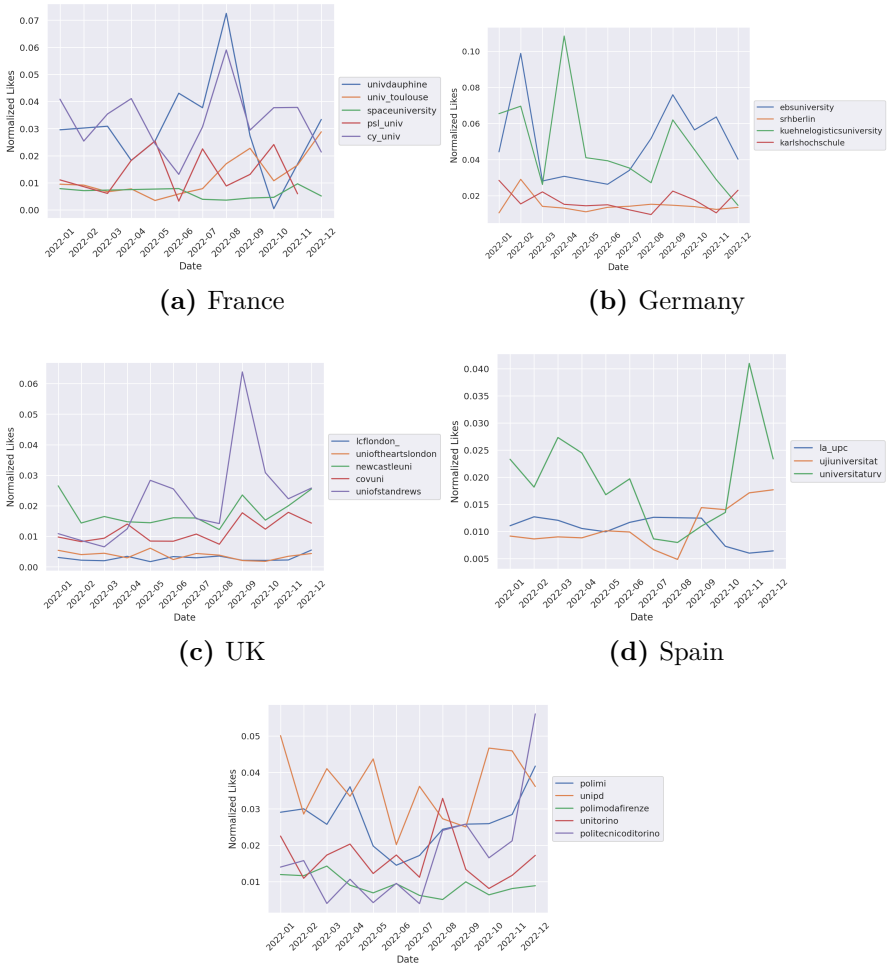



**(a)** France **(b)** Germany

**(c)** UK **(d)** Spain

**(e)** Italy

**Figure 4.13:** temporal activity trends for top5 universities according to the number of followers, diveded by country

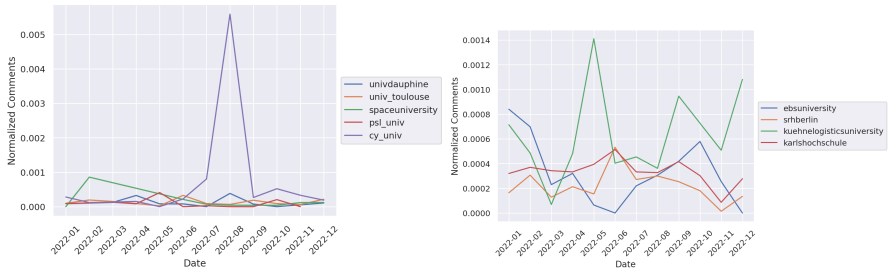



**(a)** France **(b)** Germany

**(c)** UK **(d)** Spain

**(e)** Italy

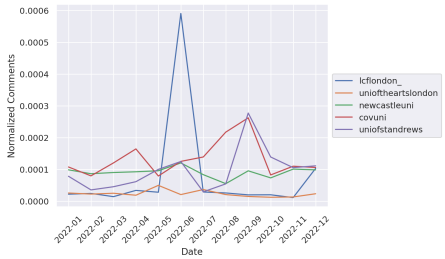
**Figure 4.14:** temporal likes trends for top5 universities according to the number of followers, diveded by country



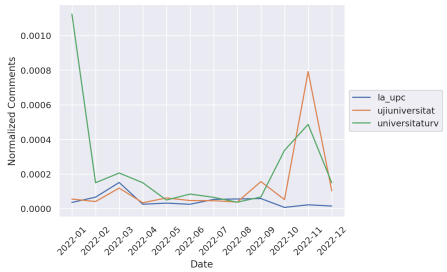

**(a)** France



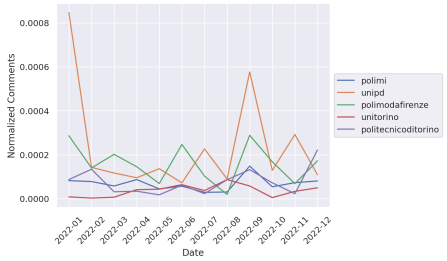

**(b)** Germany




**(c)** UK




**(d)** Spain




**(e)** Italy

**Figure 4.15:** temporal comments trends for top5 universities according to the number of followers, diveded by country

# Chapter 5

# Conclusions and Final Observations

In this concluding chapter, we will provide an overview of the results obtained from our in-depth analysis of Instagram usage and draw some key insights. The primary objective of this thesis was to understand the impact of Instagram on the homogenization or preservation of geographical and cultural identities, and now we can delve into how the data and analyses have aided us in achieving this goal.

## 5.1 Using Wikidata for Instagram Profile Profiling

A fundamental aspect that sets this research apart from others is the innovative approach employed to characterize Instagram profiles even before obtaining specific platform-related information. In this initial section, we will delve deeper into how the utilization of Wikidata enabled us to conduct a detailed and accurate analysis of the Instagram profiles involved in our study.

During the early stages of our work, we extensively relied on the Wikidata database and its semantic query language to construct a profile database containing all the necessary information for categorizing Instagram users. These profiles could encompass both individuals and organizations, such as universities, across diverse geographical and social areas of interest.

The use of Wikidata allowed us to generate lists of profiles categorized by social status, which we later further characterized by country of origin. In Chapter 3, we provided an exhaustive account of our profiling process and the dataset construction. This process involved the integration of data from Wikidata and official information acquired through Instagram APIs. Additionally, we delved into

a comprehensive examination of key attributes associated with the published posts. These attributes encompassed various facets, including post types, description lengths, and the presence of mentions or hashtags. Furthermore, we conducted a quantitative analysis of user engagement, encompassing both passive engagement in the form of "likes" and active engagement through comments.

The incorporation of Wikidata as a fundamental data source for profiling marked a significant milestone in our research methodology. This strategic choice empowered us to craft a comprehensive portrayal of Instagram users, thereby enhancing our dataset with critical insights pertaining to their identities and social affiliations.

With this groundwork established, we shall now proceed to provide an overarching summary of the pivotal findings that have emerged from our comprehensive analysis of Instagram profiles and their associated posts.

## 5.2 Key Results

Throughout this research, we analyzed 6,939 Instagram profiles and a total of 401,495 posts from five different European countries: Italy, France, Germany, the United Kingdom, and Spain, encompassing three distinct social macro-areas: politics, culture (with a specific focus on athletes, models, and actors), and academia (university profiles). In detail, we conducted a comprehensive analysis of:

- 1105 profiles of politicians, totaling 101,358 posts;
- 2576 profiles of athletes, encompassing 88,500 posts;
- 688 profiles of models, with a total of 41,446 posts;
- 2422 profiles of actors, amounting to 142,594 posts;
- 154 profiles of universities, contributing a total of 27,597 posts.

### 5.2.1 Profiles and Activity

We observed that Instagram profiles in various categories displayed significant variations in their activity and the number of followers. In characterizing the follower base of the analyzed profiles, a distinct disparity emerges between political profiles and those belonging to the entertainment figures under examination. In the case of political profiles, the majority falls into the medium-tier influencer categories (micro and mid-tier), with only a few cases surpassing the threshold of 500,000 followers or even a million. Conversely, many political figures have a following of fewer than 10,000 followers. This landscape changes significantly for entertainment figures, where most profiles are classified as mid-tier influencers,

but there are a greater number of exceptions falling into the high-tier influencer categories (macro and mega influencers), with very few profiles classified as micro or nano influencers.

Regarding academic institutions, the results reveal a significantly smaller following compared to other profile categories considered. Only three profiles exceed the threshold of 100,000 followers.

Analyses of account activity unveil further variations, particularly on a geographical level. In the case of political profiles, the United Kingdom stands out for a low frequency of posting, a behavior almost opposite to that of German politicians. Conversely, among entertainment figures, the United Kingdom leads in terms of the highest number of posts published, followed by Spain, France, and Italy.

In the case of European universities, the United Kingdom and Italy have the highest number of posts published. However, they are also the countries with academic profiles having the largest following in this profile category.

### 5.2.2 Post Analysis

Analysis of the posts revealed interesting trends in the types of posts, use of descriptions, and user interactions.

The Instagram platform allows users to publish posts in three different formats: photo, album, and video. From the analysis conducted on the post types, it is evident that politicians tend to prefer the classic format of photo posts. The exception is represented by Spain, where there is a greater inclination towards using albums. This preference for albums is common among entertainment figures, with some cases where it is comparable to the use of photos, as is the case in Italy. The video format is less frequently used by both politicians and entertainment figures. As for universities, there is a significant balance between the publication of albums and photos.

In the analysis of post descriptions on Instagram, mainly geographic differences emerge, while differences related to the social status of profiles are less obvious. The post descriptions of German profiles stand out for their significantly longer than average length. They often exceed 300 characters, indicating a preference for more detailed and informative communications. On the other hand, France and Italy maintain an average length of post descriptions between 200 and 300 characters. This suggests a more concise, but still effective, approach that may be geared toward capturing users' attention quickly. In Spain, there is a noticeable difference in post descriptions depending on the social status of the profile. Spanish politicians use post descriptions with an average length of more than 300 characters, indicating a more detailed and informative approach in their communication on Instagram. In contrast, Spanish entertainment figures maintain shorter post descriptions, with an average length around 200 characters, reflecting a more direct and concise approach

in their communication. Spanish athletes stand out for their extensive use of hashtags and mentions in their post descriptions. This suggests a tendency to make direct connections with other users and use hashtags to amplify the visibility of their content.

In our analysis of Instagram engagement, we use two metrics: the normalized number of likes over followers to value passive interaction and the normalized number of comments over followers to value active interaction. In this way, we examined how participation differs among European states and cultural personalities. These statistics offer us with an overview of the dynamics of internet engagement.

- **Politicians:** In Italy, despite having a large follower base, passive interaction is lower compared to other nations, with an average of about 40 "likes" per 1,000 followers. However, active interaction is significantly higher, with an average of about 3 comments per 1,000 followers. This suggests that, despite a more concentrated distribution of passive interactions, Italy generates broader and more engaging discussions compared to other countries.

  Spain stands out for having higher average passive interactions than other nations, with approximately 70 "likes" per 1,000 followers. Active interaction is also notable, with an average of about 4 comments per 1,000 followers. These data indicate widespread online political engagement and a different engagement strategy.

  On the other hand, the United Kingdom displays relatively low averages for both passive and active interactions, with around 30 "likes" and 1.5 comments per 1,000 followers. This situation might be attributed to the unique dynamics of British political culture and communication tactics used by politicians.

  France presents passive interactions in line with the size of the follower base, but active interaction is slightly lower compared to other nations, with around 2.5 comments per 1,000 followers.

  In Germany, passive interactions are distributed similarly to the United Kingdom, but active interaction is broader, with an average of about 4 comments per 1,000 followers. These data indicate more widespread online political activity in Germany.

- **Cultural figures:** Italian athletes have substantial passive interaction with an average of about 50 "likes" per 1,000 followers and strong active engagement with around 10 comments per 1,000 followers. Models and actors from Italy maintain moderate levels of engagement.

  Spanish athletes receive a high number of "likes" (around 60 per 1,000 followers) and have moderate comments. Spanish models and actors also enjoy significant engagement from their followers.

French athletes excel in "likes" per follower but have comparatively fewer comments. Models and actors from France attract considerable "likes."

Athletes in the UK demonstrate average engagement levels, and models and actors maintain moderate interaction with their followers. German athletes have widespread online engagement, with both high "likes" (about 70 per 1,000 followers) and comments (around 12 per 1,000 followers). Models and actors in Germany display balanced engagement.

- **Universities:** French universities stand out for having the highest passive interaction, with an average of around 20 "likes" per 1,000 followers, while German universities show similar engagement.

  Spain and France have the highest active interaction among followers of nano influencers associated with universities.

  The United Kingdom excels in active interaction among followers of nano influencers associated with universities.

### 5.2.3 Theme Recognition

After constructing and describing the dataset, in Chapter 4, we presented two types of dataset utilization. The first one involved analyzing the corpus of text descriptions in the posts to feed them into a topic recognition model with the aim of examining the topics discussed by various social figures in different European countries. For the topic recognition process, we opted to use the BertTopic model. It is an argument analysis model that employs clustering techniques and a TF-IDF-based variation to provide coherent representations of arguments in documents. It operates through three independent phases: document embedding, dimensionality reduction, and topic representation based on TF-IDF classes. This model offers a comprehensive analysis of topics within documents, enhancing our understanding of content.

From the results linked to European politicians, it becomes clear which topics are most commonly discussed in the descriptions of their Instagram posts. In France, for example, politicians use the platform somewhat lightly, sharing holiday greetings but also using it as a propaganda tool to discuss elections and political issues such as nuclear energy and healthcare. Surprisingly, the Russia-Ukraine conflict did not emerge as one of the main topics of discussion in France, unlike in the other four European countries considered. In Germany, on the other hand, the Russia-Ukraine conflict was the most discussed topic on Instagram in 2022, followed by energy-related issues, which can be considered subtopics of the conflict itself. In Italy, the predominant topic revolves around the political scenario between Italy and Europe, followed by healthcare policies, given the surge in COVID cases in the past year, and discussions on the war in Ukraine. In the United Kingdom, politicians

mainly used Instagram to discuss healthcare, focusing on staffing shortages during the most critical periods of the pandemic. Other topics addressed include policies related to the LGBTQ+ community and the Ukrainian conflict. In Spain, the most frequently mentioned topic is raising awareness about women's rights, a theme that also emerges in the posts of Italian and French politicians, albeit to a lesser extent.

Turning to the analysis of entertainment figures, it becomes evident that actors, athletes, and models primarily use Instagram to share their personal interests and activities. For example, actors utilize the platform not only to promote films, TV series, or other entertainment events they have participated in but also to provide their followers with a behind-the-scenes look at their experiences in the world of entertainment. Athletes, on the other hand, share their sporting performances in high-profile international events such as the Olympics or motorsport races, but they also use Instagram to communicate their values, promote a healthy lifestyle, and inspire others through their success stories. Models, in turn, use Instagram not only to document high-fashion events like Milan Fashion Week but also to serve as ambassadors for luxury products and brands. The platform thus becomes a tool for promoting fashion and the luxury lifestyle. What is particularly interesting is how geographical differences are gradually narrowing for these figures. Instagram now represents a crucial work tool for enhancing their image and promoting their professional activities globally. At the same time, these entertainment figures also use Instagram for entertainment and leisure purposes, increasingly aligning with the typical use of the platform by common users. This phenomenon highlights how Instagram's appeal and attraction are extending far beyond geographical boundaries, as these entertainment figures reach out to a vast global community through the platform. In this process, Instagram is becoming a meeting point between their professional and personal spheres.

Regarding the analysis of topics of interest discussed by university profiles, it becomes clear that universities in Italy and the United Kingdom prefer to promote events related to the daily life of the universities themselves, without excluding discussions on topics of global interest such as the Russia-Ukraine conflict and the passing of Queen Elizabeth. Among the most frequent topics are student health and the promotion of international study programs that pave the way for the post-academic world. Universities in Germany and France also focus on promoting international study programs and postgraduate pathways via Instagram but tend to give more attention to globally sensitive issues such as the war in Ukraine and environmental sustainability. Unfortunately, Spain was excluded from the analysis due to the lack of a substantial body of text.

### 5.2.4 Top5 Influencers

By examining the "top5 influencers" in each category and country, we were able to pinpoint those who had the most impact in the year 2022. These data provide a clear indication of the most influential profiles within each category and the extent of interactions they received.

One significant aspect is the impact of geopolitical events on the online activity of politicians. In Germany, the Russian invasion of Ukraine triggered substantial increases in followers for figures like Annalena Baerbock and Robert Habeck, both leaders of the Green Party, recording increases of 31% and 55%, respectively, between February and May 2022. These increments were closely tied to the geopolitical event. Similar situations were observed in Italy and the United Kingdom in response to key political events.

Significant variations in political communication strategies were evident. For example, in Italy, Giorgia Meloni, the current Prime Minister, gained over 500,000 followers between September and December 2022, coinciding with her victory in the October elections. In the United Kingdom, Rishi Sunak, the Prime Minister, gained over a million followers during the same period, despite the absence of official elections.

Among entertainment figures, a common trend emerged of steady follower growth, confirming the established success of these personalities online. In France, for instance, Kylian Mbappé amassed over 104 million followers, while in Spain, Sergio Ramos had nearly 59 million.

However, among entertainment figures, differences in posting strategies were apparent. Some personalities, like Dua Lipa in the United Kingdom, posted frequently, with peaks of around 50 posts per month, while others preferred a more moderate posting frequency.

In summary, these numerical data and specific profiles highlight how online dynamics can vary significantly both geographically and among different profile categories. Global events, political communication strategies, and posting preferences are just a few of the variables contributing to these differences. This research provides valuable insights for a deeper understanding of influence and online activity in diverse contexts and can be utilized for developing future communication and marketing strategies.

### 5.2.5 Final Observations

As we conclude this extensive exploration into the dynamics of Instagram usage across diverse geographical and cultural contexts, it is essential to reflect on the broader implications of our findings. Instagram, as a global social media platform, serves as a unique lens through which we can examine the convergence and divergence of identities and influences in the digital age.

Our research underscores the profound impact of geopolitical events on the online presence of political figures and how they harness Instagram to communicate with their constituents. The surge in followers during critical moments reflects not only the power of political discourse but also the platform's role as a forum for engagement with a broader audience.

In contrast, entertainment figures, such as athletes, models, and actors, showcase the global reach of Instagram. These profiles transcend geographical boundaries, demonstrating how Instagram fosters a shared cultural experience that transcends borders. It's intriguing to witness how individuals who belong to different corners of the world engage with their followers, whether by sharing personal interests or advocating for causes dear to them. This shift towards a more universal appeal showcases the potential of Instagram in creating a globalized cultural sphere.

Academic institutions, while less prominent on Instagram compared to political and entertainment figures, play a vital role in disseminating information and fostering academic discourse. Their engagement strategies differ, reflecting the diverse goals and priorities of universities across Europe. Still, they contribute to the platform's rich tapestry of content.

In summary, our research offers a multifaceted view of Instagram's role in shaping and reflecting geographical and cultural identities. It highlights the platform's adaptability to various contexts and the profound influence it wields in disseminating ideas, sparking conversations, and building communities. As Instagram continues to evolve, it will undoubtedly remain a compelling space for studying the dynamic interplay of technology, culture, and identity on a global scale. This research serves as a foundation for future studies exploring the ever-changing landscape of digital communication and its impact on societies worldwide.

### 5.2.6 Implications and Recommendations

The insights garnered from our comprehensive analysis of Instagram's impact on geographical and cultural identities carry substantial implications for various stakeholders, including policymakers, marketers, and social media platforms.

For policymakers and government agencies, understanding the role of Instagram in shaping public discourse and opinion is paramount. The significant follower surges experienced by political figures during critical events underscore the platform's potential to influence public sentiment. Policymakers must recognize the need for effective digital communication strategies and social media literacy programs to harness these tools responsibly.

Marketers and advertisers can benefit from our findings by recognizing the diverse posting strategies and engagement patterns among cultural figures. Tailoring marketing campaigns to align with these nuances can lead to more effective reach and resonance with target audiences. Moreover, understanding the global appeal

of Instagram among cultural figures highlights the platform's potential as a vehicle for international brand promotion.

For social media platforms like Instagram, our research illuminates the platform's evolving role as a global cultural arena. It underscores the importance of facilitating cross-cultural engagement and providing tools for users to transcend geographical boundaries. Instagram can further enhance its algorithms to encourage meaningful interactions and promote diverse content sharing.

In light of these implications, we recommend ongoing research in the field of social media dynamics and its impact on identities. Longitudinal studies tracking the evolving nature of online discourse can provide valuable insights into the changing landscape of digital communication.

Moreover, education and digital literacy initiatives should be promoted to help users navigate the complex world of social media, fostering responsible engagement and critical thinking.

Lastly, the platform itself, Instagram, should consider exploring features that enhance cross-cultural dialogue and promote global collaboration. Features that encourage users to explore content from different regions and engage with diverse perspectives could contribute to a more inclusive digital ecosystem.

In conclusion, our research not only sheds light on the present but also paves the way for a more informed and responsible future in the digital age. It is our hope that these findings will serve as a catalyst for further research and action to harness the potential of social media platforms like Instagram for the betterment of society and the preservation of diverse cultural identities.

## 5.3 Conclusion

In this thesis, we embarked on a comprehensive journey to explore the multifaceted impact of Instagram on geographical and cultural identities. By analyzing a diverse array of Instagram profiles and their associated posts across five European countries, we have unveiled a rich tapestry of insights that paint a complex picture of the platform's role in shaping our digital world.

Through the lens of Wikidata-driven profiling, we achieved a nuanced understanding of Instagram users across various social domains, from politics to culture and academia. This innovative approach allowed us to construct a robust dataset, enabling us to delve into the intricate details of Instagram's influence on both individual and collective identities.

Our key findings underscored the significant variations in user activity, engagement, and content preferences across different profile categories and countries. We observed how politicians, cultural figures, and universities harnessed Instagram as a dynamic tool for communication and engagement, each with their unique strategies

and objectives.

Moreover, our exploration of theme recognition exposed the diverse conversations taking place on Instagram, highlighting the impact of global events and personal interests on the topics discussed by social figures. The narrowing of geographical differences among cultural figures further emphasized Instagram's role as a global cultural arena, where personal and professional spheres converge.

Our analysis of top influencers shed light on the interplay between geopolitical events, communication strategies, and follower growth, illustrating how Instagram can be a reflection of real-world dynamics.

As we reflect on these findings, it becomes clear that Instagram, as a social media platform, wields considerable influence in shaping online discourse, transcending geographical and cultural boundaries.

In conclusion, this thesis offers a comprehensive exploration of Instagram's intricate role in influencing geographical and cultural identities. It is a call to action for further research and responsible engagement with social media platforms, with the hope that our digital world can continue to evolve in ways that celebrate diversity, foster meaningful connections, and promote a more informed and inclusive global society.

# Bibliography


[1] Jacob Cass. *OSN logos*. https://justcreative.com/best-social-media-logos/. Accessed on August 7, 2023. 2023 (cit. on p. 3).

[2] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. «The benefits of Facebook "friends:" Social capital and college students' use of online social network sites». In: *Journal of computer-mediated communication* 12.4 (2007), pp. 1143–1168 (cit. on p. 3).

[3] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. «Automatic personality assessment through social media language.» In: *Journal of personality and social psychology* 108.6 (2015), p. 934 (cit. on p. 3).

[4] Danah M Boyd and Nicole B Ellison. «Social network sites: Definition, history, and scholarship». In: *Journal of computer-mediated Communication* 13.1 (2007), pp. 210–230 (cit. on pp. 3, 6).

[5] Scott A Golder and Michael W Macy. «Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures». In: *Science* 333.6051 (2011), pp. 1878–1881 (cit. on p. 4).

[6] Danah Boyd. «Why youth (heart) social network sites: The role of networked publics in teenage social life». In: *YOUTH, IDENTITY, AND DIGITAL MEDIA, David Buckingham, ed., The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, The MIT Press, Cambridge, MA* (2008), pp. 2007–16 (cit. on p. 4).

[7] David Kirkpatrick. *The Facebook effect: The inside story of the company that is connecting the world*. Simon and Schuster, 2011 (cit. on p. 4).

[8] Jacob Cass. *LinkedIn. (n.d.). Our Story.* https://about.linkedin.com/. Accessed on August 7, 2023. 2023 (cit. on p. 4).

[9] Vincos Blog. *Social media in Italia: utenti e tempo di utilizzo 2022*. https://vincos.it/2023/08/05/social-media-in-italia-utenti-e-tempo-di-utilizzo-2022/. Accessed on August 7, 2023. 2023 (cit. on pp. 4, 5).

[10] Stacy Jo Dixon. *Number of global social media users 2017-2025*. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Accessed on August 7, 2023. 2023 (cit. on p. 5).

[11] Stacy Jo Dixon. *Number of monthly active Facebook users worldwide as of 2nd quarter 2023*. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/. Accessed on August 7, 2023. 2023 (cit. on p. 5).

[12] L. Ceci. *Number of monthly active WhatsApp users worldwide from April 2013 to March 2020*. https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/. Accessed on August 7, 2023. 2023 (cit. on p. 5).

[13] Pew Research Center. *Social Media Fact Sheet*. https://www.pewresearch.org/internet/fact-sheet/social-media/. Accessed on August 7, 2023. 2021 (cit. on p. 5).

[14] Simon Kemp. *Digital 2022: Global Overview Report*. https://datareportal.com/reports/digital-2022-global-overview-report. Accessed on August 7, 2023. 2022 (cit. on p. 5).

[15] Peter Kim. *Online Promotional Marketing Frameworks Honeycomb Model Of Social Media Structure*. https://www.slidegeeks.com/business/product/online-promotional-marketing-frameworks-honeycomb-model-of-social-media-structure-pdf. Accessed on August 7, 2023. 2010 (cit. on p. 6).

[16] Jan Kietzmann, Kristopher Hermkens, Ian McCarthy, and Bruno Silvestre. «Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media». In: *Business Horizons* 54.3 (May 2011), pp. 241–251. DOI: 10.1016/j.bushor.2011.01.005 (cit. on p. 6).

[17] Mike Thelwall and David Wilkinson. *Social media and impression management: Self-presentation in the digital age*. Routledge, 2019 (cit. on p. 6).

[18] Barry Wellman and Keith Hampton. «Living networked in a wired world: Foraging for future models in the present». In: *Knowledge, Technology & Policy* 12.3 (1999), pp. 1–24. DOI: 10.1007/s12130-999-1026-0 (cit. on p. 7).

[19] Koosha Zarei, Damilola Ibosiola, Reza Farahbakhsh, Zafar Gilani, Kiran Garimella, Noël Crespi, and Gareth Tyson. «Characterising and detecting sponsored influencer posts on Instagram». In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2020, pp. 327–331 (cit. on pp. 8–11).

[20] Davide Morante. *Anno dei Nano Influencer: Come i Marchi Collaborano con Piccole Ma Potenti Voci*. 2022. URL: https://www.shopify.com/it/blog/anno-dei-nano-influencer (cit. on p. 8).

[21] Rana Tallal Javed, Mirza Elaaf Shuja, Muhammad Usama, Junaid Qadir, Waleed Iqbal, Gareth Tyson, Ignacio Castro, and Kiran Garimella. «A First Look at COVID-19 Messages on WhatsApp in Pakistan». In: *arXiv preprint arXiv:2011.09145* (2020). Cited on pages 12, 13, 16. URL: https://arxiv.org/abs/2011.09145 (cit. on pp. 11, 12).

[22] Derek Weber and Frank Neumann. «Who's in the Gang? Revealing Coordinating Communities in Social Media». In: *arXiv preprint arXiv:2010.08180* (2020). Cited on pages 13, 14. URL: https://arxiv.org/abs/2010.08180 (cit. on pp. 11, 12).

[23] Derek Weber, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon. «A Method to Evaluate the Reliability of Social Media Data for Social Network Analysis». In: *arXiv preprint arXiv:2010.08717* (2020). Cited on page 14. URL: https://arxiv.org/abs/2010.08717 (cit. on p. 13).

[24] Matteo Cardaioli, Pallavi Kaliyar, Pasquale Capuozzo, Mauro Conti, Giuseppe Sartori, and Merylin Monaro. «Predicting Twitter Users' Political Orientation: An Application to the Italian Political Scenario». In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Cited on page 15. 2020, pp. 159–165. DOI: 10.1109/ASONAM49781.2020.9381470 (cit. on p. 13).

[25] David M Blei, Andrew Y Ng, and Michael I Jordan. «Latent Dirichlet Allocation». In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022 (cit. on p. 14).

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. «Attention Is All You Need». In: *Advances in Neural Information Processing Systems*. 2017 (cit. on p. 14).

[27] Gregor Heinrich. «LDA: Latent Dirichlet Allocation». In: *Proceedings of the 22nd international conference on Machine learning*. 2011 (cit. on p. 15).

[28] Liangjie Hong and Brian D Davison. «Discovering coherent topics using general knowledge». In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 2011, pp. 457–466 (cit. on p. 15).

[29] Xingyu Chen and Ashok N Srivastava. «A survey of natural language processing techniques for opinion mining systems». In: *Journal of Information Science* 46.6 (2020), pp. 840–861 (cit. on p. 15).

[30] Martino Trevisan, Luca Vassio, Idilio Drago, Marco Mellia, Fabricio Murai, Flavio Figueiredo, Ana Paula Couto da Silva, and Jussara M Almeida. «Towards Understanding Political Interactions on Instagram». In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. (cit. on p. 24). 2019. URL: http://hdl.handle.net/11583/2752645 (cit. on p. 16).

[31] Carlos Henrique Gomes Ferreira, Fabricio Murai, Ana Paula Couto da Silva, Jussara Marques de Almeida, Martino Trevisan, Luca Vassio, Idilio Drago, and Marco Mellia. «Unveiling Community Dynamics on Instagram Political Network». In: *ACM Conference on Web Science*. Cited on pages 24 and 25. 2020 (cit. on pp. 16, 18).

[32] Martino Trevisan, Luca Vassio, and Danilo Giordano. «Debate on online social networks at the time of COVID-19: An Italian case study». In: *Online Social Networks and Media* 23 (2021). Cited on page 27, p. 100136. ISSN: 2468-6964. DOI: 10.1016/j.osnem.2021.100136 (cit. on pp. 17, 19).

[33] Luca Vassio, Michele Garetto, Carla Chiasserini, and Emilio Leonardi. «Temporal Dynamics of Posts and User Engagement of Influencers on Facebook and Instagram». In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Cited on page 28. Association for Computing Machinery. 2021, pp. 129–133. ISBN: 9781450391283. DOI: 10.1145/3487351.3488340. URL: https://doi.org/10.1145/3487351.3488340 (cit. on p. 18).

[34] Fabio Bertone, Luca Vassio, and Martino Trevisan. «The Stock Exchange of Influencers: A Financial Approach for Studying Fanbase Variation Trends». In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Cited on page 28. Association for Computing Machinery. 2021, pp. 431–435. ISBN: 9781450391283. DOI: 10.1145/3487351.3488413. URL: https://doi.org/10.1145/3487351.3488413 (cit. on p. 19).

[35] Gabriel Altay. «Introducing the Kensho Derived Wikimedia Dataset». In: *MisinfoCon* (Feb. 2020), p. 8. URL: https://misinfocon.com/introducing-the-kensho-derived-wikimedia-dataset-c3a055f16d10 (cit. on p. 24).

[36] Emily Sundberg. *9 Very Important Ways to Use Instagram's New Album Feature*. 2017. URL: https://www.thecut.com/2017/02/the-new-instagram-album-feature.html (cit. on p. 42).

[37] Maarten Grootendorst. «BERTopic: Neural topic modeling with a class-based TF-IDF procedure». In: *arXiv preprint arXiv:2203.05794* (2022) (cit. on p. 65).

[38] Maarten Grootendorst. *BERTopic: Getting Started- 6A Representation models*. 2023. URL: https://maartengr.github.io/BERTopic/getting_started/representation/representation.html (visited on 10/03/2023) (cit. on p. 65).

[39] DW News. *Ukraine: German Foreign Minister says Russia's war is a turning point*. Accessed on: October 5, 2023. 2023. URL: https://www.dw.com/en/ukraine-german-foreign-minister-says-russias-war-is-a-turning-point/a-64783386 (cit. on p. 70).

[40] European Commission. *Speech by President Ursula von der Leyen at the Hungarian Parliament*. Accessed on: October 5, 2023. 2022. URL: https://ec.europa.eu/commission/presscorner/detail/hu/speech_22_2321 (cit. on p. 70).