Data Management Project 2022/2023

Cervini Stella Mattia Simone Montalbano Daniel

Index

1. Introduction

Social networks are nowadays incredible popular, and they started to be seen not only as a source of leisure but also as a business opportunity. Engaging the public online could be crucial for a brand or a company, to the point that they are willing to invest substantial capitals to promote their products on platforms such as Facebook, Instagram, TikTok and so on. It is common knowledge that social media mostly make their revenue by selling advertising space to corporations. However, this is not the only way in which labels can expand their presence online: sometimes brand rely on influencer to promote their trademarks and products, confident that their massive public can be attracted by what they are advertising. Focusing on Italy, there are 43.83 million social media users in January 2022. Considering that the country's population is 60.32 million ca. that is impressive. Nevertheless, Meta's social media platform are overall the most used, with Instagram counting 28.8 million users. Given that, having a good reputation is necessary for an influencer in order to get the most from their brand deals. In this perspective, the number of followers on a certain platform can be a discriminating factor: more followers, more popularity, more brand deals.

2. Goal of the project

In this project, we will try to determine if news can influence the number of followers and the performances in terms of likes and public engagement of the top 100 Instagram's influencers in Italy in 2022. Our main goal is to understand if public news may affect the follower's growth of the top 100 influencers in Italy based on follower numbers.

The final analysis to give an answer to our questions was created in Tableau Public, so this project aims to build a dataset that could be given as input to the software.

Instagram Instagram is a photo and video sharing social networking service owned by American company Meta Platforms. The app allows users to upload media that can be edited with filters and organized by hashtags and geographical tagging. Posts can be shared publicly or with preapproved followers. Users can browse other users' content by tag and location, view trending content, like photos, and follow other users to add their content to a personal feed.

Impact on businesses Instagram can help promote commercial products and services. It can be distinguished from other social media platforms by its focus on visual communication, which can be very effective for business owners. The platform can also lead to high engagement, which is due to its large user base and high growth rates. The platform can also help commercial entities save branding costs, as it can be used for free even for commercial purposes. However, the inherently visual nature of the platform can in some ways be detrimental to the presentation of content.

3. Data architecture

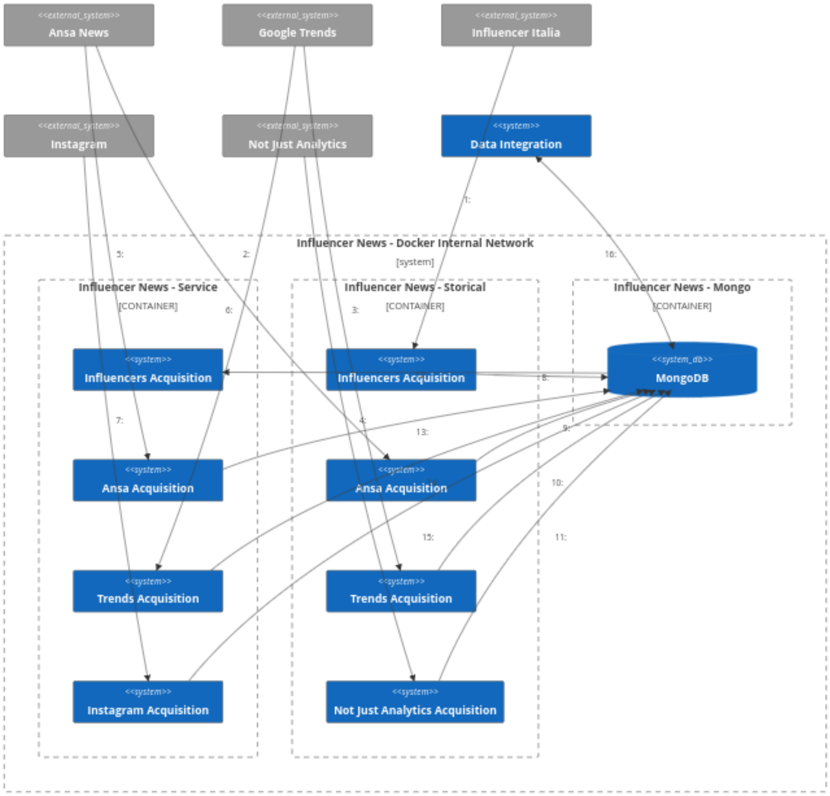
In order to achieve our research goals, we had to design and develop an infrastructure that would allow for both historical and continuous acquisition over time.

More specifically, we used a Docker Compose architecture consisting of a MongoDB container and two Python containers, built from the same Dockerfile, one dedicated to historical data acquisition and the other to real-time data acquisition.

The decision to use a Docker Compose architecture was made mainly for

- reproducibility: containers provide a consistent environment for an application, which can be easily replicated on different systems, and simplifies teamwork on different operating systems;
- reliability: since the real-time data acquisition service should remain in running state for a long period of time to acquire a large amount of data, Docker compose takes care of automatically restarting the container in case of errors. This feature combined with checks in the code, allows reducing the number of missing data in the dataset.

Data flow diagram:



The diagram above shows the flow of data throughout the project and allows us to cover the differences, in terms of data sources, between historical and real-time acquisition.

The data integration and data quality assessment phases were carried out using a Jupyter notebook that interfaced directly with the Mongo database.

4. Data acquisition

The objective of this step was to find different sources containing the data of our interest and to define a process for their acquisition. The main data sources are Instagram, the Ansa's news feed and Google Trends. Specifically, we needed the following data for each of the most followed people on Instagram in Italy:

- their Instagram's daily followers' growth;
- news and articles about them;
- an index of their popularity online.

4.1 Influencer Italia

Influencer Italia is a network for influencers, brands and communication agencies with the goal of simplifying the search and definition phase of the right influencer, for all brands and agencies that want to optimize their investments in digital campaigns and integrated communication activities.

We decided to scrape this page, using the python library BeutifulSoup, to obtain username, position, name, and surname of the top 100 influencers on the chart based on the number of followers on Instagram. This list was crucial for extracting the data from the next sources, since, usually, it is given as input to the function that executes the retrieving.

Here is the list of the top 10 (updated at 03/01/2023):

1. @khaby00
2. @chiaraferragni
3. @gianlucavacchi
4. @iammichelemorroneofficial
5. @valeyellow46
6. @fedez
7. @mb459
8. @belenrodriguezreal
9. @gianluigibuffon
10. @mrancelotti

4.2 Instagram data

Data from Instagram had two different sources, the official Instagram API and historical data provided by Not Just Analytics.

Instagram API We noticed that if you are authenticated on Instagram and you search https://www.instagram.com/username/?a=1&d=dis you obtain as answer a JSON file that contains all the basic information about the profile we were looking for. So, we copied the request as a curl, and we imported it in Postman software, where we can analyze all the params of the request and we can replicate it.

Each response contained a huge amount of data, we decided to store only the data we were interested in:

- username: Instagram's nickname of the user (string);
- followers: follower's number (number);
- following: users followed by user (number);
- full_name: name and surname (string);
- is_business_account: true if the type of the account is business (boolean);
- business_category_name and category_name: helps visitors to understand what type of business the profile is about (string);
- is_verified: true if the identity is verified by Instagram (boolean);
- profile_pic_url: user's profile image URL (string);
- post_count: number of posts published by the user (number);
- post: last 12 posts published (array of objects).

We decided to store the posts, since the request returns only the last 12, but actually we do not use them. They can still be useful for further future analysis.

By carrying out this process every day, in a medium-long period, it is possible to obtain a large amount of data on which to carry out our analysis. However, by retrieving data only with this method, we were not able to get historical daily data previous to the starting day of the project. For this reason, we had to find a way to get the historical information we needed.

Not Just Analytics Since we wanted to find the follower's daily growth of the past year (2022), we asked an API access to Not Just Analytics, a company that provides a tool for analyzing Instagram's influencers. NJL has the access to the official Instagram's API, since the company is an official partner.

The token provided us must be refreshed every 30 minutes, so we decided to store in a JSON file the history of each influencer for every day of 2022, running the acquisition script on 31/12/2022. In order to reduce the historical missing data, we also decided to store the top 100 ranking from 31/12/2022, saved in a JSON file.

For each influencer and for each day or the year 2022, an object that incorporated the following information we were interested in:

- date: date of the acquisition (string);
- follower: number of followers (number);
- following: number of following (number);
- post: number of post (number);

4.3 Ansa's News

The ANSA is an acronym that stands for "Agenzia Nazionale Stampa Associata", literally "National Associated Press Agency", and it is the first multimedia information agency in Italy and among the first in the world.

In order to retrieve what we were looking for, we noticed that we were able to create an unauthenticated request on the search page without any daily limit. So, we created a function that, given a keyword and a time range, search all the articles related to the keyword in the given time range. Once a reply is obtained, the function scrapes the response page to extract, for each article, an object containing the following information:

- title: title of the article (string);
- article_datetime: datetime of the article publication (Datetime);
- inTitle: if the keyword is in the article's title (boolean);

By carrying out this process for each influencer (which name and username serve as keywords), using a time range of one year, it is possible to obtain all articles published in 2022 about that person.

4.4 Google Trends API

Google Trends is a Google service that analyzes the popularity of top search queries in Google Search across various regions and languages and allows comparing the search volume of different search term over time.

The popularity of the search term is represented by a number from 0 to 100, with higher numbers indicating greater relative popularity. It is important to note that the output from Google Trends is not an absolute measure of the popularity of a search term, but rather a relative measure that is intended to show how the search term's popularity has changed over the indicated time compared to all other searches on Google.

To retrieve the data from Trends, we have used the python library PyTrends, an unofficial API for Google Trends that provides a simple interface for automating the download of reports from Google Trends.

Given a set of search term (e.g. the influencer's username and real name) and a time period, we obtain a list of object with the following fields:

- date: date related to the given value (Date);
- max trend: maximum popularity index in that day among all search terms, it's between 0 and 100 (number);

We retrieve the trends as weekly data and not daily, since, when fetching data on a time interval of one year, Google Trends returns it in that format. We considered that modifying the aggregation format could bring some accuracy issues, therefore we kept it like this.

Performing this process for all the influencer, providing a specific time range, makes it possible to obtain the trend popularity index for each day of 2022 about that person.

5. Data Storage & Modelling

Therefore, after the phase of Data Acquisition, we had the following data sources:

- a list of the Top 100 influencers;
- a list containing all Instagram data on all the influencers from the official Instagram API;
- a JSON containing historical Instagram data on all the influencers from NJL;
- a list containing all ANSA articles on all influencers;
- a list containing all Google Trends data on all influencers.

Considering that we obtained several JSON files from the Data Acquisition step, using a document based database (NoSQL) appeared a reasonable choice, partly because a NoSQL DB can easily store them.

We created a database called 'DB-TEST', containing a collection for each data source, and each time new data was collected, a write operation was performed to store it.

Here are some examples of queries on a single data sources:

- *news* collection: find the latest article regarding Valentino Rossi.

```
mydb = client["DB-TEST"]
mycol = mydb["news"]
lastDay = datetime(2022, 12, 31)
lastArticle = mycol.find({"username": "valeyellow46",
                          "timestamp":{"$lt": lastDay}}).limit(1)
```

Output:

```
{
 '_id': ObjectId('63c4147b51c0d54814d54150'),
 'acquisition_datetime': datetime.datetime(2023, 1, 15, 14, 57, 48, 385000),
 'username': 'valeyellow46',
 'timestamp': datetime.datetime(2022, 12, 22, 14, 27),
 'title': 'Valentino Rossi in pista con Bmw M Motorsport',
 'inTitle': True
}
```

- *instagram* collection: find the total number of followers that Chiara Ferragni had on the last day of the year.

```
lastDay = datetime(2022, 12, 31)
mycol = mydb["instagram"]
a = mycol.find({"username":"chiaraferragni", "timestamp":{"$eq":lastDay}},
               {"followers":1, "_id":0})
```

Output:

```
{
    'followers': 28333515
}
```

At this point, we had a database containing various collections, each for every data source we considered, and the Data Integration phase could begin.

6. Data Integration and Enrichment

The data integration phase, carried out using a Jupyter Notebook, aims to integrate all the data from different sources, saved in different Mongo collections, in two different way:

- JSON file;
- CSV format (Tableau required data in relational model).

The first CSV dataset contains for each Top 100 influencer the data acquired for each day of the year, more specifically each row has the following fields:

- day;
- username: influencer's Instagram username;
- trend: popularity index in that day;
- Ansa: number of articles published from the 01/01/2022;
- ansa_delta: difference of number of articles published compared to the previous day;
- followers: Instagram followers at that day;
- followers_delta: difference of Instagram followers compared to the previous day;
- following: Instagram following at that day;
- following_delta: difference of Instagram following compared to the previous day;
- post: Instagram post on the profile at that day;
- post_delta: difference of Instagram post on the profile compared to the previous day.

The second CSV dataset contains each Ansa article stored during the data acquisition phases, more specifically, each row has the following fields:

- date: publish date of the article;
- username: Instagram username of the influencer;
- title: title of the article.

The JSON dataset is a union of the two datasets: each element has the format of the first CSV file with an "ansa_articles" attribute, that contains all the articles published on that day related to that influencer (with the same fields as the second CSV dataset).

7. Data Quality

Data Quality can be defined as "fit for use". For us, this meant that the most significant data quality dimension should be *completeness.*

Since we considered Instagram's data as the main source and the other as integration/enrichment, when the first one had missing data, we simply decided to remove such data points, in order to obtain a final dataset that could be as complete as possible.

Completeness

Google Trends For this data source, an issue was that for some influencers the trend was not present, simply because there were no search for the terms we were looking for and Google did not record any data for the period and the person considered. In such cases, we inserted the value zero as trends for all the missing rows, so that, even if there were no data for a given influencer, their tuple could be inserted in the dataset, and it did not have any empty rows.

Evaluation We carried out a number of data quality evaluations on the completeness dimension on the two CSV datasets.

```
Data quality completeness evaluation - InfluencersDataset

Number of rows: 35979
Number of columns: 11

Object completeness: 98.57260273972602%

Tuple completeness (in %)
0.278% tuple with 72% of completeness
99.722% tuple with 100% of completeness

Attribute completeness (in %)
username           100.000
day                100.000
followers          100.000
following          100.000
post               100.000
followers_delta     99.722
following_delta     99.722
post_delta          99.722
trends             100.000
ansa               100.000
ansa_delta         100.000
dtype: float64

Table completeness: 99.924%
```

```
Data quality completeness evaluation - ArticlesDataset

Number of rows: 2984
Number of columns: 4

Tuple completeness (in %)
100.0% tuple with 100% of completeness

Attribute completeness (in %)
username    100.0
day         100.0
title       100.0
inTitle     100.0
dtype: float64

Table completeness: 100.0%
```

Example of accuracy

Influencer aliases When we were searching for the news about every influencer in the top 100 chart, we noticed that using only the Instagram's username as keyword was not returning the number of articles expected. For this reason, we decided to create a list of aliases for each influencer that initially contained the Instagram's username and the real name of the person. However, during the fetching of the articles, we notice that sometimes these names were still not enough or were misspelled. Therefore, we checked manually all the names in the list to reach the accuracy of 100% according to the person in the real world.

Considerations on currency/timeliness

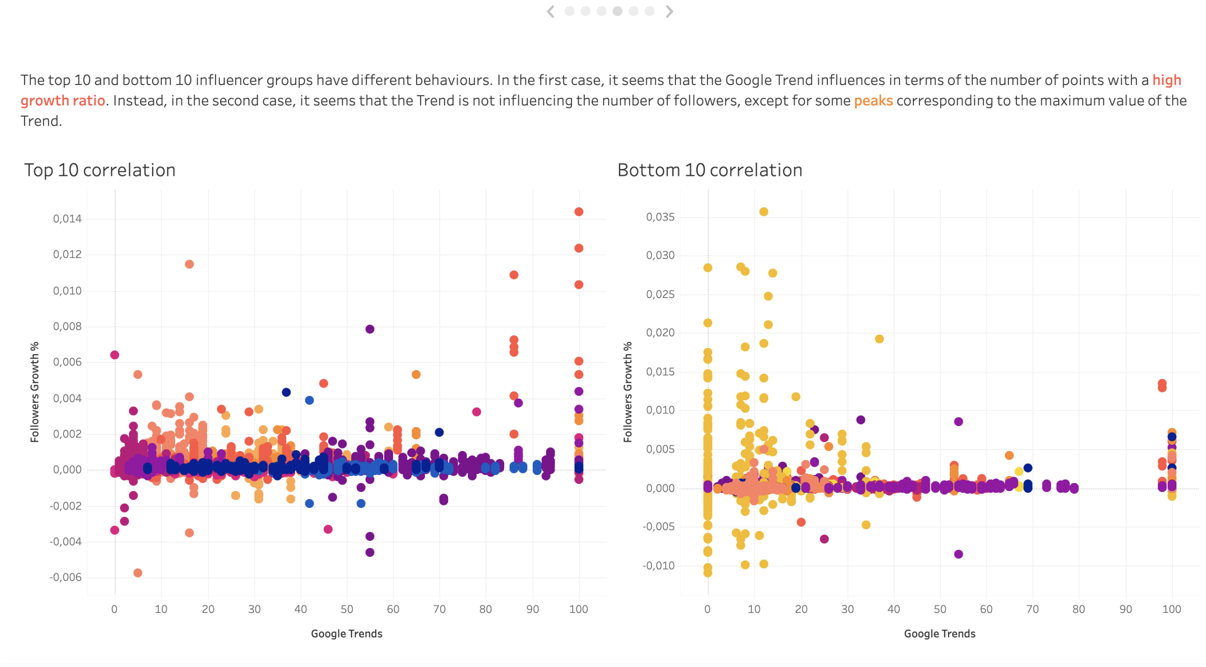
When we analyze the daily growth of an Instagram's profile, we are pretty confident that the accuracy cannot be of 100% because the data is stored once a day. This means that we are not able to catch every follower variation during the day, but we have the variation on an interval of 24 hours.

8. Data Analysis & Visualization

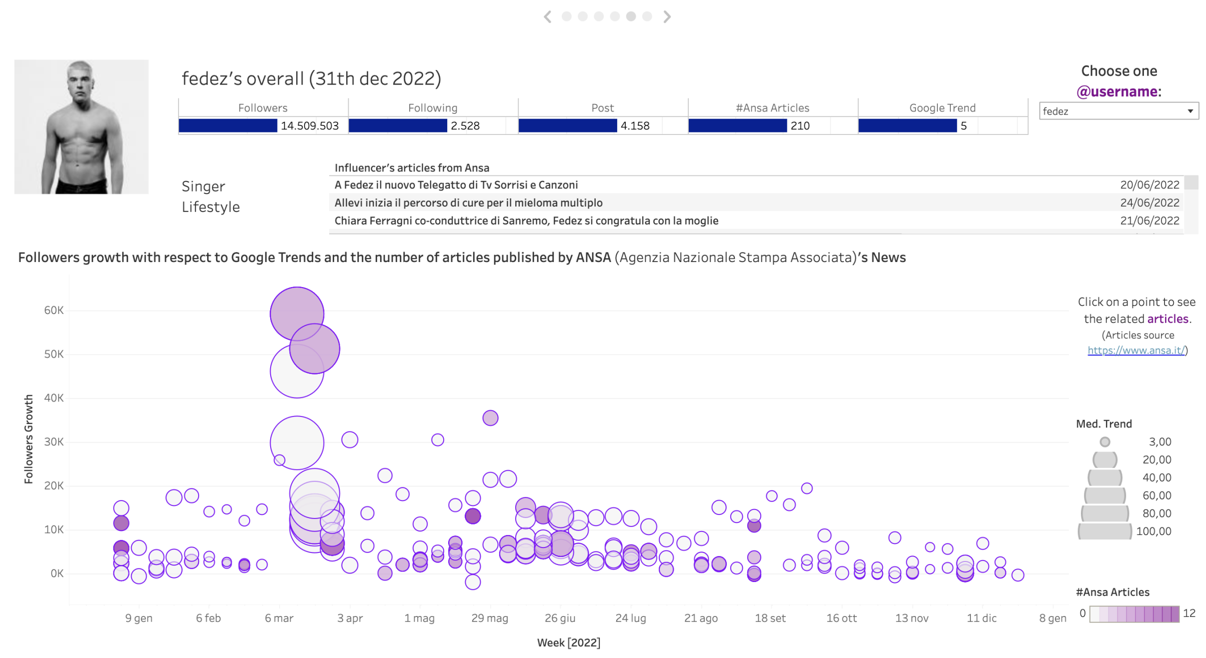
The final goal of our data management project was to build a dataset that could be given in input to Tableau Public, in order to create some data visualization to analyze the data. The final Tableau project can be found at this link.

Many questions drove the implementation of our project. In particular, we wanted to discover, for what concerns Italy, who are the most influential people on Instagram, what they do (in terms of occupation and content posted on the platform), why they are so popular and if their online presence is affected by news and trends. The last point was the most substantial part of the data visualization project, which finally lead to the creation of a Tableau story.

Here are some screenshots from the story.



The top 10 and bottom 10 influencer groups have different behaviours. In the first case, it seems that the Google Trend influences in terms of the number of points with a high growth ratio. Instead, in the second case, it seems that the Trend is not influencing the number of followers, except for some peaks corresponding to the maximum value of the Trend.



fedez's overall (31th dec 2022)

Followers	Following	Post	#Ansa Articles	Google Trend
14.509.503	2.528	4.158	210	5

Choose one @username: fedez

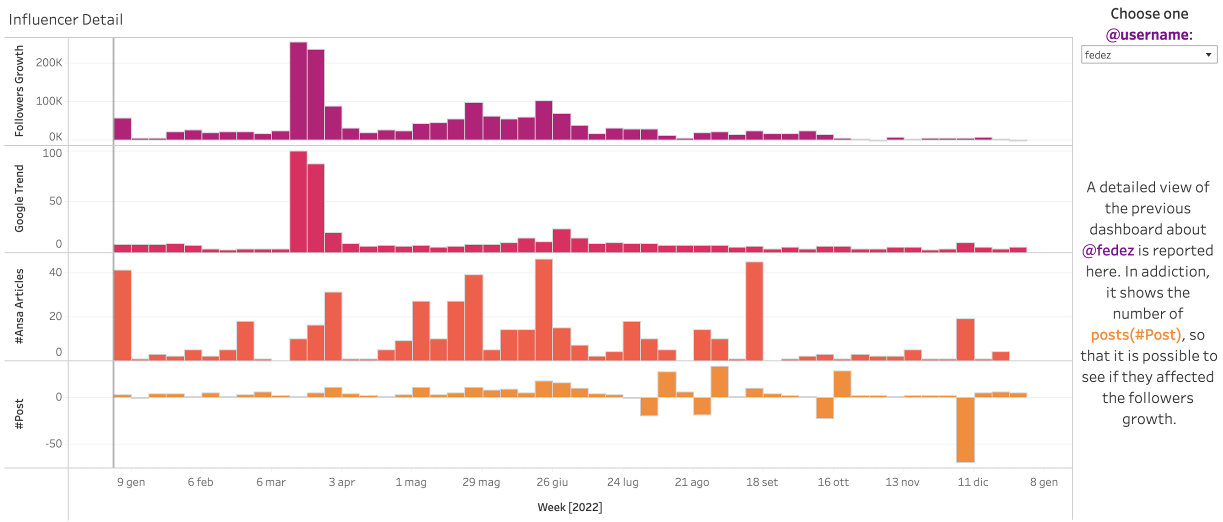
Singer
Lifestyle

Influencer's articles from Ansa	
A Fedez il nuovo Telegatto di Tv Sorrisi e Canzoni	20/06/2022
Allevi inizia il percorso di cure per il mieloma multiplo	24/06/2022
Chiara Ferragni co-conduttrice di Sanremo, Fedez si congratula con la moglie	21/06/2022

Followers growth with respect to Google Trends and the number of articles published by ANSA (Agenzia Nazionale Stampa Associata)'s News

Click on a point to see the related articles. (Articles source https://www.ansa.it/)

@fedez's details



Influencer Detail

Choose one @username: fedez

A detailed view of the previous dashboard about @fedez is reported here. In addiction, it shows the number of posts(#Post), so that it is possible to see if they affected the followers growth.

9. Conclusions & possible future developments

Focusing on the data management side, we obtained two datasets, one for the top 100 Instagram's influencers and one for the articles, with a good degree of completeness and accuracy.

By performing analyzes on Tableau dashboards, we were able to reach a relevant goal, which was to find an answer to the questions we asked ourselves at the beginning of the development: we can assert that news and trends do influence the popularity of some profiles online, sometimes positively and other negatively, and what influencers post on Instagram, as well as what they do in real life, can be the reason for their successes or setbacks.

Certainly, there are a lot of improvements that we could do and some future developments:

- storage all posts and analyze also them;
- handle in a better way the historical missing data;
- integrate new data sources.