Data Visualization Project 2022/2023

Cervini Stella Mattia Simone Montalbano Daniel

Index

The link to the project on Tableau Public can be found here.

1. Introduction

Social networks are nowadays incredibly popular. Focusing on Italy, there were 43.83 million social media users in January 2022: considering that the country's population is approx. 60.32 million, that is impressive. Nevertheless, Meta's social media platform are overall the most used, with Instagram counting 28.8 million users.

In recent years, social networks have 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, relying on the fact that their massive audiences can be captivated by what their are advertising.

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 decisive factor: more followers, more popularity, more brand deals. In this context, the role of press agencies can be relevant given that they can alter the reputation of a public figure and, consequently, their possibilities of collaborations with brands.

We wondered if it is possible to quantify the influence of news agencies on the performances, in terms of followers growth and public engagement, of the top 100 Italian Instagram influencers. To answer this question is necessary to acquire, store, integrate data from different sources, mainly regarding the performance on social media and the popularity in newspapers, and visualize the integrated data in a dashboard that allow the user to perform an analysis.

2. Data acquisition

The objective of this phase was to find different sources containing the data of our interest and to define a process for their acquisition. Specifically we needed the following data:

- Instagram daily followers growth;
- the most followed people on Instagram in Italy;
- an index of popularity in newspapers.

2.1 Instagram

Instagram is a photo and video sharing social networking service owned by the 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.

Instagram was originally distinguished by allowing content to be framed only in a square (1:1) aspect ratio of 640 pixels to match the display width of the iPhone at the time. In 2015, this restriction was eased with an increase to 1080 pixels. It also added messaging features, the ability to include multiple images or videos in a single post, and a Stories feature—similar to its main competitor Snapchat—which allowed users to post their content to a sequential feed, with each post accessible to others for 24 hours. As of January 2019, Stories is used by 500 million people daily.

Originally launched for iOS in October 2010 by Kevin Systrom and Mike Krieger, Instagram rapidly gained popularity, with one million registered users in two months, 10 million in a year, and 1 billion by June 2018. In April 2012, Facebook Inc. acquired the service for approximately US$1 billion in cash and stock. The Android version was released in April 2012, followed by a feature-limited desktop interface in November 2012, a Fire OS app in June 2014, and an app for Windows 10 in October 2016. As of October 2015, over 40 billion photos had been uploaded. Although often admired for its success and influence, Instagram has also been criticized for negatively affecting teens' mental health, its policy and interface changes, its alleged censorship, and illegal and inappropriate content uploaded by users.

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.

Instagram API We have noticed that if you are authenticated on Instagram and you visit the address `https://www.instagram.com/username/?__a=1&__d=dis` you obtain in answer a JSON file that contains all the basic information about the profile we are looking for and the latest twelve posts published.

So, for each influencer, we reproduced this Instagram request, we obtained a JSON file and we saved only the fields we were interested in:

- number of followers: number of users that follow the influencer;
- id user Instagram: unique id of the profile;
- number of following: number of users followed by the current user;
- user's full name: name and surname of the user;
- is_business_account: attribute that indicates if the current profile is business type;
- business_category_name and category_name: helps visitors to understand what type of business the profile is about;
- is_verified: whether the profile is verified by Instagram or not;
- profile image
- number of media published: numbers of posts published by the user;
- last 12 post published

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 analysis.

Not Just Analytics API Based on our goal we need to find the follower's daily growth of the past year (2022), so we could not use only the official Instagram API to get the past data: we needed a historical source of data.

We asked an API access to Not Just Analytics, a company that provides a tool for analyzing Instagram's influencers. Since NJL is an official Instagram partner, they could retrieve for us the data we needed.

We obtained for each day a JSON object that contained the following information for each influencer we were interested in:

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

2.2 Influencer Italia

Influencer Italia is a network for influencers, brands and communication agencies with the goal to simplify 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.

Here 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

2.3 Influencer and content category

In the previous Instagram API paragraph, we introduced the business_category_name and the category_name fields, which are user's profile data regarding the type of content posted.

During the analysis we noticed that this data could not always be reliable because sometimes it doesn't exists or sometimes it did not reflected the real kinds of content you can find on the posts in the influencer's profile. Since this profile's description is without constraints, the user can specify the one he prefers, even if it does not represent his real occupation.

For this reason, we grouped the influencers manually, inserting them in one if the following categories:

- Content creator
- Public figure
- Actor
- Athlete
- Singer
- Model
- TV Entertainer
- Photographer
- Art
- Athlete
- Journalist
- Designer
- DJ

We called this field influecer category" and has the goal to represent a role/occupation of the influencer. Furthermore, we observed that, despite falling into the same category, influencers could bring different types of content on the platform. For this reason, we found interesting to add also a content category that could describe deeply what an influencer posts on Instagram. The Content Categories we identified were:

- Comedy
- Lifestyle
- Sport
- Football
- Photography
- Music
- Adult content
- Food
- TV
- Makeup

For example, @gianluigibuffon and @marchisiocla8 have both Athlete as influencer category but their content categories are different, since the first one posts about Football while the second about its personal life (Lifestyle). In this context, the content category can be see a sub-category of influencer category. It is worth mentioning that these classifications are done manually and one-off, so the categories may change, especially the one regarding the content. Also, they could be inaccurate, since they are subjective. However, we tried to be as unbiased and careful as possible during the categorization process.

2.4 Ansa 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.

It was founded in Rome in 1945 in a cooperative form by the first six newspapers of liberated Italy. Today is a cooperative of 36 members of the main Italian newspapers publishers and is designed to collect and transmit information on the main events in Italy and in the world. The ANSA agencies transmit more than

3,500 news and 1,500 photos a day that are broadcasted to the Italian media, national institutions, local and international trade associations, political parties and trade unions.

In order to retrieve what we were looking for, we noticed that we were able to create an unauthenticated requests on the search page without any daily limit. So, we created a function that, given a keyword and a time range, searches 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 articles, the following information:

- title: title of the article;
- article datetime: datetime of the article publication;
- inTitle: is the keyword in the title of the article.

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.

2.5 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 to compare the search volume of different searches 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 username and the real name) and a time period we obtain a list of object with the following fields:

- date: date related of the given value;
- max trend: maximum popularity index in that day among all search terms;

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.

3. Data Storage

We decided to store all raw data in a NoSQL database, more specifically MongoDB: a cross-platform document-oriented database program that uses JSON-like documents with optional schemas. MongoDB is used to store and retrieve data in a flexible, JSON-like format called BSON (Binary JSON), which is designed to handle large amounts of data, and is particularly well-suited for storing and processing data in real-time.

We have created a single MongoDB database with several collections, one for each different data source, from which data can be retrieved via simple queries.

We have used the python library pymongo that is a useful tool for working with MongoDB in Python, and it is widely adopted in the development of web applications and other software that requires the ability to store and manipulate large amounts of data.

Example of query to retrieve all influencers:`db.influencers.find()`. Example of query to retrieve all Ansa's articles about the influencer @fedez: `db.news.find({"username":"fedez"})`.

4. Data Integration

The data integration phase aims to integrate all the data from different sources, saved in different mongo collections, in two different datasets stored in csv format.

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

- day;
- username: Instagram username of the influencer;
- trend: popularity index in that day;
- ansa: number of articles published from the 01/01/2022;
- ansa_delta: number of articles published in that 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 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.

5. Data Visualization

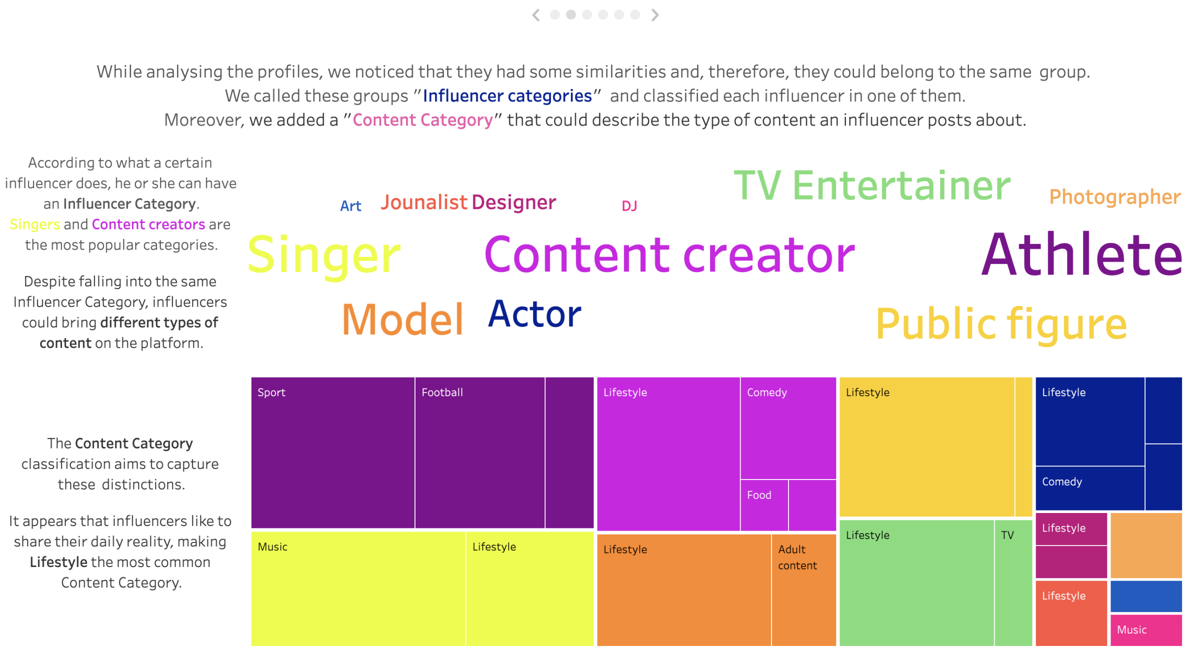
Many questions drove the implementation of our project. In particular, we wanted to discover, for what concernes Italy, who are the most influential people on Instagram, what they do (in terms of occupation and content posted on the platform), why they are so popular and if their online presence is affected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

The story is composed by five different dashboards, each one containing a different interesting view of the data we collected. To be more accurate, the dashboards contained in the story are six, but the first one serves only as an brief introduction to the whole project. For this reason, it will not be described more specifically in the next paragraphs.

In order to provide the user a consistent view throughout the artifact, all the pages share some common characteristics. For example, we created a color palette and we applied it to the elements in the different dashboards. Given that the project develops on Instagram data, we thought that an appealing choice of colors could be represented by tints that recall the logo of the app. In this way, we tried to make the visualization more engaging for the public. Also, texts that represent the same type of information (titles or descriptions) have the same font and font size so that the content could appear more coherent and harmonious.

5.1 First Dashboard: influencer and content categories

The first dashboard shows an overview of the groups in which we categorized the influencers. Since this kind of data is qualitative, we determined that the appropriate data visualizations could be a word cloud and a tree map. During the process, we also considered to represent this data via pie charts, but the both categories groups were too numerous to produce vizzes of quality.



While analysing the profiles, we noticed that they had some similarities and, therefore, they could belong to the same group.
We called these groups "Influencer categories" and classified each influencer in one of them.
Moreover, we added a "Content Category" that could describe the type of content an influencer posts about.

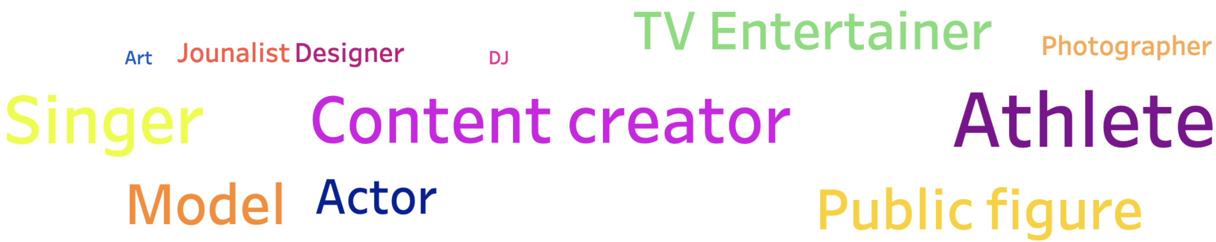
According to what a certain influencer does, he or she can have an **Influencer Category**. Singers and Content creators are the most popular categories.

Despite falling into the same Influencer Category, influencers could bring **different types of content** on the platform.

The **Content Category** classification aims to capture these distinctions.

It appears that influencers like to share their daily reality, making **Lifestyle** the most common Content Category.

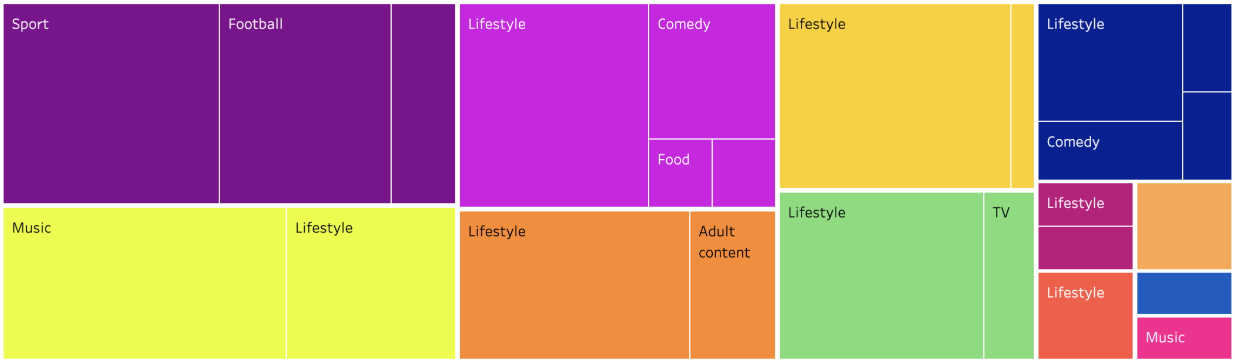
Categories word cloud This viz depict the influencer categories. A word cloud is a visual representation that, given some words, represents them in different sizes adjusted proportionally to their absolute frequency in the dataset. In this case, each category is also associated with a color, which is not very relevant in this word cloud chart but became fundamental for the tree map.



Categories tree map A treemap is a visualization method that allows to display hierarchical data in the form of nested rectangles: it can be used to represent the proportions of a whole, as well as the relationships between parts.

By creating a treemap, we wanted to emphasize the relation between the main category, that is the influencer category, and its sub-categories, which are the content categories. By representing data in this way, the distribution of influencers across different categories can be visualized in a quick and easy way. As mentioned above, this visualization is related to the previous one by color. In fact, influencer categories are mapped with the same colors in both charts, so that the user can understand which are the sub-categories for each group without the need of a legend.

The two data visualizations are also related by size and area: the larger words in the cloud map correspond to the biggest blocks in the tree map. This aspect could facilitate the perceiving of the number of elements per category and sub-category.

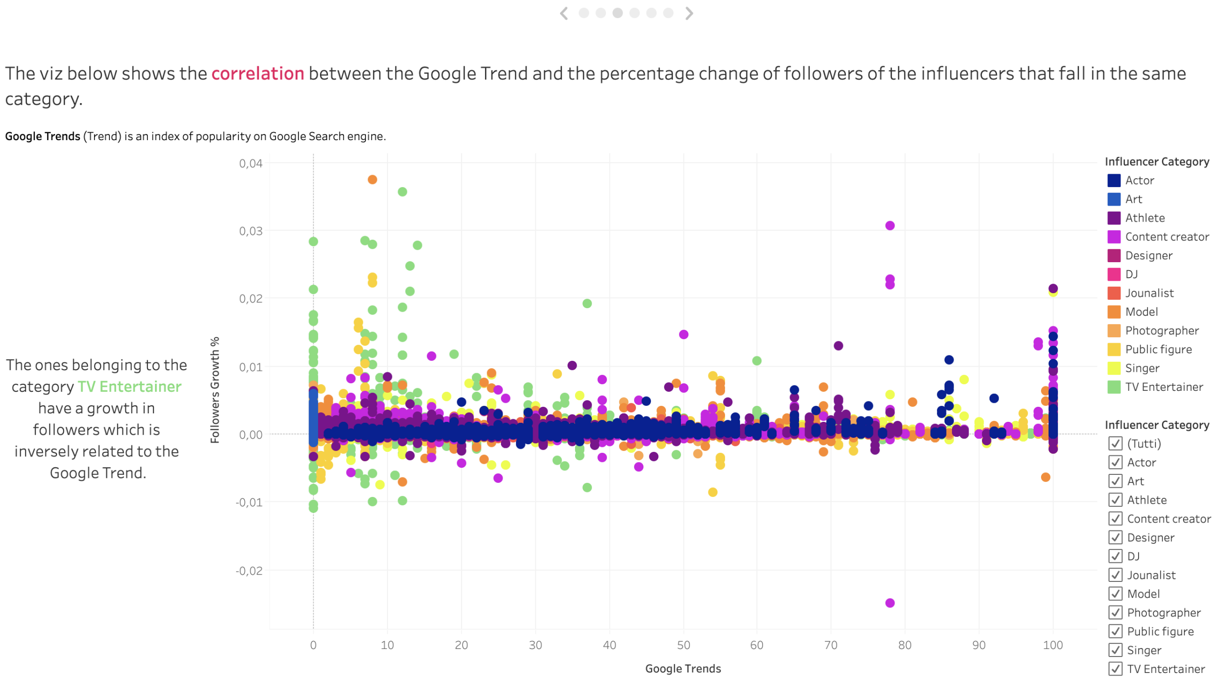


5.2 Second dashboard: correlation by categories

The second dashboard is designed to display the correlation between the followers growth in percentage and the Google Trends, for all the influencer divided by influencer category.

The data viz that can better represent such information is a scatterplot, which plots the points in a two-dimensional space. In this case, we have the Followers Growth % on the Y-axis and the Trend on the X-axis. Categories are mapped via colors, which are the same of the previous dashboard to give the story more consistency.

In this case, plotting the data is useful to understand which categories are the most influenced by the trend.



The viz below shows the correlation between the Google Trend and the percentage change of followers of the influencers that fall in the same category.

Google Trends (Trend) is an index of popularity on Google Search engine.

The ones belonging to the category TV Entertainer have a growth in followers which is inversely related to the Google Trend.

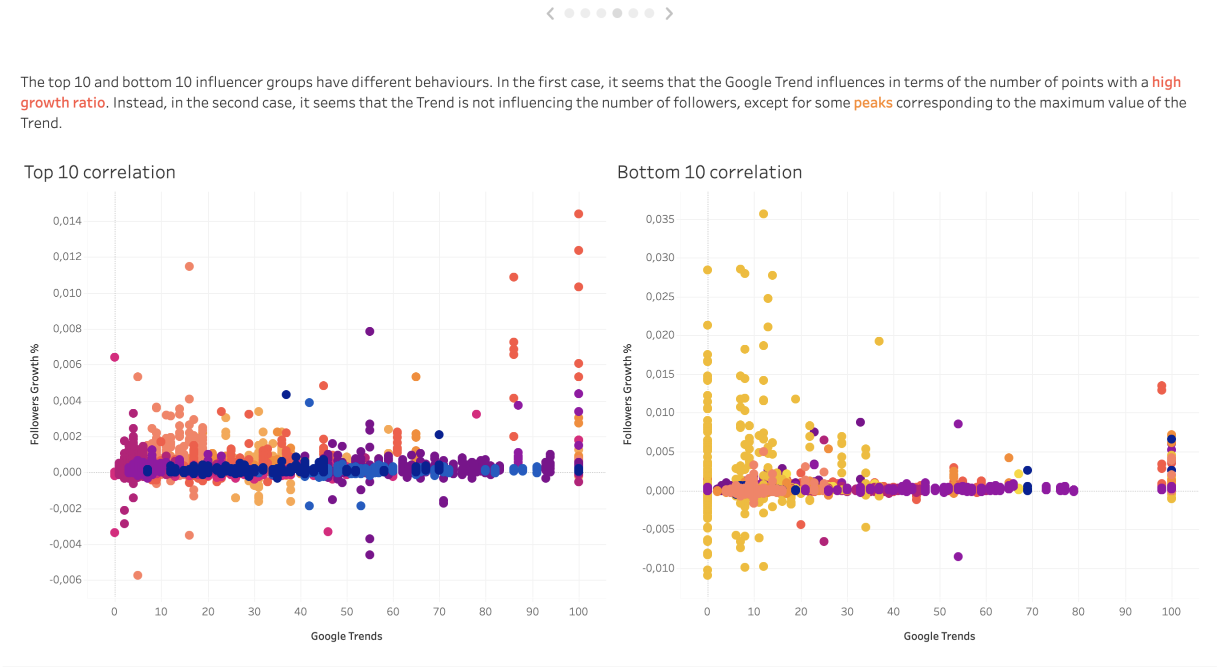
Influencer Category
☑ (Tutti)
☑ Actor
☑ Art
☑ Athlete
☑ Content creator
☑ Designer
☑ DJ
☑ Jounalist
☑ Model
☑ Photographer
☑ Public figure
☑ Singer
☑ TV Entertainer

5.3 Third dashboard: correlation by followers number

In this dashboard are displayed data similar to the previous, but this time more in depth. In fact, the dimension compared are the same but they are split in two groups of interest: top 10 and the bottom 10 influencers of the entire influencers list. This distinction aims to discover if exists a substantial difference contrast between the biggest influencers on the platform and the least popular ones.

The charts suggests that:

- the first group is not very influenced by trends, especially in the intermediate values, but it seems that the Trend influences the followers growth in terms of the number of points with a high growth ratio;
- regarding the second group, the Trend is not influencing the number of followers, except for some peaks corresponding to the maximum value of the Trend.



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.

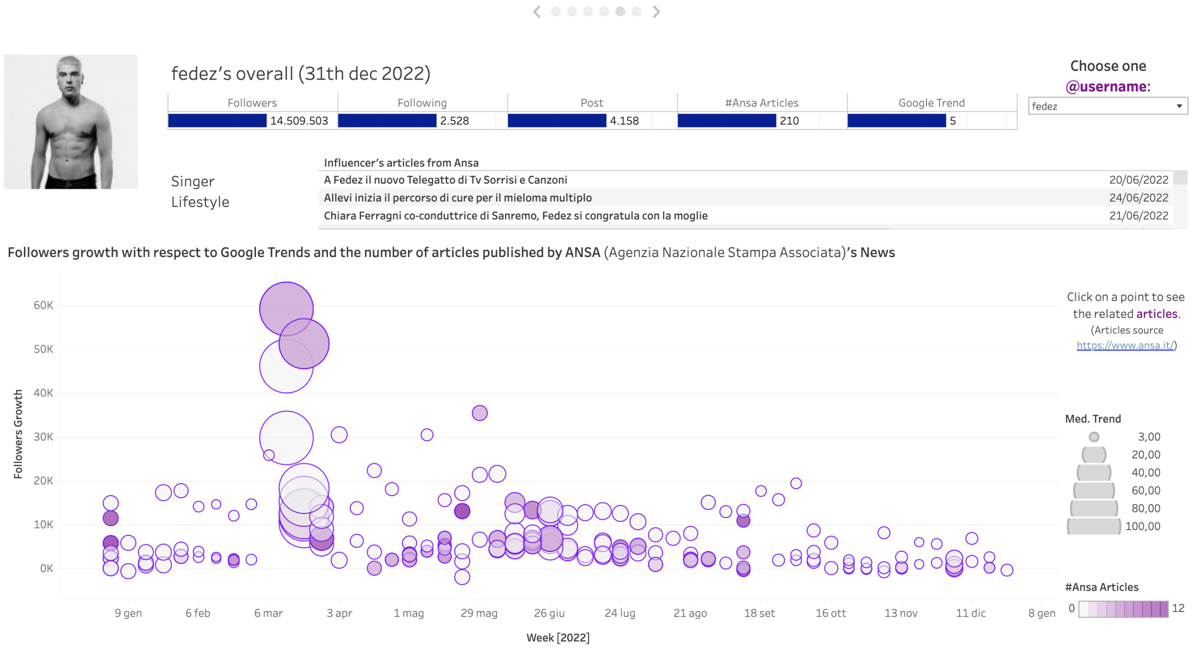
5.4 Fourth dashboard: single influencer overview

This dashboard shows a specific view containing the information related to a single influencer, which can be chosen with the appropriate selector.

We created an interactive dashboard to let the user play with the viz, by selecting a bubble in the chart. This action changes the articles list, which will be updated with the articles that were published in that week and mentioned the influencer selected. By interacting with this data visualization, the user can understand the behavior of the Instagram follower growth according to the Ansa's articles and the average Google Trends rate.

This dashboard is composed of:

- Influencer's overall section: it contains the profile's image, the category and all the statics about the Instagram profiles updated at 31/12/2022;
- Username selector: user can selected a username to update dynamically the dashboard;
- article section: it shows all the articles published during the day selected in the chart below;
- bubble chart: we chose this kind of data visualization because we needed to represent 3 measures (followers growth, numbers of Ansa articles, Google Trends value). In this way we were able to map all the measure with different graphical attributes: the point's position on the Y-axis represent the followers growth, the color depicts the number of Ansa articles published and the size that stands for the Google trends value.



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

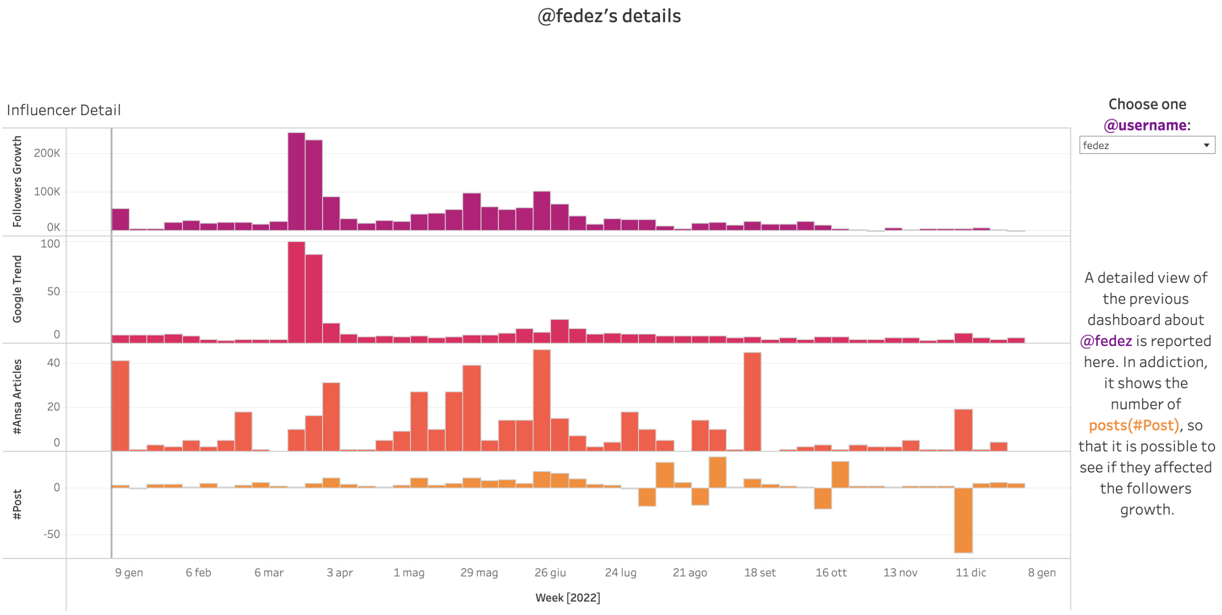
Followers growth with respect to Google Trends and the number of articles published by ANSA (Agenzia Nazionale Stampa Associata)'s News

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

5.5 Fifth dashboard: datailed overall

The last viz is closely related the one just described because it is a more detailed view of the preceding dashboard.

Our idea was to provide to the user a week-by-week comparison of followers growth, number of Ansa articles, Google trends and new post: we have added the posts to the big picture to try to understand if the creation of new contents influenced the followers growth. We created four bar charts vertically aligned so the user can easily interact with the visualization and highlighting the same week in every visualization. To let the user easily understand which chart he is analyzing, we chose four different colors according to the Instagram's color palette defined before. In this dashboard we decided to keep the username selector to provide the user a better user experience.



@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.

6. Evaluation

The following evaluation methods were applied on the dashboards:

- Heuristic evaluation
- Psychometric questionnaire
- User test

6.1 Heuristic Evaluation

During this evaluation test we have asked at 6 expert to interact with all the dashboards and to «think aloud» and detect usability problems (i.e., opportunities for improvement) interpreting the user's behavior and comments in light of some good design heuristics.

During this phase the following problems were identified: 1. dashboards are too dense and sometimes take too time to load; 2. the first dashboard is too dense by the presence of the legend, which is redundant with the word cloud; 3. in the third and fourth dashboard some axis title are not very explanatory; 4. explain the meaning of Ansa and Trend and insert references to data sources; 5. the palette used in the first two dashboards does not allow to distinguish some categories from others.

The following solutions were implemented: 1. we split the story into more dashboards of a minor dimension to improve clarity and loading times; 2. we removed the redundant legend; 3. we updated the axis title; 4. we have inserted additional labels explaining the meaning of Ansa and Trend with references to data sources; 5. we updated the palette.

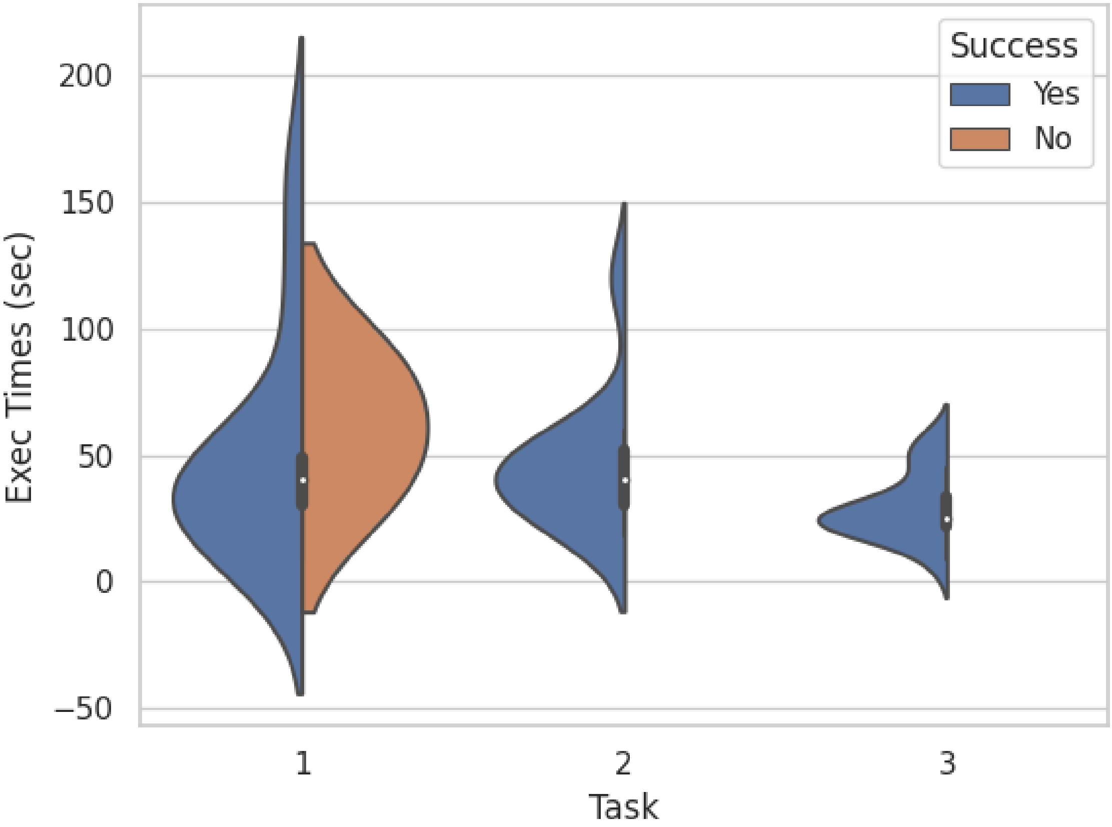
6.2 User Test

During this evaluation test, we asked 15 users to perform some tasks on all the dashboards.

We defined 3 tasks because we would avoid that the user get bored and we want the best performance possible.

Users were asked to perform the following tasks:

1. In the second dashboard between the categories "Content Creator" and "Singer" which one was more influenced by the Google Trends index? (Content Creator)
2. In the fourth dashboard, find if the peak of followers for "lorinsigneofficial" corresponds to a peak of articles published by ANSA. (Yes)
3. In the fifth dashboard, find if were published more articles on Ansa or posts on Instagram for "antoninochef" in the week of the 6th November 2022 (ANSA).



As shown by the violin plots, the major difficulties emerged during the first task, both in terms of time spent and errors: two users did not solve the task correctly, in contrast to the others where no user made a mistake. Feedback from users indicated that correlation is a complex concept and that the graph was too confusing due to the large number of points.

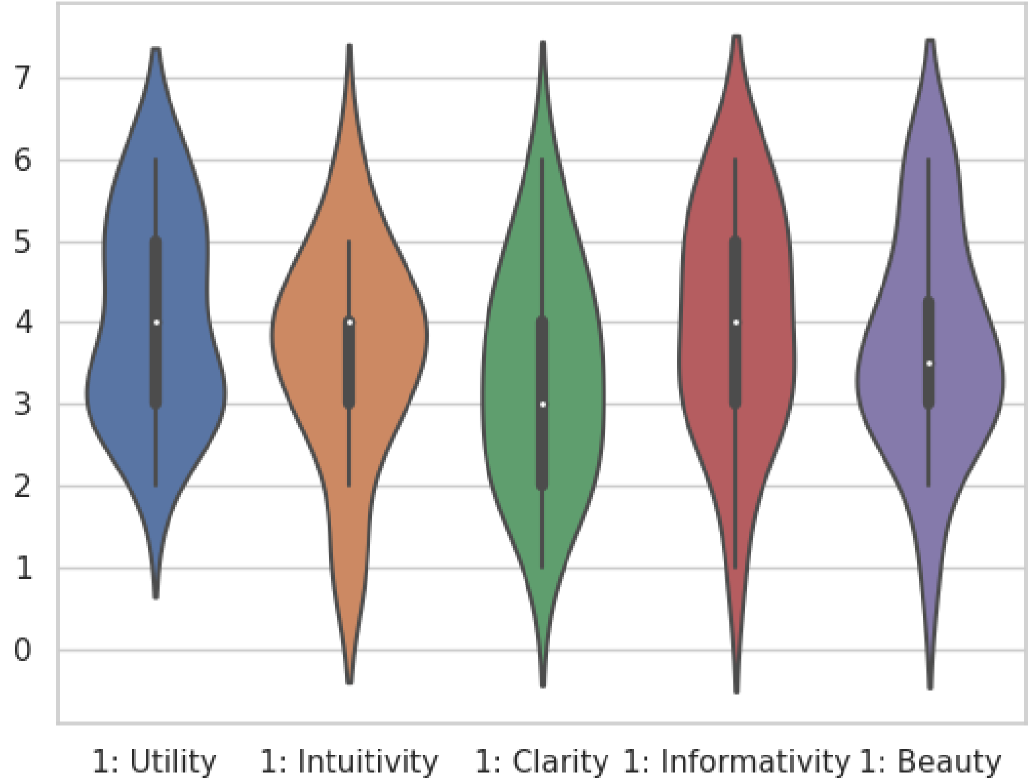
6.3 Psychometric Questionnaire

During this evaluation test, we asked 20 users to fill out a psychometric questionnaire to assess certain quality dimensions for the most important dashboards.

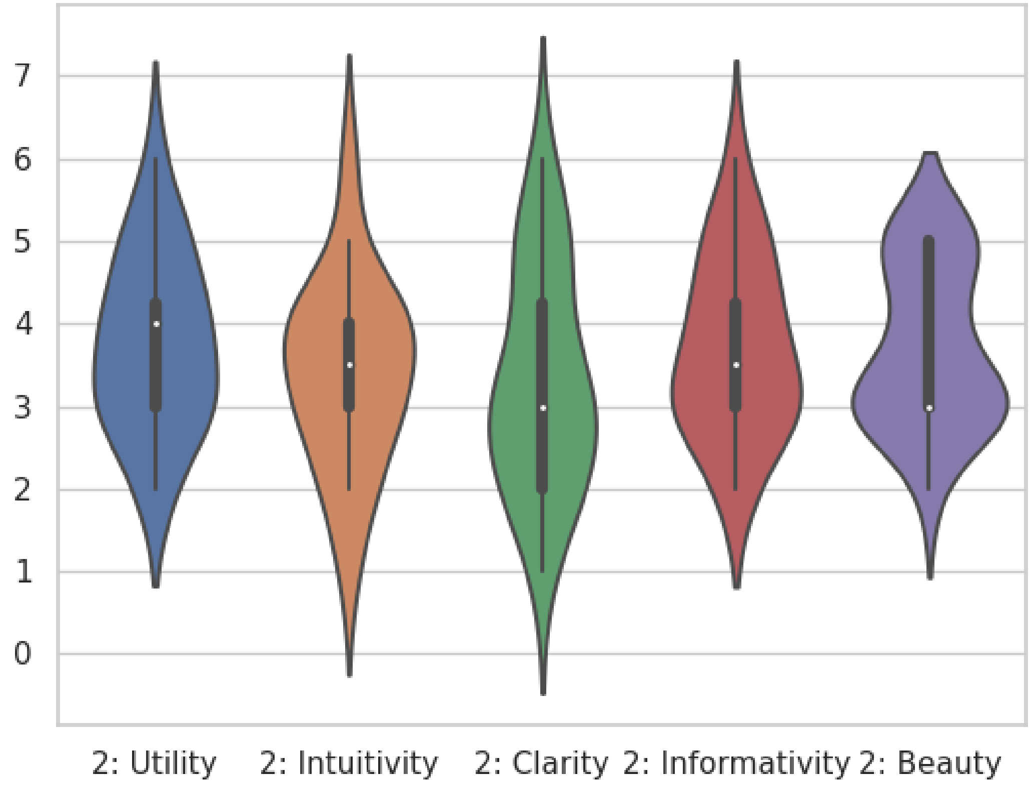
More specifically, we used the Cabitza-Locoro scale that requires to rate from 1 to 6 the utility, intuitivity, clarity, informativity and beauty of the given graphical object. The answers were collected via the 'Google Forms' tool and the results were represented through the use of violin plots.

The form deployed for this survey can be found at this link.

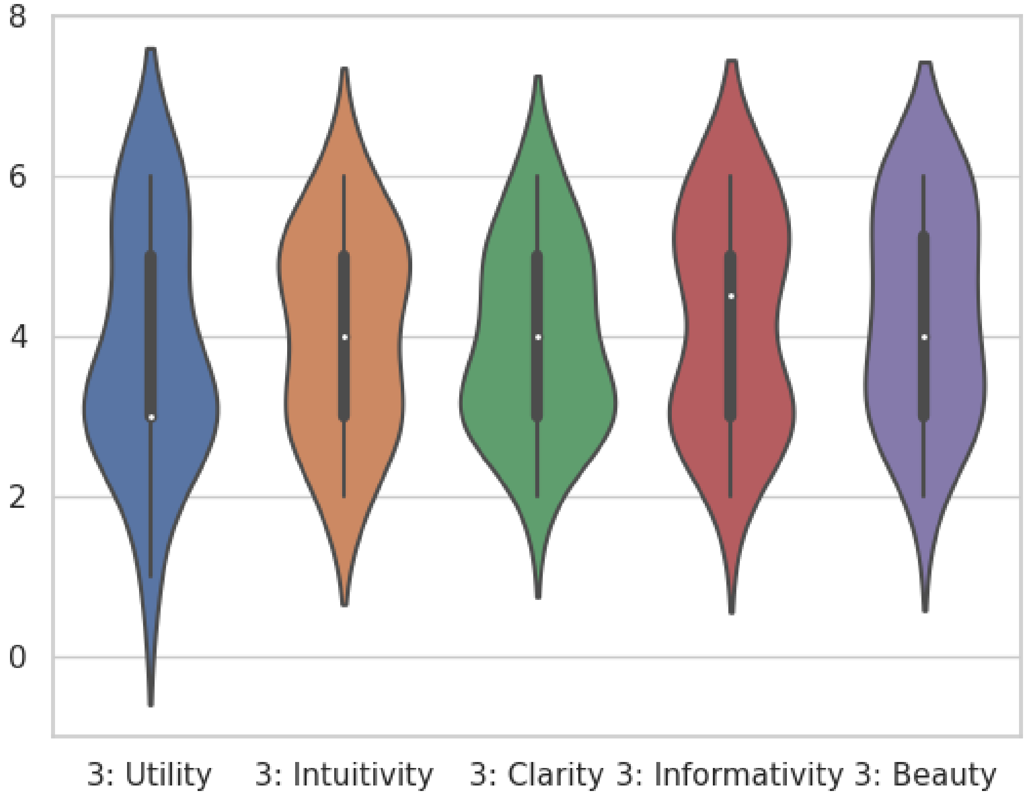
First evaluation: correlation by categories By observing the results, we can discover that the median of all the categories fall around the central values (3 and 4): this means that the the majority of people were neutral about the questions posed. Also, low values for Intuitivity, Clarity and Informativity were selected: this could be a sign of a difficulty in interpreting the chart and the data. Almost all the categories seem to have a stretched probability density function (PDF), indicating that values are not very concentrated around the median.



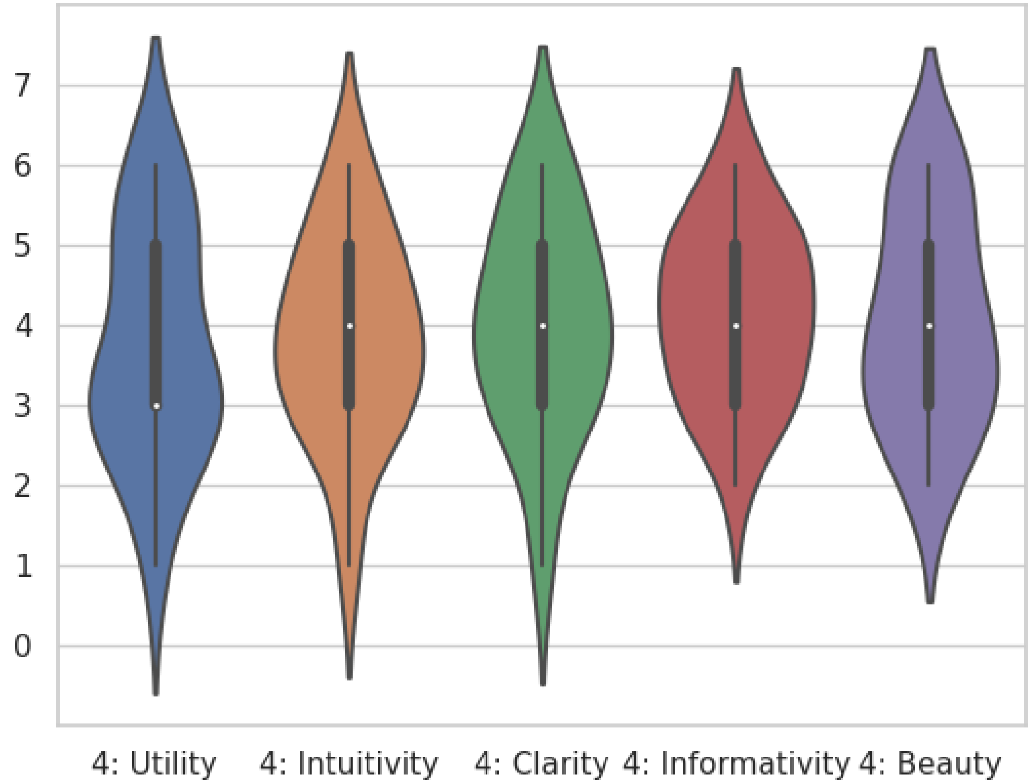
Second evaluation: correlation by followers number At first glance, it is possible to notice that the values for the PDF are not as spread as the ones from the previous chart: this could signify that the answers given by those surveyed were more similar to each other. Furthermore, as for the previous survey, the medians are between 3 and 4, but Clarity, Informativity and Beauty also received higher ratings.



Third evaluation: single influencer overview The data visualization related to this question received a general better rating with respect to the previous, with Informativity reaching a median of 5 out of 6. Also, values below 3 were not very common: this factor could indicate that people found this viz more pleasing.



Fourth evaluation: datailed overall Here too the values answered were generally higher with respect to the first two evaluations. Almost all the categories reached a median value of 4, with probability density function values condensed around the median.



7. Conclusion

Data visualization is a challenging task that involves multiple skills to produce an high-quality product.

Opinions and feedbacks from users and other people supported us in understanding which were the errors in our project and helped to find new solutions and improvements.

Through the tool we designed, 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.