POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

A financial approach for correlation with exogenous data and synergy detection in social networks

Supervisors

Prof. Luca VASSIO

Prof. Martino TREVISAN

Candidate

Fabio BERTONE

October 2022

Summary

This thesis studies the universe of Online Social Networks (OSNs) and their influencers, i.e., popular users, by applying instruments that typically belongs to the financial fields, technical analysis in particular.

Two aspects of OSNs has been investigated. The first is the correlation between social network dynamics (e.g. fanbase evolution) and other exogenous dynamics (e.g. search engines queries). The second is the synergy between couples of influencers, meant as the highly correlated movement of dynamically normalized social network metrics of two influencers during a variable length time interval.

First, this work provide a basic understanding of the fundamental financial concepts on top of which our reasoning is built. The first is the so called "Efficient Market Hypothesis" (EMH). A market is said to be efficient if at time $t + \epsilon$ it fully reflects the information available up to time t. The second one is the study of the Bollinger bands, an instrument belonging to the technical analysis of the financial markets. The relative position of the signal with respect to these bands allows to dynamically normalize signals that are of different scale and volatility, giving a mean for meaningful comparisons.

Given these two ingredients, in this thesis it has been developed the analogy between the OSNs world and the stock market. In this view, the fanbase cardinality of an influencer can be seen as the price of a stock, and the followers, seen as buyers, can purchase such stock by the act of following. The objective is to test if, and to which extent, followers dynamic is efficient in the sense of EMH. In order to assess that, the choosen source of exogenous information is the Google Trends Search Volume Index (SVI), measuring the amount of queries submitted to the Google search engine (normalized on a scale from 0 to 100) having as a keyword the influencer. The endogenous variable, instead, is the fanbase cardinality of the given influencer. Fanbase cardinality and SVI are considered in the same time period.

Once the data about these two measures over time has been collected, they are correlated through our efficiency measure. The latter is based on the average distance between the two dynamically normalized signals: the lower the distance, the higher the efficiency. In a perfectly efficient case, the SVI and the percentage increase of followers overlaps, while in the opposite case they are totally unrelated.

This study shows that influencer with a relatively small (for being in the Italy top 100 ranking, e.g. 1 million followers) are more efficient than the ones with a very big fanbase (e.g. 10-20 million followers). Taking the parameter number of followers constant, the singer category is more efficient than the VIP and athlete.

The second aspect covered by this thesis is sinergy between influencers. To be clear, in this work when we use the word 'sinergy' the meaning is that two influencers are in sinergy in a certain time interval, when their dynamically normalized fanbase cardinalities (number of followers) change following trends that are highly (Pearson) correlated in such time interval.

Selecting the top 10 scoring couples of influencers, we found that 9 out of 10 have an actual reason to be correlated. For example, infuencers can be married, engaged or even YouTube partners that often do videos/streaming events together.

Differently from efficiency, sinergy applies equally both to small and big fanbase cardinalities.

This work can be considered as a pioneeristic one, meant to show that concepts from finance and instrument from technical analysis can successfully be employed in the study of social network macro dynamics, uncovering hidden behaviours and synergies.

Acknowledgements

To my number one supporters,
my parents

To a dear friend
that reminded me how I like programming
when I needed it the most

To everybody that has been with me
during this hard and amazing adventure

Table of Contents

List of Tables

List of Figures

Acronyms

OSN
Online Social Network

EMH
Efficient Market Hypothesis

SVI
Search Volume Index

SMA
Simple Moving Average

BB
Bollinger Bands

Chapter 1

Introduction

1.1 The Rise of Social Networks

Interactions among individuals shape how our societies unfold and the graph of such interactions can reveal a lot about our social organisation and its evolution in time. That is why social networks have attracted a great deal of attention to understand the mechanisms underlying their evolution and provide valuable information on the microscopic determinants of social dynamics, for instance, individuals' search strategies or the schemes to allocate time in socially charged activities.

The evolution of social networks is shaped by the interplay of complex mechanisms operating at different scales. Indeed, individuals have a heterogeneous propensity to engage in social interactions, featuring heavy-tailed distributions of activity and degree. Also, people allocate their social interactions toward similar alters, for instance connecting to a friend of a friend. At the same time, individuals may seek novel connections outside of their inner circle of contacts, based on shared interests or experiences. Moreover, social networks are intrinsically dynamical systems that evolve in time as links between nodes are continuously created and destroyed [1].

In the last two decades, Online Social Networks (OSNs) have become increasingly popular and are nowadays part of everyday life and a fundamental means of communication.

1.2 The Efficient Market Hypothesis

The primary role of the capital market is allocation of ownership of the economy's capital stock. In general terms, the ideal is a market in which firms production-investment decisions, and investors can choose among the securities that represent ownership of firms' activities under the assumption that security at any time *fully*

reflect all available information. A market in which prices always fully reflect available information is called *efficient* [2].

We start our analysis from the efficient-market hypothesis (EMH) from the financial field. The EMH is a cornerstone yet debated hypothesis about financial economics proposed in 1970 by Eugene Fama. Essentially, it states that: (i) Current prices of stocks incorporate all available information and expectations, and (ii) current prices of stocks are the best approximation of their intrinsic value.

Some investors do believe that the market is efficient, others do not [3]. In an inefficient market, there is a period of time, following a news or financial statement, during which an asset could be mispriced, i.e., its current price does not coincide with its intrinsic value [4]. Thus, trying to predict its intrinsic value, e.g., by means of fundamental analysis techniques, could drive investors to bet in such a way to anticipate the equilibrium. Conversely, in an efficient market, prices change (almost) instantaneously according to market news and similar relevant external factors. In this work, our stocks are influencers, which we put in relationship with the Google Trends search volume index (SVI).

1.3 Technical Analysis of the Financial Markets

Technical Analysis consists in studying stock price graphs. Usually, technical analysts employ a bunch of technical indicators, that are a subset of the several available, that are usually in one of the following forms: trends indicators, momentum indicators and volatility indicators. These indicators mainly consists in oscillators and moving averages.

One cornerstone assumption of the technical analysis is that it is based entirely on prices. This means, technical analyst do believe in the efficient market hypothesis. Thus, no balance sheet nor finalcial ratios - belonging to the counterpart (fundamental analysis) - are considered. Instead of the just mentioned instruments, technicians (also called chartists) exclusively rely on the use of historical data.

The three main assumption of technical analysis are the following. The first, that is a direct implication of the EMH, is that "market discounts everything" - with everything meaning external information and fundamentals. The second is that prices moves in trends (most technical strategies are based on this assumptions). The last is that history repeats itself.

"Technical analysis could be applied in New York in 1850, in Tokyo in 1950, and in Moscow in 2150. This is true because price action in financial markets is a reflection of human nature, and human nature remains more or less constant. Technical principles can also be applied to any freely traded entity in any time frame. A trend-reversal signal on a 5-minute bar chart is based on the same indicators as one on a monthly chart; only the significance is different. Shorter

time frames reflect shorter trends and are, therefore, less significant" [5]. This, as we also do believe, states that technical analysis is a general (or generic, versatile) tool.

1.4 The Efficiency Measure

As explained previously, a market is said to be efficient if at time $t + \epsilon$ it fully reflects the information available up to time t.

Given that, we need now a way to detect - or better, quantify - the existence of such efficiency property. The two main ingredients are: a way to represent the exogenous world ("source signal") and a "response signal" to be monitored in correlation with the source signal.

In our work, the source signal will be the Google Trends SVI (Search Volume Index) - presented exhaustively in chapter 2 - and the response signal will be the fanbase (number of followers) percentage increase. Based on the average absolute difference between the source and the response signal, after they have been dynamically normalized, we will compute the efficiency score.

1.5 The Concept of Synergy

The study of how social network influencers are connected is not new neither exhaustively investigated so far. Indeed, most of the studies among social network entity connections has been done with the focus on the followers behaviours, mainly in the marketing field, rather than directly on influencers.

In our work, a time series oriented approach has been developed. The correlation between time series has been extensively studied [6] with different approaches that we will present in the dedicated chapter.

After a review of the related work in chapter 2, in chapter 4 we are going to explain our novel method to spot *synergy* among influencers. The word synergy has different meanings in different specific context. To be clear, from now on in the text when we use this word we mean the following. Synergy exists between two influencers, and in a certain time interval, when their fanbase cardinality (number of followers) changes following trends that are highly (Pearson) correlated in such time interval.

Our method can be consider a novelty since it involves a time series approach, rather than clustering or graph ones, and exploit concept like Bollinger Bands [7] that usually belongs to the analysis of financial markets.

1.6 Objective of this Thesis

This thesis has the ambitious objective of demonstrating that financial instruments taken from technical analysis can be exploited in a wider range of time series analysis belonging to different domains, in our case to online social networks dynamics.

Chapter 2

Related Work

In this chapter we are going to present a collection of examples showing the power of Google Trends SVI (Search Volume Index) as an instrument for generic forecasting use cases as well as a technical indicator for financial-related purposes.

We will conclude the chapter with an overview of some studies on connections in social networks.

2.1 Google Trends and Forecasting

The ability of Google Trends data to forecast the number of new daily cases and deaths of COVID-19 has been examined in an array of papers including [8]. The analysis includes the computations of lag correlations between confirmed cases and Google data, Granger causality tests, and an out-of-sample forecasting exercise. This evidence shows that Google-augmented models outperform the competing models for most of the countries.

Also, a method to improve the one-step-ahead forecasts of the Spanish unemployment monthly series has been presented [9]. To do so, we use numerous potential explanatory variables extracted from searches in Google (GoogleTrends tool). Two different dimension reduction techniques are implemented (PCA and Forward Stepwise Selection) to decide how to combine the explanatory variables or which ones to use. The results of a recursive forecasting exercise reveal a statistically significant increase in predictive accuracy of 10–25%.

Another test conducted about the usefulness of Google Trends data has been predicting monthly tourist arrivals and overnight stays in Prague during the period between January 2010 and December 2016 [10]. First, they analyzed whether Google Trends provides significant forecasting improvements over models without search data. Second, we assess whether a high-frequency variable (weekly Google Trends) is more useful for accurate forecasting than a low-frequency variable

(monthly tourist arrivals) using Mixed-data sampling (MIDAS). Our results stress the potential of Google Trends to other more accurate prediction in the context of tourism: we find that Google Trends information, both two months and one week ahead of arrivals, is useful for predicting the actual number of tourist arrivals.

Another use case that has been explored is forecasting residential real estate price changes from online search activity. According to [11], the intention of buying a home is revealed by many potential home buyers when they turn to the Internet to searh for their future residence. Their findings are economically meaningful and suggest that abnormal search intensity for real estate in a particular city can help predict the city's future abnormal housing price change. Below we report some plots that compare the actual price and the Google Trends SVI taken from their paper.

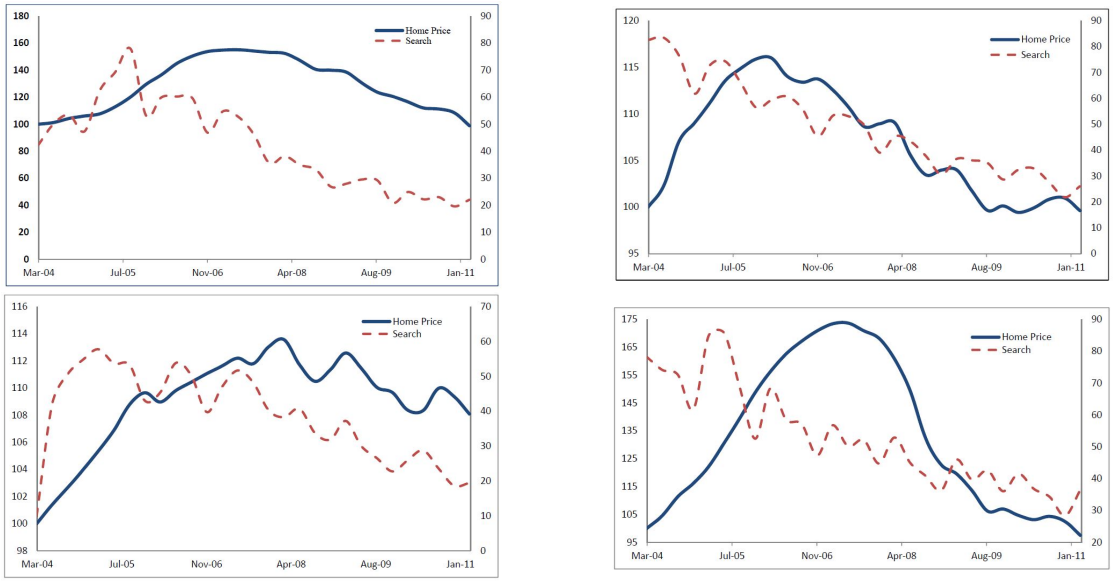


Figure 2.1: House Pricing and Google Trends SVI for houses in Boise, Boston, Des Moines and Miami [11]

With the increasing popularity of tourism activities, the forecasting of tourist volume has become an important research issue in the field of tourism management. However, the traditional statistical data cannot reflect the changes in tourism demand in real time. In order to make up for this shortcoming, scholars have found that web search data and big data technologies can provide a new way to forecast tourism demand which can expose user behavioral intentions in real time. Accordingly, it has been tried [12] to make a prediction of the number of China inbound foreign tourists based on Google Trends data, and by applying Random Forest, the model has higher accuracy than without Google Trends.

Another study [13] utilizes relevant Google Trends of specific search terms related to COVID-19 pandemic along with European Centre for Disease prevention and Control (ECDC) data on COVID-19 spread, to forecast the future trends of daily new cases, cumulative cases and deaths for India, USA and UK. From the plot below you can easily detect the high correlation between them.

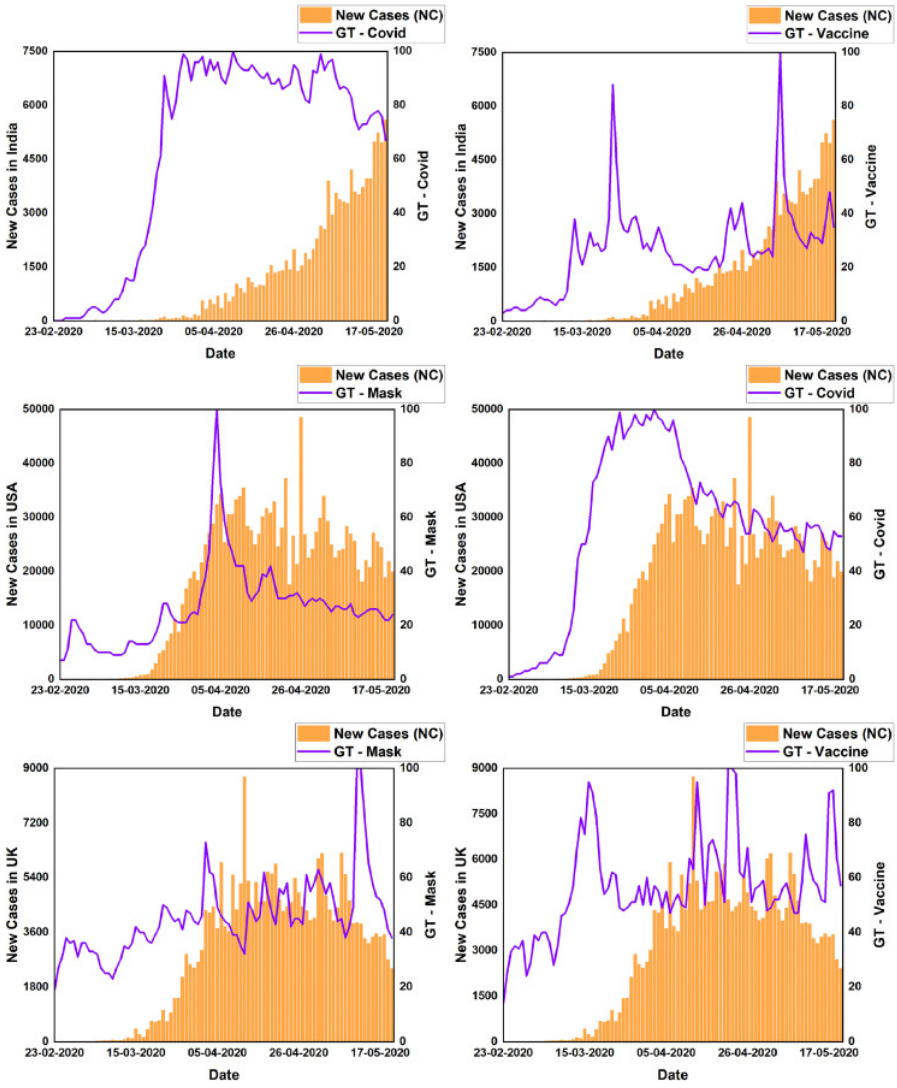


Figure 2.2: Covid-19 cases and related words searched in Google Trends [13]

E-commerce is becoming more and more the main instrument for selling goods to the mass market. This led to a growing interest in algorithms and techniques able to predict products future prices, since they allow us to define smart systems able to improve the quality of life by suggesting more affordable goods and services.

The joint use of time series, reputation and sentiment analysis clearly represents one important approach to this research issue. The primary aim is to predict the future price trend of products generating a customized forecast through the exploitation of autoregressive integrated moving average (ARIMA) model. It has been experimented [14] the effectiveness of the proposed approach on one of the biggest E-commerce infrastructure in the world: Amazon. They used specific APIs and dedicated crawlers to extract and collect information about products and their related prices over time and, moreover, we extracted information from social media and Google Trends that they used as exogenous features for the ARIMA model. We fine-estimated ARIMA's parameters and tried the different combinations of the exogenous features and noticed through experimental analysis that the presence of Google Trends information significantly improved the predictions.

Google Trends has been also employed as a useful tool for fashion consumer analytics [15]. Their work show the importance of being able to forecast fashion consumer trends and then presents a univariate forecast evaluation of fashion consumer Google Trends to motivate more academic research in this subject area. Using Burberry—a British luxury fashion house—as an example, they compare several parametric and nonparametric forecasting techniques to determine the best univariate forecasting model for "Burberry" Google Trends.

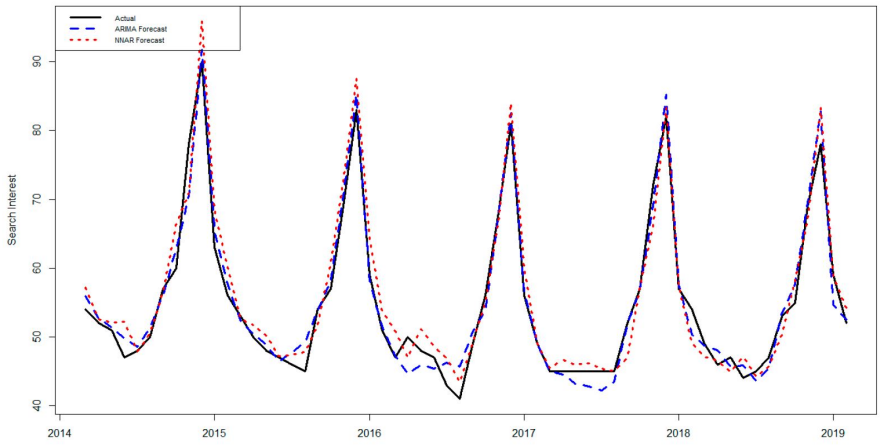


Figure 2.3: Out-of-sample forecasts for "Burberry" fashion consumer Google Trends at h = 1 month-ahead [15]

Google Trends data are an efficient source for analysing internet search behaviour and providing valuable insights into community dynamics and health-related problems. It has published an article [16] that aimed to evaluate if Google Trends data could help monitor the COVID-19 vaccination trend over time. The satisfactory result are reported in the plot below.

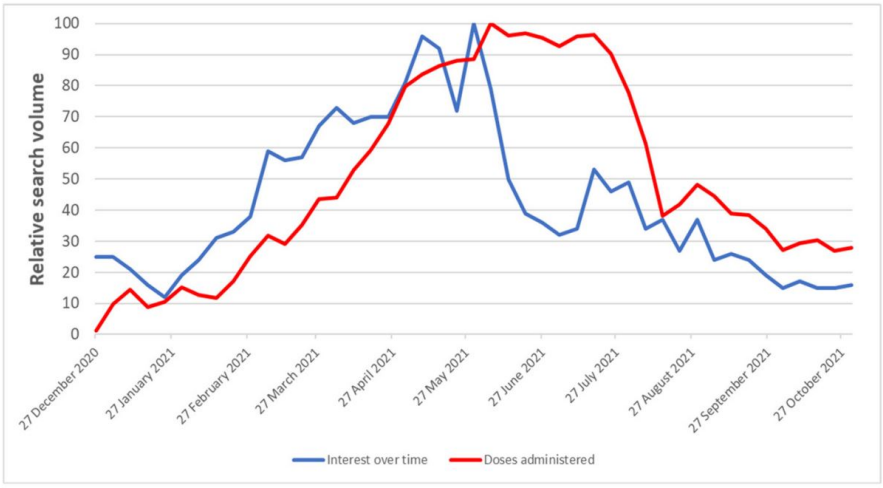


Figure 2.4: Relationship between Google Trends data and the number of COVID-19 vaccinations [16]

2.2 Google Trends SVI as Technical Indicator

Financial decisions are among the most significant life-changing decisions that individuals make. There is a strong correlation between financial decision making and human behavior. The relationship between what people think and how stock market moves has been investigated [17]. The data range from 2010 to 2015 of some of business, political and financial events which directly impact the local stock market in Pakistan is analyzed. The data was collected from search engine Google via Google trends. The association between internet searches regarding the political or business events and how the subsequent stock market moves is established. It was found that increase in search of these topics may lead to stock market fall or rise.

From the end of the 2000s, scholars and practitioners have started to include web interest metrics, such as Google Trends, to complement technical analyses on stocks. For example, Choi et al. [18] studied weekly search volume data for various search terms from 2004 to 2010. They found a link between search volume data and financial market fluctuations, observing that weekly transaction volumes of S&P 500 companies are correlated with the search volume of the corresponding company names.

Increasing transaction volumes of stocks coincides with an increasing search volume and vice versa. Preis et al. [19], instead, make a similar hypothesis, using Google Trends as an external measure. They claim that, even if Google Trends might not be useful to predict the future, it certainly helps in predicting and describing the present. For example, the volume of queries on a particular brand

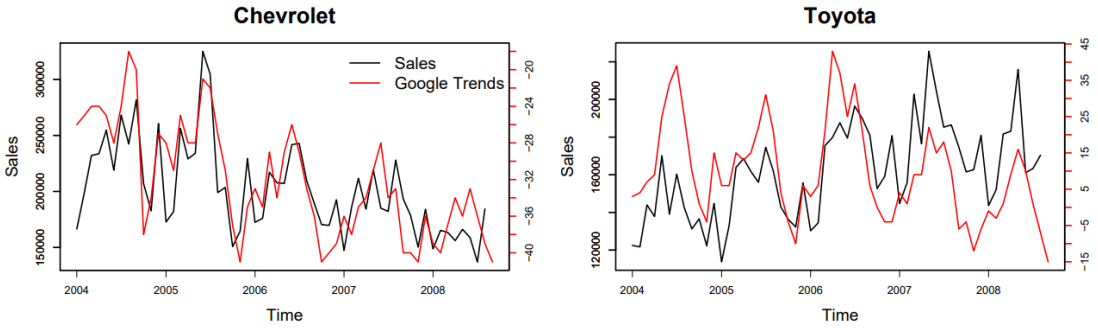


Figure 2.5: Sales and Google Trends SVI for Chevrolet and Toyota [19]

of cars during the second week in June helps to estimate (and predict) the June sales report for the brand, a number that might be available much later. Such kind of prediction is also referred to as *nowcasting*. In the context of social networks, the fusion of data from different social networks has already been studied [20]. However, less work has been done in studying the contribution of data coming from external sources, with a focus on Twitter only [21, 22].

2.3 Studies on Connections in Social Networks

The study of how social network influencers are connected is not new neither exhaustively investigated so far. Indeed, most of the studies among social network entities connections has been done with the focus on the followers behaviours, mainly in the marketing field, rather than directly on influencers.

Viable techniques for studying the connection between influencers could be exploit graph algorithms[23], or using clustering techniques in order to spot similarities. In our work, we use a more time series oriented approach. The correlation between time series has been extensively studied [6] with different approaches that we will present in the next section. Other ways of investigating relationships related to time series analysis are clustering [24] and classification [25].

One of the interesting works that model influencers connections by means of graphs is the one from Kim et al. [23]. They used the number of comments between influencers as weights of the edges, and the fact that follower A follows B as an edge itself. The study, that has been conducted in the fields of brand marketing research, founds that influencers tend to have a large number of followers who are potential customers of brands, make reciprocal relationships with other influencers, and share common followers with other influencers. They also reveal that influencers who are connected each other tend to share common followers. Thus, having a partially overlapped fanbase, connected influencers will show similar trends in the number

of followers change over time.

Let's overview now the some time series correlation studies. Papadimitriou et al. [6] addressed the problem of capturing and tracking local correlations among evolving time series. Their approach is based on comparing the local auto-covariance matrices (via their spectral decompositions) of each series. One of the reason to use that instrument, instead of the Pearson correlation, is the claim that this last concept is less effective in capturing complex non-linear relationships. In our work, indeed, we didn't limit our correlation analysis to the naive application of the Pearson correlation, but we put before that step the Bollinger Bands's %B application for taking into account trends and filter out noise due to volatility.

Clustering time-series data has been used in diverse scientific areas to discover patterns which empower data analysts to extract valuable information from complex and massive datasets. Time series data is one of the popular data types in clustering problems and is broadly used from gene expression data in biology to stock market analysis in finance [24].

Time-series classification techniques can be essentially divided into two main branches: feature based (FB) and distance based (DB). FB perform a feature extraction procedure before the classification phase. In general, from the original signal $v(t)$ a moving window k of fixed length n is considered to obtain a time-series z_k and a set x of p features is calculated over it: to give some examples, commonly chosen features are mean, variance, maximum, minimum, entropy. DB methods avoid the feature extraction phase in favor of the definition of suitable distances [25].

2.4 Previous publications by our research group

The study for this thesis project derives from the wider work on Online Social Networks (OSNs) carried out by the research group SmartData@Polito[1]; this center focuses on Big Data technologies, Data Science (from data management, to data modeling, analytics, and engineering), and Machine Learning methodologies applied to several domains of knowledge, finding solutions for both theoretical problems and helping companies toward applications.

An initial work of Data Analytics applied to OSNs is the study by Trevisan et al. [26]: a first research regarding how people behaved and interacted with politicians and personalities on Instagram before the European Elections of May 2019. A custom crawler was used to collect the data used for the analyses: it downloaded and stored data and meta-data about the profiles of top public Italian figures (i.e. influencers), the related activities (i.e. their posts) and the interactions

[1] `https://smartdata.polito.it`

(i.e. users' likes and comments in the first 24 hours after posting time), at the turn of two months. The study focuses on checking if interactions across political figures follow general patterns, and if there are any differences with those ones across profiles of different categories of influencers, such as music, sport and show entertainment.

Another paper related to politics and OSNs is the one provided by Ferreira et al. [27], in which the researchers' goal was to study communities of co-commenters[2] to reveal characteristics and dynamics of interactions on the Instagram environment, to highlight common trends as well as particularities, the level of engagement and coordination.

The COVID-19 pandemic profoundly changed economy, culture, politics, but, above all, the society and, as a consequence, it was important and interesting to study its impact on OSNs. The study by Trevisan et al. [28] was focused on understanding the effects on social life of the total lockdown imposed during the first six months of the year 2020, offline but also online, because OSNs represented an alternative solution to physical meetings.

Vassio et al. [29, 30] conducted researches on how influencer postings attract interactions (number of 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 across time, from their peak to the conclusion of a post-life. The researchers looked into the activities of Italian influencers and their followers on Facebook and Instagram for more than five years (from 2016 to 2021).

[2]Commenters that comment on the same post.

Chapter 3

Correlation with exogenous data: Influencers' Efficiency

3.1 Methodology

In this section we are going to present two fundamental concepts for our work: Bollinger bands and our formulation of the concept of efficiency.

Bollinger Bands are a financial technical indicator whose purpose is to provide a relative definition of high and low prices, based on volatility and past history, proposed by Jhon Bollinger in the 1980s.

Three curves over time characterize the Bollinger bands:

- A Simple Moving Average (SMA) looking back at T time units;
- Two bands, upper and lower, respectively obtained by adding and subtracting C times the standard deviation (also computed looking back T time units) of the quantity of interest measured by the signal to the SMA.

By definition, prices are high at the upper band and low at the lower band. Typically, Bollinger bands are used in conjunction with other indicators to understand if the price of a certain asset (typically a stock) is overpriced/underpriced with respect to its intrinsic value. An extensive dissertation about the topic can be found in [7]. Bollinger bands have already been applied to other contexts, such as to identify the start and end of demand for pediatric intensive care in real-time [31]. The time window T, typically days in finance, is months in our study, and the signal under analysis is the absolute variation of followers or search volume in place of the variation in price.

The $\%B$ is a derived indicator that quantifies the position of the signal relative to the two bands. Formally, it is defined as:

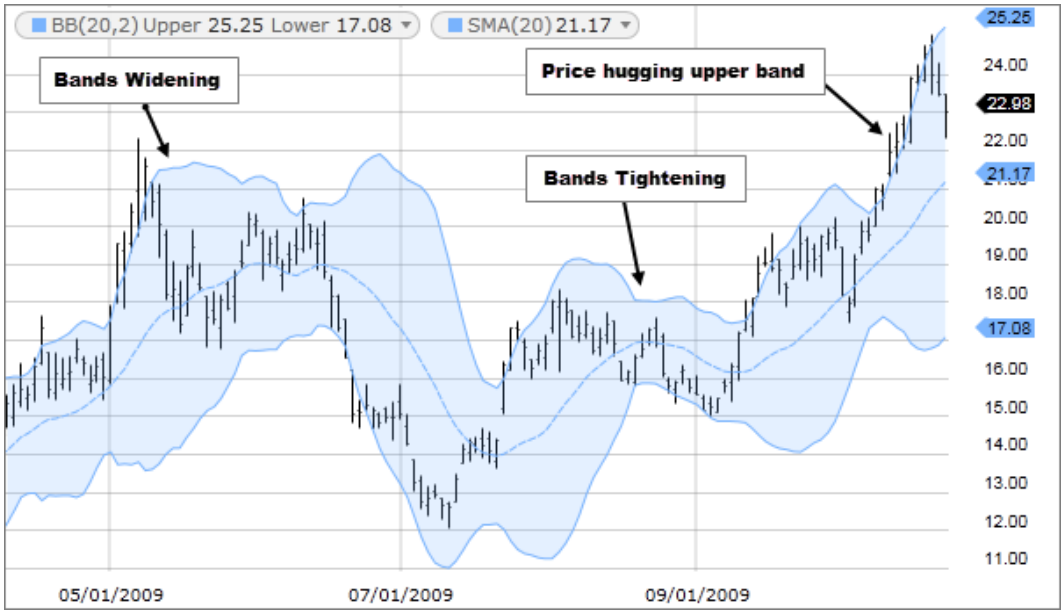


Figure 3.1: Bollinger Bands with the typical 20 days moving window and 2 standard deviations shift https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/bollinger-bands

$$\%\mathrm{B} = \frac{Signal(t) - LowerBand(t)}{UpperBand(t) - LowerBand(t)}$$

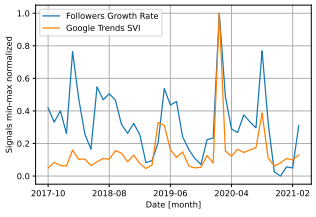


Figure 3.2: Static normalization for Google Trends SVI and Followers Growth Rate

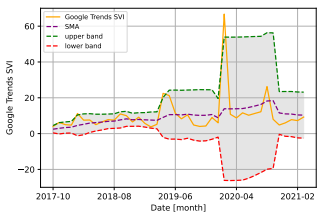


Figure 3.3: Bollinger bands for the Google Trends SVI

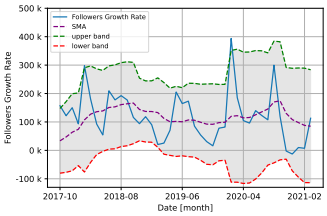


Figure 3.4: Bollinger bands for the Followers Growth Rate

Thanks to a dynamic normalization that exploits the %B, we can get rid of the differences brought by the different order of magnitude and the different volatility over time. Intuitively, a %B close to 1 (or even exceeding it) indicates that the asset (the influencer in this case) is undergoing an intense short-term increasing

trend, while, conversely, a %B close to 0 indicates a decreasing one. When %B is ≈ 0.5 no quick variations are occurring.

In the following, we will show how short-term phenomena co-occur with similar intensity both in the SVI and in the fanbase trends. In particular, we observe that, while the long-term trends may diverge, short-term ones, pinpointed using %B, are often similar. To quantify them, we define an *Efficiency* measure that quantifies how %B curves are close. We define the Efficiency as the complement of the average absolute difference between the two curves, i.e.:

$$\text{Efficiency} = 1 - \frac{\sum_t |\%B(t)_{Followers} - \%B(t)_{SVI}|}{\sum_t 1}$$

where $\%B(t)_{Followers}$ is the curve for the followers growth rate and $\%B(t)_{SVI}$ is the curve for Google Trends SVI. The denominator simply counts the number of observations. An efficiency of 1 indicates that %B signals overlap, meaning that the fanbase trends closely follow those in the search volume. This corresponds to the case where the EMH hypothesis hold.

3.2 Case study: Instagram influencers

3.2.1 History and Figures

Instagram was founded in 2010. It is a mobile application for Smartphone which is freely available in the Application Store (App Store) and Google Play. Being mainly a photo-sharing application, Instagram has excelled as an effective communication and marketing tool to display products with visual descriptions. Hence, it becomes a useful social networking platform instantly to individuals and companies. Moreover, the acquisition of Instagram by Facebook has potentially made the application more attractive and appealing to millions of users [32].

Today Instagram is a multifaceted platform that has established itself as a popular buzzword among the youths. Instagram has also spurred a revolution in branding and advertising by providing a platform for the most prominent brands to endorse their products. Instagram has moved beyond photography and its biggest upgrade so far came last year with the worldwide availability of Instagram Stories and Instagram Live – a feature that allows you to broadcast video to your followers in real-time.

According to the Digital 2021 Global Statshot Report[1], Instagram is used by 1.3 billion users worldwide. In this large and complex ecosystem, a limited portion of social profiles emerges and reaches a large base of followers. One of the main goals

[1] `https://datareportal.com/reports/digital-2021-april-global-statshot`

of these so-called *influencers* is to increase their fanbase engaging users through the content they offer. In many cases, social celebrities monetize their social presence, offering brands a practical way for marketing [33, 34]. As such, users in OSNs can be roughly divided into two non-exclusive categories: *regular users*, that consume the content of the *influencers* they follow.

Launch date	6 October 2010
HQ	Menlo Park, California
People	Adam Mosseri (Head of Instagram), Kevin Systrom (co-founder), Mike Krieger (co-founder)
Business type	Subsidiary
Owner	Facebook
Industry	Social media

Table 3.1: Instagram Overview

Over the years, Instagram has grown linearly in the number of users and exponentially in the amount of revenue.

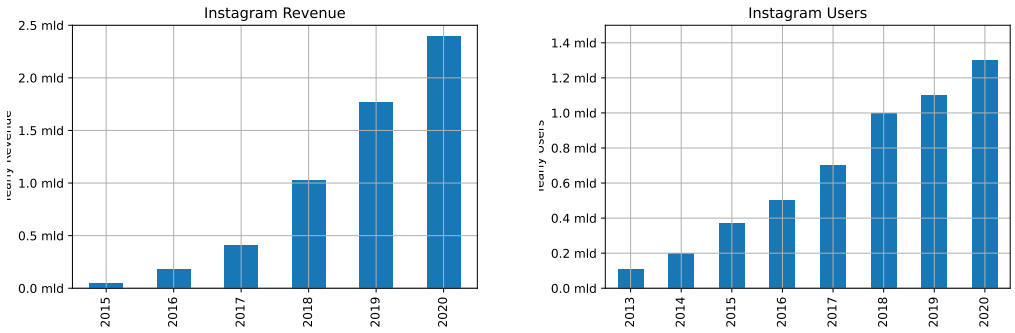


Figure 3.5: Instagram revenue and users growth

3.2.2 Instagram Dataset Exploration

The dataset under analysis is a sample of 60 influencers taken from the italian top 100 influencers ranking.

Let's examine to the followers distributon.

Another important aspect to consider of the dataset is how influencers are splitted by gender, age and category: regarding the gender we have 55% males and 45% females, while the distribution for the other two attributes is represented in the pie charts in next page.

Name	followers	likes	posts	gender	category	birth year
Chiara Ferragni	23.4 M	31000 M	14425	F	vip	1987
Gianluca Vacchi	20.0 M	3933 M	3197	M	vip	1967
Fedez	12.5 M	8526 M	3819	M	singer	1989
Michele Morrone	12.0 M	2554 M	568	M	vip	1990
Valentino Rossi	10.3 M	2248 M	1241	M	athlete	1979
Belen Rodriguez	10.0 M	5736 M	7427	F	vip	1984
Gianluigi Buffon	9.8 M	1484 M	808	M	athlete	1978
Mario Balotelli	9.0 M	1152 M	894	M	athlete	1990
Andrea Pirlo Official	8.5 M	784 M	504	M	athlete	1979
Marzia Kjellberg	8.2 M	4913 M	1666	F	vip	1992
Diletta Leotta	7.8 M	2939 M	2071	F	vip	1991
Elettra Lamborghini	6.7 M	1230 M	329	F	singer	1994
Mariano Di Vaio	6.4 M	6460 M	7797	M	vip	1989
Carlo Ancelotti	5.3 M	272 M	329	M	athlete	1959
Marco Verratti	5.3 M	416 M	878	M	athlete	1992
Alessia Marcuzzi	5.1 M	1378 M	5522	F	vip	1972
Giulia De Lellis	5.0 M	2797 M	1601	F	vip	1996
Emma Marrone	4.9 M	1826 M	4370	F	singer	1984
Michelle Hunziker	4.9 M	1079 M	1258	F	vip	1977
Claudio Marchisio	4.9 M	1311 M	1943	M	athlete	1986
Stefano De Martino	4.5 M	1359 M	2259	M	athlete	1989
Leonardo Bonucci	4.4 M	1192 M	2035	M	athlete	1987
Melissa Satta	4.4 M	1338 M	4230	F	vip	1986
Cecilia Rodriguez	4.4 M	1617 M	5960	F	vip	1990
Stephan El Shaarawy	4.1 M	639 M	788	M	athlete	1992
Francesco Totti	4.0 M	489 M	199	M	athlete	1976
Monica Bellucci	4.0 M	506 M	652	F	vip	1964
Alessandro Del Piero	3.9 M	738 M	780	M	athlete	1974
Giorgio Chiellini	3.9 M	504 M	989	M	athlete	1984
Valentina Ferragni	3.9 M	2554 M	4135	F	vip	1992
federica nargi	3.9 M	1156 M	2671	F	vip	1990
Benedetta Rossi	3.9 M	907 M	3244	F	vip	1972
Laura Pausini	3.6 M	1005 M	4477	F	singer	1974
Sferaebbasta	3.5 M	1203 M	337	M	singer	1992
Alessandra Amoroso	3.5 M	836 M	2357	F	singer	1986
Frank Matano	3.3 M	634 M	1479	M	vip	1989
Clio Zammatteo	3.2 M	1424 M	3433	F	vip	1982
Luciana Littizzetto	3.1 M	1064 M	4633	F	vip	1964
Antonino Cannavacciuolo	2.9 M	181 M	678	M	vip	1975
Elisabetta Canalis	2.9 M	759 M	4351	F	vip	1978
Ludovica Pagani	2.7 M	1674 M	553	F	vip	1995
Christian Vieri	2.6 M	690 M	3005	M	athlete	1973
Taylor Mega	2.6 M	695 M	760	F	vip	1993
Elodie	2.5 M	446 M	174	F	singer	1990
J-Ax	2.4 M	316 M	611	M	singer	1972
Vanessa Incontrada	2.4 M	331 M	1425	F	vip	1978
Ghali	2.3 M	503 M	37	M	singer	1993
Tiziano Ferro	2.1 M	396 M	670	M	singer	1980
Gue Pequeno	2.0 M	378 M	120	M	singer	1980
Lorenzo Jovanotti	1.9 M	788 M	5434	M	singer	1966
Barbara D'Urso	2.8 M	759 M	6165	F	vip	1957
Alessandro Borghese	1.9 M	305 M	2612	M	vip	1976
Giuseppe Conte	2.0 M	293 M	1307	M	athlete	1964
Rosario Fiorello	2.0 M	95 M	586	M	vip	1960
Rudy Zerbi	2.0 M	617 M	3435	M	vip	1969
Paolo Maldini	2.0 M	147 M	251	M	athlete	1968
Aurora Ramazzotti	2.2 M	596 M	1821	F	singer	1996
Salmo	2.2 M	967 M	1910	M	singer	1984
Mara Venier	2.3 M	489 M	2705	F	vip	1950
Matteo Salvini	2.3 M	3252 M	11106	M	vip	1973

Table 3.2: Instagram Dataset

The social network metrics - likes, posts and interactions - shows an exponential behavior when sorting influencers by increasing number of followers.

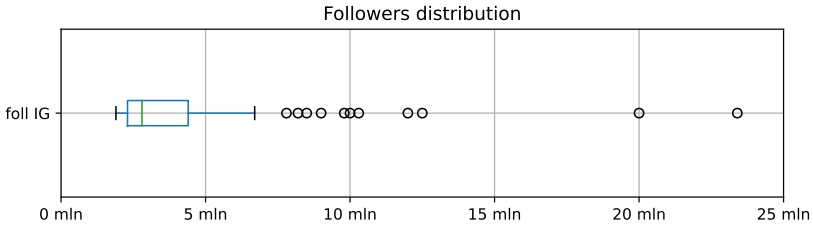


Figure 3.6: Followers Distribution (Instagram)

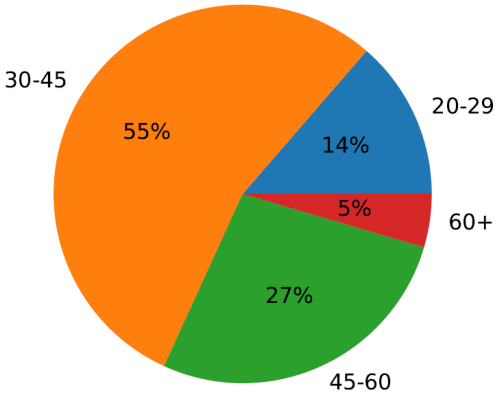


Figure 3.7: Influencers by Age

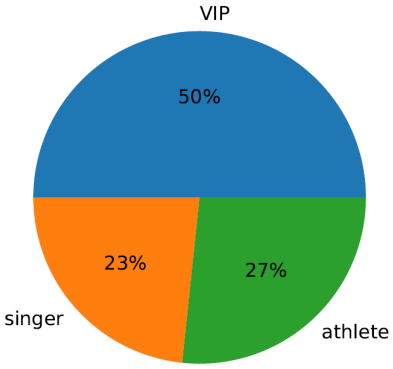


Figure 3.8: Influencers by Category

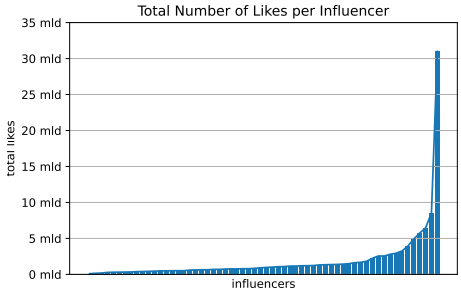


Figure 3.9: Likes distribution (Instagram)

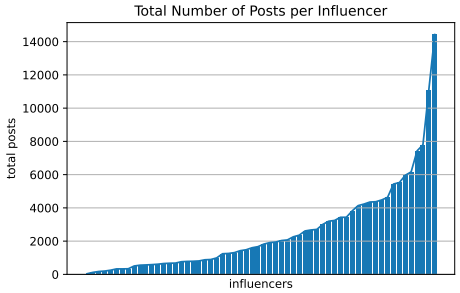


Figure 3.10: Total posts distribution (Instagram)

3.2.3 Results

We first exemplify our approach by showing the time series for the Italian singer Elettra Lamborghini. We report in Figure 3.2 the two original signals that we put on the same scale with range [0,1] using a static min-max normalization. They follow different long-term trends, with the Google Trends SVI (orange line) showing

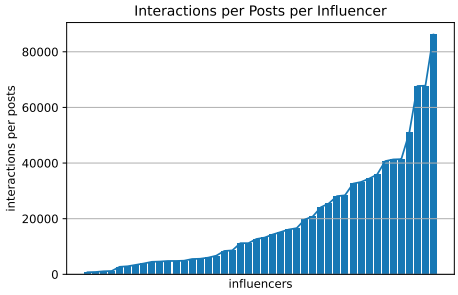


Figure 3.11: Interactions per post distribution (Instagram)

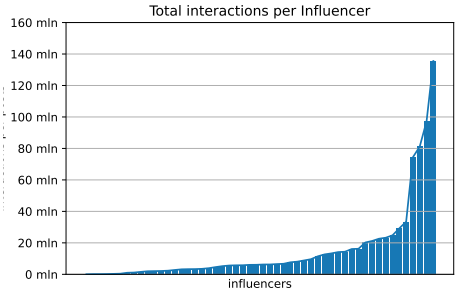
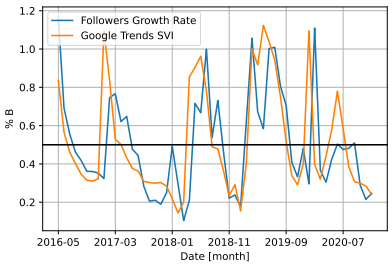


Figure 3.12: Total interactions distribution (Instagram)

an overall increasing trend, while the Follower variation (blue line) exhibits seasonal cycles and, in general, more variability. However, the curves have some simultaneous peaks that we aim to pinpoint in the following. These peaks often coincide with events that boost the popularity of the artist. For example, the peaks in February 2020 are due to the singer participation to the popular *Italian song festival of Sanremo*. In Figure 3.3 and Figure 3.4, we show the signals together with their Bollinger bands for the Google Trends SVI and Follower variation, respectively. The dashed purple line depicts the moving average (SMA) computed over 9 months (T), while the grey area delimits the Bands ($C = 2$). In this preliminary work the parameters T and C have been manually tuned to obtain the best results, and we leave automatic tuning as future work. We observe how the bands dynamically adjust their range according to the variability of the underlying signal.

The %B metric indicates the relative position of the signal with respect to the range of the bands. We show %B for Elettra Lamborghini and two other influencers. Focusing on the first picture, we observe how the %B time series mostly overlap. If we compare them with the original (but normalized) signals in Figure 3.2, the role of the Bollinger bands and %B is clear. They allow us to mine the short-term trends, which we find to co-occur more frequently than long-term shifts. Indeed, the Pearson correlation coefficient on the original signals is 0.63, while the %B are 0.68 correlated. Similar considerations hold for the singer Elodie and the actress Michelle Hunziker. Measuring the Efficiency of the signals, we obtain 0.86 for Elettra Lamborghini, 0.88 for Elodie and 0.86 for Michelle Hunziker.

Figure 3.13: Elodie (singer)



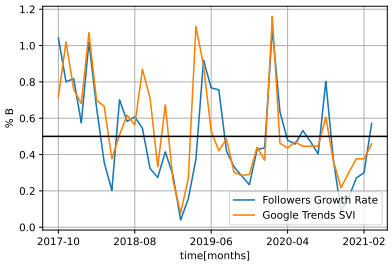
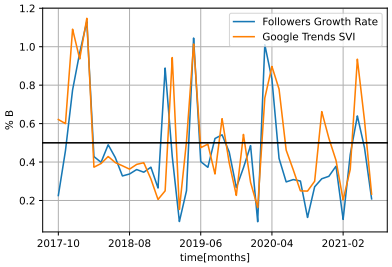


Figure 3.14: Elettra Lamborghini (VIP)

Figure 3.15: Michelle Hunziker (VIP)



3.3 Case study: Facebook influencers

3.3.1 History and Figures

Facebook is a California-based social media giant that has evolved from a simple online networking site to a social networking powerhouse. What started out just as a plain website for Harvard students has now become the internet sensation with billions of users worldwide. It has come a long way since its debut in 2004

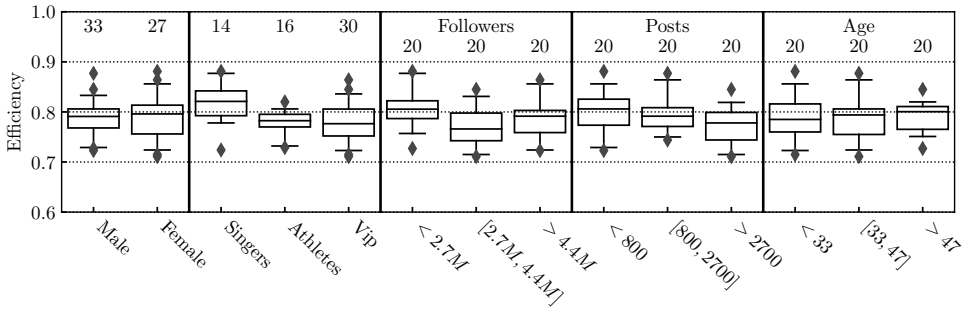


Figure 3.16: Efficiency distribution among different categories of influencers

for arriving at today, where it is the popular social network with 2.8 billion user around the globe.

Launch date	4 February 2004
HQ	Menlo Park, California
People	Mark Zuckerberg (CEO, co-founder), Sheryl Sandberg (COO), Chris Cox (CPO)
Business type	Public (NASDAQ:FB)
Industry	Social networking

Table 3.3: Instagram Overview

Facebook is a mobile-first app, like Instagram and YouTube, as shown by the pie chart below.

Over the years, Facebook has grown linearly in the number of users and exponentially in the amount of revenue.

3.3.2 Facebook and Instagram Compared

Facebook is a California-based social media giant that has evolved from a simple online networking site to a social networking powerhouse. What started out just as a plain website for Harvard students has now become the internet sensation with billions of users worldwide. It has come a long way since its debut in 2004 for arriving at today, where it is the popular social network with 2.8 billion user around the globe.

While Facebook remains the leading social network by total number of users, Instagram reaches the younger generation, and appeals to diverse societies more prevailingly than other social networking services. It has been reported that youngsters today spend more time on Instagram than Facebook. This is likely

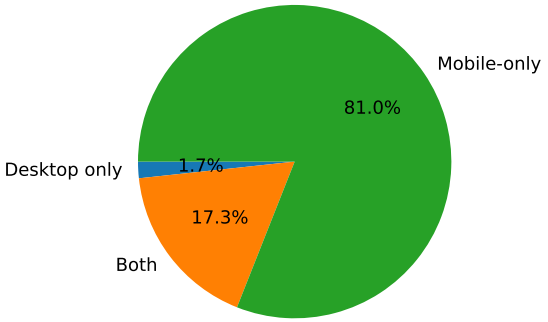


Figure 3.17: Device distribution in Facebook usage

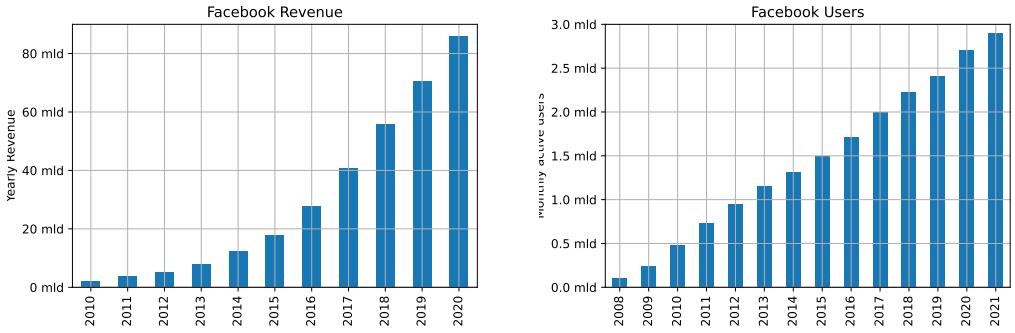


Figure 3.18: Facebook revenue and users growth

because young mobile users are extremely driven to take photos or pictures using their mobile phones, and share them with others instantly. As such, the sharing of images rather than words alone has made communication with friends and broader groups of users who share similar interests more ideal, convenient and fascinating [1].

Despite the common goal, Facebook and Instagram are two separate social media platforms with different approaches to networking. Facebook is mostly a closed-knit community of people who know each other while Instagram lets you build and join communities of people who share your common interest such as pets, photography, fashion, movies, technology, travel, etc.

Facebook has been around much longer than mostly every social networking platform out there, including Instagram. This makes Facebook clearly a dominant

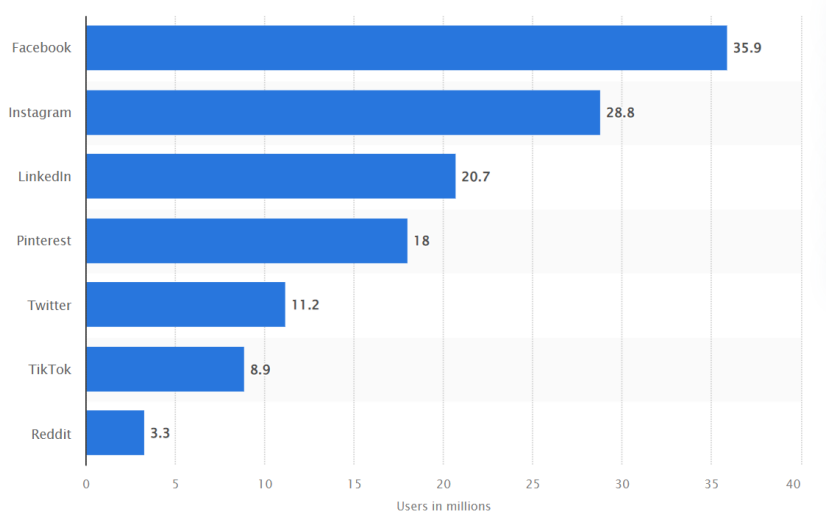


Figure 3.19: Number of users of leading social networks in Italy in March 2021 (Statista)

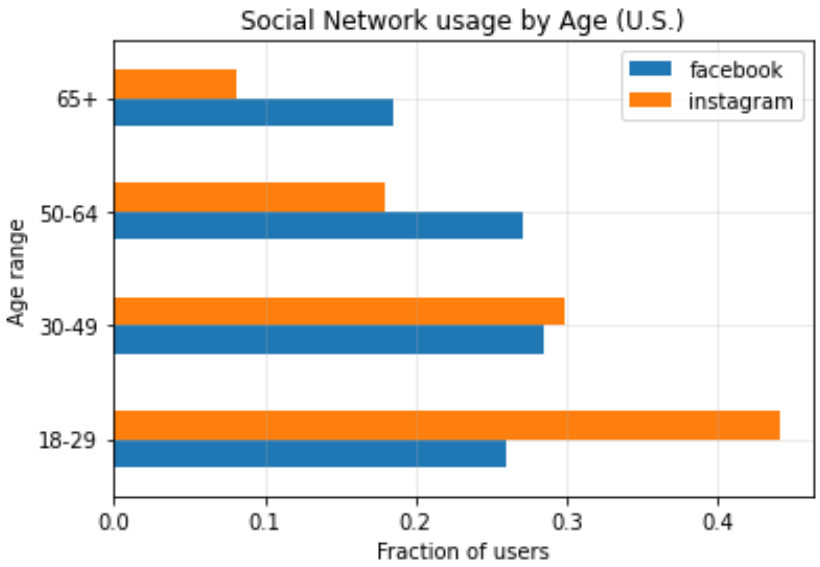


Figure 3.20: Social Media Use in 2021 by Age in US (PEW RESEARCH CENTER)

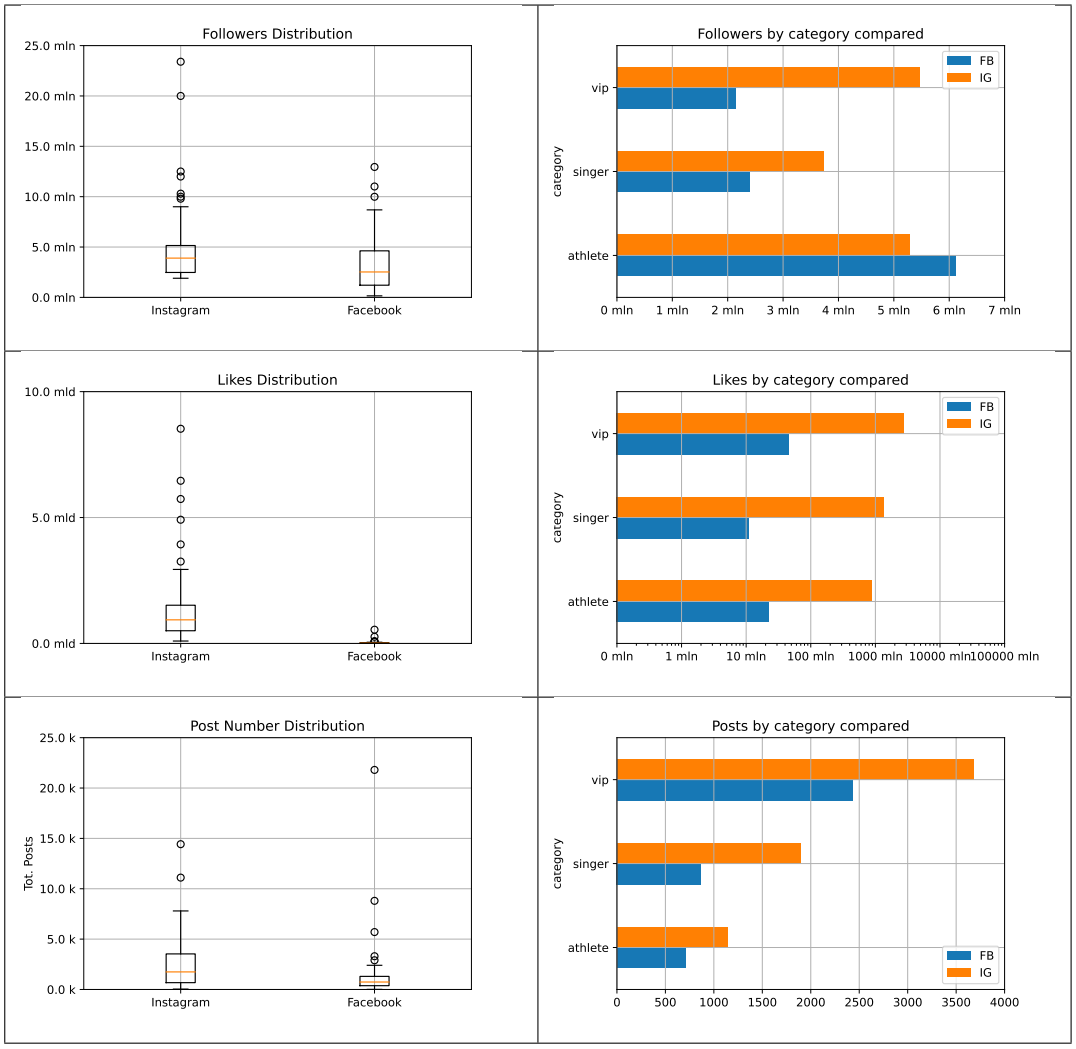
player among the social networking community with a much larger consumer reach and active users. Instagram, although rapidly growing, has relatively lesser user database in terms of consumer reach but it's quite popular among youths. Facebook is generally targeted at all age groups.

3.3.3 Differences among Facebook and Instagram metrics

By inspecting the numbers both from overall statistics and the two influencers ranking we discover that the two social platforms - Facebook and Instagram - has some significant differences. For each quantity - followers, likes and posts - we will report both the general behavior than the representation split by category.

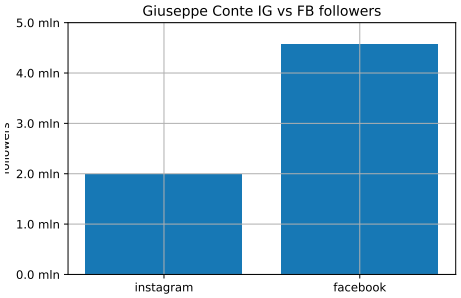
Overall, we can notice that:

- Influencers on has more followers on Instagram
- Influencers on has more likes on Instagram
- Influencers on tends to publish more contents on Instagram



Another change, from one platform to the other, is the ranking of the influencers. Sometimes it is negligible, sometimes not. We report below, first, the distribution of likes, followers and posts by category of influencers; then, also some example of significant ranking shift.

Figure 3.21: Giuseppe Conte (VIP)



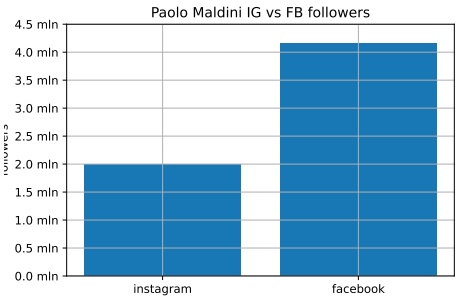


Figure 3.22: Paolo Maldini (athlete)

Figure 3.23: Matteo Salvini (VIP)

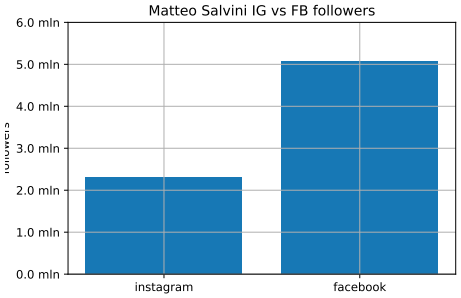
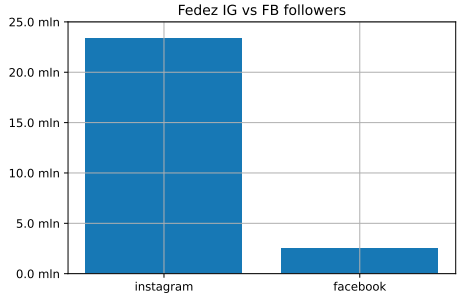


Figure 3.24: Fedez (singer)



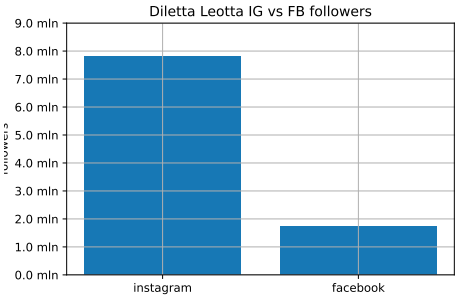
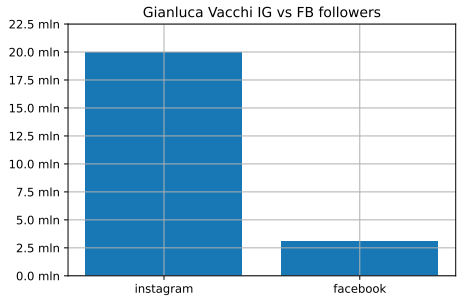


Figure 3.25: Diletta Leotta (VIP)

Figure 3.26: Gianluca Vacchi (VIP)



3.3.4 Dataset Exploration

Let's examine to the followers distributon in Facebook for the same sample analysed for Instagram.

Name	followers	likes	posts
Gianluca Vacchi	3.09 M	11.74 M	1300
Fedez	2.55 M	20.51 M	486
Michele Morrone	2.50 M	10.68 M	448
Valentino Rossi	12.95 M	74.33 M	1200
Belen Rodriguez	5.00 M	63.87 M	2400
Gianluigi Buffon	13.3 M	25.28 M	853
Mario Balotelli	9.99 M	115500	8
Andrea Pirlo Official	11.01 M	17.68 M	344
Diletta Leotta	1.74 M	18.58 M	754
Elettra Lamborghini	1.29 M	5.84 M	595
Mariano Di Vaio	2.91 M	47.47 M	2100
Carlo Ancelotti	8.60 M	9.01 M	271
Marco Verratti	7.66 M	2.94 M	134
Alessia Marcuzzi	1.92 M	2.95 M	643
Giulia De Lellis	777600	26.23 M	5700
Emma Marrone	3.50 M	24.85 M	2900
Michelle Hunziker	2.00 M	18.67 M	1300
Claudio Marchisio	3.08 M	33.77 M	1300
Stefano De Martino	1.25 M	20.98 M	932
Leonardo Bonucci	4.75 M	28.29 M	1400
Melissa Satta	707900	786600	373
Cecilia Rodriguez	465000	2.52 M	1200
Stephan El Shaarawy	8.69 M	5.09 M	139
Francesco Totti	3.52 M	12.05 M	304
Monica Bellucci	2.06 M	26.14 M	2400
Alessandro Del Piero	8.41 M	24.43 M	731
Giorgio Chiellini	7.73 M	14.78 M	969
federica nargi	411800	455800	256
Benedetta Rossi	7.89 M	263.56 M	8800
Laura Pausini	7.77 M	25.73 M	2000
Alessandra Amoroso	3.00 M	13.08 M	749
Frank Matano	3.55 M	1.88 M	105
Luciana Littizzetto	1.09 M	1.75 M	588
Antonino Cannavacciuolo	2.72 M	6.83 M	507
Ludovica Pagani	242000	2.62 M	511
Christian Vieri	153000	392400	377
Taylor Mega	352300	139300	32
Elodie	783500	1.06 M	436
J-Ax	2.21 M	6.23 M	380
Vanessa Incontrada	1.80 M	12.37 M	1100
Ghali	540400	3.09 M	513
Tiziano Ferro	2.93 M	13.55 M	769
Lorenzo Jovanotti	2.78 M	12.37 M	1300
Barbara D'Urso	1.42 M	18.07 M	3300
Alessandro Borghese	1.25 M	3.22 M	1200
Giuseppe Conte	4.58 M	89.18 M	2300
Rosario Fiorello	2.16 M	1.89 M	525
Rudy Zerbi	321000	1.08 M	914
Paolo Maldini	4.16 M	3.23 M	112
Aurora Ramazzotti	333500	102800	18
Salmo	954200	2.12 M	264
Matteo Salvini	5.06 M	545.18 M	21800

Table 3.4: Facebook Dataset

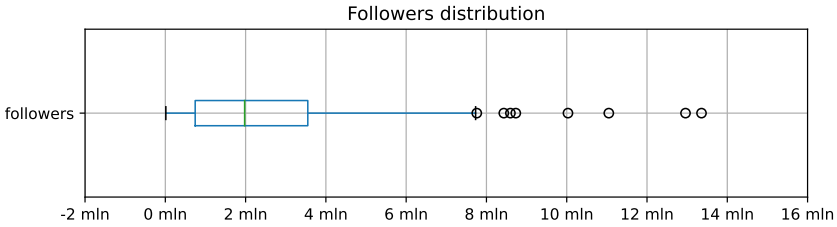


Figure 3.27: Followers Distribution (Facebook)

Also here, the social network metrics - likes, posts and interactions - shows an exponential behavior when sorting influencers by increasing number of followers.

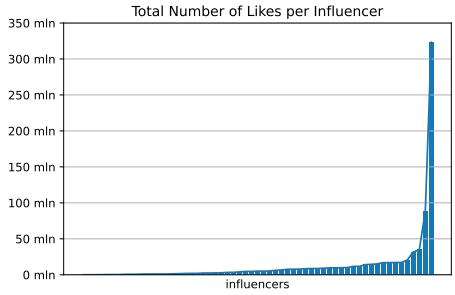


Figure 3.28: Likes distribution

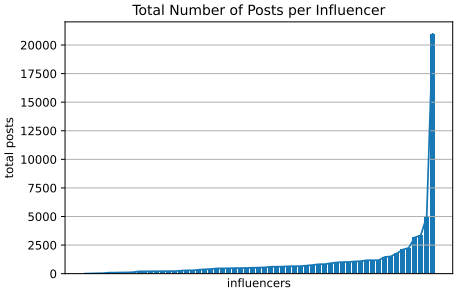


Figure 3.29: Total posts distribution

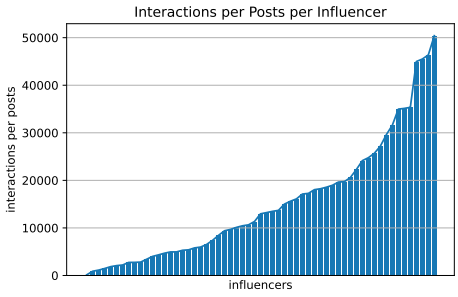


Figure 3.30: Interactions per post distribution (Facebook)

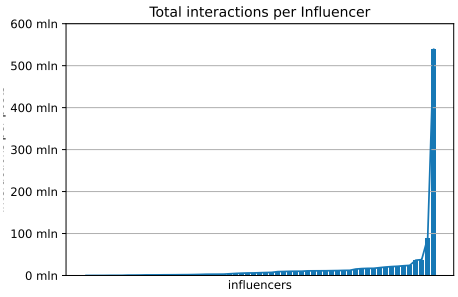


Figure 3.31: Total interactions distribution (Facebook)

3.3.5 Correlation with Popularity

In this section we show the different influencers' characteristics that are correlated to the total number of follower both for Facebook and Instagram. The first table is for Instagram, while the second is for Facebook.

	followers	M	F	singer	sport	vip	age
followers	1.000	0.007	-0.007	-0.173	0.040	0.110	-0.165
M	0.007	1.000	-1.000	0.023	0.545	-0.502	0.125
F	-0.007	-1.000	1.000	-0.023	-0.545	0.502	-0.125
singer	-0.173	0.023	-0.023	1.000	-0.332	-0.551	-0.190
sport	0.040	0.545	-0.545	-0.332	1.000	-0.603	0.059
vip	0.110	-0.502	0.502	-0.551	-0.603	1.000	0.108
age	-0.165	0.125	-0.125	-0.190	0.059	0.108	1.000

Table 3.5: Instagram correlation matrix

	followers	M	F	singer	sport	vip	age
followers	1.000	0.327	-0.327	-0.156	0.547	-0.377	0.154
M	0.327	1.000	-1.000	-0.127	0.545	-0.400	0.156
F	-0.327	-1.000	1.000	0.127	-0.545	0.400	-0.156
singer	-0.156	-0.127	0.127	1.000	-0.387	-0.502	-0.216
sport	0.547	0.545	-0.545	-0.387	1.000	-0.602	0.160
vip	-0.377	-0.400	0.400	-0.502	-0.602	1.000	0.037
age	0.154	0.156	-0.156	-0.216	0.160	0.037	1.000

Table 3.6: Facebook correlation matrix

3.3.6 Efficiency Characteristics and Examples

Despite the differences listed in the previous section, our parallel still holds also in the Facebook ecosystem. After all, dynamics of followers are similar, and the website has its influencers too.

In terms of efficiency, Facebook is slightly below Instagram, but the figures are comparable.

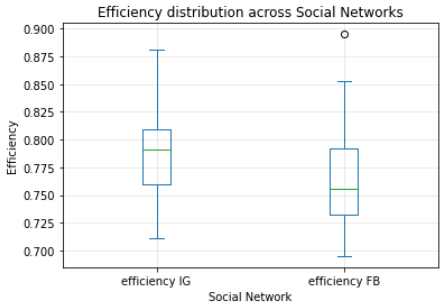
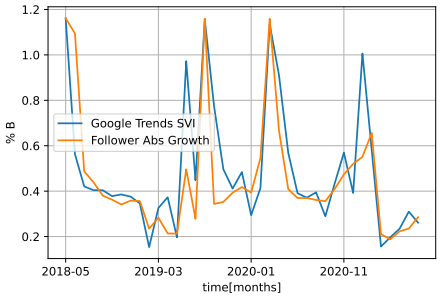


Figure 3.32: Box Plot of Efficiency distribution across Social Networks

Figure 3.33: Giuseppe Conte (VIP)



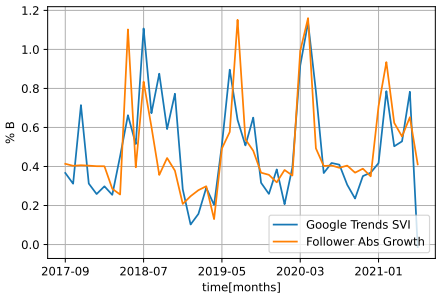


Figure 3.34: Christian Vieri (Athlete)

3.4 Case Study: YouTube Influencers

3.4.1 History and Figures

Today, YouTube is the largest user-driven video content provider in the world; it has become a major platform for disseminating multimedia information. A major contribution to its success comes from the user-touser social experience that differentiates it from traditional content broadcasters [35].

Launch date	April 2005
HQ	San Bruno, California
People	Susan Wojcicki (CEO), Scott Silver (VP, Engineering), Sundar Pichai (Alphabet CEO)
Business type	Subsidiary
Owner	Alphabet

Table 3.7: Instagram Overview

Over the years, Facebook has grown linearly in the number of users and exponentially in the amount of revenue.

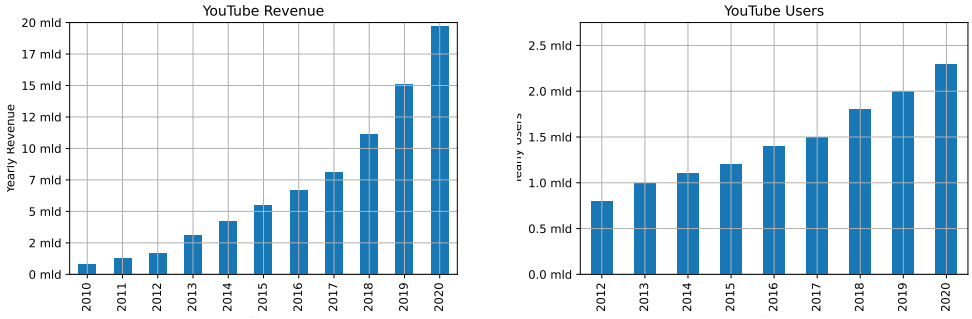


Figure 3.35: Youtube revenue and users growth

In 2021, figures about YouTube has continued to grow and became impressive:

- YouTube has more than 2 billion logged-in monthly users
- 74% of adults in the U.S. use YouTube
- YouTube is the world's second-most visited website (Right after its parent company, Google)
- It's also the world's second-most used social platform (right behind Facebook)
- People watch more than a billion hours of video on YouTube every day
- Viewers aged 18 and over spend 41.9 minutes on YouTube daily, on average

3.4.2 Dataset Exploration

In this section we will show the main characteristics of the dataset under analysis in this chapter.

The following table represents figures that come both from Instagram and Youtube. The first three columns are Instagram followers, likes and posts; the next one tells if a person is either a singer or a youtuber, while the last two columns are numbers from Youtube (where the figures are missing is because the person doesn't have a personal channel but, for example, publish video on the channel of his/her record company).

Name	followers	likes	posts	category	subscibers	videos
Elodie	2.5 M	446 M	174	singer	359 k	21
J-Ax	2.4 M	316 M	611	singer	1100 k	78
Sferaebbasta	3.5 M	1203 M	337	singer	1890 k	72
Alessandra Amoroso	3.5 M	836 M	2357	singer	878 k	144
Ghali	2.3 M	503 M	37	singer	2640 k	96
Elettra Lamborghini	6.7 M	1230 M	329	singer	1270 k	23
Gue Pequeno	2.0 M	378 M	120	singer	1040 k	182
Casa Surace	1.2 M	79 M	2700	youtuber	1010 k	294
Laura Pausini	3.6 M	1005 M	4477	singer	-	-
Salmo	2.2 M	967 M	1910	singer	1330 k	150
Emma Marrone	4.9 M	1826 M	4370	singer	624 k	60
Benedetta Rossi	4.1 M	104 M	3200	youtuber	2840 k	801
I Pantellas	2.2 M	41 M	673	youtuber	5500 k	486
link4universe	79800	2 M	4500	youtuber	421 k	1167
Tiziano Ferro	2.1 M	396 M	670	singer	1320 k	122
Anita Stories	110400	1 M	548	youtuber	1290 k	986
Me Contro Te	1.6 M	69 M	1400	youtuber	6080 k	1777
Fedez	12.5 M	8526 M	3819	singer	2040 k	147
fanpageIT	1.8 M	222 M	11200	youtuber	2660 k	29518
Maryna	712700	23 M	616	youtuber	1090 k	319
Gli Autogol	3.5 M	657 M	4500	youtuber	2040 k	572
Francesca Presentini	285000	5 M	525	youtuber	1170 k	384
Clio Makeup	3.2 M	131 M	1900	youtuber	1360 k	1026

Table 3.8: Youtube Dataset

Below are reported the followers, posts and interactions distributions, as well as the split between youtubers and singers in terms of cardinality. In addition, the last two boxplots highlight the difference in terms of number of YouTube videos and subscribers between singer and youtubers.

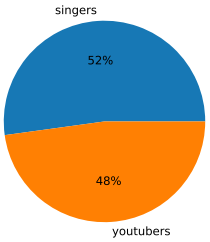


Figure 3.36: Youtubers/Singers split

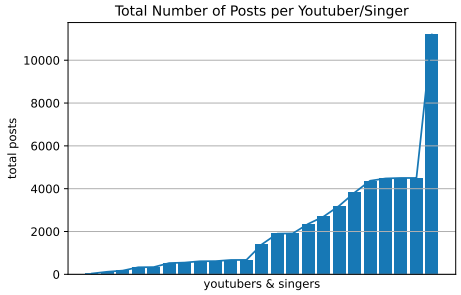


Figure 3.37: Total posts distribution

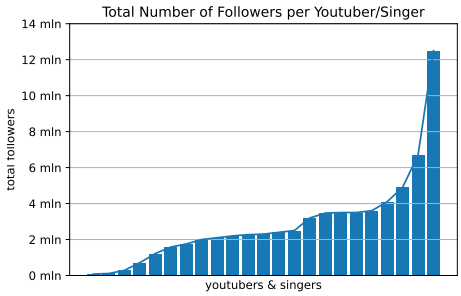


Figure 3.38: Followers distribution (YouTube)

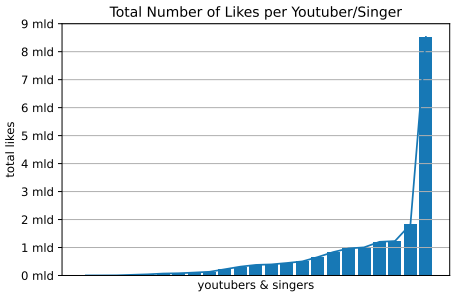


Figure 3.39: Total interactions distribution

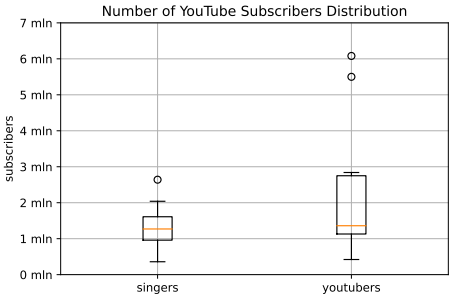


Figure 3.40: YouTube Subscribers distribution

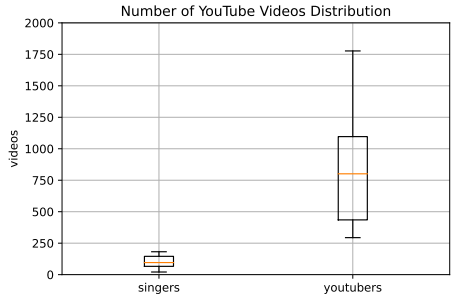


Figure 3.41: YouTube Videos distribution

3.4.3 Examples of High Efficiency

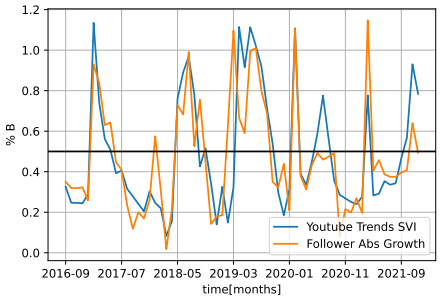


Figure 3.42: Elodie (singer)

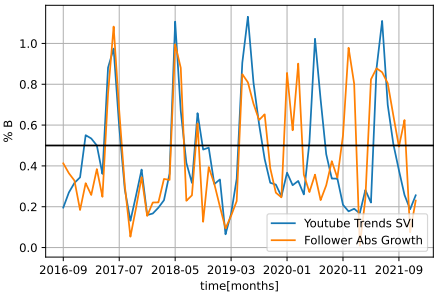
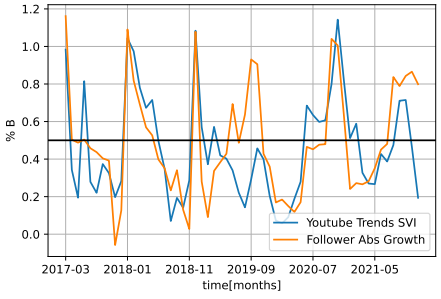


Figure 3.43: J-Ax (singer)

Figure 3.44: Sfera Ebbasta (singer)



Chapter 4

Synergy among Influencers

We developed a novel method to spot *synergies* among influencers. The word synergy has different meanings in different specific context. To be clear, from now on in the text when we will use this word we will mean the following. Synergy exists between two influencers, and in a certain time interval, when their fanbase cardinality (number of followers) changes follow trends that are highly (Pearson) correlated in such time interval.

Our method can be consider a novelty since it involves a time series approach, rather than clustering or graph ones, and exploit concept like Bollinger Bands [7] that usually belongs to the analysis of financial markets signals.

4.1 Data

As in any analytical process, it all starts by collecting, understanding and preprocessing data.

We monitored the activities triggered by top Italian influencers on the two aforementioned social networks. To this end, we built lists of the most popular Italian influencers, including different categories, like politicians, musicians, and athletes. Those marked as Italian are the ones that communicate on the online social platform mainly using the Italian language.

To get popular profiles, we exploited the online analytics platform `hypeauditor.com` for IG, and `www.socialbakers.com` and `www.pubblicodelirio.it` for FB. The analysis has been restricted to the influencers with at least 10,000 followers on June 1, 2021. The lists of influencers we used are publicly available.[1]

For each monitored profile, we downloaded the corresponding metadata, i.e., the profile information, and all the generated posts, using the CrowdTangle tool and its

[1] `https://mplanestore.polito.it:5001/sharing/KhoYSXAHR`

API[2]. CrowdTangle is a content discovery and social analytics tool owned by Meta[3], which is open to researchers and analysts worldwide to support research, upon subscription of a partnership agreement. Notice that, on IG, users can *like* posts, whereas on FB, they can *react* to posts with a thumbs up or other five pre-defined emojis. Thus, for each post, we collected the *number of likes/reactions* the post received, hereinafter referred to as *interactions*, which CrowdTangle provide in an anonymized manner. Moreover, we also collect statistics about number of comments per post for FB and IG and number of times posts are shared for FB.[4] Finally, we have stored the data, which takes around 110 GB of disk space, on a Hadoop-based cluster, and we have used PySpark for scalable processing.

For each influencer, we downloaded all the data related to the posts published between January 1, 2016 and June 1, 2021. Filtering the not relevant profiles, less than 1k followers, our source is a table with 1.6k influencers and 6 years of data related to post publications.

For each post we have information about the poster account, as well as some metrics. Among the metrics we have likes, comments and - most important for us - the number of followers. From this abundant wealth of data, we filter out three fields that are the relevant ones for our purposes: the influencer name, the fanbase cardinality at the time of the post, and the date. For the sake of description of the dataset we will include also interactions.

name	subscriberCount	commentCount	favoriteCount	date
CALCIATORIBRUTTI	2464046	308	85302	2020-12-31
negramaroofficial	647436	108	6579	2020-12-31
Gio Evan	813443	174	31692	2020-12-31
Antonella Clerici	841190	1555	39277	2020-12-31
SSC Napoli	2564607	701	53640	2020-12-31
Vasco Rossi	1650031	3516	61307	2020-12-31
Mara Venier	2169740	1182	19233	2020-12-31
FEDEZ	11501367	4978	1075043	2020-12-31
Mara Venier	2169740	2035	30353	2020-12-31

Table 4.1: Synergy dataset tail

[2]`https://github.com/CrowdTangle/API`

[3]`https://www.facebook.com/formedia/tools/crowdtangle`

[4]On Instagram it is not possible to share/repost a post on the feed.

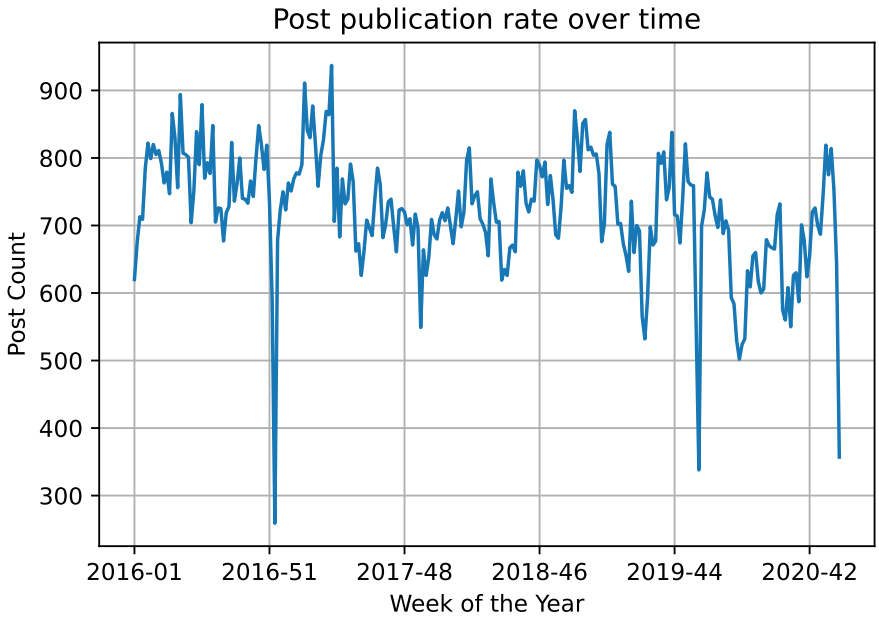


Figure 4.1: Post count time series from 2016 to 2021

4.2 Methodology

The second step of our analytical procedure is transforming the raw wealth of data in a form that is more suitable for our investigation and, lastly, into results and insights.

The steps in our "pipeline" are described in the following list:

- Filtering and keeping only the field of interest (name, subscriber count, date)
- Sampling the data weekly - taking the last value in case of more than one value in a week - in order to avoid noise
- Computing the week-to-week percentage increase in followers
- Applying the %B to the signal computed in the step before
- Identifying sliding windows of length W in time where a couple of influencers has a Pearson correlation greater or equal than a threshold P
- Grouping the above found "synergistic intervals"

The validation process that lead to determine if a *synergistic couple* is an real match (not due to chance) is manual. It consists in searching for status - married,

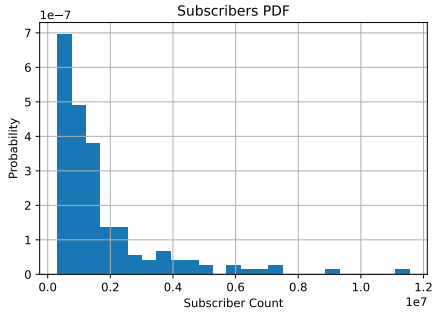


Figure 4.2: Subscriber Count Probability Density Function

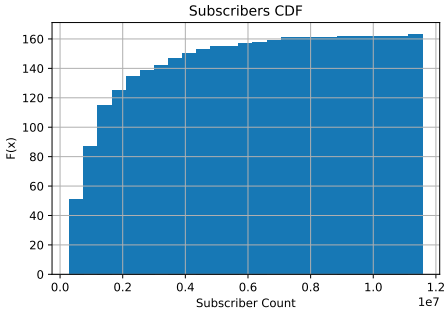


Figure 4.3: Subscriber Count Cumulative Probability Function

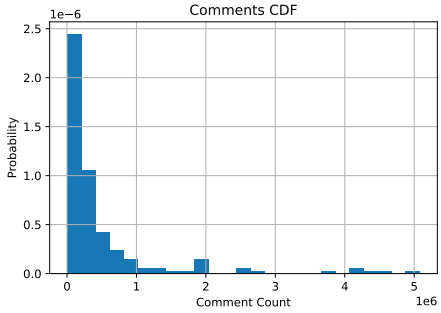


Figure 4.4: Comment Count Probability Density Function

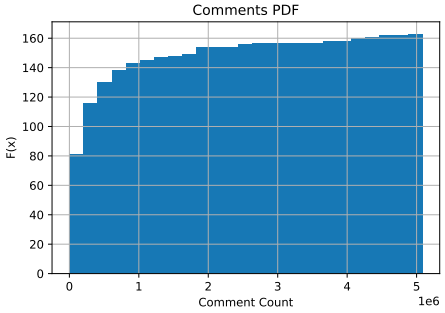


Figure 4.5: Comment Count Cumulative Probability Function

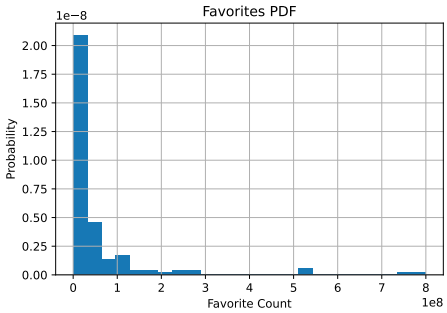


Figure 4.6: Favorite Count Probability Density Function

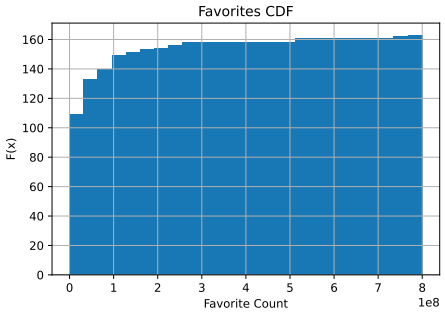


Figure 4.7: Favorite Count Cumulative Probability Function

engaged -, interactions or checking that the couple members belong to the same category of influencers.

This manual validation makes difficult the process of hyper-parameter tuning. Despite that, after several runs of the algorithm, we have found satisfactory results with the following configuration:

- Applying the %B along 8 weeks and using a vertical shift of 2 standard deviations
- Sliding window of width W of 8 weeks
- Pearson correlation threshold P of 0.7

After having found and merged the consecutive synergistic intervals we to sort by relevance the couples of influencers. The relevance measure, that we will simply call *score* from now on, is a simple computation that takes in account both the length and the strength of the intervals. In a general form, we can write the score formula as follows:

$$\text{Score(x,y)} = f(\sum_{i=0}^{NI}\sum_{l=0}^{Li} 1)g(\frac{\sum_{i=0}^{NI}\sum_{l=0}^{Li} corr_{x,y}}{\sum_{i=0}^{NI}\sum_{l=0}^{Li} 1})$$

where NI is the number of separated intervals, L_i is the length of the i-th interval. The first addendum is just the total length of the union of the intervals, while the second is the average correlation. Both terms are mediated by two scale functions f and g. Our best results have been obtained with f f being the identity function and g being the natural logarithm. Thus, the applied formula is:

$$\text{Score(x,y)} = (\sum_{i=0}^{NI}\sum_{l=0}^{Li} 1)\log(\frac{\sum_{i=0}^{NI}\sum_{l=0}^{Li} corr_{x,y}}{\sum_{i=0}^{NI}\sum_{l=0}^{Li} 1})$$

4.3 Results

In the section before we have presented our score function. Sorting in a descending fashion we take under validation the top10 couples. We have that 9/10 couples are actual matches. In our classification a match can be of three different types:

- Gold: very intense interaction (e.g. partnerships, relationships, marriages)
- Silver: not negligible interaction (e.g. flirt, tastes, participating to the same reality show in the same edition)
- Bronze: belonging to the same category

In the following table we will use the acronyms AC for average correlation and IUL for intervals union length.

Before getting to view the final plots for the three gold matches, it is absolutely worth to display the important effect in transforming the signal from absolute number of follower to %B, passing through the percentage difference.

person1	person2	AC	IUL	score	match kind
St3pNy	Surry	0.876	108	4.103	gold
Luca Vezil	Valentina Ferragni	0.854	75	3.690	gold
Chiara Ferragni	Juventus	0.860	70	3.655	bronze
Chiara Ferragni	FEDEZ	0.889	60	3.640	gold
Gianni Morandi	Luciana Littizzetto	0.862	68	3.639	silver
Lorenzo Ostuni	Valentino Bisegna	0.871	63	3.609	silver
Alessia Marcuzzi	Stefano De Martino	0.879	59	3.584	silver
Giorgio Muratore	Surry	0.874	59	3.566	bronze
BastardiDentro	St3pNy	0.870	60	3.562	-
Alessia Marcuzzi	Melissa Satta	0.893	53	3.546	bronze

Table 4.2: Top 10 detected synergies

As mentioned in the introduction and confirmed by the results in the chapters about influencers' efficiency, Bollinger Bands are tremendously helping in performing a dynamic normalization. We show below the process taking into account one of the three gold matches.

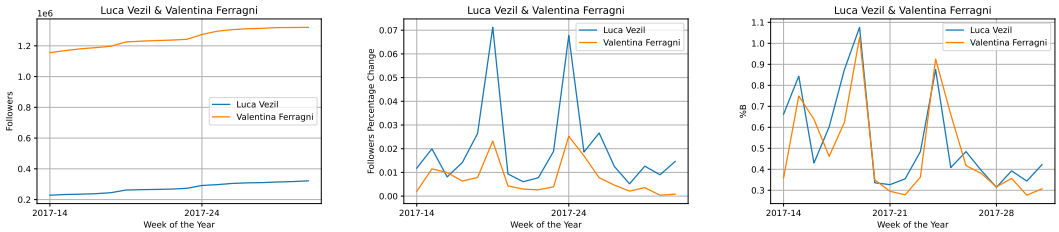
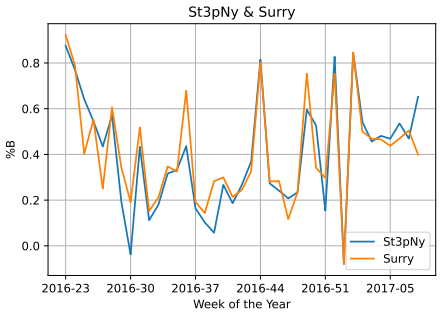


Figure 4.8: Followers (left), Follower percentage variation over time (centre) and Follower percentage variation (in %B terms) over time (right)

Figure 4.9: Stepny and Surry (youtubers)



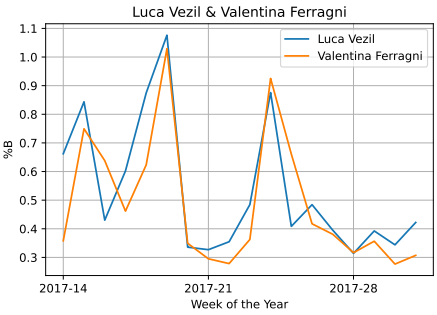
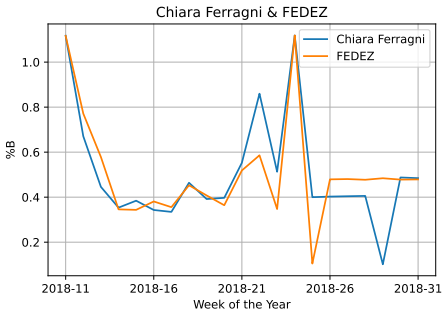


Figure 4.10: Valentina Ferragni and Luca Vezil (engaged)

Figure 4.11: Chiara Ferragni and Fedez (married)



Chapter 5

Conclusions and Future Work

5.1 Takeaway of results

Qui rispondi alle domande che avevi fatto in Introduzione

5.2 Alternatives for exploiting Efficiency

In our analysis, the exogenous factors has been always collectively represented by the Google Trends SVI. This has demonstrated to be a powerful methodology to effectively summarize the interest of the external world in a certain influencer in a given point in time. Given that, this is just one way in which one can quantify the exogenous factors affecting a person.

Other methods can be the followings:

- News scraping
- Social network posts scraping

Regarding news scraping, once one have collected the data, two main kind of analysis can be performed: studying the effects of either the absoulte count of news or the separated count of "good" and "bad" news (in terms of sentiment analysis of the news title and content) on the variation of the involved influencer(s) fanbase cardinality.

Social network post scraping can lead to kind of analysis more focused on the inner working of the OSN dynamics. One example can be measuring the interest about the last song of a given artist by counting and analyzing the posts containing

a tag for either the title of the artist in a time interval in the near future of the release date.

5.3 Future work using the Sinergy

The time-series "financial" approach we adopted during our discussion has proved to give satisfactory results.

Yet, other viable techniques for studying the connection between influencers could be exploited

- Graph algorithms[23]
- Clustering techniques (in order to spot similarities)
- Time series clustering [24]
- Time series classification [25]

Spotting synergies among influencers may provide significant possibilities for social networks interactions. First, developers can enhance the recommendation about who to follow taking the union of the not-followed halves of synergistic influencers couples/groups a person follows. Second, following the same similarity logic of above, one can also develop purchase recommendation. Similar examples can be found with a little bit of imagination.

Bibliography

[1] Enrico Ubaldi, Raffaella Burioni, Vittorio Loreto, and Francesca Tria. «Emergence and evolution of social networks through exploration of the Adjacent Possible space». In: *Communications Physics* 4.1 (Feb. 2021). DOI: 10.1038/s42005-021-00527-1. URL: https://doi.org/10.1038%2Fs42005-021-00527-1 (cit. on pp. 1, 22).

[2] E.F. Fama. «Efficient Capital Markets: A Review of Theory and Empirical Work». In: *Journal of Finance* 25.2 (1970), pp. 383–417 (cit. on p. 2).

[3] Burton G Malkiel. «The efficient market hypothesis and its critics». In: *Journal of economic perspectives* 17.1 (2003), pp. 59–82 (cit. on p. 2).

[4] Patricia M Fairfield, J Scott Whisenant, and Teri Lombardi Yohn. «Accrued earnings and growth: Implications for future profitability and market mispricing». In: *The accounting review* 78.1 (2003), pp. 353–371 (cit. on p. 2).

[5] Pring M. J. *Technical analysis explained: The successful investor's guide to spotting investment trends and turning points*. Ed. by McGraw Hill (cit. on p. 3).

[6] Spiros Papadimitriou, J. Sun, and Philip Yu. «Local Correlation Tracking in Time Series». In: Dec. 2006, pp. 456–465. DOI: 10.1109/ICDM.2006.99 (cit. on pp. 3, 10, 11).

[7] J. Bollinger. «Bollinger on Bollinger Bands». In: (2002) (cit. on pp. 3, 13, 36).

[8] Dean Fantazzini. «Short-term forecasting of the COVID-19 pandemic using Google Trends data: Evidence from 158 countries». In: 59 (2020), pp. 33–54 (cit. on p. 5).

[9] Rodrigo Mulero and Alfredo Garcia-Hiernaux. «Forecasting Spanish unemployment with Google Trends and dimension reduction techniques». In: *SERIEs* 12 (Apr. 2021) (cit. on p. 5).

[10] Tomas Havranek and Ayaz Zeynalov. *Forecasting Tourist Arrivals with Google Trends and Mixed Frequency Data*. Tech. rep. 2018 (cit. on p. 5).

[11] Eli Beracha and M. Babajide Wintoki. «Forecasting Residential Real Estate Price Changes from Online Search Activity». In: *Journal of Real Estate Research* 35.3 (2013), pp. 283–312 (cit. on p. 6).

[12] Yuyao Feng, Guowen Li, Xiaolei Sun, and Jianping Li. «Forecasting the number of inbound tourists with Google Trends». In: *Procedia Computer Science* 162 (2019). 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence, pp. 628–633. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.12.032. URL: https://www.sciencedirect.com/science/article/pii/S1877050919320423 (cit. on p. 6).

[13] Prasanth Sikakollu, Uttam Singh, Arun Kumar, P.H.J. Chong, and Vinay Tikkiwal. «Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach». In: *Chaos Solitons & Fractals* 142 (Oct. 2020). DOI: 10.1016/j.chaos.2020.110336 (cit. on p. 7).

[14] Salvatore Carta, Andrea Medda, Alessio Pili, Diego Reforgiato Recupero, and Roberto Saia. «Forecasting E-Commerce Products Prices by Combining an Autoregressive Integrated Moving Average (ARIMA) Model and Google Trends Data». In: *Future Internet* 11.1 (2019). ISSN: 1999-5903. DOI: 10.3390/fi11010005. URL: https://www.mdpi.com/1999-5903/11/1/5 (cit. on p. 8).

[15] Emmanuel Silva, Hossein Hassani, Dag Madsen, and Liz Gee. «Googling Fashion: Forecasting Fashion Consumer Behaviour Using Google Trends». In: *Social Sciences* 8 (Apr. 2019), p. 111. DOI: 10.3390/socsci8040111 (cit. on p. 8).

[16] Andrea Maugeri, Martina Barchitta, and Antonella Agodi. «Using Google Trends to Predict COVID-19 Vaccinations and Monitor Search Behaviours about Vaccines: A Retrospective Analysis of Italian Data». In: *Vaccines* 10.1 (2022). ISSN: 2076-393X. DOI: 10.3390/vaccines10010119. URL: https://www.mdpi.com/2076-393X/10/1/119 (cit. on pp. 8, 9).

[17] Farrukh Ahmed, Dr. Raheela Asif, Dr. Saman Hina, and Muhammad Muzammil. «Financial Market Prediction using Google Trends». In: *International Journal of Advanced Computer Science and Applications* 8.7 (2017) (cit. on p. 9).

[18] Hyunyoung Choi and Hal Varian. «Predicting the present with Google Trends». In: *Economic record* 88 (2012), pp. 2–9 (cit. on p. 9).

[19] Tobias Preis, Daniel Reith, and H Eugene Stanley. «Complex dynamics of our economic life on different scales: insights from search engine query data». In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368.1933 (2010), pp. 5707–5719 (cit. on pp. 9, 10).

[20] Jiawei Zhang. «Social network fusion and mining: a survey». In: *arXiv preprint arXiv:1804.09874* (2018) (cit. on p. 10).

[21] Seth A Myers, Chenguang Zhu, and Jure Leskovec. «Information diffusion and external influence in networks». In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, pp. 33–41 (cit. on p. 10).

[22] Marçal Mora-Cantallops, Salvador Sánchez-Alonso, and Anna Visvizi. «The influence of external political events on social networks: The case of the Brexit Twitter Network». In: *Journal of Ambient Intelligence and Humanized Computing* (2019), pp. 1–13 (cit. on p. 10).

[23] Seungbae Kim, Jinyoung Han, Seunghyun Yoo, and Mario Gerla. «How Are Social Influencers Connected in Instagram?» In: Sept. 2017, pp. 257–264. ISBN: 978-3-319-67255-7. DOI: 10.1007/978-3-319-67256-4_20 (cit. on pp. 10, 44).

[24] Sr Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Wah. «Time-series clustering - A decade review». In: *Information Systems* 53 (May 2015). DOI: 10.1016/j.is.2015.04.007 (cit. on pp. 10, 11, 44).

[25] Gian Antonio Susto, Angelo Cenedese, and Matteo Terzi. «Time-Series Classification Methods: Review and Applications to Power Systems Data». In: Jan. 2018, pp. 179–220. ISBN: 9780128119686. DOI: 10.1016/B978-0-12-811968-6.00009-7 (cit. on pp. 10, 11, 44).

[26] 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*. 2019. URL: http://hdl.handle.net/11583/2752645 (cit. on p. 11).

[27] 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*. 2020 (cit. on p. 12).

[28] 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), p. 100136. ISSN: 2468-6964. DOI: https://doi.org/10.1016/j.osnem.2021.100136 (cit. on p. 12).

[29] 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*. ASONAM '21. Virtual Event, Netherlands: 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. 12).

[30] Luca Vassio, Michele Garetto, Emilio Leonardi, and Carla Fabiana Chiasserini. «Mining and modelling temporal dynamics of followers' engagement on online social networks». In: *Social Network Analysis and Mining* 31 (Jan. 2022), p. 012012. DOI: 10.1007/s13278-022-00928-2. URL: https://doi.org/10.1007/s13278-022-00928-2 (cit. on p. 12).

[31] Christina Pagel, Padmanabhan Ramnarayan, Samiran Ray, and Mark J Peters. «A novel method to identify the start and end of the winter surge in demand for pediatric intensive care in real time». In: *Pediatric Critical Care Medicine* 16.9 (2015), pp. 821–827 (cit. on p. 13).

[32] Hiram Ting, Winnie wong poh ming, Ernest De Run, and Sally Choo. «Beliefs about the Use of Instagram: An Exploratory Study». In: *International Journal of Business and Innovation* 2 (Jan. 2015), pp. 15–31 (cit. on p. 15).

[33] Duncan Brown and Nick Hayes. *Influencer marketing*. Routledge, 2008 (cit. on p. 16).

[34] Morgan Glucksman. «The rise of social media influencer marketing on lifestyle branding: A case study of Lucie Fink». In: *Elon Journal of undergraduate research in communications* 8.2 (2017), pp. 77–87 (cit. on p. 16).

[35] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. «The YouTube Social Network». In: (Jan. 2012) (cit. on p. 31).