Big Data Research 24 (2021) 100189

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Social Media Data and Users' Preferences: A Statistical Analysis to Support Marketing Communication

Elisa Arrigo, Caterina Liberati *,[1], Paolo Mariani

University of Milano-Bicocca, p.zza Ateneo Nuovo 1, 20126 Milan, Italy

ARTICLE INFO

Article history:
Received 16 February 2020
Received in revised form 22 August 2020
Accepted 21 December 2020
Available online 7 January 2021

Keywords:
Social media marketing
Facebook likes
Marketing communication
Dimensions reduction
Users profiling

ABSTRACT

Differently from traditional transaction data, social media data are difficult to investigate due to their volume, variety and velocity. Indeed, the knowledge extraction from social media data raises several issues especially for what concerns statistical exploration and synthesis of complex information. Our work aims to study users' preferences, stated on a social media platform, in order to aid businesses to make their marketing communication decisions. We rely our analysis upon 5685 Italian Facebook users interested in pharmaceutical products and health. The data have been collected at the end of 2014 and are focused on Likes actively expressed by the subjects on specific categories of interests (TV Channels and Magazines). Through a factorial analysis we uncovered significant associations between marketing communication Media and users' profiles. This allows sketching out a marketing strategy in twofold actions: first, identifying the target group to reach and, then, the nearest suitable channel where to develop the marketing communication.

1. Introduction

In September 2019, worldwide over 2.45 billion users monthly were active in Facebook (FB) with a 8% increase year over year [1]. In Italy, Internet users continue to increase, by reaching a penetration rate of 78.4% of the Italian population, and a percentage equal to 56% refers to Italians who have an account on Facebook [2]. Customers interact with brands through social media for several reasons such as, for example, entertainment, brand engagement, access to customer service and content, product information, and promotions [3]. Consequently, social media represent for companies a relevant marketing tool that affects marketing strategies and practices [4]. At first, the social media advent has surely changed the marketing communication media landscape since new channels have been complemented and, sometimes, substituted the traditional ones [5]. However, this does not represent the unique opportunity provided by social media to companies. Previous research [6] has also shown how Facebook *Likes* can be used to automatically and accurately predict customer personalities and profiles.

Therefore, the aim of this paper is to investigate and synthesize FB data to support marketers involved in setting a marketing communication media plan in understanding the traditional media usage patterns of their customers. The usage of the Facebook *Likes* – actively set by the users – to uncover subjects' preferences, is the innovative contribution of this paper. Naturally, this proposal is strongly affected by the business' capacity to deal with big data, still far to be operative. In order to make data dimensions manageable for any quantitative exploration, we propose to shrink the standard data matrix, that generally has the monitored individuals as units, into psychographic segments (idealtypes). Such an approach makes it possible to extract value from big data and also allows to get reliable results, although the direct link to the micro level is dropped. Moreover, it opens up new opportunities to gain insights about clients (or potential ones), adapting the schema to the new data environment. To the authors' knowledge, no study has yet fully explored the synthesis process of Facebook data for marketing communication purposes and the present research aims at filling this research gap.

The paper is organized as follows: after the Introduction, Section 2 analyses the theoretical background on social media marketing and marketing communication media environment. Section 3 clarifies the data and Section 4 describes the empirical analysis and the results. Section 5 contains the discussion of results and, finally, Section 6 draws the conclusions with limitations.

* Corresponding author at: Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, p.zza Ateneo Nuovo 1, 20126 Milan, Italy.
E-mail address: caterina.liberati@unimib.it (C. Liberati).

[1] Authors' contribution: Arrigo E. investigated the marketing implications of the work and wrote the final version of the paper. Liberati C. developed the analytical framework, performed the computations and wrote the final version of the paper. Mariani P. conceived the initial idea and supervised the project.

https://doi.org/10.1016/j.bdr.2021.100189
2214-5796/

2. Theoretical background

2.1. Social media marketing and Facebook likes

Social media marketing refers to the process of gaining customers' attention and acceptance through social media platforms such as Facebook or Twitter [7]. Social media can be defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of User Generated Content [8, p. 61]. By monitoring social media and collecting data about customer desires and competing offers, companies can obtain a multifaceted vision of the market and a good understanding of current or potential issues inside [9,10]. Recent studies have shown that social media provide marketers with several business opportunities by representing an informal source for understanding customers' preferences, competitors' activities, market trends and product feedbacks [11].

Within social media platforms, companies have the opportunity to listen to consumers directly during their online conversations without any filters and to develop a deep customer knowledge. Social media intelligence [12–15,9] allows deriving actionable information from social media, by creating corresponding decision-making frameworks, and providing solutions for existing and new applications that could benefit from the crowd through the Web [16]. In particular, social media intelligence enables linking social media data to strategic management decisions and business performance by covering the traditional competitive intelligence steps of data collection and evaluation of gathered information.

Facebook provides users with a form of many-to-many communication, which enables them to broadcast information simultaneously to their network. Since broadcasting on Facebook can be associated with low-quality interactions and ambiguity by leading users to experience frequent information overloads, Facebook designers have conceived some tools and features such as *Likes*, comments, and post length and type to simplify the users' identification of relevant information [17].

Facebook employs the rating mechanism of *Likes* to provide users with an easy measure of popularity by indicating the number of people that positively responded to a shared information or post [17]. In fact, the Facebook's system to register likes, comments and shares, measures brand-consumer interactions [18,19]. Moreover, many scholars have considered the number of *Likes* of each post on Facebook as a measure of consumer response [20] also an affective response [21].

Likes are considered, also, as one of the most popular electronic forms of communication since the one-click social plugins, such as the *Like* button through which customers can share their interest or convey their attitude about various contents, usually facilitates the development of an electronic Word-Of-Mouth (eWOM) [22]. Namely, eWOM refers to any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet [22]. One-click social plugins are thus buttons through which customers can show and share their interest about a content; they differ from any other online eWOM forms since only with a click customers can share content relevant for them with their network of contacts [23,24] increasing the popularity of their posts [25]. Once a user likes a company page or post, this information appears in his friends' feeds.

2.2. Integrated Marketing Communication (IMC) media environment

In recent years, the marketing communication media environment has changed dramatically by becoming global thanks to digital technologies that have also enabled firms to reach both new markets and the existing ones with different channels. New technologies have also fragmented the media exposure via unparalleled expansion of communication channels [26,27]. At the same time, they have transformed the way in which customers process firms' marketing communications; in fact, customers are empowered to choose when, why, and how to enter in contact with company communications. More precisely, they can decide whether to receive marketing or corporate communications and on which media.

Social media represent for companies an efficient channel to deliver marketing and institutional communications to the online community and are even considered as a hybrid element of the promotion mix [28]. In fact, social media bring together aspects of the traditional marketing communication mix to a highly magnified form of Word-Of-Mouth among customers where companies cannot manage either the content or the frequency of this communication.

Nowadays the Internet advent has modified the nature of communication models [29] and some studies have explored how social media disrupt traditional channels and media [5] [30]. The new social channels represent surely a valuable addition to the set of marketing channels that companies can use; however, they have posed some problems to solve such as measurement or brand control issues. While in TV, print, outdoor and radio, marketing communications are one-way and controlled, within social media companies cannot fully control what customers are saying about their brand since social media are completely outside their control [29]. Social media, smartphones, and tablets have, thus, forced firms to rethink many traditional practices in their marketing communication plan.

Companies have been pushed to put an increasing attention on searching the best integration between traditional and digital media inside their marketing communication plan to drive sales and brand building [31]. Customer buying behavior is shorter in length and more complex; customers do not passively receive brand information only through mass media and store it in memory for later use as in the past. On the contrary, they actively seek information when needed through Internet and social media [32].

Although marketers try to encourage customers to engage with their brands on social media, research has shown that these latter platforms cannot represent the unique source of marketing communications for a brand because they cannot be as effective as traditional media in attracting new customers and increasing brand penetration [32]. Moreover, although social media and online communications may be more influential that mass communications, traditional mass media are still considered very important to stimulate them [33]. In fact, an Integrated Marketing Communication plan requires the coordination and integration of all marketing communication tools and media within a company into a seamless program which maximizes the impact on consumers at a minimal cost [34].

With regard to the pharmaceutical market, which Facebook data used for the analysis refer to, it is important to underline that, in Italy, the promotional activities (such as advertising) on prescription drugs are banned.[2] It is required an approval by a regulatory body appointed by the Ministry of Health on the marketing communication content pertaining the Over-The-Counter (OTC), before the marketing communication be disseminated. In order to comply with these regulations, pharmaceutical companies delineate marketing communication policies addressed at promoting OTC products or their corporate brand. Surely, the key features

[2] Pharmaceutical companies are subject to Italian Law regulations, such as Leg. Decree No. 219/2016, which establishes that only Over-The-Counter products can be promoted through marketing communications.

of social media have added complexity to pharmaceutical drug marketing policies since companies can reach quickly and cheaply many customers online with interactive and promotional activities and customers can add content as well. Moreover, about 17% of Facebook posts located in searches for popular pharmaceutical brands are advertisements for illegal online pharmacies; therefore, it becomes necessary monitoring what information consumers are exposed[3] to when searching online [35].

3. The data

As hinted in the Introduction, the 52th Report of the Censis, [2] about the usage of the media, noticed that the 78.4% of the Italian population declared to surf in Internet regularly. Also people that have been using Facebook are increased up to 56% among Italian citizens, ensuring to FB a preeminent position in the rank of the most used platforms to connect users with other people, companies, institutions, groups and so on. In particular, by means of Facebook and its APIs[4] available to software developers, it is possible to access much of users' data, upon authentication procedure to the social platform:

- information directly disclosed by user: as the description of the user account in terms of socio-demographics, or location or the date when a particular event is posted
- information and contents that other people provide: for example when they share a photo
- information about people and groups a person is connected with, or about the groups he/she likes to share content with
- purchases or economic transactions
- geographic locations collected by pc, phones or other devices via GPS, Bluetooth or Wi-Fi
- information on websites' visits or the use of third-party apps, such as advertising and measurement services.

As such regards, several articles have investigated the impact of Facebook in the modern society (in a broad sense): some are focused on users' characteristics [36–38] others on the role of the platform in the social interactions [39]. The growing interest about FB to build a potential segmentation and to infer about personality traits or users' behavior remains of constant relevance [6]. Among the data (most suitable) for analyzing users' activities, *Likes* represent a quantitative alternative to any other ways to express a reaction to a content [40].

Our research was conducted at the end of 2014 on Italian Facebook users interested in pharmaceutical products and health. In collaboration with Cubeyou [41], we collected the interactions (i.e. likes) assigned by people to pages of pharmaceutical companies or institution related to health and wellness. Of course, such huge amount of information could not be handled and processed with standard computing engineering. Therefore, the raw data were stored on a cloud platform with servers active on Amazon Web Services infrastructure. More than 5 Terabyte (distributed) database were gathered and updated daily via Hadoop2.[5] In this case, the synthesis process stored information into three main tables: the Behavioral Table, that contained each user by Facebook page interactions, the User Demographic Table that collected unstructured data about users' profiles, the Demographic Table that stored unstructured data about Facebook pages (Fig. 1). In each table, records were extracted with queries based on users' keys and behaviors.

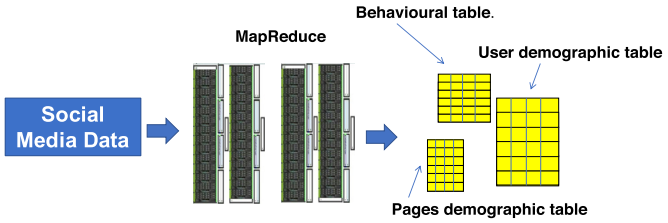

Fig. 1. MapReduce process flow.

Finally, the built matrix had size 5685 rows (Facebook users with at least one like to the monitored pages) and 140 dummy columns (pages visited/liked). Beside information collected by the social platform, the subjects were segmented according to classification in [6], which distinguish users into 19 alternative psychographic profiles: Pet Lovers, Outdoor Enthusiast, Techies, Car Lovers, Book Lovers, Social Activist, Gamers; Movie Lovers, Politically Active, Sport Lovers, Fashion Lovers, Music Lovers, Travel Lovers, Public Figures Followers, Food Lovers, Home Decorators, Beauty and Wellness Aware, Business People and House-keepers.[6]

4. Empirical analysis and results

As it is well known, FB data are very rich and comprehensive compared to the other social networks. Indeed, records of FB users range among the socio-economic attributes that contain information as age, gender and education, to social status, that stores data referred to the user-claimed friendship, to individual interests that collect subjects likes relative to different domains as music, TV channels, magazines or activities. For sake of brevity, we focus our attention to users' *Likes* for TV channels and magazines but the analysis could be easily extended to all the other categories.

The observed sample shows a balanced gender distribution (male 45.63% female 54.37%) if we exclude the missing values (Table 1). The subjects are mostly engaged and aged lowest thru 40. As expected, mature age classes have little relevance due to the peculiarity of the medium which is particularly used by young people [2]. Regarding the education, the sample shows a clear presence of graduated subjects (65.60%) with respect to ones that hold a high school degree (Table 1). Such prevalence is even more noticeable if we consider the valid percentage (instead of the frequency) which includes the missing values.

For what concerns the users' interests for the television broadcasting (Table 2 in Appendix A.1), it emerges a clear subjects' preference for the entertainment, followed by a propensity to watch thematic TV channels and news.[7] Same picture if we consider magazines, albeit the categories have different weights. Also in this case, the percentage of missing is relevant.

Lack of information occurs very often with social media data. Referring to our study, we chose to evaluate only preferences explicitly declared by the users via *Like* button. Missing data in our

[3] On July 25th, 2017 the Italian Ministry of Health has also provided guidelines for regulating OTC advertising on social media such as Facebook, Instagram, and YouTube.

[4] An application Programming Interface (API) is a set of tools, definitions and protocols for compiling and integrating application software. It allows your products or services to communicate with other products or services without having to know how they are implemented.

[5] Hadoop is an open-source software designed to handle extremely high volumes of data in any structure. It has two components: 1) the Hadoop Distributed File System, which supports data in structured relational form, in unstructured form, and in any form in between and 2) the MapReduce paradigm for managing applications on multiple distributed servers and to perform parallel computations.

[6] Cubeyou collected all the interactions between people and brands, products and services (shares, likes, tweets, pins, posts) on several social media platforms and classified them according to 1,500 categories. They developed a classification algorithm in collaboration with Cambridge University that created the 19 psychographic profiles.

[7] The measure of intensity of the user interest toward a category has been carried out by means of sum of the likes collected by the single TV channels or magazines.

X (5685 x 60)

id	Psyc. Profile 1	...	Psyc. Profile 18	Mag 1	..	Mag 18	Tv 1	..	Tv 19	Gender	Age class	Education	Rel. Status
1	1	..	0	1	..	0	0	..	0	F	<=20	High School	Single
2	0	..	1	0	..	0	0	..	1	M	31-40	High School	Engaged
3	0	..	1	0	..	0	1	..	1	M	>=61	Graduated	Single
..	..	..	..	..	..	..	..	..	..	..	..	..	..

X_{red} (19 x 37)

	Mag 1	..	Mag 18	Tv 1	..	Tv 19	F	M	<=20	..	>=61	High school	Graduated	Single	Engaged
Psyc. Profile 1	n_{11}	..	n_{118}	n_{119}	..	n_{137}	n_{138}	n_{139}	n_{140}	..	n_{146}	n_{146}	..	..	n_{149}
Psyc. Profile 2	n_{21}	..	..	..	..	..	..	..	..	..	..	..	..	..	..
.....		..	..	..	..	..	..	..	..	..	..	..	..	..	..
Psyc. Profile i	n_{i1}	..	n_{i18}	n_{i19}	..	n_{i37}	n_{i38}	n_{i39}	n_{i40}	..	n_{i46}	n_{i46}	..	..	n_{i49}
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

Fig. 2. Dimensions reduction process.

Table 1
Percentage distributions of demographics variables.

Demographics		Frequency	Percent	Valid percent
Gender	Female	2702	47.50%	54.37%
	Male	2268	39.90%	45.63%
	missing	715	12.60%	
Education	Graduate	3727	65.60%	80.41%
	High school	908	16.00%	19.59%
	missing	1050	18.50%	
Age	Lowest thru 20	171	3.00%	4.90%
	21-25	796	14.00%	22.60%
	26-30	968	17.00%	27.50%
	31-40	901	15.80%	25.60%
	41-50	447	7.90%	12.70%
	51-60	190	3.30%	5.40%
	61 thru highest	42	0.70%	1.20%
	missing	2170	38.20%	
Relationship status	Engaged	2119	37.3%	68.60%
	Single	970	17.1%	31.40%
	missing	2596	45.7%	

case might indicate subjects' indifference, dislike, or unawareness about existence of some Facebook web pages. Given the impossibility to distinguish among those attitudes, we restrict our analysis to the stated preferences.

In such regards, the sparsity of the matrix has been overcome by means of a dimensions reduction process that organizes users' preferences per psychographic profiles. The initial matrix X was 5685 (subjects' records) $\times$ 60 (variables). Specifically, 4 columns of X collect information about four social-demographic characteristics of the users (gender, age, education and relationship status), the others are individuals' attributes relative to their lifestyle values and behavioral traits. More in detail, such attributes were dummy variables: 19 were subjects' psychographic descriptors and 37 were preferences' indicators for the monitored media. According to the standard coding they show value 1 in case the user has been assigned to a specific psychographic segment[8] or in case the user likes a TV channel/magazine, 0 elsewhere. Visual inspection of X reveals a severe presence of null records that would effect any synthesis. Such an issue was solved working with a reduced matrix of instances X_{red} whose rows do not correspond to the single individuals but to the psychographic segments. Based on that, we build a two-way contingency table composed by the 19 alternative segments (rows) and 37 variables (columns) obtained as sum of the likes for each category (Fig. 2).

In order to study and visualize the associations between the levels of a two-way contingency table, we employed the Correspondence Analysis (CA) on X_{red} (outlined in Appendix). CA provides a geometric representation of the rows and columns as points in a low-dimensional space, according to the chi-square metric [42,43]. In our case the chi-square statistic for testing independence between rows and columns produces a value of $\chi^2 = 1423.263$ (648 degrees of freedom), which is highly significant (Table 3 in Appendix A.3). The CA solution based on eigenvalue-eigenvector decomposition, provided 18 uncorrelated factors. For our analysis, we retained the first two axes (f_1, f_2) disregarding a residual amount of inertia (29.5%). Each factor accounts for a separate part of variance: the first axis f_1 explains 40.8% of the total inertia, the second f_2 an additional 29.6% (Table 3 in Appendix A.3). Fig. 3 displays the symmetric map of the principal coordinates of the rows and columns.[9] The first (horizontal) dimension reflects a clear contrast between Sport and News Media (on the right) and Fashion and Design Media (on the left). Actually, all the periodicals devoted to the style and new trends lie very close to each other on the negative side of the dimension, while all the other magazines are in the opposite side of the factor. In particular, the 17.6% of f_1 variance is explained by Sky Sport (0.10) and Internazionale (0.076), whereas Donna Moderna (0.166) Cosmopolitan Italia (0.063) Grazia (0.053) Casa Facile (0.052) and Tu Style (0.05) weight for an additional 38.40%. Also the quality of points representation, that provides additional richness to the interpretation of the relationships in the contingency table, confirms such pattern (Table 4 in Appendix A.3). As for the media, the psychographic segments associated to the factor are coherent with the description provided: on the right, we see Sport Lovers and Gamers that account for 18.9% of the f_1 variance (as from contributions in Table 5 in Appendix A.3), instead Fashion Lovers, Beauty and Wellness Aware and Housekeepers, located at the opposite left-hand side of this axis, contribute to the explanation of the first principal component for an additional 58.6%. Based on such considerations f_1 can be interpreted as *Users' Interests*. The second dimension discloses differences between Journalistic in-depths (positive pole) and Entertaining Programs/Leisure Magazines (negative pole). Also this contrast is strong: it accounts for most of the 29.6% of f_2 variance. As we already noted by inspecting Table 4, Internazionale (0.10) Artribune (0.099) National Geographic Magazine (0.059) on one side, Sky Sport (0.208) Real Time (0.095) DMax Italia (0.074) on the other side, contribute for more than 63% to the factor's interpretation. For what concerns the segments, their distribution along the second factor mirrors a coherent pattern as we already observed for the former dimension: Sports Lovers, Gamers and Car Lovers oppose to Politically Active, Book Lovers and Social Actives. Also in this case such opposition explains more than 78%

[8] As illustrated in sect. 3, the psychographic segments have been built using further personal information, external to FB choices, produced by the subjects.

[9] Fig. 3 is the map of the principal coordinates of the rows and columns of X_{red}. The graphical representation has been split into two plots for guarantee an easy inspection of the labels.

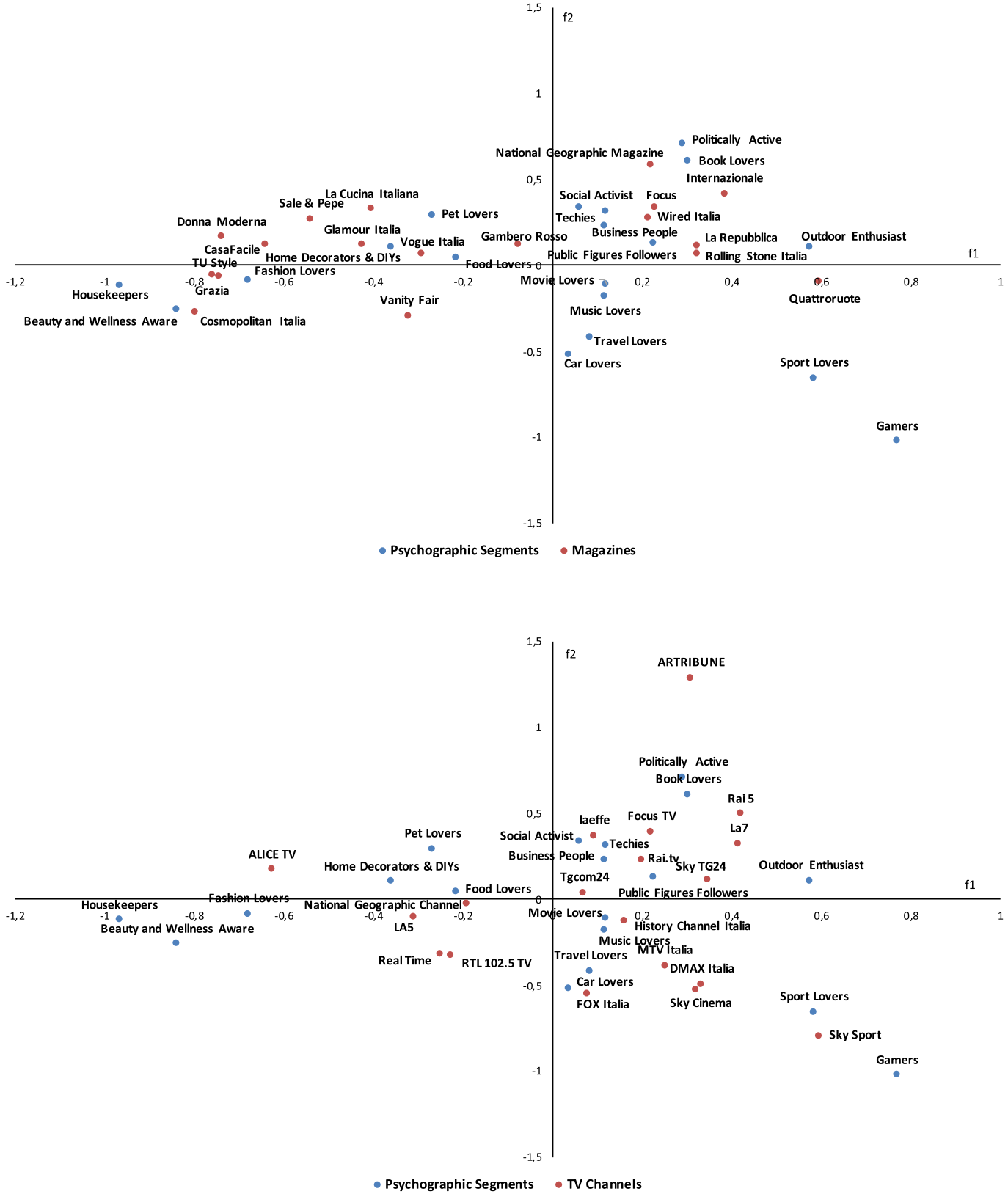

Fig. 3. Normalized 2-dimensional plots of Psychographic Segments vs Magazine (upper panel) and Psychographic Segments vs TV channels (lower panel). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

of the factor variance (Table 5). That led to name f_2 as *Genres Media*. The map of the association between users profiles and media is shown in Fig. 3.

In order to obtain more insights about the observed configuration, we projected categories of additional variables on the already existing solution plane. According to the CA framework, the role of those supplementary variables is to support and complement the interpretation of the original categories associations based on the active variables. Indeed, the geometric orientation of the axes is not affected by the presence of such new variables. To display the supplementary points, we used the transition formulae [44] which are based on the singular value decomposition of the original two-way contingency table (in our case X_{red} of dimensions $R \times C$). Illustrative rows (r^+) or columns (c^+) can be obtained easily:

$$r_s^+ = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{C} \frac{f_{ij}^+}{f_{i\cdot}^+} c_s(j) \qquad (1)$$
$$c_s^+ = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{R} \frac{f_{ij}^+}{f_{\cdot j}^+} r_s(i)$$

where λ_s is the eigenvalue corresponding to the s-th factor axis, $f_{ij}^+/f_{i\cdot}^+$ is the supplementary row profile, $f_{ij}^+/f_{\cdot j}^+$ is the supplementary column profile, $c_s(j)$ indicates the coordinate of the category j on axis s, whereas $r_s(i)$ is the coordinate of the category i on the same axis (Eq. (8)–(9) in Appendix A.2).

In our analysis, we used as illustrative the socio demographic variables available for the FB users: Gender, Education, Relationship Status and Age. Results, depicted in Fig. 4, show very interesting polarizations coherent with the points' configurations already illustrated.

The interpretation of the new maps is similar to the ones already analyzed: the proximity of an illustrative variable to a profile/media means strong association between those points. Moreover, the more distant lie the supplementary variables from the origin[10] the more important are in the characterization of the points' configurations. For example, the quadrant where users appreciate mainly news and cultural in-depth, it is also the one with a high effect of males young and single. Directly opposed quadrant, where users are big fans of entertainment shows and fashion magazines, is characterized by females profiles generally mature poorly-educated and engaged. Inspecting the remaining quadrants there aren't any supplementary variables: this occurs when no characterization, measured by the further information, is relevant. In other words, the socio-demographic indicators do not show any peculiar association with the active profiles.

5. Discussion

The objective of this research was to synthesize Facebook data in order to support marketing communication managers in forecasting the traditional media usage patterns of their customers. The carried out analysis on Facebook *Likes* combined with different psychographic segments has provided two maps, which four quadrants display the associations between customers' profiles, socio-demographics data and the most suitable magazines and TV channels. Thus, for instance, female gender with a high school's education title appears near to magazines and watching TV channels focused on Fashion and Design for leisure and entertainment purposes; while graduate customers, mainly masculine, are near to a journalistic content of magazines and TV channels. Such results could be useful for companies to direct future media choices into an Integrated Marketing Communication Plan. In fact, it is well known that for the development of an effective marketing communication strategy, a company should be able to reach its target audience directly; however, this has become very difficult due to the media expansion [26]. Moreover, a customer audience segmentation comes necessary since it is impossible to generate a communication flow attractive and effective for all customers: customers are often split up according to demographic, geographical and other factors [34]. It is very important to know not only the size but also some features of the audiences provided by different media, traditional and social they are. According to [29], Internet has surely modified the nature of marketing communication models by multiplying the opportunities of customers' touch points. However, in this new scenario, traditional media such as press and television still are maintaining a key role [33] due to the wide and segmented audience they offer. The explosive growth in Facebook and social media has produced huge quantities of texts, images and videos that some scholars define as "digital footprint" [45]. Digital footprint provides companies with an alternative channel to gain an understanding of their customers useful in the formulation of a marketing communication strategy. In this paper, social media data and, more precisely, Facebook *Likes* have allowed to model the customer profiles' usage of magazines and TV channels by providing a marketing manager with a better understanding of the best magazine of TV channel to deliver promotional communications directed to the target audience.

The contribution of this study is multiple: it provides a practical strategy to extract knowledge from social media data whose synthesis raises several statistical issues [46]. One of these difficult challenges is related to the veracity of the social media data that often are of low quality due to the presence of massive missing data. Without a pre-treatment, the inferences provided might be unreliable and noisy [15]. In this regard, our approach is able to do meaningful synthesis that speed-up the computational process and the interpretation of the empirical evidence, due to the overcoming of the sparse matrix. Moreover, the pre-treatment proposed that transforms fine-micro data in aggregate profiles (as open data) prevents any issues related to the inference of individuals' identity [47], according to the guidelines of General Data Protection Regulation [48]. Concerning a marketing perspective, this research contributes to the academic research on data driven decision-making since it provides empirical evidence that social media analytics can have an important impact on improving marketing communication media choices of companies. Moreover, findings contribute also to marketing communication academic literature where some studies [5] [30] argue that social media disrupt traditional channels and media. Our study reinforces previous research by suggesting that social media do not necessarily substitute traditional media but, on the contrary, they are proven useful to orientate the marketing communication investments on magazines or TV channels.

The study offers also practical insights and guidance for managers who are engaged in marketing communication and social media. First, especially fast moving consumer goods' companies that use both traditional and social channels for their marketing communications should consider the opportunities provided by social media analytics in terms of whole IMC plan. In fact, social media data and big data are considered very relevant in the current market place, however as illustrated having social data is not enough; more complex is elaborating them. Therefore, marketing managers in order to leverage social media data should acquire the best skills and competences, such as consumer data scientists, able to extract and exploit customer insights. Furthermore, concerning marketing communication decisions, it is not a problem of removing from the IMC plan some communication channels; instead, the key aspect is to develop a perfect integration among all market-

[10] The origin of the bi-plot always represents the average profile values.

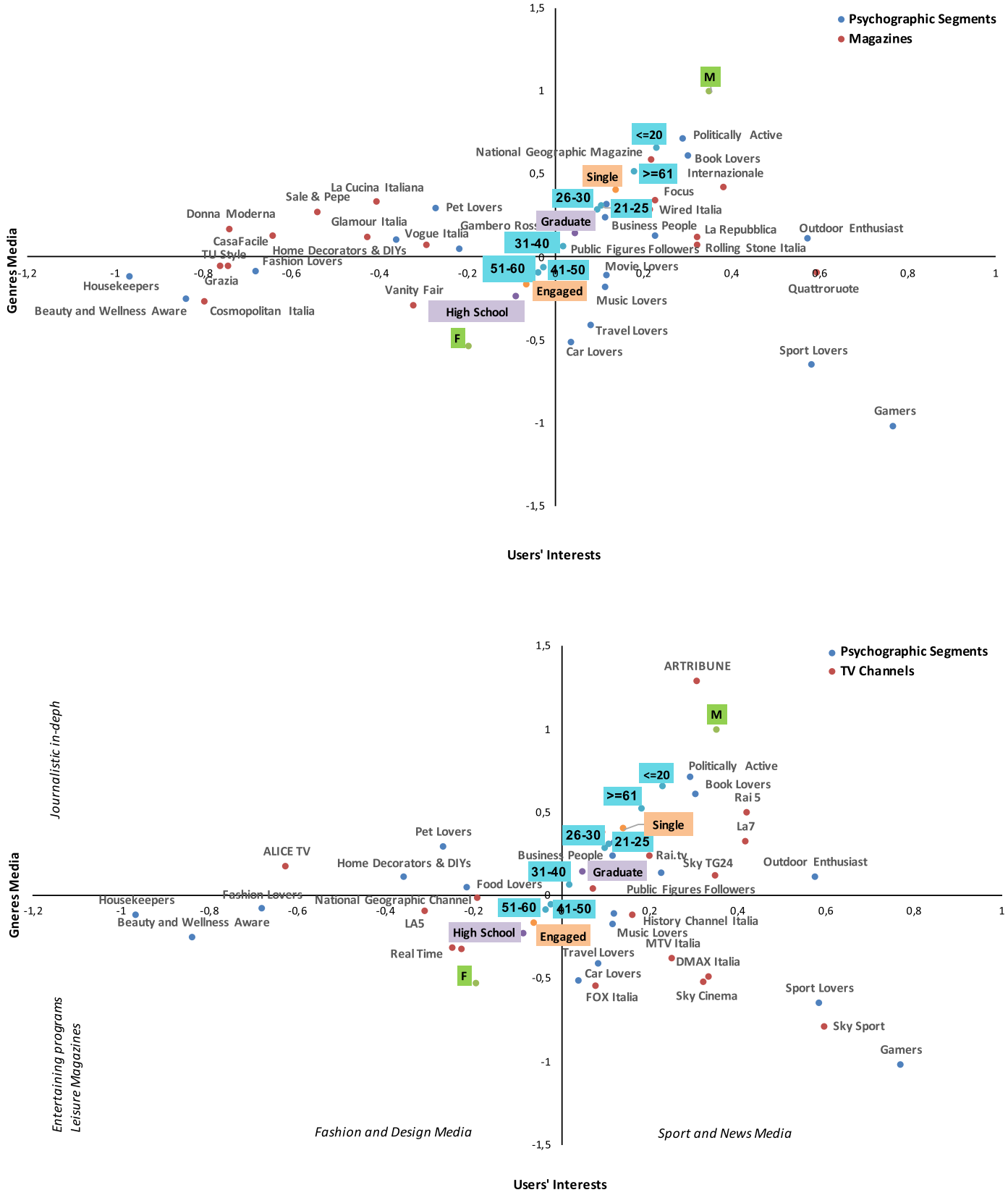

Fig. 4. Normalized 2-dimensional plots of Supplementary Variables with Psychographic Segments and Magazine (upper panel), with Psychographic Segments and TV channels (lower panel).

ing communication tools and channels and even in this case social media analytics could be useful.

6. Conclusions

This study has adopted a multivariate approach for predicting consumer decision-making styles pertaining to magazines and TV channels and based on digital footprints referred to *Likes* on Facebook. To the best of our knowledge, this is the first study that synthesizes Facebook data for marketing communication purposes. It represents an application of social media data pertaining to the Italian Facebook users that have expressed their appreciation to web pages of wellness websites through a Like button. The rating mechanism of *Likes* is considered as expression of affective responses of customers [20,21] thus findings based on customers' *Likes* can offer valuable insights for companies to develop marketing strategy.

Some limitations should be kept in mind when reviewing findings. Firstly, findings are not generalizable since data pertain to Italian Facebook users interested in pharmaceutical and health products. Although, the proposed approach allows comparisons of the findings with additional segments studied with respect of the same variables (media) or investigations about additional variables (external information) characterizing the existing profiles.

Future research could try to understand how to assess missing data in order to make them actionable information. Furthermore, promotional activities are important driving factors of success for companies pertaining to all industries and despite in our paper Facebook Likes refer to pharmaceutical products' webpages, a similar analysis could be replicated in other sectors such as for instance, food or consumer electronics. Otherwise, a similar analysis could be carried out by considering data collected through the social buttons of other social platforms such as Twitter or Instagram.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

A.1. Descriptive statistics

A.2. Correspondence analysis outline

Correspondence Analysis is an adaptation to categorical data of Principal Component Analysis (PCA), which is a method for identifying dimensions explaining maximum variance in metric data [43]. As demonstrated by [42], the geometric representation of a rectangular table that describes the similarities among the rows (and among the columns) has to be computed according to chi-square distance. Variability of a two-way frequency matrix, defined by chi-square metric, has to be represented in a continuous space generated by a vector decomposition that holds all the relative distances among rows (columns) [44]. The binary Correspondence Analysis provides such decomposition.

In a two-way contingency table as X_{red}, with R rows and C columns, we denote f_{ij} the joint relative frequency referred to the i-th row $(i = 1, \ldots, R)$ and associated to column j $(j = 1, \ldots, C)$ such that

$$\sum_{i=1}^{R}\sum_{j=1}^{C} f_{ij} = 1 \tag{2}$$

$$\sum_{i=1}^{R} f_{ij} = f_{\cdot j}$$

$$\sum_{j=1}^{C} f_{ij} = f_{i\cdot}$$

So, a general row i of X_{red} is considered as a point in the space $\mathbf{R}^R$ with coordinates $\{f_{ij}/f_{i\cdot}, j = 1, \ldots, C\}$ weights $\{f_{i\cdot}, i = 1, \ldots, R\}$ and centroid of the rows set is the point $\{f_{\cdot j}, j = 1, \ldots, C\}$ [49]. Then, the proximities between two rows points i, i' is measured using the chi-square distance:

$$d^2(i, i') = \sum_{j=1}^{C} \frac{1}{f_{\cdot j}}\left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}}\right)^2 \tag{3}$$

Consistently with the geometric approach, the dispersion of the set of rows (and symmetrically, of the set of columns) around its centroid is measured through the inertia:

$$\phi^2 = \sum_{i=1}^{R} f_{i\cdot} d^2(i, centroid) = \sum_{j=1}^{C} f_{\cdot j} d^2(j, centroid) = \frac{\chi^2}{n} \tag{4}$$

being n the total instances.

Table 2
Number of Likes per TV channel and Magazine.

Category	Magazine	Like	Missing	Category	Magazine	Like	Missing
Cultural	Internazionale	838	4847	News	La Repubblica XL	421	5264
	Focus	627	5058		Vogue Italia	707	4978
Entertainment	Grazia	333	5352	Thematic	Sale & Pepe	399	5286
	Cosmopolitan Italia	315	5370		Nat. Geographic Mag.	385	5300
	TU Style	268	5417		Wired Italia	318	5367
	Vanity Fair	266	5419		Gambero Rosso	299	5386
	Rolling Stone Italia	261	5424		La Cucina Italiana	294	5391
	Glamour Italia	195	5490		Quattroruote	292	5393
	Donna Moderna	629	5056		Casa Facile	288	5397
Category	TV Channel	Like	Missing	Category	TV Channel	Like	Missing
Cultural	Laeffe	353	5332	News	La7	428	5257
	History Channel Italia	350	5335		Tgcom24	867	4818
	Sky Arte	228	5457		SkyTG24	475	5210
	Focus TV	263	5422		Sky Sport	693	4992
Entertainment	Real Time	1469	4216	Thematic	ALICETV	361	5324
	DMAX Italia	603	5082		Sky Cinema	306	5379
	MTV Italia	946	4739		Nat. Geographic Ch.	622	5063
	RTL102.5TV	253	5432		Artribune	156	5529
	FOX Italia	557	5128				

CA describes the discrepancy of the observed system from the independence model, by displaying approximations between rows onto the axes of maximum dispersion (factorial axes). The factorial axes can be obtained by performing a PCA on the table $\widetilde{X}_{red}$, whose general term is:

$$\widetilde{x}_{ij} = \frac{f_{ij} - f_{i\cdot} \cdot f_{\cdot j}}{f_{i\cdot} \cdot f_{\cdot j}} \tag{5}$$

According to this framework, in the row space, the inertia of the s-th axis corresponds to the eigenvectors u_s, ($\|u_s\|_{D_C} = 1$) of the matrix $\widetilde{X}'_{red} D_R \widetilde{X}_{red} D_C$ associated with the eigenvalues λ_s (in decreasing order), where D_R is the diagonal matrix with general term $f_{i\cdot}$ and D_C is the diagonal matrix with general term $f_{\cdot j}$. In the column space, the inertia of the s-axis corresponds to the eigenvectors v_s, ($\|v_s\|_{D_R} = 1$) of the matrix $\widetilde{X}_{red} D_C \widetilde{X}'_{red} D_R$ associated with the same eigenvalues λ_s (in decreasing order).

Therefore, the vectors of the row scores are

$$r_s = \widetilde{X}_{red} D_C u_s = \sqrt{\lambda_s} v_s \tag{6}$$

and the vectors of column scores are

$$c_s = \widetilde{X}'_{red} D_R v_s = \sqrt{\lambda_s} u_s \tag{7}$$

These scores, or principal coordinates [43], lead to both sets of distances (those between rows and those between columns) corresponding to the distances defined in equation (3). It is possible to visualize in a reduced factor space both row (column) points simultaneously on the base of the base of the transition relationships that link the coordinates of one point in one space (the row-space for example) to those of all the points of the other space (the column-space in our example). Transition formulae (also known as quasi-barycentric coordinates), actually, allow the transition from the set of row coordinates to the set of column coordinates (and vice-versa):

$$r_s = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{C} \frac{f_{ij}}{f_{i\cdot}} c_s(j) \tag{8}$$

$$c_s = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{R} \frac{f_{ij}}{f_{\cdot j}} r_s(i) \tag{9}$$

where $c_s(j)$ indicates the coordinate of the category j on axis s, whereas $r_s(i)$ is the coordinate of the category i on the same axis.

A.3. Correspondence analysis output

Table 3
Summary of eigenvalues decomposition of X_{red} matrix.

Dimension	Eigenvalue	Inertia	Chi Square$_{gdl=648}$	Sig.	Proportion of inertia	
					Accounted for	Cumulative
1	0.150	0.023			0.408	0.408
2	0.128	0.016			0.296	0.703
3	0.071	0.005			0.091	0.794
4	0.053	0.003			0.051	0.845
5	0.046	0.002			0.039	0.883
6	0.039	0.002			0.028	0.911
7	0.030	0.001			0.017	0.928
8	0.029	0.001			0.015	0.943
9	0.028	0.001			0.014	0.957
10	0.024	0.001			0.010	0.968
11	0.021	0.000			0.008	0.975
12	0.018	0.000			0.006	0.981
13	0.017	0.000			0.005	0.986
14	0.016	0.000			0.004	0.991
15	0.015	0.000			0.004	0.994
16	0.012	0.000			0.003	0.997
17	0.009	0.000			0.002	0.999
18	0.009	0.000			0.001	1.000
Total		0.056	1423.263	0.000	1.000	1.000

Table 4
Overview column points.

Column	Mass	Score in dimension		Measures for interpretation			
		1	2	Relative contribution		Quality of representation	
				1	2	1	2
Internazionale	0.077	0.386	0.411	0.076	0.101	0.419	0.405
Vogue Italia	0.039	−0.289	0.067	0.022	0.001	0.431	0.019
Donna Moderna	0.046	−0.737	0.161	0.166	0.009	0.873	0.035
Focus	0.047	0.230	0.337	0.016	0.042	0.282	0.517
La Repubblica XL	0.024	0.326	0.110	0.017	0.002	0.299	0.029
Sale & Pepe	0.024	−0.538	0.263	0.047	0.013	0.669	0.136
National Geographic Magazine	0.023	0.221	0.580	0.007	0.059	0.114	0.667
Grazia	0.014	−0.757	−0.061	0.053	0.000	0.809	0.004
Cosmopolitan Italia	0.015	−0.794	−0.276	0.063	0.009	0.796	0.082
Wired Italia	0.025	0.217	0.276	0.008	0.015	0.242	0.331
Gambero Rosso	0.012	−0.075	0.116	0.000	0.001	0.019	0.039
La Cucina Italiana	0.015	−0.403	0.324	0.016	0.012	0.433	0.239
Quattroruote	0.011	0.596	−0.100	0.025	0.001	0.401	0.010
CasaFacile	0.019	−0.638	0.122	0.052	0.002	0.780	0.024

(*continued on next page*)

Table 4 (*continued*)

Column	Mass	Score in dimension		Measures for interpretation			
		1	2	Relative contribution		Quality of representation	
				1	2	1	2
TU Style	0.014	−0.741	−0.063	0.050	0.000	0.814	0.005
Vanity Fair	0.015	−0.319	−0.295	0.010	0.010	0.210	0.153
Rolling Stone Italia	0.013	0.325	0.068	0.009	0.000	0.407	0.015
Glamour Italia	0.010	−0.423	0.116	0.011	0.001	0.379	0.024
ARTRIBUNE	0.008	0.310	1.280	0.005	0.099	0.060	0.874
Real Time	0.116	−0.247	−0.323	0.047	0.095	0.326	0.475
MTV Italia	0.044	0.253	−0.387	0.019	0.051	0.248	0.493
Tgcom24	0.061	0.072	0.035	0.002	0.001	0.048	0.009
Sky Sport	0.042	0.597	−0.794	0.100	0.208	0.356	0.535
National Geographic Channel	0.031	−0.190	−0.026	0.007	0.000	0.175	0.003
DMAX Italia	0.039	0.334	−0.495	0.029	0.074	0.319	0.596
Rai. tv	0.039	0.200	0.228	0.010	0.016	0.147	0.163
FOX Italia	0.025	0.080	−0.549	0.001	0.058	0.016	0.634
Sky TG24	0.029	0.350	0.112	0.024	0.003	0.386	0.034
La7	0.021	0.418	0.320	0.024	0.017	0.471	0.235
ALICE TV	0.017	−0.624	0.169	0.043	0.004	0.662	0.042
History Channel Italia	0.013	0.162	−0.126	0.002	0.002	0.233	0.120
Laeffe	0.017	0.094	0.368	0.001	0.018	0.040	0.526
Sky Cinema	0.014	0.323	−0.527	0.010	0.031	0.223	0.503
Focus TV	0.014	0.222	0.387	0.004	0.016	0.201	0.516
Rai 5	0.010	0.422	0.492	0.012	0.019	0.331	0.383
RTL 102.5 TV	0.011	−0.225	−0.328	0.004	0.009	0.163	0.295
LA5	0.011	−0.308	−0.103	0.007	0.001	0.320	0.030
Active Total	1.000			1.000	1.000		

Table 5
Overview row points.

Row	Mass	Factor coordinates		Measures for interpretation			
		1	2	Relative contribution		Quality of representation	
				1	2	1	2
Pet Lovers	0.025	−0.267	0.288	0.012	0.016	0.243	0.241
Outdoor Enthusiast	0.017	0.577	0.106	0.038	0.002	0.357	0.010
Techies	0.052	0.120	0.311	0.005	0.039	0.072	0.410
Car Lovers	0.043	0.039	−0.520	0.000	0.091	0.004	0.675
Book Lovers	0.044	0.305	0.604	0.027	0.124	0.163	0.544
Social Activist	0.084	0.063	0.334	0.002	0.073	0.028	0.678
Gamers	0.013	0.771	−1.024	0.053	0.109	0.283	0.425
Movie Lovers	0.088	0.121	−0.116	0.009	0.009	0.122	0.096
Politically Active	0.047	0.294	0.707	0.027	0.183	0.130	0.641
Sport Lovers	0.060	0.586	−0.657	0.136	0.201	0.442	0.472
Fashion Lovers	0.055	−0.678	−0.090	0.169	0.004	0.781	0.012
Music Lovers	0.084	0.118	−0.184	0.008	0.022	0.115	0.235
Travel Lovers	0.049	0.085	−0.420	0.002	0.067	0.023	0.484
Public Figures Followers	0.092	0.229	0.125	0.032	0.011	0.535	0.136
Food Lovers	0.087	−0.214	0.040	0.026	0.001	0.616	0.018
Home Decorators & DIYs	0.039	−0.358	0.100	0.033	0.003	0.488	0.032
Beauty and Wellness Aware	0.043	−0.836	−0.261	0.199	0.023	0.844	0.070
Business People	0.045	0.117	0.229	0.004	0.018	0.069	0.228
Housekeepers	0.035	−0.965	−0.123	0.218	0.004	0.817	0.011
Active Total	1.000			1.000	1.000		

References

[1] Zephoria, Digital marketing, Available at https://zephoria.com, 2020.
[2] Censis, Cinquantesimo Rapporto sulla situazione sociale del Paese 2016, Franco Angeli, 2018.
[3] S. Shawky, K. Kubacki, T. Dietrich, S. Weaven, A dynamic framework for managing customer engagement on social media, J. Bus. Res. (2020).
[4] M.J. Valos, V.L. Maplestone, M.J. Polonsky, M. Ewing, Integrating social media within an integrated marketing communication decision-making framework, J. Mark. Manag. 33 (2017) 1522–1558.
[5] Y. Pan, I.M. Torres, M.A. Zúñga, Social media communications and marketing strategy: a taxonomical review of potential explanatory approaches, J. Internet Commer. 18 (2019) 73–90.
[6] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci. USA 110 (2013) 5802–5805.
[7] Y.-M. Li, C.-Y. Lai, L.-F. Lin, A diffusion planning mechanism for social marketing, Inf. Manag. 54 (2017) 638–650.
[8] A.M. Kaplan, M. Haenlein, Users of the world, unite! the challenges and opportunities of social media, Bus. Horiz. 53 (2010) 59–68.
[9] E. Arrigo, Deriving competitive intelligence from social media: microblog challenges and opportunities, Int. J. Online Mark. 6 (2016) 49–61.
[10] A. Geissinger, C. Laurell, C. Öberg, C. Sandström, N. Sick, Y. Suseno, Social media analytics for knowledge acquisition of market and non-market perceptions in the sharing economy, J. Knowl. Manag. (2020).
[11] R.M. Rojas, A. Garrido-Moreno, V.J. García-Morales, Fostering corporate entrepreneurship with the use of social media tools, J. Bus. Res. 112 (2020) 396–412.
[12] W. He, W. Zhang, X. Tian, R. Tao, V. Akula, Identifying customer knowledge on social media through data analytics, J. Enterp. Inf. Manag. 32 (2019) 152–169.
[13] W.W. Moe, D.A. Schweidel, Social Media Intelligence, Cambridge University Press, 2014.
[14] A. Arora, A. Srivastava, S. Bansal, Business competitive analysis using promoted post detection on social media, J. Retail. Consum. Serv. 54 (2020) 101941.
[15] S. Stieglitz, M. Mirbabaie, B. Ross, C. Neuberger, Social media analytics - challenges in topic discovery, data collection, and data preparation, Int. J. Inf. Manag. 39 (2018) 156–168.
[16] J. Choi, J. Yoon, J. Chung, B. Coh, J.M. Lee, Social media analytics and business intelligence research: a systematic review, Inf. Process. Manag. 57 (2020) 102279.

[17] K. Koroleva, G.C. Kane, Relational affordances of information processing on Facebook, Inf. Manag. 54 (2017) 560–572.
[18] C. Gerlitz, A. Helmond, The like economy: social buttons and the data-intensive web, New Media Soc. 15 (2013) 1348–1365.
[19] R. Hinson, H. Boateng, A. Renner, J.P. Basewe Kosiba, Antecedents and consequences of customer engagement on Facebook: an attachment theory perspective, J. Res. Interact. Mark. 13 (2019) 204–226.
[20] D.H. Kim, L. Spiller, M. Hettche, Analyzing media types and content orientations in Facebook for global brands, J. Res. Interact. Mark. 9 (2015) 4–30.
[21] T. Maree, G. van Heerden, Beyond the "like": customer engagement of brand fans on Facebook, Eur. Bus. Rev. (2020).
[22] C. Ding, H.K. Cheng, Y. Duan, Y. Jin, The power of the "like" button: the impact of social media on box office, Decis. Support Syst. 94 (2017) 77–84.
[23] K. Swani, G. Milne, B.P. Brown, Spreading the word through likes on Facebook: evaluating the message strategy effectiveness of fortune 500 companies, J. Res. Interact. Mark. 7 (2013) 269–294.
[24] I. Heimbach, O. Hinz, The impact of sharing mechanism design on content sharing in online social networks, Inf. Syst. Res. 29 (2018) 592–611.
[25] S. Banerjee, A. Chua, Identifying the antecedents of posts' popularity on Facebook fan pages, J. Brand Manag. 26 (2019) 621–633.
[26] P.J. Kitchen, I. Burgmann, Integrated marketing communication: making it work at a strategic level, J. Bus. Strategy 53 (2015) 34–39.
[27] R. Felix, P.A. Rauschnabel, C. Hinsch, Elements of strategic social media marketing: a holistic framework, J. Bus. Res. 70 (2017) 118–126.
[28] E. Pantano, C.V. Priporas, G. Migliano, Reshaping traditional marketing mix to include social media participation, Eur. Bus. Rev. 31 (2019) 162–178.
[29] E.M. Payne, J.W. Peltier, V.A. Barger, Omni-channel marketing, integrated marketing communications and consumer engagement, J. Res. Interact. Mark. 11 (2017) 185–197.
[30] C. Laurell, C. Sandström, Comparing coverage of disruptive change in social and traditional media: evidence from the sharing economy, Technol. Forecast. Soc. Change 129 (2018) 339–344.
[31] K. Mukherjee, Social media marketing and customers' passion for brands, Mark. Intell. Plann. 38 (2019) 509–522.
[32] R. Batra, K. Keller, Integrating marketing communications: new findings, new lessons, and new ideas, J. Mark. 80 (2016) 122–145.
[33] V. Kumar, J. Choi, M. Greene, Synergistic effects of social media and traditional marketing on brand sales: capturing the time-varying effects, J. Acad. Mark. Sci. 45 (2017) 268–288.
[34] P. De Pelsmacker, M. Geuens, J. Van den Bergh, Marketing Communications: A European Perspective, Pearson Education, 2017.
[35] J. Tyrawski, D.C. DeAndrea, Pharmaceutical companies and their drugs on social media: a content analysis of drug information on popular social media sites, J. Med. Internet Res. 17 (2015) e130.
[36] R. Caers, T. De Feyter, M. De Couck, T. Stough, C. Vigna, C. Du Bois, Facebook: a literature review, New Media Soc. 15 (2013) 982–1002.
[37] N.B. Ortiz Alvarado, M. Rodríguez Ontiveros, C. Quintanilla Domínguez, Exploring emotional well-being in Facebook as a driver of impulsive buying: a cross-cultural approach, J. Int. Consum. Mark. (2020) 1–16.
[38] J. Mellon, C. Prosser, Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users, Res. Polit. 4 (2017) 2053168017720008.
[39] H. Ditchfield, Behind the screen of Facebook: identity construction in the rehearsal stage of online interaction, New Media Soc. 22 (2020) 927–943.
[40] M. Brettel, J.C. Reich, J.M. Gavilanes, T.C. Flatten, What drives advertising success on Facebook? An advertising-effectiveness model, J. Advert. Res. 55 (2015) 162–175.
[41] D. Parlangeli, Cos'é cubeyou, l'azienda dal ceo italiano che è stata bloccata da Facebook, Available at https://www.wired.it/internet/social-network/2018/04/10/cose-cubeyou-facebook/. (Accessed 18 September 2020), 2018.
[42] J. Benzècri, Analyse des Donnèes, Dunod, Paris, 1973.
[43] M. Greenacre, Theory and Applications of Correspondence Analysis, Academic Press, New York, 1984.
[44] L. Lebart, A. Morineau, M. Piron, Statistique Exploratoire Multidimensionnelle, Dunod, Paris, 1997.
[45] Y. Chen, Y. Chen, Y.J. Hsu, J.H. Wu, Predicting consumers? Decision-making styles by analyzing digital footprints on Facebook, Int. J. Inf. Technol. Decis. Mak. 18 (2019) 601–627.
[46] M.W. Berry, Large-scale sparse singular value computations, Int. J. Supercomput. Appl. 6 (1992) 13–49.
[47] V. Weerakkody, Z. Irani, K. Kapoor, U. Sivarajah D. Y. K, Open data and its usability: an empirical view from the citizen?s perspective, Inf. Syst. Front. 19 (2017) 285–300.
[48] G. della Privacy, General data protection regulation, Available at https://www.garanteprivacy.it/regolamentoue, 2018.
[49] M. Bécue-Bertaut, J. Pagès, A principal axes method for comparing contingency tables: Mfact, Comput. Stat. Data Anal. 45 (2004) 481–503.