First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.2974

A Multivariate Approach to Facebook Data for Marketing Communication

Arrigo, Elisa[a]; Liberati, Caterina[a] and Mariani, Paolo[a]
[a] Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Italy.

Abstract
The aim of this paper is to propose a method to explore and synthesize social media data in order to aid businesses to make their communication decisions. The research was conducted at the end of 2014 on 5607 Italian Facebook subjects interested in drugs and health. In this study, we refer to the pharmaceutical market that is characterized by strict legal constraints, which prevent any promotional activities (such as advertising) of companies on prescription drugs. Thus, pharmaceutical businesses tend to promote their corporate brand instead of a single product brand. In such context, social media offer the opportunity to gather customers' information about their attitudes and preferences, helpful to address marketing activities. Through a multivariate statistical approach on Facebook data, we have highlighted the associations existing between TV channels and users' profiles. Therefore, depending on the value proposition to promote, every business could choose, first, the target group to reach and, then, the nearest suitable channel where to develop the corporate brand communication.

Keywords: *Social Media; Business Analytics; Marketing Communication; Facebook; Binary Correspondence Analysis; Pharmaceutical Industry.*

Elisa Arrigo (E.A.), Caterina Liberati (C.L.) and Paolo Mariani (P.M.) share the final responsibility for this paper, however E.A. wrote 1, 2, 5 sections; C.L., 1, 4, 5 sections and P.M wrote 1, 3, 5 sections.

1. Introduction

According to the 49th Censis Report, the percentage of Italian population who accessed a media at least once in 2015 was equal to 96.7% for television, 83.9% for radio, and 52.9% for newspapers. Likewise, Internet users continue to increase, by reaching a penetration rate of 70.9% of the Italian population and a percentage equal to 50.3% pertains to Italians who accessed Facebook, the most famous social network, at least once during the last year. In the light of such evidence, businesses should try to integrate successfully both traditional and social media in order to realize the best marketing communication strategy.

In this study we refer to the pharmaceutical market, that is characterized by strict legal constraints which prevent any promotional activities (such as advertising) of companies on prescription drugs. Thus, pharmaceutical businesses promote their corporate brand instead of the product brand. In such context, social media offer the opportunity to gather customers' information about their attitudes and preferences helpful to address marketing activities. The aim of this study is to propose a method to explore and synthesize social media data in order to aid businesses to make their communication decisions; and more precisely, to orientate their media choices into a marketing communication plan.

2. Theoretical Background

The 21st century has seen a big change in the media landscape due to the introduction of social media and the consequent increase in the variety of channels that businesses can employ to reach their customers. In fact, the digital environment has enlarged the set of communication media that businesses had used for 50-100 years: television, radio, newspapers, magazines and outdoor. Actually, traditional media are not disappeared from firms' media choices and, on the contrary, some of them are experiencing a new boom such as television and radio through the launch of satellite and other digital formats.

Social media have emerged as a new powerful marketing tool useful to achieve business purposes by providing businesses with new ways to communicate, collaborate and share content with customers (Lee, 2014). They encompass a wide range of tools and technologies and 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" (Kaplan & Haenlein, 2010, p. 61). Although few accepted classifications exist to distinguish them, there are several social media formats and platforms such as blogs, social networks, virtual social worlds, collaborative projects, content communities, virtual game worlds, etc. In particular, social networks are applications in which users create personal profiles accessible to others in the exchange of personal content and communication. Facebook has over one billion registered accounts

worldwide and among these many refer to companies that have created their own Facebook page where customers can sign up to become fans. Social media represent for businesses both an efficient channel to display ads, commercial and institutional communications and to collect data and information on customers' lifestyles, needs and problems encountered with existing products. Social media have empowered customers to publish opinions and spread new content online that is argued to be particularly valuable in business strategies for both marketing communication and intelligence purposes (Lee, 2014; Kietzmann *et al*., 2011). From a marketing communication perspective, the relevance of digital marketplace lies exactly in the interaction between consumers and the online community and in the immediate, interactive and low-cost communications. With a shift towards a multimedia environment, where traditional and digital media are available, the nature of marketing communication model has dramatically changed and businesses have lost the full control on their marketing communications. In the past, marketing communication was created by businesses and pushed towards the customers, which have a passive role of recipient; businesses could control the content, timing and frequency of the communication flow. Instead, nowadays, customers interact online with each other by taking part in conversations about brands and products and freely expressing their opinions and, in doing so, they alter the original marketing communication flow and become themselves a source of communication towards the online community (Fill, 2009). From a marketing intelligence perspective, market research analysts have recognized social media as an excellent base for tracking the behavior of customers. Generally, these latter create a virtual identity and, being aware of the fact that they are unidentifiable, say what they really think and produce information that is considered quite truthful. By continuously scanning social media, firms can acquire users' data at any moment, which greatly improves the availability of information about customer experiences (Bose, 2008) and allows monitoring the customer evolution over a specific period. In fact, digital technologies and access to social media only via login subscription enable businesses to collect enormous quantities of customer data with which to build customer web analytics (Zeng *et al*., 2010). Although social media appear to offer significant opportunities to businesses in terms of customer knowledge acquisition, since the most of social data is free, available in large quantities and in real time, the amount of these data is overwhelming and, consequently, the search of the desired information can be very complex and expensive to find. Then, in this study, we propose a method to reduce the complexity of analysis of social media data in order to orientate the media decisions of businesses within their marketing communication strategy.

3. The Data

The research was conducted at the end of 2014 on 5607 Italian Facebook subjects interested in drugs and health[1]. We collected all the possible interactions among people and brands, products and services (i.e. shares, likes, tweets, pins, posts, etc.). 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 20 servers active on Amazon Web Services (AWS) infrastructure. More than 5 Terabyte (distributed) database were gathered and updated daily via Hadoop2[2]. In our 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 *Pages Demographic* Table that stored unstructured data about Facebook pages (Fig. 1). In each table, records were extracted with queries based on users' keys and behavior.

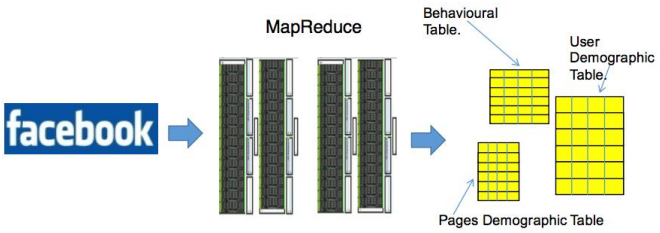

Figure 1. MapReduce Process Flow

Finally, the built matrix had size 5607 rows (Facebook users) and 140 dummy columns (pages visited/liked). In order to reduce the dimensionality of the data, we employed the subjects' classification of Kosinski *et al.* (2013), 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[3].

[1] The research was jointly conducted with Cubeyou, which is a company that delivers customer insights based on Social Media data. We monitored only websites of top pharmaceutical companies and Italian Public health istitutions.

[2] 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 (HDFS), 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.

[3] The Kosinski *et al.* (2013) approach transformed the unstructured data in meaningful customer information as personality traits, user location, hobbies and interests. After coding process, all the information is synthesized via a singular vector decomposition. The classification of the istances into a specific psychografic group is obatained by means of a logistic regression.

4. Empirical Analysis

Facebook dataset is generally considered a very complex example of unstructured information. Its volume and variety of contents makes impossible to deal with it without considering any manipulation for reducing the size of the data. Before considering any technique or model, the question that should be faced with is: "How do we elaborate this data set?" We run a pre-processing step to squeese the dimensions of the original matrix: we summed users and correspondent likes in order to obtain a contingency table of 19 psychographics profiles and 140 topics (i.e. Celebrites, TV shows, TV channels, Magazines, On-line sources). Due to the fact we focused our attention on the choice of TV channels for addressing a media campaign, we further reduced the number of the columns of the matrix to 20. In order to uncover and visualize the associations between the levels of a two-way contingency table we employed a Correspondance Analysis (Benzècri, 1973; Grenacre, 1984). The technique provides a geometric representation of the rows (psychographic-profiles) and columns (TV channels) as points in a low-dimensional space, according to the chi-square metric. In our case the hypothesis of independence between rows and columns is rejected (χ^2=613.679, p-value=0.000) and the first two factors, retained for our analysis, generate a principal factor (f_1, f_2) plane that explains 68.50% of the total inertia. For sake of brevity, in Tables 1-2 we show only the main statistics relative to those TV channels and profiles that crucially weight in characterizing the two axis. According to the contributions[4] displayed in Table 1, the first factor contrasts *Action type TV* (positive pole) vs *Generalist TV* (negative pole).

[4] Contribution (or absolute contribution) measures the proportion of variance provided by each element (row/column) in explaining a principal axis.

Table 1. Main statistics on selected columns.

Column	Mass	Coordinates		Contributions		Quality of representation	
		Axis 1	**Axis2**	**Axis 1**	**Axis 2**	**Axis 1**	**Axis2**
Real Time	0.204	0.130	0.446	0.028	0.346	0.078	0.877
Sky Sport	0.074	0.653	-0.501	0.259	0.159	0.552	0.310
DMAX Italia	0.068	0.386	-0.203	0.082	0.024	0.659	0.175
Rai.tv	0.068	-0.450	-0.266	0.112	0.041	0.540	0.180
Sky TG24	0.051	-0.302	-0.457	0.038	0.092	0.246	0.537
ALICE TV	0.029	-0.501	0.620	0.060	0.096	0.278	0.406
Laeffe	0.030	-0.490	-0.042	0.058	0.000	0.557	0.004

Table 2. Main statistics on selected rows.

Row	Mass	Coordinates		Contribution		Quality of representation	
		Axis 1	**Axis2**	**Axis 1**	**Axis 2**	**Axis 1**	**Axis2**
Techies	0.050	-0.451	-0.310	0.084	0.041	0.539	0.243
Gamers	0.016	0.879	-0.807	0.104	0.092	0.456	0.367
Politically Active	0.043	-0.740	-0.500	0.194	0.093	0.588	0.256
Sport Lovers	0.069	0.637	-0.437	0.229	0.113	0.631	0.284
Fashion Lovers	0.050	0.146	0.543	0.009	0.125	0.059	0.773
Beauty and Wellness Aware	0.043	-0.036	0.650	0.000	0.154	0.003	0.813
Housekeepers	0.034	-0.125	0.810	0.004	0.190	0.018	0.731

In particular, the variance of f_1 is explained for 34.10% by Sky Sport (0.259) and DMAX Italia (0.082), whereas Rai.tv (0.11.2) and Laeffe (0.052) weight 16.40%. Also the quality of representation[5] of the points that provides additional richness to the interpretation of the relationships in the contingency table, confirm such picture (Tab. 1). The psychographic profiles associated to the factor are coherent with the description provided: Sport Lovers and Gamers show positive coordinates and relevant contributions with the principal axis 1 (Tab.2), instead Politically Active and Techies are located and associated with the opposite side of the same factor. The second axis discloses differences between *Entertaining TV* (positive pole) and *Newcast* (negative pole). It is reasonably clear from the inspection of the contributions that Real Time (0.346), ALICE TV (0.096) on one side, and Sky TG24 (0.092) on the other side, are the relevant elements for interpreting the factor f_2 (Tab.1). The reading of the axis is also upholded by the quality of representation relative to those channels that are always greater than 40%. It is interesting to highlight a coherent association of those results with the psychograpic profiles Fashion Lovers, Beauty and Wellness Aware and Housekeepers. A complete representation of the associations is visualized in Figure 2.

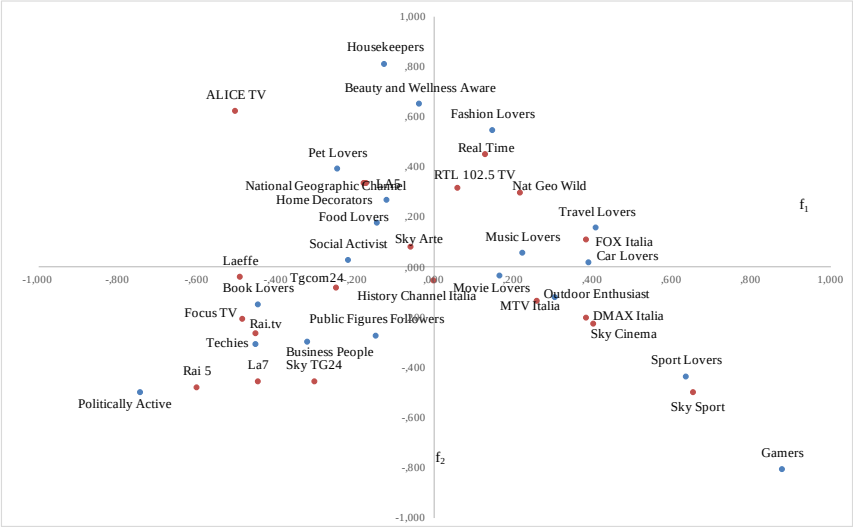

Figure 2. Normalized 2-dimensional plot of psychographic profiles and TV channels

[5] The quality of representation (or relative contribution) measures the proportion variance provided by each principal axis in explaining a single point.

The interpretation of the map is straightforward: the proximities among channels and profiles indicate high similarities (associations). For example, in the first quadrant of the map (where f_1 and f_2 are both positive), we find RTL 102.5 TV, Real Time and Fox Italia characterized by an emotional hedonistic broadcast style. Such result is a useful suggestion for a business that has to select a suitable channel for reaching target audience as Fashion Lovers, Travel Lovers, Music Lovers, Car Lovers.

5. Conclusion

As before illustrated, the four quadrants allow to highlight the associations between channels and users' profiles. Depending on the product, every pharmaceutical business can choose the target group and consequently the nearest suitable channel where to deliver the corporate brand communication. Together with the discussed provided opportunities, it is important to point out limitations of Facebook data: differently from survey data, that are a collection of conscious customers answers, pages without likes are not necessary a users' aware choice. These situations could occur both for lack of users' visits or for a reasoned decision of them. In our study it is likely that the users were exposed to all the TV channels (due to their high popularity). In the light of such considerations, the hope is that the usage of social media, always growing, will be done through a rigorous research design that illustrates gains (and limitations) of the results. Having more data does not necessarily mean having more information, since the knowledge extraction process is not only an automatic computational syntesis. The future research could focus on analyzing other media as magazines or celebrities and other statistical exploration.

References

Benzècri, J. (1973). *Analyse des Donnèes*, Dunod, Paris.

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. *Industrial Management & Data Systems*, 108(4), 510–528.

Censis (2015). 49 Rapporto sulla Situazione Sociale del Paese. Milano: FrancoAngeli.

Fill, C. (2009). *Marketing communications. Interactivity, communities and content*. Fifth Ed., London: Prentice Hall.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, New York.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68.

Kietzmann, J. H., Hermkens, K., & McCarthy, I. P. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241-251.

Kosinski, M., Stillwell D. & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *Proceedings of the National Academy of Sciences*, 110, 5802-5805.

Lee, I. (Ed.) (2014). *Integrating Social Media into Business Practice, Applications, Management, and Models*. IGI Global, Hershey, PA: Business Science Reference, USA.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *Intelligent Systems*, IEEE, 25(6), 13-16.