Computational Statistics
https://doi.org/10.1007/s00180-022-01261-0

ORIGINAL PAPER

Missing values and data enrichment: an application to social media liking

Paolo Mariani[1] · **Andrea Marletta**[1] · **Matteo Locci**[1]

Received: 10 February 2022 / Accepted: 9 July 2022

Abstract
In the big data context, it is very frequent to manage the analysis of missing values. This is especially relevant in the field of statistical analysis, where this represents a thorny issue. This study proposes a strategy for data enrichment in presence of sparse matrices. The research objective consists in the evaluation of a possible distinction of behaviour among observations in sparse matrices with missing data. After selecting among the multiple imputation methods, an innovative technique will be presented to impute missing observations as a negative position or a neutral opinion. This method has been applied to a dataset measuring the interaction between users and social network pages for some Italian newspapers.

Keywords Missing values · Data enrichment · Multiple imputations · Social network data

1 Introduction

The treatment of missing values is still a neglected phase in the field of quantitative analysis. In not statistical contexts, row elimination is the most abused solution. This method consists in the deletion of each observation containing missing values, but it may be misleading (Acock 2005). Appropriate missing data processing is more complex than row elimination. Firstly, some preliminary analysis are needed to understand and recognise the nature and the mechanism underlying the lack of information. This relationship aims to evaluate the link between the observed value

✉ Andrea Marletta
andrea.marletta@unimib.it

Paolo Mariani
paolo.mariani@unimib.it

Matteo Locci
m.locci2@campus.unimib.it

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, Italy

Published online: 09 August 2022

and the missing one. This leads to the well-known classification of Little and Rubin, which differentiates MCAR (Missing Completely At Random) data, MAR (Missing at Random) data and MNAR (Missing Not At Random) (Little 1988). Only after the identification of these mechanisms, the best solution to solve the missing values problem may be sought. If the complete case analysis has not been considered as a valid alternative, it is necessary to proceed with the imputation of the missing observation. In this study, the proposed imputation methods could be considered as a data enrichment technique because the initial dataset will be enriched based on decisional rules using a data-driven approach.

Another important step in the treatment of missing values, is to consider the starting structure of data. In this study, this technical issue is strictly related to big data and in particular to social media data. For this category of information, data can assume a different configuration: for example they could be represented as a network using graph analysis. But in this study, social media data are considered as a rectangular matrix in which each row has been represented by a user and each column the expression of an interaction in the network.

In this study, an innovative analysis path to discern the missing value from a behaviour for some individuals has been proposed using several statements. The term behaviour is here intended as synonym of motivation leading the individual to make an action. Firstly, it was hypothesised the presence of a behaviour behind the missing observation. Such hypothesis could represent a tentative to find a possible solution to the Missing Denominator problem, as described in Tufekci (2014). According to the author, it is important to consider also the behaviour of users who see a social network content without reacting to it. Secondly, a two-step procedure was implemented to impute the missing values. In the first step, the substitution of missing values was implemented using a threshold based on the number of 1 values in the sparse matrix, where 1 stands for the presence of an appreciation. In the application, a 0 value is considered as expression of a negative opinion only when for that row the percentage of 1 values is higher than the selected threshold. Alternatively, if the percentage of 1 values is lower than the threshold, a missing observation is imputed as a neutral behaviour. The second step was pursued using a multiple imputation technique known as the MIMCA method (Multiple Imputation with Multiple Correspondence Analysis) (Audigier et al. 2017). This procedure was applied to social media data from the official pages of 7 Italian newspapers.

This paper is organised as follows. Sect. 2 briefly describes the rising interest in big and social data. In Sect. 3, the procedure to discern between negative and neutral opinion through multiple imputation is presented. In Sect. 4, an application of the procedure to impute missing "Likes" for social media pages on Italian newspapers is shown. Sect. 5 concludes this work.

2 Social media and big data

The capability to read and transform a great amount of data into useful information for the business is pivotal for the management in terms of decision making. In this context, the possibility to treat big data represents an extremely important

issue. Effectively, the demand for specialised figures related to big data is constantly rising. The advent of big data led to a revolution in many fields. The use of big data observed a substantial increase at the business level; the advisory companies offer the transformation of data in available knowledge for the customers. For example, some authors used online data as a new source of information for default prediction (Crosato et al. 2021).

Many contributions have been presented during the last years to define the big data using the 3V model (Volume, Velocity and Variety), and the number of Vs successively increased to 7V models (adding to the 3V model Value, Veracity, Validity and Visualisation) (Liberati and Mariani 2018). These features show that big data have principally to contain a great amount of statistical observations subject to continuous updates without losing the properties of validity and efficiency. Social media seem to be the typical category of data where these characteristics are present. Moreover, in this field to measure the possibility to interact among users is one of the most interesting aims to pursue (Angelone 2021). Since this study is focused on missing values, social media data were taken into account because they are often a source of incompleteness (Kossinets 2006). This implies a loss of information that may increase the risk of bias in statistical results. One of the main sources of incompleteness is a non-response (Stork and Richards 1992; Robins et al. 2004). Two different kinds of non-responses can be distinguished in the treatment of missing data: unit non-response where the vector of responses for a respondent is fully missing, and item non-response where only some vector elements are missing (Huisman 2009).

Big data analysis and social media mining may be challenging. The main issues are related to the quality of data collected and to the integration of multiple datasets. The quality of information generated from big data is dependent on the quality of data collected and the robustness of the measures or indicators used. In particular, social media data often present biased information, especially in relation to opinions and sentiments about specific products and services. Indeed, online reviews generally include overly negative comments and feedback, since users tend to feel freer to express their dissatisfaction online, rather than in other contexts (Dalla Valle and Kenett 2018). Social data represent a source of uncertainty because of their collection method, besides robust statistical techniques are necessary to reduce missing data uncertainty. For this reason, it is not so easy to understand the mechanism of generation data and apply the standard classification of Little and Rubin (Little 1988). Since it is not present the option to express a negative opinion about a social media page, if the missing value is represented by the lack of appreciation for a content, this missingness could be interpreted as a neutral or a negative opinion. A possible solution could be represented by comparing the distribution of "Likes" and its complementary. The hypothesis is that the behaviour of the complementary observations is similar to those of real observations (Mariani et al. 2020). Alternatively, the absence of a "Like" can be measured on the basis of a placed "Like" for the same user for a similar social page (Mariani et al. 2019). This last hypothesis is the starting point for the approach proposed in the next section.

3 Methodological content

Since the proposed approach could be seen as a path of data enrichment, a general definition of this concept was reported. Moreover, the usual methodologies to discern a missing value from a behaviour was presented. A rule based on a threshold value was proposed. This technique yields satisfactory results, but a second imputation for the remaining observations is necessary. For this reason, the MIMCA approach was shown as a solution for filling in the missing observations. The application of this method led to an informative improvement measurable through a specific indicator.

3.1 Data Enrichment

The term data enrichment is usually adopted to define all the processes using different sources to validate and integrate information and raw data in a business database (Touya 2010). For example, the record linkage technique could be intended as a method to enrich data, and because of information coming from external sources, the analyst can acquire new values to fill the starting table (Fellegi and Sunter 1969). In this context, the aim is not to integrate the data with new information but to use the initial database to enrich data, assuming that the missing observations could be the result of a behaviour.

The social network scope is a field where this situation is commonplace due to sparse matrices made of 0 and 1 (Mariani et al. 2019). Therefore, if each zero could be intended as a missing observation, the application of a statistical approach based on the imputation of these values could be represented as an application of data enrichment.

The proposed approach to fill up the missing observations could be divided into a two-step procedure: in the first step, the choice of a threshold value based on starting data will give the possibility to impute some values. This rule is not effective for all the observations: if the data-driven approach result is close to the threshold value, there are no conditions to assign a correct imputation for some of the missing values. For the remaining missing observations, the second step of the procedure will be applied. which consists in using a MIMCA approach (Josse et al. 2012). The use of the two-step procedure makes it possible to consider a missing value from an imputed observation as the expression of a behaviour. This will lead to a construction of an index of data enrichment intended as the informative advantage for the final version of the database.

3.2 How to discern between missing values and users' behaviour

Before presenting the innovative approach, it could be useful to discuss some of the most common tools used in the field of missing values. The statistical techniques about missing observations may be divided into two macro-categories: the complete-case approach and the imputation approach. On the one hand, the complete-case

approach consists in the deletion of all records whenever missing values are present. Even if this represents the simplest way to manage missing observations, the risk of losing a significant amount of data is exceedingly high. On the other hand, imputation techniques use the record with missing values after the substitution of them with plausible values (Baraldi and Enders 2010; Donders et al. 2006; Van Buuren 2018).

In modern procedures for missing data, this phenomenon is considered from a probabilistic point of view. It is possible to treat N as a set of random variables with a joint probability distribution without specifying a particular distribution. In the statistical literature, the distribution for N is called the missing mechanism and is considered as a mathematical system that is useful to describe the schemes of missing values and to catch the relationships between the lack of information and the values of missing objects (Little 1988). In the social network context, on the basis of the classification already cited, data are defined as MCAR if the missing value is correlated with the value of the mechanism and not to the observed data. In this case, there are no systematic differences between the missing and observed values, and the missing data is a random subset of the original data set. The assumption that the generating mechanism of missing values is considered to be MAR implies that the missing responses are predictable based on the observed data for the other variables. Alternatively, when missing data are MNAR, information in the dataset is not sufficient to predict the unobserved values. The MNAR data are not ignorable because the mechanism of the missing data must be modelled. Thus, it is necessary to include a model to explain why the data are missing and which are imputed values (Kossinets 2006; Sharma et al. 2016).

Using Rubin's definitions, it is possible to describe the relationship between the data and the generating mechanism but not the causal relationships. If the data are MCAR, the root causes of the missing values are contained in the independent component Y. If the data are MAR, some causes could be correlated with X. Finally, if the data are MNAR, some causes are residual in the relationship with Y beyond those considered in the link between X and Y (Schafer and Graham 2002).

Statistical inference based on missing data usually involves some assumptions about the mechanism of the missing values. The validity of these hypotheses requires a preliminary evaluation. For example, inference based on likelihood is valid only if the missing value mechanism is ignorable based on the assumption of MAR data. Therefore, the MAR test is not generally performed as it requires unavailable information for the missing values. On the other hand, as the assumption on the MCAR data supposes missing values independent of the observed and unobserved data, it can be tested using only the observed data. The MCAR test is widely used in real data applications because simple methods to account for missing values are valid only for MCAR data (Little 1988).

In the Little test for MCAR, let y_i (with $i = 1, \ldots, n$) be modelled as the n-dimensional normal variables with the same average μ and covariance matrix Σ in which the components of y_i are missing. When the normality assumption is not valid, the Little test operates asymptotically for the random quantitative vectors y_i and not for categorical variables. This test determines if there is a significant distinction between the different kinds of missing values.

The Little test has two important limits. Firstly, this procedure is applicable even if data are not distributed as Multivariate Normal. Thanks to the Central Limit Theorem, it is possible to obtain an asymptotic distribution of the test statistic based on a Chi-squared with $\sum_{g=1}^{G} p_g - p$ degrees of freedom, where G is the number of hypothesised pattern of missing data. Nevertheless, it is necessary that these data are quantitative, so Little's test is not adapted to categorical or binary data. The second limit regards the scarce power of this test for MNAR data. For these reasons, in this study, the Lin and Bentler was used to verify the mechanism of missing values (Lin and Bentler 2012).

Let Y be the starting data matrix composed by K column vectors $Y_1, \ldots, Y_K$, where each element Y_j is decomposable as $\{Y_{\text{obs},j}, Y_{\text{mis},j}\}$ and distributed as $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. Let M be the indicator matrix for missing values $M_1, \ldots, M_K$, where $M_j \sim Bernoulli(\psi_j)$. The parameters$\psi_j$ represent the probabilities of missing values in the population for the correspondent variables Y_j. If data are MCAR, the joint density of M in the population could be modelled as:

$$f(M|\psi, Y) = f(M_1, \ldots, M_K|\psi, Y) = f(M_1, \ldots, M_K|\psi) = \prod_{j=1}^{K} \psi_j,$$

where K is the number of variables and ψ is the vector of probabilities of missing values in the population.

Since the distribution of variable of columns of matrices Y and M is unknown, two independent samples are defined as $(y_{1j}, \ldots, y_{nj}) \sim Y_j$ and $(m_{1j}, \ldots, m_{nj}) \sim M_j$ where the element m_{ij} is defined as follows:

$$m_{ij} = \begin{cases} 1, \text{ if the value } y_{ij} \text{ is missing} \\ 0, \text{ else} \end{cases}$$

The probability of missing data for each variable Y_j in the population is estimable as:

$$\hat{\psi}_j = \frac{1}{n} \sum_{i=1}^{n} m_{ij}, \tag{1}$$

where n is the dimension size.

Let $\psi_o^{(g)}$ be the observed probability of missing data in the population and $\psi_e^{(g)}$ the expected probability of missing data in the population. The hypothesis test is:

$$\begin{cases} H_0 : \psi_o^{(g)} = \psi_e^{(g)} \\ H_1 : \psi_o^{(g)} \neq \psi_e^{(g)}, \end{cases} \tag{2}$$

for $g = 1, \ldots, G$. Let p_o the vector of the observed probabilities $(p_o^{(1)}, \ldots, p_o^{(G)})^T$ and p_e the vector of the expected probabilities $(p_e^{(1)}, \ldots, p_e^{(G)})^T$. The test statistic to evaluate the null hypothesis is given by:

$$X^2 = n(p_o - p_e)^T D^{-1}(p_o - p_e),$$

with $G-K$ degrees of freedom and where $D = diag(p_e^{(1)}, \ldots, p_e^{(G)})$. If the null hypothesis is rejected, then data are not MCAR.

The Lin and Bentler test and the Little test are quite similar because the form of the test statistic is the same. The only difference is that the Lin and Bentler test is based on the estimate of the probability assuming the missing value mechanism directly. Both tests are based on the hypothesis of the Normal Multivariate distribution for the data, but the Lin and Bentler test could be also applied to qualitative data. Finally, it is important to note that the Lin and Bentler test is more powerful than the Little test in presence of MNAR data (Lin and Bentler 2012).

3.3 The choice of a threshold value

In a sparse matrix with 0 and 1 where 0 values represent a missing observation and 1 values an appreciation for a variable, the threshold value method is based on a precise assumption: The higher the occurrence of 1 values, the lower the probability that 0 values may be imputable as missing observation.

The first step of this technique consists in computing the total number of 1 values for each considered variable. Secondly, this sum is divided for the total number of variables in order to obtain for each statistical unit in the dataset the proportion of variables with 1 values. This proportion has been defined as l_i, for $i = 1, \ldots, n$, and $l_i \in (0, 1]$. Specifically, if $l_i = 1$ then the i-th observation has 1 values for each variable. On the other hand, if $l_i \approx 0$, then the i-th observation has 0 values except for one variable. For the detailed study, all the observations with $l_i = 0$ have been deleted from the analysis.

The choice of the threshold for the disambiguation is determined from the average value of the proportion of 1 values for each row. Formally, let c the threshold, it could be defined as:

$$c = \frac{1}{n} \sum_{i=1}^{n} l_i.$$

The disambiguation step is based on a plausible hypothesis. If the proportion of 1 values is higher than the threshold value, it is reasonable to hypothesise that the 0 values are not the result of a missing observation. Indeed, they could express a voluntary lack of positive opinion. On the contrary, if the proportion of 1 values is lower than the threshold value, then it is possible to hypothesise that the 0 values are missing observation due to a neutral opinion for that variable.

In particular, the decisional rule for the disambiguation is:

- if $l_i > c$, it is supposed that the i-th observation is interested in the content of the variables. Operationally, the 0 values are imputed as an expression of a negative opinion of the content of the variables;
- if $l_i < c$, it is supposed that the i-th observation is not interested in the content of the variables. Therefore, the 0 values are not imputed as an expression of a negative opinion of the content of the variables but they are still missing values;

- if $l_i \approx c$, further information is necessary to investigate the behaviour of the i-th observation and additional control is needed to solve the uncertainty. For the moment, the 0 values are still not imputed.

3.4 Multiple imputation with multiple correspondence analysis

The first step of "data enrichment" about missing values realized using the threshold method produces an evident information gain. However, the analysis could be not considered completely satisfactory. Many missing values are still present since the behaviour of individuals with a proportion l_i very close to the threshold has not been specified. In order to further enrich the data, an available alternative consists in introducing a second imputation technique. In particular, whether the variables considered are qualitative, a possibility is represented by Multiple Imputation with Multiple Correspondence Analysis (or MIMCA).

Multiple imputation using MCA allows the user to impute data sets with incomplete categorical variables. The principle of MI with MCA, as well as all the other multiple imputation techniques, consists in creating M different data sets to reflect the uncertainty on imputed values. In the context of MI with MCA, each of these data sets is obtained with an algorithm called *iterative MCA*, which is useful to impute qualitative data. In few words, the iterative MCA algorithm consists in recoding the incomplete data set as an incomplete disjunctive table Z, randomly imputing the missing values, estimating the principal components and loadings from the completed matrix and then, using these estimates to impute missing values according to the following reconstruction formula:

$$\hat{Z} = \hat{U}\hat{\Lambda}\hat{V}^T + M.$$

Let consider the left singular vectors $\hat{U}$, the diagonal matrix of singular values $\hat{\Lambda}$ and the right singular vectors $\hat{V}$. The matrix Z has a final version obtained as $Z = W * Z + (\mathbb{I} - W) * \hat{Z}$, where $*$ is the Hadamard product and W is a matrix of weights where $w_{ij} = 1$ if z_{ij} is missing and $w_{ij} = 0$ else. In this context, it should be underlined that MCA is configured as a singular values decomposition applied on the triplet of matrices $(Z - M, \frac{1}{K} D_\Sigma^{-1}, R)$. The matrix Z represents the disjunctive table, M is a matrix whose rows are equal to the vectors of the means of each component of Z, D_Σ is a diagonal matrix with the proportions of individuals characterised by a specific category and R is the matrix of uniform weights assigned to individuals. After the first step of imputation, the procedure of iterative MCA is repeated many times until a convergence criterion is reached. In many cases, due to overfitting problems, a regularized version of this algorithm is used (Josse et al. 2012).

Multiple imputation with Multiple Correspondence Analysis is based on regularised iterative MCA. In order to consider the uncertainty concerning the imputed values, M data sets are created. In this regard, there are two classical approaches: the Bayesian approach and the bootstrap approach. MI with MCA is based on a bootstrap approach. In few words, the algorithm of MI with MCA consists in drawing randomly M sets of weights for individuals sampling

by replacement from the original data set. Then, M single imputations are performed: at first, the regularised iterative MCA algorithm is used to impute the incomplete disjunctive table using the previous weighting for the individuals. Next, coin flipping is used to obtain categorical data and mimic its distribution (Audigier et al. 2017).

First of all, MI with MCA is part of the family of joint modelling MI method, which means that it is more computationally efficient than conditional models. In fact, this MI technique is based on Multiple Correspondence Analysis and then the number of parameters estimated is small. Another advantage of MI with MCA is the goodness of estimation even if the number of individuals is small (this behaviour is directly linked to the small number of parameters to be estimated). Lastly, MI with MCA well represents less frequent categories in the step of imputation. This last is another property that derives from Multiple Correspondence Analysis.

The number of components is chosen with a repeated cross-validation. Cross validation consists in searching the number of dimensions S minimising the prediction error. In other terms, missing values are added at random to the data set. Then, missing values of the incomplete disjunctive table Z are predicted using regularized iterative MCA. Lastly, the mean squared error is calculated. This procedure is repeated k times for each number of components considered. The optimal value of S corresponds to the minimum mean of MSEs.

Once the value of S components is obtained, the MIMCA approach can be applied. As for the threshold method, the aim is to impute, if possible, the missing values as a negative or neutral position.

In the last step of the MIMCA approach, it is necessary to impute M datasets and for each missing cell, datasets with only an imputation are considered. Furthermore, in order to make not influential the bootstrap simulation error, a threshold d is assumed to define the final decision rule:

- if the proportion of the imputation as negative opinion is higher than $50\% + d$, then the missing value is imputed as an expression of negative opinion;
- if the proportion of the imputation as negative opinion is lower than $50\% - d$, then the missing value is imputed as an expression of neutral behaviour;
- if the proportion of the imputation as negative opinion is between $50\% - d$ and $50\% + d$, then the missing value is not imputed.

The last phase of the new procedure provides the introduction of an index for the measurement of the advantage obtained by the two-step imputation. Once realized the disambiguation of the missing data, the Infomative Earning (IE) index has been obtained through the following formula:

$$IE_i = \frac{IV_i}{(IV_i + NIV_i)} * 100$$

where IV is the number of imputed values and NIV is the number of not imputed values. The index measures in percentage the negative opinion respect to the total of the missing values before the imputation procedure.

The full procedure of Data Enrichment has been represented using a flow in Fig. 1.

4 Application on italian newspaper social pages

4.1 Social media and liking

In the context of social network, two forms of interaction are possible. The first consists in an active presence using some specific tools as a "Like", the post sharing or the comments. The second one is a passive interaction based on the vision of a content, the click or a visualisation of a post (Ekström and Östman 2015). Usually, the attention is focused on the first form of interaction, in particular with the behaviour related to the appreciation. Placing a "Like", users show approval for a content stating a preference and showing positive feedback (Sumner et al. 2018). The second interaction form suggests the potential presence of other conducts measurable on the basis of definite hypotheses. The active presence on a social network opens to the possibility that the user knows other social pages on similar themes hypothesising a passive presence (Jiang et al. 2013).

Several articles have investigated the impact of social network in the modern society: some are focused on users' characteristics (Arrigo et al. 2021; Caers et al. 2013; Mellon and Prosser 2017; Ortiz Alvarado et al. 2020) others on the role of the platform in the social interactions (Ditchfield 2020). The growing interest in social network to build a potential segmentation and infer personality traits from users' behaviour remains of constant relevance (Kosinski et al. 2013). Among the most suitable data for analysing users' activities, "Likes" represent a quantitative alternative to any other ways to express a reaction to a content (Brettel et al. 2015).

In order to analyze the role of the "Likes" in this context, a dataset has been considerer containing information forby users that expressed at least one "Like" for a set of social media pages, websites, and forums concerning healthcare. The research was conducted on 1, 000 Italian subjects considering all interactions between people and brands and between products and services on social network[1]. The application is about social network pages with regard to Italian newspapers. Each variable of the dataset is a binary variable $(0-1)$ that represents the presence or absence of a "Like." The 7 Italian newspapers[2] are:

- La Repubblica
- Corriere della Sera
- Il Fatto Quotidiano

[1] The research was conducted at the end of 2014 on Italian social network users interested in pharmaceutical products and health. Cubeyou collected the interactions (i.e. likes) assigned by people to pages of pharmaceutical companies or institution related to health and wellness.

[2] La Repubblica, Corriere della Sera, Il Fatto Quotidiano, Il Messaggero e La Stampa are general-interest newspapers. Il Sole 24 Ore is a financial and economic newspaper. La Gazzetta dello Sport is a sport newspaper.

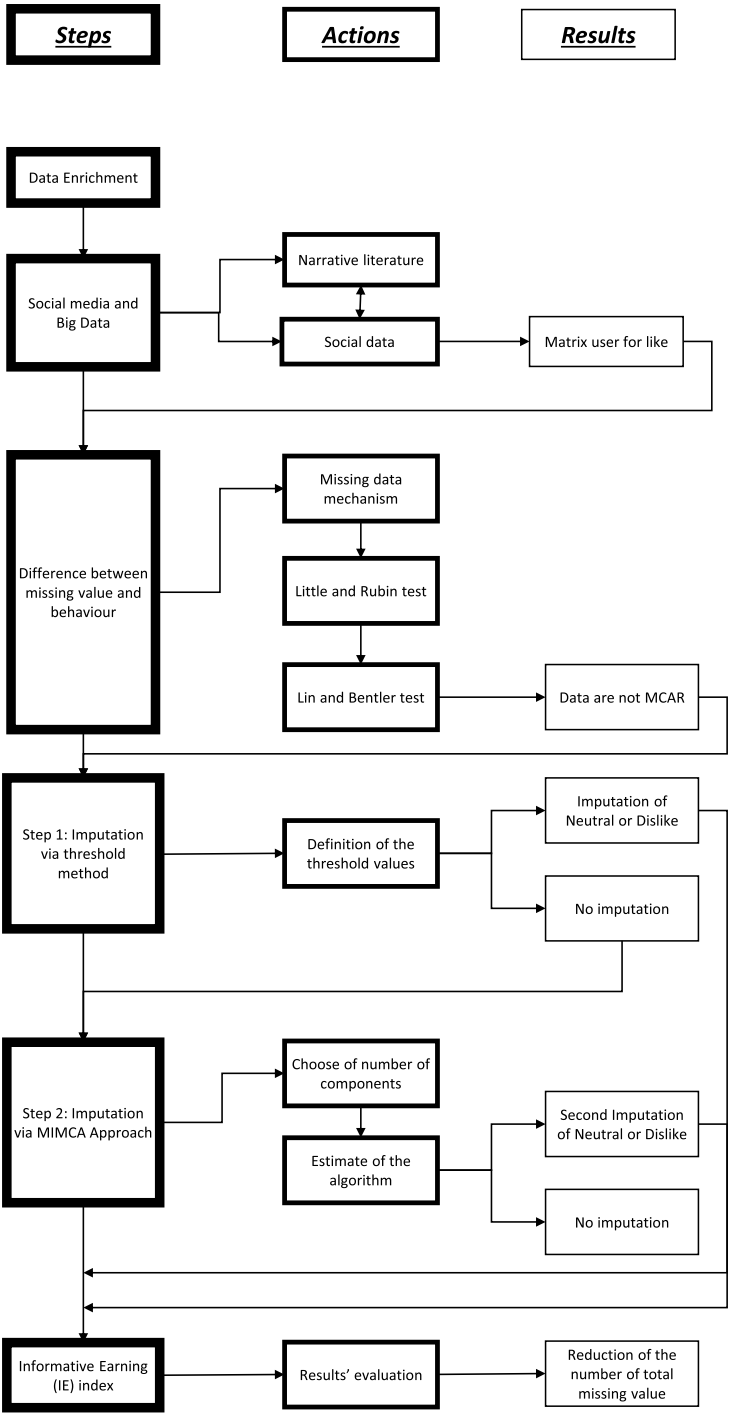

Fig. 1 Flow of the entire procedure of Data Enrichment

Table 1 Head of the dataset

	la Repubblica	Corriere della Sera	Il Fatto Quotidiano	Il Sole 24 Ore	La Gazzetta dello Sport	Il Messaggero	La Stampa
1	1	1	1	0	0	0	0
2	1	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0

Table 2 "Like" distribution for user

Number of "Like"	0	1	2	3	4	5	6	7
Frequency	504	174	127	78	60	24	16	17

- Il Sole 24 Ore
- La Gazzetta dello Sport
- Il Messaggero
- La Stampa

In 2020, 83.7% of the Italian population declared to surf the Internet regularly. Besides, people who have been using social network are increased up to 68% of Italian citizens, ensuring social networks a preeminent position in the rank of the most used platforms to connect users with other people, companies, institutions and groups (Wearesocial 2021).

In order to clarify the form of the initial dataset, Table 1 displays the first five rows. For example, since each row represents a user, the first user placed a "Like" to la Repubblica, Corriere della Sera and Il Fatto Quotidiano generating a 1 value; conversely, he/she did not place a "Like" to Il Sole 24 Ore, La Gazzetta dello Sport, Il Messaggero and La Stampa generating a 0 value.

This kind of behaviour is subject to disambiguation. The hypothesis is the presence of a difference among the users that did not place a like for a selected page. The two-step procedure presented in the previous section allows to impute a plausible value for a negative (Dislike) or neutral (Nothing) judgement. As reported in Table 1, the third, the fourth and the fifth users have only 0 values. This means that these users did not place any "Like" for the considered pages. It could be interesting to show the "Like" distribution, in order to make more robust the analysis for the detection of answers for the research question.

Table 2 shows the "Like" distribution for user with the aim of linking the number of "Likes" to the absolute frequency for single user. Only 17 subjects placed "Like" to all selected pages, while 504 users did not place any "Like" at all. Since the proposed procedure is data-driven and based on the "Like" placed, these users have been removed from the dataset. For this reason, the final number of analysed users is equal to 496.

Table 3 "Like" distribution for social page

	La Repubblica	Corriere della Sera	Il Fatto Quotidiano	Il Sole 24 Ore
"Like"	299	244	268	158
Missing Values	197	252	228	338
Missing Values (%)	39.7	50.8	46	68.1

	La Gazzetta dello Sport	Il Messaggero	La Stampa	Total
"Like"	116	66	86	**1237**
Missing Values	380	430	410	**2235**
Missing Values (%)	76.6	86.7	82.7	**64.7**

Bold indicate the total number of observations in the row

Table 4 Joint "Like" distribution for user and social page

"Like"	La Repubblica	Corriere della Sera	Il Fatto Quotidiano	Il Sole 24 Ore	La Gazzetta dello Sport	Il Messaggero	La Stampa
1	46	34	51	15	19	5	4
2	84	48	72	20	20	6	4
3	64	52	47	19	25	5	12
4	50	54	45	39	18	14	20
5	23	23	20	22	10	8	14
6	15	16	16	16	7	11	15
7	17	17	17	17	17	17	17

A second "Like" distribution could be obtained for the social pages. The idea is to find the total number of users that placed a "Like" for a page. The complementary distribution returns the proportion of subjects with 0 values. This allows to classify the social pages on the basis of the number of "Likes" to understand the most appreciated Italian newspapers. This distribution is represented in Table 3.

La Repubblica was the most liked social page, while Il Messaggero was the least popular with 86.7% of 0 values. Considering all social pages, there is a prevalence of missing values equal to 64.7%.

The "Likes" distribution for social page is useful to hypothesise the presence of a general pattern for missing values. There is no clear distribution of 0 values as for univariate, multivariate or monotone pattern.

The last distribution of "Likes" is joint for users and social pages. The users have been grouped for the number of Likes placed and on the basis of this feature, grouped for social page "Liked". Results are presented in Table 4.

Among the users who placed a unique "Like", Il fatto quotidiano was the most appreciated social page. By only considering the users with 2 or 3 Likes, the most appreciated page is La Repubblica. For users with 4 Likes, the highest frequency is

Table 5 Distribution of "Like" for the first five users

	Number of "Like"	Number of missing values	Proportion of observed "Like"
1	3	4	0.43
2	1	6	0.14
3	5	2	0.71
4	3	4	0.43
5	1	6	0.14

for the social page Corriere della Sera. It is confirmed the presence of 17 users that placed "Like" to all considered pages.

4.2 First imputation through threshold values

As reported in the methodology section, a first imputation of 0 values has been realised using a rule based on a threshold value. Before starting with the imputation, it is necessary to detect the nature of the generating mechanism for the missing values. The hypothesis to verify is the presence of MCAR data. To realise this purpose, a Lin and Bentler test has been applied on the described dataset. The null hypothesis of MCAR data is refused if:

$$X^2 > X^2_{f,1-\alpha}$$

where f are the degrees of freedom for the test statistic and α the significance level. For the Lin and Bentler test, $f = G - K$, where G is the number unique patterns in the dataset and K is the number of variables. For this application $G = 74$ and $K = 7$, therefore f is equal to 67. For $\alpha = 0.05$, the quantile of Chi-squared is equal to 90.53[3]. Since $X^2 = 2129.7$, the null hypothesis of MCAR data is refused. The obtained result is reassuring because the refuse of the null hypothesis of MCAR data is equal to refuse the hypothesis of independence of missing values from observed and not observed data. This hypothesis is aligned with the presence of behaviour associated to the missing values.

Given this hypothesis, it is possible to start with the first imputation. Table 5 shows the number of "Likes", missing values and the proportion of "Like" on the total of considered pages for the first five users of the dataset.

The first user placed 3 "Likes" over 7 pages, therefore the proportion of observed "Like" is $l = 3/7 = 0.43$. The second and the fifth user placed a unique "Like" over 7 pages with a proportion equal to $l = 0.14$. Therefore, for this application $l_i \in [0.14, 1]$, for $i = 1, \ldots, n$.

[3] This is the value for $f = 70$, but however it is a good approximation respect to the value of X^2.

Table 6 Head of the dataset after the first imputation

	La Repub-blica	Corriere della Sera	Il Fatto Quotidi-ano	Il Sole 24 Ore	La Gazzetta dello Sport	Il Messaggero	La Stampa
1	1	1	1	0	0	0	0
2	1	3	3	3	3	3	3
3	1	1	1	1	1	2	2
4	1	0	1	0	1	0	0
5	3	3	3	1	3	3	3

If this operation is repeated for all the users, it is possible to extract the average of the proportions of observed "Likes" obtaining the threshold value for the disambiguation. The c threshold value is equal to 0.36.

On average, each user placed a "Like" to 36% of considered pages. Since the social pages are 7, the assumed values for l_i are (0.14, 0.29, 0.43, 0.57, 0.71, 0.85, 1). The decisional rule for the disambiguation is specified as follows:

- if $l_i < 0.29$, that is to say, if the i-th user placed a unique "Like", the missing value will be imputed with a "Nothing". It is plausible to assume that the user does not know the other social pages;
- if $l_i > 0.43$, that is to say, if the i-th user placed at least 4 "Likes", the missing value will be imputed with a "Dislike". It is plausible to assume that the user is interested in the category and know the social pages with 0 values;
- if $l_i \in \{0.29, 0.43\}$, that is to say, if the i-th user placed 2 or 3 "Likes", the missing value will be not imputed. The available information is not sufficient to make clear disambiguation.

In Table 6, the imputation of the 0 values is reported for the first 5 rows, using the proposed decisional rule. Code 2 stands for "Dislike", the negative opinion about the considered page. Code 3 is for "Nothing", the neutral behaviour for the social page. In the table, some 0 values are still present for the missing observations not already disambiguated.

Table 7 displays the final results of the first imputation method with a threshold value. It is possible to note that 1288 missing values have been imputed. These values have been imputed as 244 "Dislike" and 1044 "Nothing". The most "disliked" social page is Il Messaggero, where Il Corriere della Sera is the least "disliked". About the "Nothing", the page with more imputation is La Stampa, the one with less imputation is Il Fatto Quotidiano. The prevalence of "Nothing" compared to "Dislike" is related to the threshold c. There are still 947 not disambiguated missing values, but the achieved informative advantage is remarkable. The percentage of missing values has been decreased from 64.4% to 27.8%.

In the next subsection, a second imputation method, the MIMCA approach, will be applied in order to further reduce the not disambiguated values.

Table 7 First imputation using the threshold rule

	la Repubblica	Corriere della Sera	Il Fatto Quotidiano	Il Sole 24 Ore
"Like"	299	244	268	158
"Dislike"	12	7	19	23
"Nothing"	128	140	123	159
Missing Values	57	105	86	156

	La Gazzetta dello Sport	Il Messaggero	La Stampa	Total
"Like"	116	66	86	**1237**
"Dislike"	65	67	51	**244**
"Nothing"	155	169	170	**1044**
Missing Values	160	194	189	**947**

Bold indicate the total number of observations in the row

Table 8 Confusion matrix

		Predictions		
		"Like"	"Dislike"	"Nothing"
Actuals	"Like"	296	15	60
	"Dislike"	60	16	0
	"Nothing"	8	0	303

4.3 Imputation of missing values with MIMCA and index of enrichment

Before performing MI with Multiple Correspondence Analysis, it is important to validate it. More specifically, the aim is to build a confusion matrix and summarise it with an index of accuracy. The following steps have been followed. A validation set has been created; in particular, 30% of the observed cells have been set to "missing value". In order to create a validation set similar to the original data set, the proportion of each category ("Like", "Dislike" and "Nothing") has been maintained. In this way, the validation set is composed of 758 cells with 49% of "Likes", 10% of "Dislike" and 41% of "Nothing".

After drawing the validation set, the optimal number of components useful for MI with MCA has been determined following the procedure described in the previous paragraph. In particular, a repeated cross validation has been performed with 100 iterations and for a number of components between 0 and 5.

Finally, MI with MCA can be used to impute missing categories. The number of multiple data sets generated is equal to 100. The category to be imputed is selected by the majority rule. In other terms, among 100 imputations for each cell of the validation set, the category imputed at least 34 times is selected. In order to evaluate the performances of MI with MCA, in Table 8 the confusion matrix is projected. As can be seen from the table, 296 "Like" over 371 are correctly imputed. Also the imputation of "Nothing" works well, with only 8 errors.

Differently, the number of "Dislike" correctly imputed is less than the number of "Like" and "Nothing" (only 16 "Dislike" are correctly imputed).

In order to summarise the confusion matrix, an index of accuracy can be derived. It is sufficient to divide the trace of the matrix, equal to 615, and the total number of cells of the validation set, equal to 758:

$$Accuracy = \frac{615}{758} = 0.8114 = 81.14\%$$

In other terms, MI with MCA correctly imputes more than 81% of the cells considered. Then, the performance of this technique is satisfactory.

The process of data enrichment about cells without a category observed or imputed can be completed after that the goodness of MI with Multiple Correspondence Analysis has been proved. As reported in the previous paragraph, the application of MI with MCA is based on a two-step procedure. The first step is to select the optimal number of components to consider in the analysis. The second one consists in estimating the algorithm of interest. Differently from the validation step, in this context "Dislike" and "Nothing" are the only categories of interest. In fact, the aim of the study is to specify the behaviour of individuals who do not like the Facebook pages considered. In order to achieve this goal, different scenarios with $M = 25, 50, 75, 100$ data sets have been imputed with MIMCA and, for each cell with a missing value, only those where a "Dislike" or a "Nothing" has been imputed are considered. Best results in prediction terms have been obtained for $M = 100$, for this reason only this scenario has been here reported. As specified in the previous paragraph, in order to minimize the simulation error due to the application of a bootstrap procedure, a d threshold has been introduced equal to 10%. In particular, considering only the data sets where a "Dislike" or a "Nothing" has been imputed, the following imputation rule for each cell has been obtained:

- a "Dislike" is imputed when the proportion of "Dislike" imputed is greater than or equal to 60%;
- a "Nothing" is imputed when the proportion of "Dislike" imputed is less than or equal to 40%;
- when the proportion of "Dislike" is between 40% and 60%, then neither a "Dislike" nor a "Nothing" is imputed.

All the elements useful for the analysis are ready. Now, the first step consists in selecting the optimal number of components to consider in MCA. In this regard, the graph presented in Fig. 2 has been derived. As can be seen from the graph, the optimal number of components is $S = 1$. This result is similar to that of the validation step. In fact, apart from cells of validation set, data are the same. After this step, MI with MCA can be applied. The main results are summarised in Table 9.

As can be noted from the table, few missing values are still present. In fact, in some cases the number of "Dislike" imputed in a specific cell is very similar

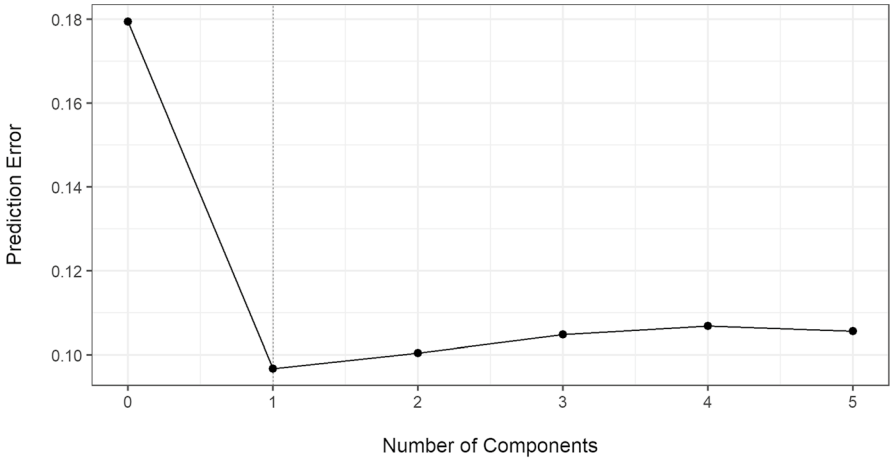

Fig. 2 Prediction Error for each component in the final application of MI with MCA

Table 9 Distribution of "Like", "Dislike" and "Nothing" after MIMCA

	La Repubblica	Corriere della Sera	Il Fatto Quotidiano	Il Sole 24 Ore
"Like"	299	244	268	158
"Dislike"	24	8	57	47
"Nothing"	149	235	139	197
Missing Values	24	9	32	94

	La Gazzetta dello Sport	Il Messaggero	La Stampa	Total
"Like"	116	66	86	**1237**
"Dislike"	221	255	200	**812**
"Nothing"	155	169	170	**1214**
Missing Values	4	6	40	**209**

Bold indicate the total number of observations in the row

to the number of "Nothing" (or even is the same). In particular, this behaviour is manifested when the proportion of "Dislike" (or "Nothing") is between 40% and 60%. Even if there are some cases where missing values are not imputed, MI with MCA works well. In fact, the proportion of missing values is now equal to 6.4%.

This reduction of missing observations could be also measured using the Informative Earning (IE) index, obtained as the ratio between the imputed values and the total missing observations. The IE index for the considered data is equal to the sum of the observations imputed as expression of negative opinion and neutral position equal to 2026 over the total of 0 values equal to 2235, registering a value of 90.6%.

5 Final remarks

In the statistical literature, the issue of missing values is a well-discussed topic. In this paper, the research question is about the possibility to face this problem in a context of sparse matrices with only 0 and 1 values, where the zeros are considered the missing observations. First of all, it is necessary to detect the nature of the zeros, if they could be considered as a simple voluntary non-response or as a consequence and symptom of a behaviour. If this second hypothesis is confirmed, then the attention could be moved on how to express this missing behaviour. For the application presented in this study, the ambit is represented by social networks where the starting matrix is composed by users and social pages. In this matrix, 1 values correspond to the presence of a "Like" for a social page given by a user and 0 values the absence of "Like. The final aim of the study is to enrich the starting matrix using a data-driven approach based on multiple imputation techniques.

To understand the generating mechanism of the 0 values, a Lin and Bentler test has been applied and the hypothesis of MCAR data has been refused, therefore it is plausible to transform the 0 values as expression of a behaviour. Once defined the mechanism, a double imputation has been effected using a decisional rule based on a threshold value and method related to the Multiple Correspondence Analysis.

The application regards 1000 users for a social networks that placed "Like" for a group of 7 social pages representing some Italian newspapers. Before the use of the two imputation techniques, the starting matrix was composed prevalently by 0 values (64.7%). Using the proposed approach, the missing values has been converted into two new categories, expression of a negative opinion or a neutral position respect to the considered social page. The imputation as a negative opinion presumes that the user has a knowledge about the content of the social pages, therefore an absence of a "Like" could correspond to a potential "Dislike". On the other hand, a neutral position about the page could be the result of a lack of interest for that topic. After the imputation, the 0 values have been reduced to 6.4%, with a value for IE+ index of 90.6%.

This approach is presented as a plausible solution in the case of sparse matrices with only 2 choices. Since it is a data-driven approach, the disambiguation could be affected by effects featured by the internal structure of data. The presented approach could be object of enhancements and future studies. More specifically, these improvements could regard the hypothesis of MAR data and the mechanism of missing values. Moreover, the used thresholds for the disambiguation rules in the MIMCA method have been defined by the authors according to a data-driven approach.

Future research will explore several directions. Firstly, a generalisation of the procedure could be applied in different contexts. For example, this procedure could be extended in the field of business management evaluating the users' behaviour about some published post on the own social page. In alternative, from a methodological point of view, the disambiguation process could regard more than two possible alternatives, solving the problem of the not imputed values. Thirdly, similar considerations could regard the introduction of this procedure for data represented by an

adjacent matrix in which the missing observation is the absence of link between two subjects, using for example a dual representation or different weights for the "Likes" as in Fersini et al. (2017). Finally, a different point of view could be accomplished comparing these results with approaches based on recommender systems.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. No funding was received to assist with the preparation of this manuscript.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Acock AC (2005) Working with missing values. J Marriage Fam 67(4):1012–1028
Angelone R (2021) Ricerche di marketing, strumenti e tecniche. PKE srl
Arrigo E, Liberati C, Mariani P (2021) Social media data and users' preferences: a statistical analysis to support marketing communication. Big Data Res 24:100189
Audigier V, Husson F, Josse J (2017) Mimca: multiple imputation for categorical variables with multiple correspondence analysis. Stat Comput 27(2):501–518
Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J School Psychol 48(1):5–37
Brettel M, Reich JC, Gavilanes JM, Flatten TC (2015) What drives advertising success on facebook? an advertising-effectiveness model: measuring the effects on sales of likes and other social-network stimuli. J Adv Res 55(2):162–175
Caers R, De Feyter T, De Couck M, Stough T, Vigna C, Du Bois C (2013) Facebook: a literature review. New Media Soc 15(6):982–1002
Crosato L, Domenech J, Liberati C (2021) Predicting sme's default: Are their websites informative? Econ. Lett. 204:109888
Dalla Valle L, Kenett R (2018) Social media big data integration: a new approach based on calibration. Expert Syst Appl 111:76–90
Ditchfield H (2020) Behind the screen of facebook: Identity construction in the rehearsal stage of online interaction. New Media Soc 22(6):927–943
Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG (2006) A gentle introduction to imputation of missing values. J Clin Epidemiol 59(10):1087–1091
Ekström M, Östman J (2015) Information, interaction, and creative production: The effects of three forms of internet use on youth democratic engagement. Commun Res 42(6):796–818
Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64(328):1183–1210

Fersini E, Pozzi FA, Messina E (2017) Approval network: a novel approach for sentiment analysis in social networks. World Wide Web 20(4):831–854
Huisman M (2009) Imputation of missing network data: some simple procedures. J Soc Struct 10(1):1–29
Jiang J, Wilson C, Wang X, Sha W, Huang P, Dai Y, Zhao BY (2013) Understanding latent interactions in online social networks. ACM Trans Web (TWEB) 7(4):1–39
Josse J, Chavent M, Liquet B, Husson F (2012) Handling missing values with regularized iterative multiple correspondence analysis. J Class 29(1):91–116
Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci USA 110(15):5802–5805
Kossinets G (2006) Effects of missing data in social networks. Soc Netw 28(3):247–268
Liberati C, Mariani P (2018) Big data meet pharmaceutical industry: An application on social media data. In Classification,(Big) Data Analysis and Statistical Learning, pages 23–30. Springer
Lin JC, Bentler PM (2012) A probability based test for missing completely at random data patterns. In meeting of the National Council on Measurement in Education, Vancouver, Canada
Little RJA (1988) A test of missing completely at random for multivariate data with missing values. J Amer Statist Assoc 83(404):1198–1202
Mariani P, Marletta A, Mussini M, Zenga M, Grammatica E (2020) A missing value approach to social network data:dislike or nothing? Comput Manag Sci 17(4):569–583
Mariani P, Marletta A, Missineo N (2019) Missing values in social media: an application on twitter data. In ASA2019, Statistics for Health and Well-being
Mellon J, Prosser C (2017) Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. Res Politics 4(3):2053168017720008
Ortiz Alvarado NB, Rodríguez Ontiveros M, Quintanilla Domínguez C (2020) Exploring emotional well-being in facebook as a driver of impulsive buying: A cross-cultural approach. J Int Consum Mark 32(5):400–415
Robins G, Pattison P, Woolcock J (2004) Missing data in networks: exponential random graph (p) models for networks with non-respondents. Soc Netw 26(3):257–283
Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Method 7(2):147
Sharma R, Magnani M, Montesi D (2016) Effects of missing data in multilayer networks. Soc Netw Anal Min 6(1):1–19
Stork D, Richards WD (1992) Nonrespondents in communication network studies: Problems and possibilities. Group Org Manag 17(2):193–209
Sumner EM, Ruge-Jones L, Alcorn D (2018) A functional approach to the facebook like button: An exploration of meaning, interpersonal functionality, and potential alternative response buttons. New Media Soc 20(4):1451–1469
Touya G (2010) A road network selection process based on data enrichment and structure detection. Trans GIS 14(5):595–614
Tufekci Z (2014) Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In Eighth international AAAI conference on weblogs and social media
Van Buuren S (2018) Flexible imputation of missing data. CRC Press, London
Wearesocial. Global digital report. In Available at https://wearesocial.com/digital-2021 (Accessed 18 May 2021), (2021)