Missing values in Social Media: an application on Twitter data

Paolo Mariani, Andrea Marletta, Nicholas Missineo

University of Milano-Bicocca

Statistics for Health and Well-being

1. Brief introduction to mismatch in Italian Labour market
2. Methodology: An exploratory approach on social media data combining missing values and Principal Component Analysis
3. Application and results
4. Conclusions and future work

Mismatch in Italian Labour market

The 2018 annual report of Excelsior ed Anpal provides a lack of alignment regarding the 26% of the labour contracts predicted by the productive system. The difficulty of recruiting is more evident for young people: 1.267.000 contracts have been predicted for under 30 years employees and the 28% is believed as not easy to find (Unioncamere, 2018).

During the last three-month period employed increased: 60.000 workers more, 44.000 with a long-term contracts. Italy did not recovered after the economic crisis and Italian workers are divided in two groups:

- those able to keep the safeguard of the past, this are on average old and close to the retirement and they are addressed to decrease.
- younger and precarious workers and their number is increasing.

Under occupation and involuntary part-time are more spread, while the wages are locked or in decrease.

Comparison in time and space

A very important index for the labour market is the number of worked hours of the total amount of the workers in a year. In 2008 in Italy the worked hours were 45,8 billions, in 2018 they reduced to 43,6.

Since the number of workers is similar to pre-crisis levels, the intensity of working is lower with respect to ten years ago.

The total unemployment rate decreased to 10,2%, but it is still twice the OECD average 5,2%, and over the average value of EU countries (6,5%).

The young unemployment rate comprise 33%, three times to the rate of the average value of OECD and EU countries (OECD, 2019).

The average income of the families decreased during last years and the increasing of the employed is only due to the increase of precarious workers.

Reddito di cittadinanza

One of the measured approved by the Italian government to improve the situation of the economy is the introduction of guaranteed minimum income for specified categories of citizens.

The guaranteed minimum income (Reddito di cittadinanza in Italian language) is a form of sustain for families in difficult condition composed by two parts: an economic part, providing a minimum income and a contribution for the house location; a project for the insertion in the labour market of the person receiving the minimum income.

This measure lasts 18 months (plus a renewal of more 18 months). The Italian government approved this measure on 17th January 2019. It is a selective measure, only oriented to those who present a determined profile of difficulty. It is not universal and it requires a precise commitment.

Goals of the study

In this work the aim is to understand the perception of this economic measure for Italian citizens and stakeholders before the introduction using social media data.

The work involves Twitter users about the perception of the Italian guaranteed minimum income on the basis of different categories of users:

- verified Twitter users as politician, institutional authorities and other official stakeholders
- not verified users as citizens and other subjects not directly involved in the process of realization of this measure

Moreover, an analysis of the KPI will be conducted using their presence and absence through the use of the complementary values.

Social media data and Twitter

Social networks are identified as an on line informative system allow the realization of virtual social interactions. They are websites or technologies permitting to share textual contents, images, videos and interactions among users (Finger, 2013).

Social media data are data collected from social network. Twitter is one of the most spread and well-known and differently from Facebook or Instagram, it has been used to share news, official contents about economics and political issues.

This is why for this study, Twitter data have been provided and statistical units are represented by Twitter record.

Twitter elements and complementary values

Each record represent a tweet, retweet or quote made by a Twitter user. A tweet is a written post on Twitter with a maximum of 280 characters. A retweet is a reproduction of a written post by another user. A quote is a comment on a tweet or a retweet.

To evaluate the effective communicative capability of a Twitter record, a comparison has been implemented among the KPIs and the respective complementary frequencies.

The difference between the distribution of KPIs and the complementary frequencies has been carried out using a chi-squared test and a principal component analysis. A difference between these two distributions could lead to a presence of bots or a dislike effect (Mariani et al., 2019).

Here two indicators are taken into account: the number of likes and the number of a retweets obtained by each record.

Like or retweet

Social networks are communication media based on the interaction between users, in which the participants tend to express in a clear way their subjectivity.

Possible way of interaction:

- a like expresses an appreciation for a post
- the retweet does not imply an approval but it equals to show the attention for that post to the followers

Since a tweet is more effective if it receives more visualizations, like and retweets make bigger the attention on it. The hypothesis is that the behaviour of the complementary observations was dissimilar to those of real observations.

Steps of the procedure

The procedure to apply the proposed approach is the following:

1. to create 3 tables (tweet, retweet and quote) with number of likes and retweets
2. to create 3 complementary tables
3. to compute the chi-squared on the two distributions (real and complementary)
4. to use a principal component analysis to reduce the dimensions underlining the positioning of variables and observations

Dataset description

Data have been collected in April 2019 using the official API Twitter for the entire Italian territory searching for #redditodicittadinanza. Using this scraping method, 4797 records and 63 covariates have been obtained.

These records have been divided into three categories:

- 945 tweets
- 3774 retweets
- 78 quotes

Covariates have been classified in three class:

- variables of the element (text of the element, date of publication, type of element, . . .)
- variables of the user (characteristics of the user who published the element)
- variables of the follower (features of the user who interact the element)

Comparison between real data and complementary data

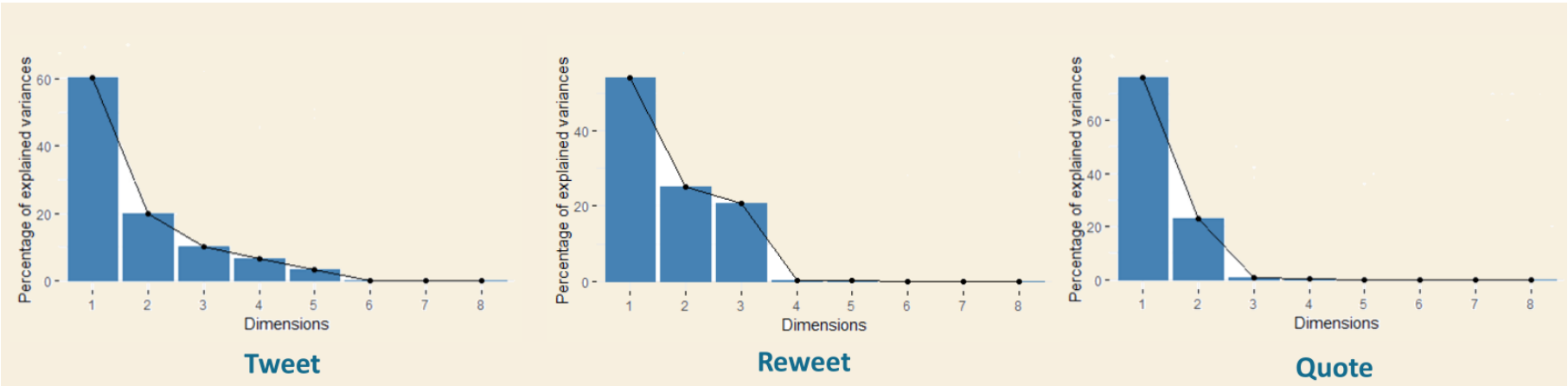
For each category of Twitter element a chi-squared test has been implemented to verify the hypothesis H_0 of the same distribution among the real data and the complementary one.

Element	$\chi^2_{oss,\alpha}$	$\chi^2_{(p-1)(n-1),\alpha}$	Result
Tweet	14869.01	594.12	Rejected H_0
Retweet	304134.40	6258.45	Rejected H_0
Quote	532.84	117.43	Rejected H_0

Principal component analysis

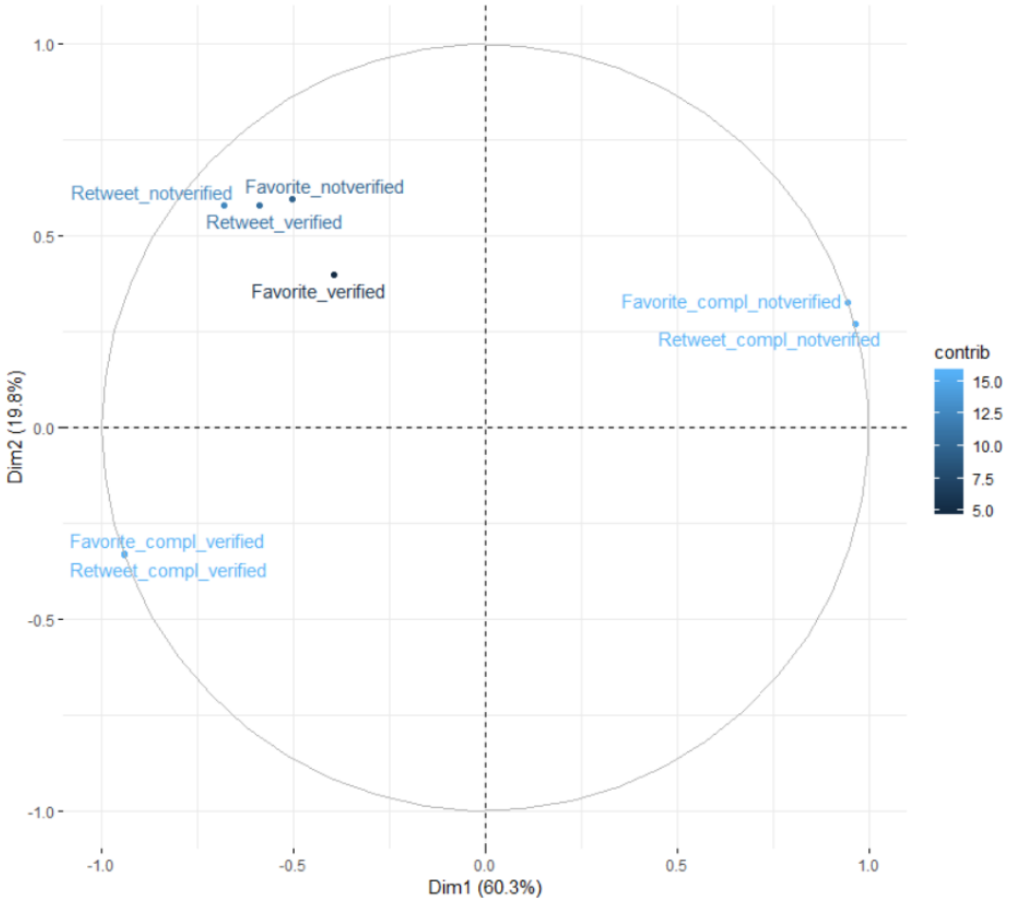
Once the hypothesis of equal distribution between real and complementary frequencies for all the considered elements, it is possible to extract latent dimensions using a principal component analysis.

A scree plot is displayed to choose the K number of the components of principal component analysis. For each category $K = 2$.

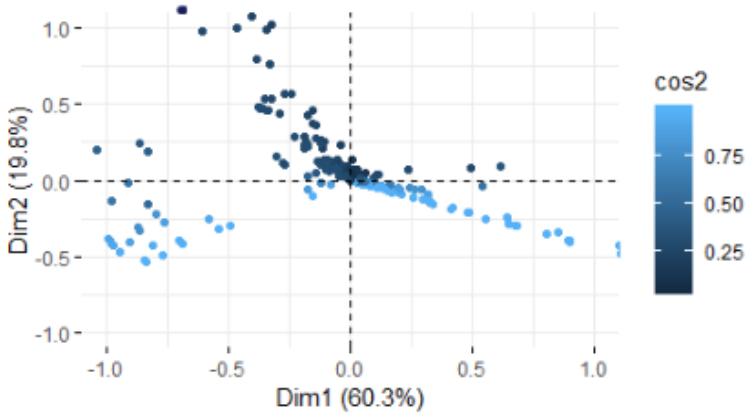


Correlation circle for Tweets

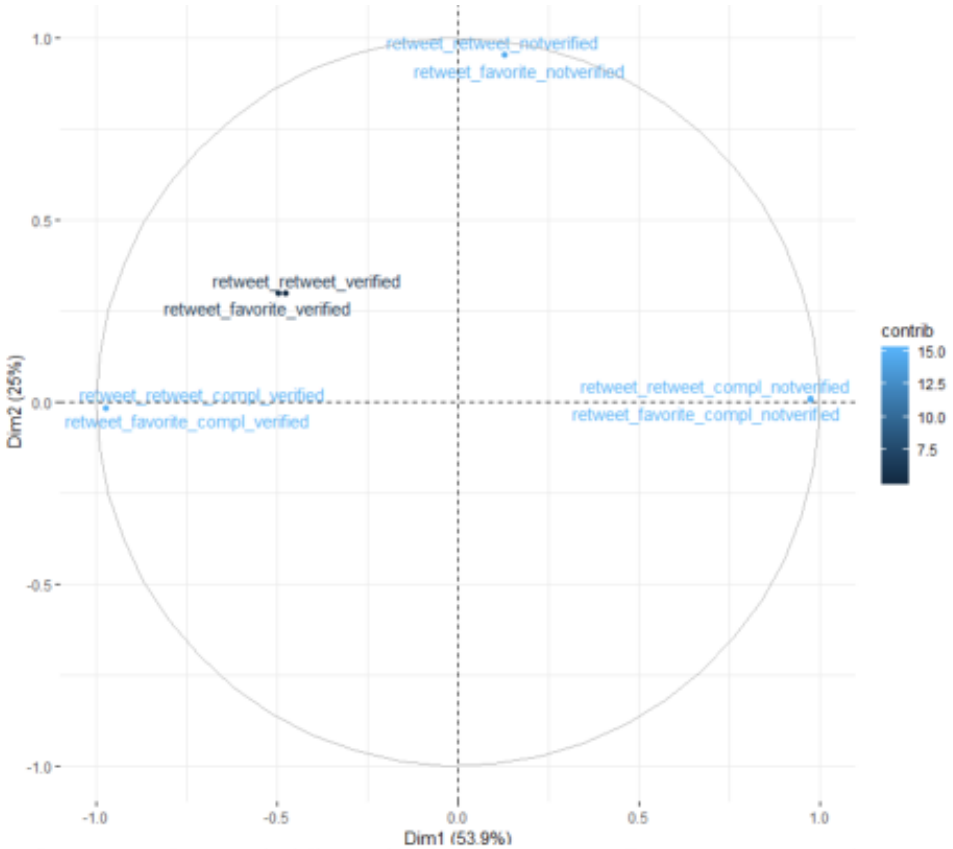
The representation of real and complementary data is not diametrically opposed.

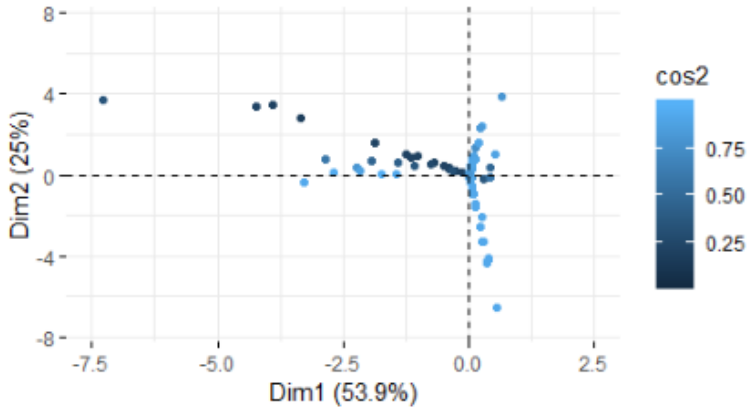


Three groups of users are detected from the plot of the individual contributes.

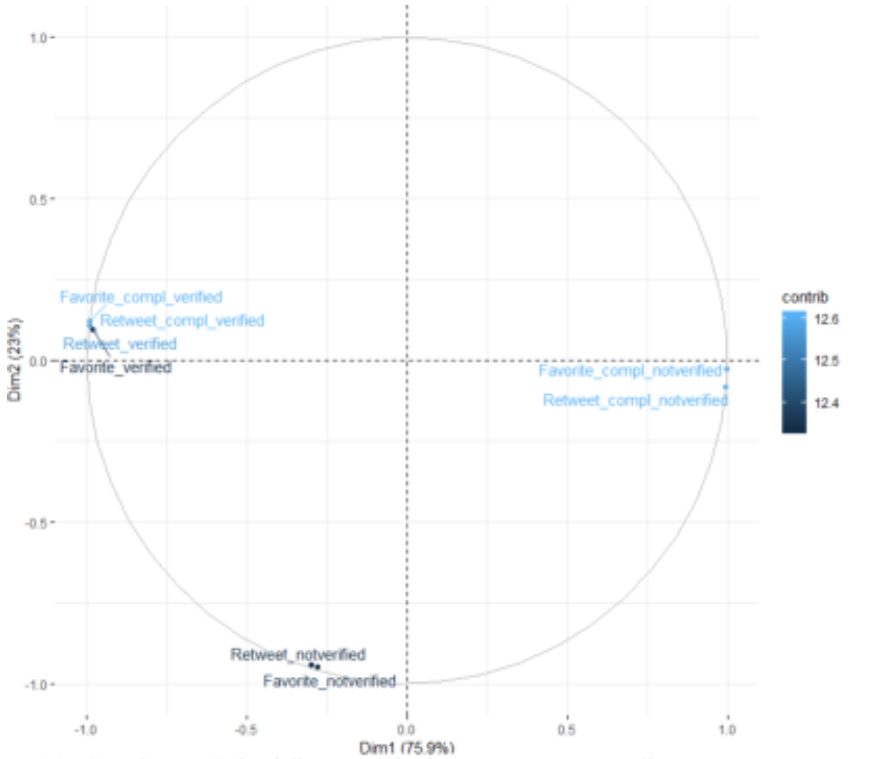


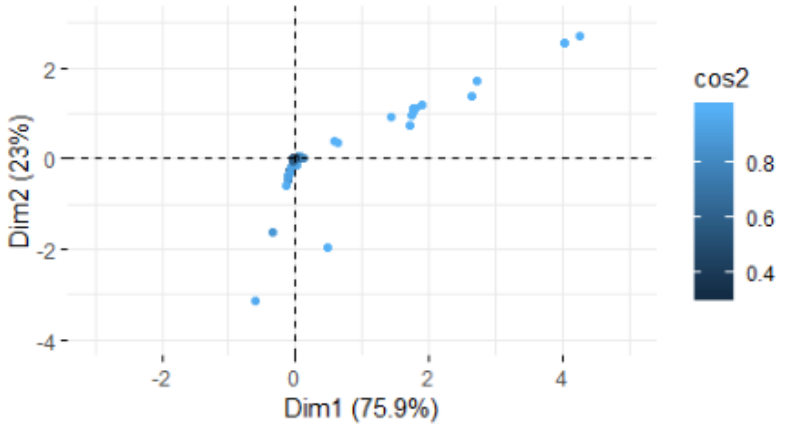
Correlation circle for Retweets





Correlation circle for Quotes





Conclusions and Future Research

- An exploratory analysis on the impact of guaranteed minimum income has been carried out after the approval and before the application using social media data
- Twitter elements as tweets, re-tweets and quotes and their complementary frequency distributions allowed to verify the presence of a behaviour in the lack of expression for the observations
- Variables for real frequencies and complementary observations are not diametrically opposed
- This supports the initial hypothesis of a behaviour between users, probably due to a potential presence of a dislike effect or bots among observations

Future Research

- Enlarging the period of scraping and compare using other keywords
- Building a predictive model to classify verified and not verified users
- Using of more accurate Text Mining models for classification of the Twitter elements
- Adding texts from other sources to find textual pattern of verified users
- Recognizing bot users looking for sentiment analysis