SIS 2017

Statistics and Data Science: new challenges, new generations

PROCEEDINGS OF THE CONFERENCE
OF THE ITALIAN STATISTICAL SOCIETY
28-30 June 2017 Florence (Italy)

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS

SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA
Sede: Salita de' Crescenzi 26 - 00186 Roma
Tel +39-06-6869845 - Fax +39-06-68806742
email: sis@caspur.it web:http://www.sis-statistica.it

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

Organi della società:

Presidente:
- Prof.ssa Monica Pratesi, Università di Pisa

Segretario Generale:
- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

Tesoriere:
- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

Consiglieri:
- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore
- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre
- Prof.ssa Francesca Bassi, Università di Padova
- Prof. Eugenio Brentari, Università di Brescia
- Dott. Stefano Falorsi, ISTAT
- Prof. Alessio Pollice, Università di Bari
- Prof.ssa Rosanna Verde, Seconda Università di Napoli
- Prof. Daniele Vignoli, Università di Firenze

Collegio dei Revisori dei Conti:
- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

SIS2017 Committees

Scientific Program Committee:
Rosanna Verde (chair), Università della Campania "Luigi Vanvitelli"
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara "G. D'Annunzio"
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:
Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:
Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Social emotional data analysis. The map of Europe

Analisi emozionale dei Social Network. La mappa dell'Europa

Felicia Pelagalli, Francesca Greco and Enrico De Santis

Abstract In this paper we present an investigation of the emotional content conveyed by words in online conversations captured on Twitter. A multivariate technique applied to co-occurence of words together with Correspondence Analysis is adopted in order to find clusters of meaningful words detecting emotional categories that provide meaning to everyday events. Specifically, given the current historical period, where the European Union has to gain trust in its citizens, a corpus of 155000 tweets selected through the Italian keywords "Europa" and "EU" is analyzed. Results show clearly how the textual content is structured according to the different emotional expressions.

Abstract *In questo articolo è presentata un'analisi testuale che esplora il contenuto emozionale delle parole nelle conversazioni su Twitter. È stata adottata una tecnica di analisi multivariata applicata alla co-occorrenza delle parole assieme all'analisi delle corrispondenze al fine di raggruppare le parole in cluster di significato e individuare le categorie e le emozioni che danno senso agli eventi – ossia, i significati attribuiti agli eventi dagli attori partecipanti a un determinato contesto. Dato il particolare periodo storico in cui versa l'Unione Europea, che si trova a dover guadagnare la fiducia dei propri cittadini, è stato preparato ed analizzato un corpus di 155000 tweet selezionati attraverso le keyword "Europa" ed "EU". I risultati mostrano chiaramente come il contenuto testuale è strutturato secondo le differenti espressioni emozionali del fenomeno.*

Felicia Pelagalli
Culture s.r.l., Piazza Capranica, 95 00186 ROMA, Italia,
Scuola di Specializzazione in Psicologia della Salute, Sapienza Universit degli Studi di Roma, Via degli Apuli, 1 - 00185 Roma, Italia, e-mail: feliciapelagalli@yahoo.it

Francesca Greco
Dipartimento di Psicologia Dinamica e Clinica, Sapienza Universit degli Studi di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italia, e-mail: francesca.greco@uniroma1.it

Enrico De Santis
Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada, e-mail: enrico.desantis@ryerson.ca

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Key words: Text Mining, Social Data Mining, Multivariate Analysis, Correspondence Analysis, Clustering.

1 Introduction

With the spread of social networks and micro-blogging platforms, statistical methodologies boosted with machine learning techniques find their natural habitat in the sea of available online data. In fact, related techniques enable us to perceive the feeling that runs through the network. An overwhelming quantity of conversations are exchanged, mostly through words in a written form. If from one side it can be possible grasping the opinions underlying the online social exchanges, from the other it is clearly interesting to have a measure of the emotional significance that gives meaning to social phenomena. Now more than ever, this knowledge can help institutions and community managers to realize people needs and problems. It is the emotion that drives us in making relation with the objects of a given context on the basis of affective symbolizations and social representations. Hence, in conveying emotions, words show the functioning structure of the mind-brain, according to a dual logic [1]: i) the asymmetrical conscious thought which allows entering in a relationship with a context or event; ii) the symmetrical emotional thinking that the context or the events immediately arouses within us. Thus, the content analysis of conversations has to catch and externalize the emotional "density" conveyed by words or chains of words, through suitable knowledge models substantiated by statistical techniques, such as the multivariate analysis. In fact, the latter, as an unsupervised technique, can find recurrences, relations between nodes of a network or can help grouping words in meaningful clusters, detecting emotional categories that provide meaning to everyday events. According to this framework, the linguistic communication can be interpreted not only on the basis of its semantic elements but also through the emotional framework that yields value to a given text. This context fits with the co-occurrence analysis of words, used as the first step of our investigation, to find associative links among words. In this study we analyze online conversations trying to discover how they are organized within the current social context and upon a given object represented by a set of keywords. Specifically, the corpus consists in 155000 tweets gathered, in the time period ranging from January 11, 2017, to February 11, 2017, trough the Twitter API, filtering the stream by the Italian keywords "*Europa*" and "*UE*". The corpus is analyzed through a pipeline of statistical and learning techniques briefly described in next section. Specifically, in order to obtain a thematic analysis based on the co-occurrence of lexical units upon the corpus at hand, a mapping of the latter in the Vector Space Model (VSM) [2] is performed. The k-means algorithm is then adopted obtaining a suitable partition through the cosine dissimilarity measure between word vectors. Finally, the Boolean contingency matrix, describing documents membership to the retrieved clusters, is analyzed with the well-known Correspondence Analysis (CA) technique.
The current paper is organized as follows. In Sec. 2 we provide a brief summary

of the adopted methodology, while in Sec. 3 main results are discussed. Finally, conclusion are drawn in Sec. 4.

2 Material and methods

To finalize the herein proposed investigation, data is cleaned and pre-processed. In particular, instead of raw words, lemmas as main categories are used. Subsequently, the the most common words and the very rare words are filtered out. Lemmatization and filtering allows to obtain a more compact VSM, reducing even the sparsity of the model. We note that in the current section the formal terms "document" and "context" are interchangeable, such as "term"', "word" or "lexical units". Following

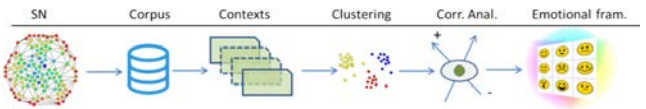

Fig. 1 Schematic diagram of the adopted methodology for measuring the emotional structures underlying the online conversations.

the diagram of Fig. 1 the analysis presented is centered on the VSM [2], a particular vector or distributional model of meaning. VSM is based on a co-occurrence matrix, i.e. the word-document matrix, that is a way of representing how often words co-occur. From a methodological point of view the VSM embeds information retained within a corpus in a vector space representation, substantiating the distributional hypothesis according to which words that occur in similar contexts tend to have similar meanings. Lets define the term-document matrix $\mathbf{X} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_D]$ where the content of each document vector $\mathbf{d}_j = [w_1, w_2, ..., w_V]$ is represented as a vector in the term space of dimension V that is usually the dimension of the vocabulary. A standard weighting scheme, used in the current work for w_i, is the the tf-idf (term frequency-inverse document frequency) [3], that provides higher weights to terms or words that are frequent in the current j-th document but rare overall in the collection.

In order to measure the similarity between two documents $\mathbf{d}_p$ and $\mathbf{d}_q$ enabling the cluster analysis, a well-suited similarity measure is used. It is the cosine similarity, that is $sim(\mathbf{d}_p, \mathbf{d}_q) = \cos(\mathbf{d}_p, \mathbf{d}_q) = \frac{\mathbf{d}_p \cdot \mathbf{d}_q}{\|\mathbf{d}_p\| \|\mathbf{d}_q\|}$.

The k-means algorithm is a *partitional* clustering algorithm [4, 5] based on squared error optimization approach. Specifically, given a set of objects (word vectors) $\mathbf{X} = \{\mathbf{d}_j\}_{i=1}^{D} \in \mathbb{R}^V$, where V is the dimension of data vectors, it finds a suitable partition $P = \{\mathscr{C}_1, \mathscr{C}_2, ..\mathscr{C}_k\}$ so that the sum of the squared distances between objects in each cluster and the respective representative element is minimized:

$$\arg\min_{P} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathscr{C}_i} \left\| \mathbf{x}_j - \mathbf{c}_i \right\|^2, \tag{1}$$

where $\mathbf{c}_i$ is the representative of the i-th cluster $\mathscr{C}_i$. Belonging to the family of the NP-Hard problems, a complete analytical solution is not know and k-means as greedy algorithm, can only converge to a local minimum.

CA is a statistical method useful for data visualization that is applicable to cross-tabular data such as counts, compositions or any ratio-scale data. In this work, it is performed on the Boolean contingency matrix describing the partition P [6]. Let $\mathbf{P}$ denote a $q_r \times q_c$ data matrix with non negative elements that sum up to 1, i.e. $\mathbf{1}_{q_r}^T \mathbf{P} \mathbf{1}_{q_c} = 1$, where in general $\mathbf{1}_q$ is a q-dimensional vector of ones and T is the transpose operator. The CA is formulated as the following least-squares problem:

$$\min_{\mathbf{A},\mathbf{B}} \left\| \tilde{\mathbf{P}} - \mathbf{D}_r^{1/2} \mathbf{A} \mathbf{B}^T \mathbf{D}_c^{1/2} \right\|^2, \tag{2}$$

where $\tilde{\mathbf{P}} = \mathbf{D}_r^{-1/2} \left(\mathbf{P} - \mathbf{r}\mathbf{c}^T \right) \mathbf{D}_c^{-1/2}$, $\mathbf{r} = \mathbf{P}\mathbf{1}_{q_c}$, $\mathbf{c} = \mathbf{P}^T \mathbf{1}_{q_r}$, $\mathbf{D}_r$ and $\mathbf{D}_c$ are corresponding diagonal matrices. The column coordinate matrices $\mathbf{A}$ and $\mathbf{B}$ are of rank k that is the dimensionality of the approximation. By imposing $\mathbf{B}^T \mathbf{D}_c \mathbf{B} = \mathbf{I}_k$, it is possible obtaining a solution through the well-known Singular Value Decomposition: $\tilde{\mathbf{P}} = \mathbf{U} \Lambda \mathbf{V}^T$, where Λ is a diagonal matrix with in descending order the singular values on the leading diagonal and $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices. A least-squares approximation of $\tilde{\mathbf{P}}$ is obtained by selecting the first k columns of $\mathbf{U}$ and $\mathbf{V}$ and the corresponding singular values in Λ. Finally, the coordinate matrices are $\mathbf{A} = \mathbf{D}_r^{-1/2} \mathbf{U} \Lambda$ and $\mathbf{B} = \mathbf{D}_c^{-1/2} \mathbf{V}$, so that $\mathbf{A}^T \mathbf{D}_r \mathbf{A} = \Lambda^2$. Given the coordinate matrices the row coordinates are referred to as principal coordinates whereas the column coordinates are standard coordinates. The two sets of coordinates are also known as biplot and the inner-product $\mathbf{D}_r^{1/2} \mathbf{A} \mathbf{B}^T \mathbf{D}_c^{1/2}$ in (2) approximates the data. If the matrix $\mathbf{P}$ constitutes a contingency table, $\tilde{\mathbf{P}}$ is the matrix of standardized residuals, i.e. the matrix of standardized deviations from the independence model. Hence, a low-dimensional approximation of these standardized residuals is given by the biplot coordinates in $\mathbf{A}$ and $\mathbf{B}$. In other words, it can be shown that this biplot will approximate, by euclidean distances on the plot, chi-square distances in $\mathbf{P}$. Chi-square distance is mathematically the euclidean distance inversely weighted by the marginal totals.

3 Results

As concerns the cluster analysis the cardinality k of the partition P is set to 5. In Tab. 1 are reported the explained variances for each principal components that hereinafter are named "factors". In Fig. 2 we can appreciate the emotional *map* of the Europe coming out from Italian tweets. It shows how discovered clusters are placed in the factorial space, whereas in Tab. 2 is reported the factors–clusters matrix that

summarizes our main findings. The emerging *map* shows on the horizontal plane a sharp contrast between the "political power" and the "populist protest". The cluster of words $\mathscr{C}_1$ sees the chill and sooty European institutional places that are perceived as a remote center of power in which citizens do not definitely recognize themselves. The theme is the election of Antonio Tajani as president of the European parliament and Pittella defeat. Congratulations words, but even disappointment and irritation for who does not feel represented (*dividere*, *urtare*, *sensibilità*, *impera*). On the opposite side, a strong sense of helplessness regarding the big problems, such as immigrants and the economic crisis. $\mathscr{C}_3$ is characterized by the UE plan proposed in order to stop the sea blockade in front of Libyan territories. We have also tweets where the Italian Economy ministry is perceived as "unable", while the former prime minister Matteo Renzi together with Angelino Alfano (current Italian foreign minister) are considered "hypocrites". Another emerging contrast on the vertical plane is the "success of the economic power" and "people problems". From $\mathscr{C}_2$ it emerges a two-speed Europe and the "economic power"' represented by Germany with the chancellor Anghela Merkel and the president of the European Central Bank Mario Draghi. It is a strong power (*velocità*, *vincere*) that cohabits/forgets the human tragedies (*permettere*, *vergognarsi*). On the opposite side, $\mathscr{C}_5$ refers to the necessity of funds for places hit by the earthquake. Furthermore, it shows clearly the arising of new political movements, such as the one referred to Marine Le Pen in France, evidencing tension, betrayal, isolation and risks for Europe. Finally, in $\mathscr{C}_4$ (in a middle position on the map) we find the ambivalence fear/anguish related to the dichotomy opening–closing, where closing seems to prevail together with the fantasy of closing themselves off in the localism to avoid chaos. This is a cluster full of fears that undermine the Altiero Spinelli's project for a united Europe. $\mathscr{C}_4$ is close the origin of axes on the factorial map, in fact it contains basic emotions that seem to span all the facets of the underlying discourse.

Table 1 Explained variance for each factor.

Ind	Eigenvalues	%	Cumul. %
1	0.1538	29.54	29.5438
2	0.1308	25.14	54.683
3	0.1214	23.32	77.9995
4	0.1145	22.00	100

4 Conclusion

The current paper presents an analysis of a huge corpus of tweets in Italian language based on a set of statistical techniques, specifically a Cluster analysis and a Correspondence Analysis. Unlike the current sentiment analysis techniques, the proposed

Table 2 The factor-clusters matrix.

	Factor 1	Factor 2	Factor 3	Factor 4	
$\mathscr{C}_1$	–	–	changing	–	Problems related to the political power unable to drive the changing.
	–	-0,2485	0,2177	-0,5145	
$\mathscr{C}_2$	hopes	power	changing	–	The success is related to the economic power represented by A. Merkel and M. Draghi.
	-0,5119	0,6132	0,2352	0,0558	
$\mathscr{C}_3$	alert	–	changing	injustice	Alert generated by the changing related to balance of powers.
	0,3088	-0,1367	0,2492	0,3808	
$\mathscr{C}_5$	alert	–	changing	injustice	The idea about the union with the social power of foreign countries because of the loss of identity.
	0,4653	0,3534	-0,5362	-0,1396	
$\mathscr{C}_5$	hopes	problems	product	–	The European genesis has a cost that causes problems: economic request for *help* and the rejection to *give*.
	-0,5718	-0,4382	-0,4911	0,2101	

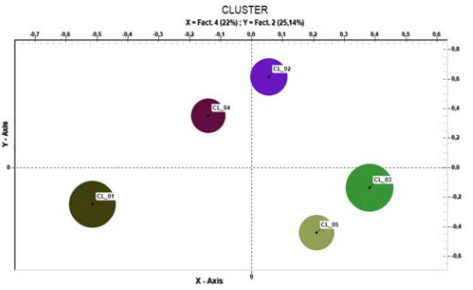

Fig. 2 The map of the Europe.

methodology takes into account the conversations on social networks like structured corpora, in which the relationships between words can be described beyond the evaluative bias (positive/negative or agree/disagree), giving rise to a dense structure of meaning. Results show clearly how the textual content is structured according to the different emotional expressions.

References

[1] Matte Blanco. *L'inconscio come insiemi infiniti*. Biblioteca Einaudi, 2000.
[2] Gerard Salton. The smart retrieval system – experiments in automatic document processing, 1971.
[3] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. documentation*, 28(1):11–21, 1972.
[4] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM (JACM)*, 51(3):497–515, 2004.
[5] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.*, 31(8):651–666, June 2010.
[6] M. van de Velden, A. Iodice D'Enza, and F. Palumbo. Cluster correspondence analysis. *Psychom.*, 82(1):158–185, 2017.