Studies in Classification, Data Analysis, and Knowledge Organization

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia · Biagio Aragona
Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications

Springer

Editors
N. Carlo Lauro
Department of Economy and Statistics
University of Naples Federico II
Naples
Italy

Biagio Aragona
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Enrica Amaturo
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Marina Marino
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Maria Gabriella Grassia
Department of Social Sciences
University of Naples Federico II
Naples
Italy

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-319-55476-1 ISBN 978-3-319-55477-8 (eBook)
DOI 10.1007/978-3-319-55477-8

Library of Congress Control Number: 2017937698

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Data Science is a multidisciplinary approach based mainly on the methods of statistics and computer science suitably supplemented by the knowledge of the different domains to meet the new challenges posed by the actual information society. Aim of Data Science is to develop appropriate methodologies for purposes of knowledge, forecasting, and decision-making in the face of an increasingly complex reality often characterized by large amounts of data (big data) of various types (numeric, ordinal, nominal, symbolic data, texts, images, data streams, multi-way data, networks, etc.), coming from disparate sources.

The main novelty in the Data Science is played by the role of the KNOWLEDGE. Its encoding in the form of logical rules or hierarchies, graphs, metadata, and ontologies, will represent a new and more effective perspective to data analysis and interpretation of results if properly integrated in the methods of Data Science. It is in this sense that the Data Science can be understood as a discipline whose methods, result of the intersection between statistics, computer science, and a knowledge domain, have as their purpose to give meaning to the data. Thus, from this point of view, it would be preferable to speak about DATA SCIENCES.

The Data Science and Social Research Conference has represented an interdisciplinary event, where scientists of different areas, focusing on social sciences, had the opportunity to meet and discuss about the epistemological, methodological, and computational developments brought about by the availability of new data (big data, big corpora, open data, linked data, etc.). Such a new environment offers to social research great opportunities to enhance knowledge on some key research areas (i.e. development, social inequalities, public health, governance, marketing, communication).

Along, the conference has been a crucial issue to discuss critical questions about what all this data means, who gets access to what data, and how data are analysed and to what extent.

Therefore, aim of the conference, and of the present volume, has been to depict the challenges and the opportunities that the "data revolution" poses to Social Research in the framework of Data Science, this in view of building a SOCIAL DATA SCIENCE … Let us own data science!

Naples, Italy

N. Carlo Lauro
Professor Emeritus of Statistics

Contents

Part III On-Line Data Applications

Part IV Off-Line Data Applications

The Sentiment of the Infosphere: A Sentiment Analysis Approach for the Big Conversation on the Net

Antonio Ruoto, Vito Santarcangelo, Davide Liga, Giuseppe Oddo, Massimiliano Giacalone and Eugenio Iorio

Abstract In the Network Society the use of hashtags has become a daily routine for the participation on the Big Conversation Iorio and Ruoto (Nessun tempo, 2015). Designated by a 'hash' symbol (#), a hashtag is a keyword assigned to information that describes it and aides in searching. Hashtags are now central to organize information on Social Networks. Hashtags organize discussion around specific topics or events and they are becoming an integrated part of the Infosphere, the whole informational environment constituted by all informational entities. The sentiment analysis of Hashtags shared on the Big Conversation can return a possible snapshot about the sentiment shared by users. Scope of this work is to present an application of sentiment analysis on the Italian hashtags of mainly social networks as part of the 'Infosphere'. This analysis returns a semantic sentiment report about the hashtags shared by the users of the social networks, that can produce a semantic sentiment trend about users. This approach could be applied to every language simply changing the sentiment thesaurus used.

A. Ruoto · D. Liga
iInformatica S.r.l.s., Corso Italia 77, Trapani, Italy

A. Ruoto · E. Iorio
University of Naples 'Suor Orsola Benincasa', Naples, Italy

M. Giacalone
Department of Economics and Statistics, University of Naples 'Federico II', Naples, Italy

G. Oddo · V. Santarcangelo (✉)
Centro Studi S.r.l., Zona Industriale, Buccino, Italy
e-mail: santarcangelo@dmi.unict.it

V. Santarcangelo
Department of Mathematics and Computer Science, University of Catania, Catania, Italy

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_20

1 Introduction

This research paper aims to determine a methodology for investigating and mapping the emotional level of the Italian Big Conversation on the Net as a part of the larger Infosphere level. More specifically, this paper will provide an example of this methodology applied to the Instagram platform, and in particular to Italian-speaking users, with the following objectives:

- Assess what direction the generated emotional forces are going;
- Understand what are the main emotional projections being transmitted;
- Assess what are the ten most important emotional polarizations.

Sentiment analysis of text is a well-known technique in Natural Language Processing. The idea is that some words hold positive or negative meanings. For example the word 'good' might have a positive score '+2' and the word 'terrible' negative '−3'. Each word in a social network caption, added by its owner, is converted into a corresponding score, and all we have to do is to sum them up to get the overall mood. Sentiment analysis is based on sentiment thesaurus, that is customized for the target language. For the Italian language there are two thesaurus: AIN Thesaurus Santarcangelo et al. (2015) and SentiStrenght. The first is the most complete thesaurus for Italian language also for the number of words considered, for the difference between adjectives and intensifiers and also because it includes a semantic information for each word Santarcangelo et al. (2015). In the Philosophy of Information (PI) the term Infosphere denotes the entirety of the Information environment. Coined from the term 'biosphere', this neologism has been first introduced in the literature by Luciano Floridi, one of the most influent academics of the PI. According to Floridi, the infosphere is 'the semantic space composed by the entirety of documents, agents, and their operations', where 'documents' refers to any kind of data, information and knowledge that has been codified and achieved in any semiotic form; 'agents' refers to any system that is able to interact with an independent document (e.g. individuals, organisations, robot softwares); 'operations' refers to any action, interaction and transformation that an agent can carry out, and that at the same time can be regarded as a 'document'. All the above constitutes an environment in which organisms are able to develop, as if they were interconnected cells Floridi (2011). Considering todays Network Society, in which we all live, it is quite clear that the Big Conversation on the Net is currently an essential part of the infosphere. The widespread of the Internet, and the access to mobile devices, digital media, along with a wide range of social platform, fostered the development of interactive and horizontal networks of communication that are able to connect local and global spheres at any time. The communication system of the industrial era revolved around mass media, characterized by unidirectional one-to-many mass dissemination of information. Conversely, the basis of our Network Society revolves around a global system of horizontal communication capable of generating multimodal exchanges of information and many-to-many interactions (both synchronously and asynchronously). Following the consolidation of Manuel Castells mass self-communication paradigm, individuals have increasingly got used to such new means of communication and they have generated their own mass communication system, made of sms, blog, social networks and messages exchanged on all sorts of instant messaging platforms. It is a huge, multilinguistic and multicultural communication environment Castells (2009). In this sense, the Net has become a sort of place where everybody can say something; an infinite room in which a privatization of the public sphere occurs, an intimism that alters the conventional ideas of 'public' and 'private'. If we consider the public sphere like a room in which public opinion develops, the analysis of the current Big Conversation on the Net (influenced by events and agenda-setting topics) can reveal interesting aspects of todays major transformations, providing an insightful prospective on the future mechanisms involved in the generation of common sentiment and collective/individual imagination. However, it should be considered that public opinion derived from the Big Conversation on the Net is giving way to a sort of emotional opinion. In particular, the effects of information overload and emotional sharing are going to turn the public debate into a debate that seems to be more emotional than sensible Lovink (2010), Carr (2010). Users do not pay attention to information unless it promotes their interest, enthusiasm, fear, anger, disgust. As a consequence, it seems that the only effective message is the emotional one, which originates from something strongly emotional. In this way, within the mass self-communication, the amplification of the emotional sphere becomes the foundation on which information can be spread; and this spread follows the rules of the emotional contagion. To a certain extent, it is the confirmation of what the philosopher David Hume argued almost three centuries ago: reason is the slave of passion, and not the reverse Westen (2008). Therefore, if we become aware of the fact that behind social behaviours there are processes that are able to alter social emotion, an action of emotional intelligence would help us understand the extent of such alteration, the way in which it influences common sentiment and the role it plays in the creation of collective and individual consciousness.

2 Methodology

In order to map the emotions circulating within the Big Conversation on the Net, this paper resorts to OSINT methodology, gathering information provided by publicly available sources. More specifically, considering a period of almost 6 years (from 10th October 2010, launch date of Instagram in Italy, to 10th January 2016), we have classified the most commonly used Italian adjectives that appear in the hashtags of Italian-speaking users of Instagram (with a frequency that is greater than or equal to 100.000) according to an 'emotional scoring system'. Each hashtag has been assigned an 'emotional score' on the basis of the AIN Thesaurus Pilato et al. (2015), one of the most exhaustive Italian thesaurus for the Sentiment Analysis. This thesaurus assigns to each adjective an emotional polarity score which ranges from negative to positive, according to the following scoring system:

- + 2 (very positive)
- + 1.5
- + 1 (positive)
- + 0.5
- 0 (neutral)
- −0.5
- −1 (negative)
- −1.5
- −2 (very negative)

In addition to this, each hashtag has been classified according to the emotion classification system introduced by the American psychologist Robert Plutchik (2002) (Figs. 1, 2, 3 and 4).

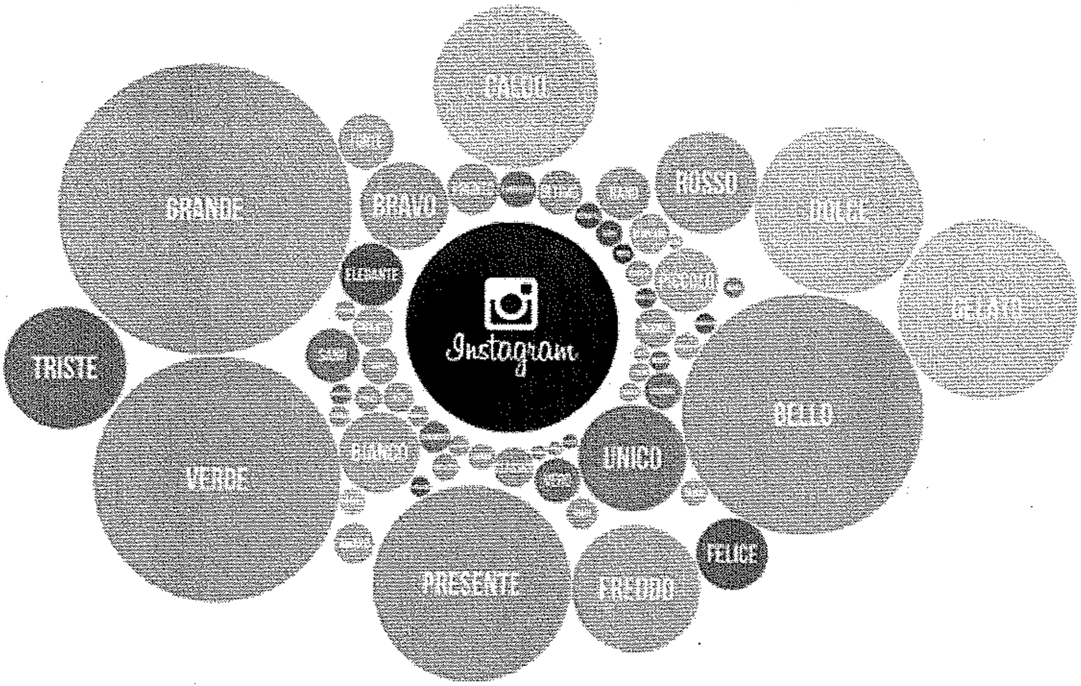

Fig. 1 Instagram sentiment hash cloud

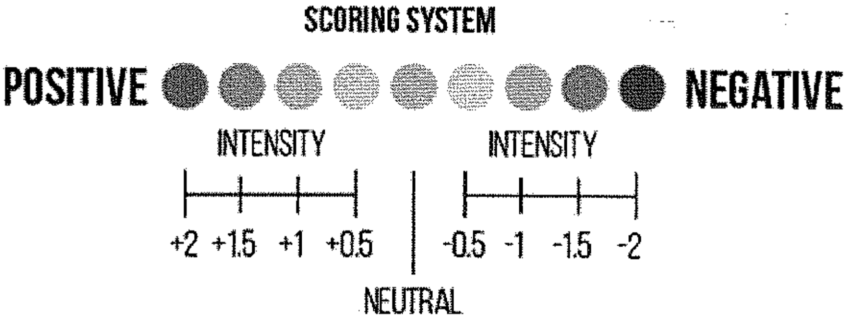

Fig. 2 AIN scoring system

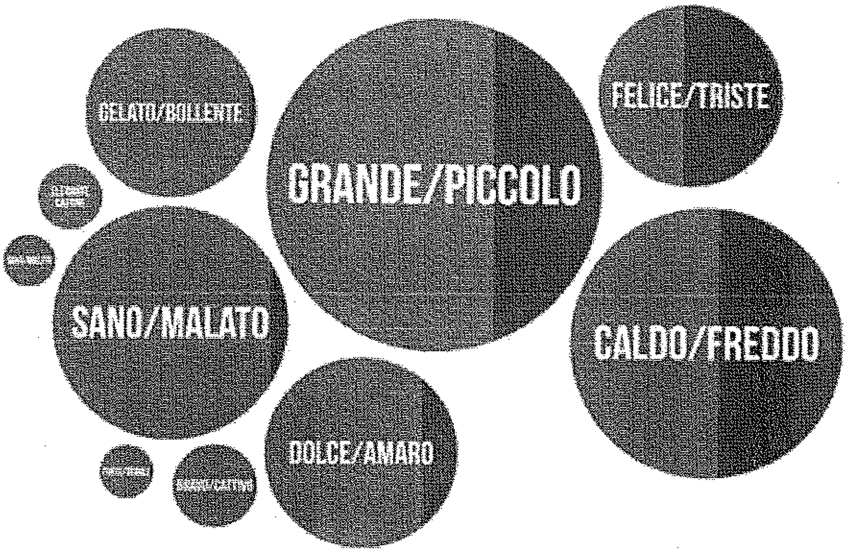

Fig. 3 Instagram top sentiment polarizations

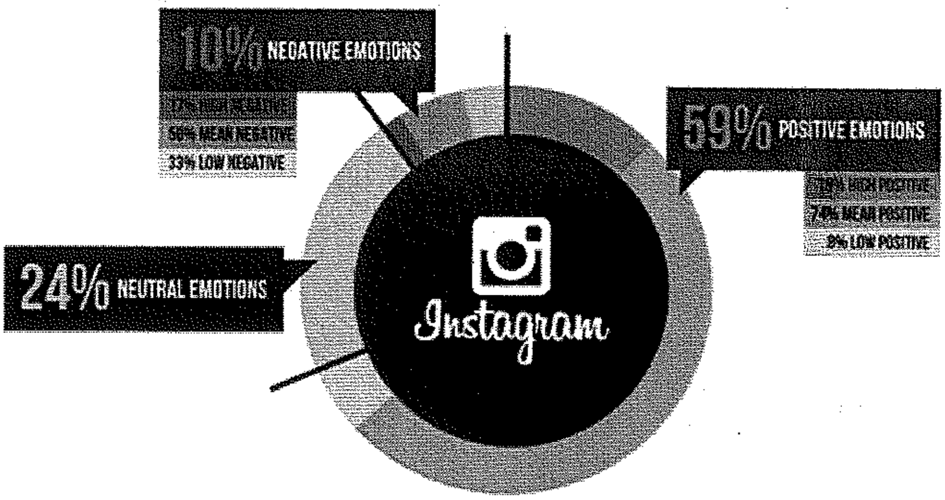

Fig. 4 Instagram results

3 Discussion on the Results

In short, the study suggests that Italian users of Instagram use the social network to:

- transmit emotional polarizations that are usually positive: of all the examined hashtags, those that have been used more frequently show more positive emotional polarity, whilst on average very few hashtags have a negative emotional polarity;
- generates a social environment in which the most represented emotional categories are those linked to the spheres of 'anticipation' and 'joy'; conversely, emotions linked to the ideas of 'anger' and 'disgust' are almost non-existent.

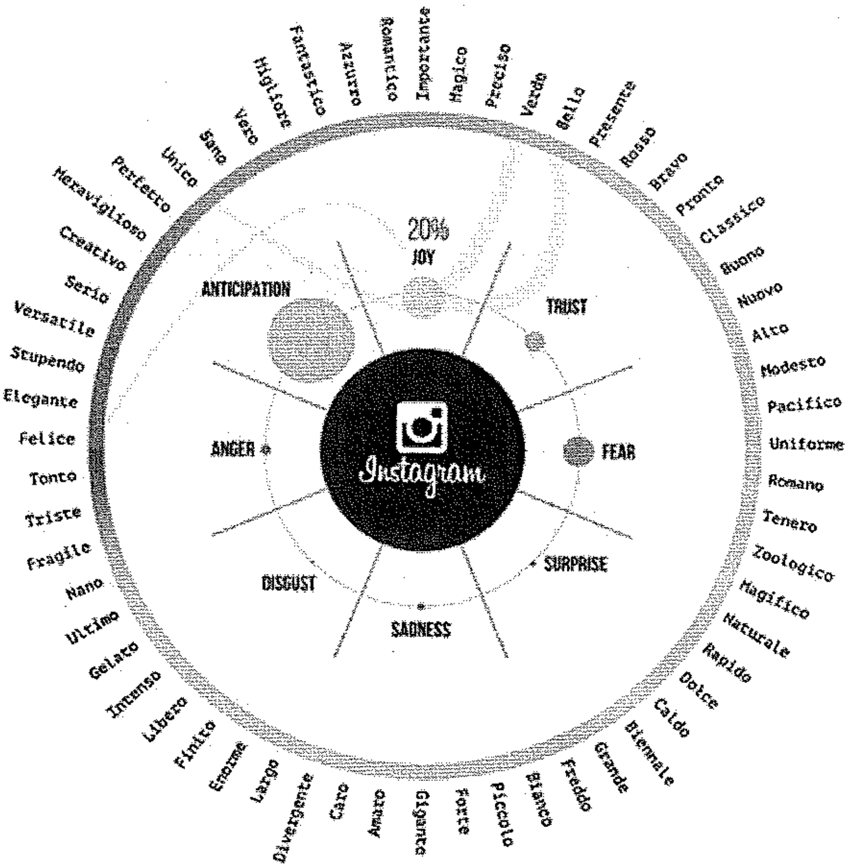

Fig. 5 Hashtag Plutchik analysis

These results should be interpreted taking into account the following key observation: the form in which contents are presented and the architecture of the social networking environment are more important than contents themselves. In other words, communication processes turn into storytelling. Due to the nature of the means, the narratives transmitted by Instagram users seem to serve one specific purpose: the digital packaging of the Self. In this sense, the users content creation leads, more or less consciously, to a sort of self-marketing in which users main objective (declared or not) is to paint himself in the best light possible in the market of social relationships. In a society in which you exist only if you appear, this general trend seems to respond to a need for people to 'perform themselves' (Figs. 5, 6).

4 Conclusions

It is not surprising that the emotions expressed in hashtags of our study tend to be positive. It is simply a seductive strategy, a sort of negotiation: by transmitting a positive emotion and narrative, users receive more positivities (e.g. likes). In this way, it seems that Instagram leaves little room for a truly critical debate on reality. Further

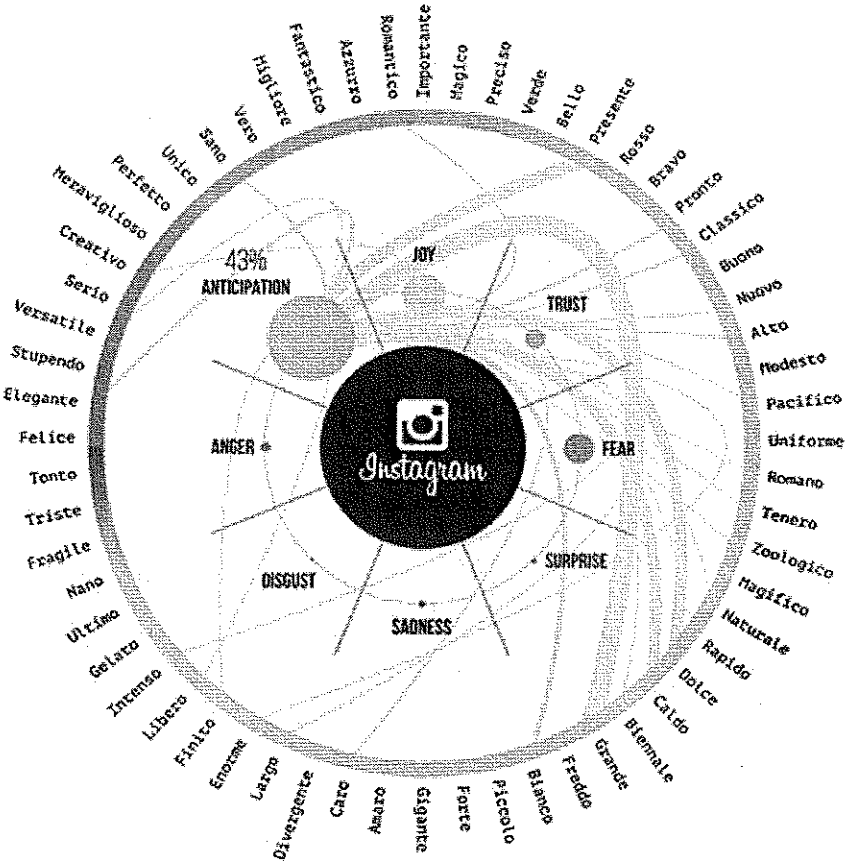

Fig. 6 Hashtag Plutchik analysis—anticipation

evidence of this is also provided by the predominance of a homophile reasoning in users' dynamics. In this regard, likeability ends up being not only the true yardstick when developing a narrative strategy aimed to take centre stage on social networks, but also a major driving force in the creation of common sentiment.

References

Carr, N. (2010). The shallows: What the Internet is doing to our brains. W. Norton & Company

Castells, M. (2009). *Communication power*. New York: Oxford University Press, Oxford.

Floridi, L. (2011). *Information a very short introduction*. Oxford: Oxford University Press.

Iorio, E., & Ruoto, A. (2015). *Nessun tempo*. Nessun Luogo: La comunicazione pubblica italiana all'epoca delle Reti.

Lovink, G. (2010). *Networks without a cause: A critique of social media*. Cambridge: Polity.

Pilato, M., Santarcangelo, G., Santarcangelo, V., & Oddo, G. (2015). AIN Thesaurus, RCE MULTIMEDIA

Plutchik, R. (2002). *Emotions and life: Perspectives from psychology, biology, and evolution*. Washington, DC: American Psychological Association.

Santarcangelo, V., Pilato, M., et al. (2015). *An opinion mining application on OSINT for the reputation analysis of public administrations.* Bari: Choice and preference analysis for quality improvement and seminar on experimentation.

Santarcangelo, V., Oddo, G., Pilato, M., Valenti, F., & Fornaro, C. (2015). Social opinion mining: an approach for Italian language. SNAMS2015 at FiCloud2015.

Westen, D. (2008). The political brain: The role of emotion in deciding the fate of the nation. PublicAffairs