

Studies in Classification, Data Analysis, and Knowledge Organization

Managing Editors

Wolfgang Gaul, Karlsruhe, Germany
Maurizio Vichi, Rome, Italy
Claus Weihs, Dortmund, Germany

Editorial Board

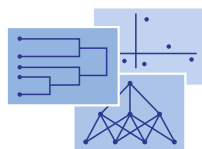
Daniel Baier, Bayreuth, Germany
Frank Critchley, Milton Keynes, UK
Reinhold Decker, Bielefeld, Germany
Edwin Diday, Paris, France
Michael Greenacre, Barcelona, Spain
Carlo Natale Lauro, Naples, Italy
Jacqueline Meulman, Leiden, The Netherlands
Paola Monari, Bologna, Italy
Shizuhiko Nishisato, Toronto, Canada
Noboru Ohsumi, Tokyo, Japan
Otto Opitz, Augsburg, Germany
Gunter Ritter, Passau, Germany
Martin Schader, Mannheim, Germany

More information about this series at <http://www.springer.com/series/1564>

Domenica Fioredistella Iezzi ·
Damon Mayaffre · Michelangelo Misuraca
Editors

Text Analytics

Advances and Challenges



 Springer

Editors

Domenica Fioredistella Iezzi
Department of Enterprise Engineering Mario
Lucertini
Tor Vergata University
Rome, Italy

Damon Mayaffre
BCL
University of Côte d'Azur and CNRS
Nice, France

Michelangelo Misuraca
Department of Business Administration
and Law
University of Calabria
Rende, Italy

ISSN 1431-8814

ISSN 2198-3321 (electronic)

Studies in Classification, Data Analysis, and Knowledge Organization

ISBN 978-3-030-52679-5

ISBN 978-3-030-52680-1 (eBook)

<https://doi.org/10.1007/978-3-030-52680-1>

Mathematics Subject Classification: 97K80

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

The Statistical Analysis of Textual Data is a broad field of research that has been developed since the 1950s. The subjects that significantly contributed to its development are Linguistics, Mathematics, Statistics, and Computer Science. Over time, the methodologies have been refined, and the applications have been enriched with new proposals from the segmentation of texts, to the development of linguistic resources, the creation of lexicons, concordance analysis, text classification, sentiment analysis. The fields of application are the most varied, ranging from Psychology to Sociology, Marketing, Economics, Medicine, and Politics. Noteworthy, the Internet has become an inexhaustible source of data, also providing through social media a cross section of the changing society. Big data is the word that most echoes among the data scientists.

This volume aims to collect methodological and applicative contributions to text analysis, previously discussed during the JADT18 conference which took place in Rome from 12 to 15 June 2018. This biennial conference, which has continuously gained importance since its first occurrence in Barcelona (1990), is open to all scholars and researchers working in the field of textual data analysis; ranging from lexicography to the analysis of political discourse, from information retrieval to marketing research, from computational linguistics to sociolinguistics, from text mining to content analysis. After the success of the previous meetings, the three-day conference in Rome continued providing a workshop-style forum through technical paper sessions, invited talks, and panel discussions.

This book, composed of 23 papers, is divided into four macro-parts: (1) techniques, methods, and models parts, (2) dictionaries and specific languages, (3) multilingual text analysis, and (4) applications.

As Umberto Eco said in *The Name of the Rose* (1980) “The beauty of the cosmos is given not only by unity in variety but also by variety in unity.” This latter claim outlines the traits of the present book, a variety in unity of analysis, techniques, and methods for the interpretation of the textual phenomenon in a variety of applications to several domains.

Those papers represent a virtual showcase of the wealth of this research field, a classical music concert where each instrument has its role and is indispensable for the general harmony.

The first part (*Techniques, Methods and Models*) is composed of six papers.

Iezzi and Celardo traced the timeline of text analytics, illustrating indices and techniques that have had a strong impact on this field of research. They showed the long way from the past to the present demarcating what could be the future scenarios.

Misuraca and Spano explained how to prepare a set of documents for quantitative analyses and compared different approaches widely used to extract information automatically, discussing their advantages and disadvantages.

Felaco presented a joint use of text analysis and network text analysis in order to study the narrations.

Vanni, Corneli, Longree Mayaffre, and Precioso compared statistical analysis and deep learning approaches to textual data.

Fronzetti and Naldi described concentration indices for dialogue dominance phenomena in TV series.

Naldi introduced the methods to measure interactions in personal finance forums.

The second macro-area (*Dictionaries and Specific Languages*) consists of six papers focusing the attention on dictionaries in particular areas and on the search for specific forms.

Iezzi e Bertè analyzed the judicial measures issued by the Court of Audit, from 2010 to 2018 in matters of responsibility and pension. At this aim, the authors proposed a dictionary to support the accounting magistracy in drafting the final measures the judge's decision-making process.

Revelli reviewed the scholastic Italian modeled by teachers in the first 150 years after unification, with a view to assessing the strengths and weaknesses of applying lexicometric parameters to a linguistic variety that was targeted at an inexperienced audience. Battisti and Dolcetti analyzed emotional textual analysis based on the study of the association of dense words that convey most of the emotional components of texts.

Romano, Baldasserini, and Pavone proposed a specific dictionary for public administration. They presented the preliminary results on 308 sentences of the Court of Cassation.

Pincemin, Lavrentiev, and Guillot-Barbance designed several methodological paths to investigate the gradient as computed by the correspondence analysis (CA) first axis. Rossari, Dolamic, Hütsch, Ricci, and Wandel analyzed the discursive functions of a set of French modal forms by establishing their combinatory profiles based on their co-occurrence with different connectors.

Multilingual Text Analysis part is composed of four papers.

Moreau examined the access to the signs in French Sign Language (LSF) within a corpus taken from the collaborative platform Ocelles, from a multilingual French bijective/LSF perspective.

Farina and Billero described the work done to exploit the LBC database for the purpose of translation analysis as a resource to edit the bilingual lexical sections of our dictionaries of cultural heritage (in nine languages).

Henkel observed the frequency of the conditional perfect in English and French in a corpus of almost 12 million words corpus consisting of four 2.9 million-word comparable and parallel subcorpora, tagged by POS and lemma, and analyzed using regular expressions.

Shen underlined the need for multilingual monitoring and on the current or future developments of text mining, for three major languages (French, English, and Chinese), in crucial areas for the world's future, and to describe the specificity and efficiency of anaphora.

Applications part is composed of seven papers.

Lebart presented a brief review of several endeavors to identify latent variables (axes or clusters) from an empirical point of view.

Gerli examined the extended abstracts of the EU-funded research projects (2007–2013) realized within the broad domain of the Social Sciences and Humanities (SSH). The aim is to verify how the emergence of a European research funding affects the directions and processes of scientific knowledge production.

Sanandres proposed a latent Dirichlet allocation (LDA) topic model of Twitter conversations to determine the topics shared on Twitter about the financial crisis in the National University of Colombia.

Celardo, Vallerotonda, De Santis, Scarici, and Leva presented a pilot project of the Italian Institute of Insurance against Accidents at Work (INAIL) about mass media monitoring in order to find out how the press deals with the culture of safety and health at work.

Santelli, Ragozini, and Musella analyzed the information included in the open-ended question section of Istat survey “Multiscopo, Aspetti della vita quotidiana” (Multi-purposes survey, daily life aspects), released in the year 2013, regarding the description of the tasks performed individually as volunteers.

Bitetto and Bollani reflected on the valorization of the wide availability of clinical documentation stored in electronic form to track the patient's health status during his care path.

Cordella, Greco, Meoli, Palermo, and Grasso explored teachers' and students' training culture in clinical psychology in an Italian university to understand whether the educational context supports the development of useful professional skills.

Rome, Italy
Nice, France
Rende, Italy

Domenica Fioredestella Iezzi
Damon Mayaffre
Michelangelo Misuraca

Contents

Techniques, Methods and Models

Text Analytics: Present, Past and Future	3
Domenica Fioredistella Iezzi and Livia Celardo	
Unsupervised Analytic Strategies to Explore Large Document Collections	17
Michelangelo Misuraca and Maria Spano	
Studying Narrative Flows by Text Analysis and Network Text Analysis	29
Cristiano Felaco and Anna Parola	
Key Passages : From Statistics to Deep Learning	41
Laurent Vanni, Marco Corneli, Dominique Longrée, Damon Mayaffre, and Frédéric Precioso	
Concentration Indices for Dialogue Dominance Phenomena in TV Series: The Case of the Big Bang Theory	55
Andrea Fronzetti Colladon and Maurizio Naldi	
A Conversation Analysis of Interactions in Personal Finance Forums . .	65
Maurizio Naldi	
Dictionaries and Specific Languages	
Big Corpora and Text Clustering: The Italian Accounting Jurisdiction Case	77
Domenica Fioredistella Iezzi and Rosamaria Berté	
Lexicometric Paradoxes of <i>Frequency</i>: Comparing VoBIS and NVdB	91
Luisa Revelli	

Emotions and Dense Words in Emotional Text Analysis: An Invariant or a Contextual Relationship?	101
Nadia Battisti and Francesca Romana Dolcetti	
Text Mining of Public Administration Documents: Preliminary Results on Judgments.	117
Maria Francesca Romano, Antonella Baldassarini, and Pasquale Pavone	
Using the First Axis of a Correspondence Analysis as an Analytic Tool	127
Bénédicte Pincemin, Alexei Lavrentiev, and Céline Guillot-Barbance	
Discursive Functions of French Modal Forms: What Can Correspondence Analysis Tell Us About Genre and Diachronic Variation?	145
Corinne Rossari, Ljiljana Dolamic, Annalena Hütsch, Claudia Ricci, and Dennis Wandel	
Multilingual Text Analysis	
How to Think About Finding a Sign for a Multilingual and Multimodal French-Written/French Sign Language Platform?	161
Cédric Moreau	
Corpus in “Natural” Language Versus “Translation” Language: LBC Corpora, A Tool for Bilingual Lexicographic Writing	167
Annick Farina and Riccardo Billero	
The Conditional Perfect, A Quantitative Analysis in English-French Comparable-Parallel Corpora	179
Daniel Henkel	
Repeated and Anaphoric Segments Applied to Trilingual Knowledge Extraction	199
Lionel Shen	
Applications	
Looking for <i>Topics</i>: A Brief Review	215
Ludovic Lebart	
Where Are the Social Sciences Going to? The Case of the EU-Funded SSH Research Projects	225
Matteo Gerli	
Topic Modeling of Twitter Conversations: The Case of the National University of Colombia	241
Eliana Sanandres, Raimundo Abello, and Camilo Madariaga	

Analyzing Occupational Safety Culture Through Mass Media Monitoring	253
Livia Celardo, Rita Vallerotonda, Daniele De Santis, Claudio Scarici, and Antonio Leva	
What Volunteers Do? A Textual Analysis of Voluntary Activities in the Italian Context	265
Francesco Santelli, Giancarlo Ragozini, and Marco Musella	
Free Text Analysis in Electronic Clinical Documentation	277
Antonella Bitetto and Luigi Bollani	
Educational Culture and Job Market: A Text Mining Approach	287
Barbara Cordella, Francesca Greco, Paolo Meoli, Vittorio Palermo, and Massimo Grasso	
Author Index	299
Subject Index	301