JADT 2022

PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT 2022
PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA
Volume 2
Michelangelo MISURACA, Germana SCEPI, Maria SPANO 2022

2022 Vadistat press in coedizione Edizioni Erranti

Associazione VADISTAT - Per Simona Balbi
Dipartimento di Scienze Economiche e Statistiche
Università di Napoli "Federico II"
Complesso Universitario Monte Sant'Angelo
Cupa Nuova Cintia 21 - Napoli

Edizioni Erranti
Editoria Indipendente - Libero Sapere
Via Caloprese 23 - Cosenza

www.edizionierranti.org
info@edizionierranti.org

Isbn 979-12-80153-31-9

Edited by
Michelangelo MISURACA
Germana SCEPI
Maria SPANO

JADT 2022

PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Naples, Italy - July 06-08, 2022)

VOLUME 2

VADISTAT PRESS
2022

The editing process of the proceedings has been actively supported by:
Alessandra Belfiore, Nicola d'Alesio, Luca D'Aniello, Agostino Gnasso

Scientific Committee

Ramón Álvarez Esteban (Univ. de León, ES)
Valérie Beaudouin (Telecom ParisTech, FR)
Mónica Bécue (Univ. Polyt. de Catalunya, ES)
Sergio Bolasco (Univ. di Roma La Sapienza, IT)
Lou Burnard (Oxford Univ., UK)
Isabella Chiari (Univ. di Roma La Sapienza, IT)
Anne Dister (Univ. Saint-Louis / UCLouvain, BE)
Jules Duchastel (UQÀM, CA)
Serge Fleury (Univ. Sorbonne Nouvelle - Paris 3, FR)
Cédrick Fairon (UCLouvain, BE)
Serge Heiden (ENS de Lyon, FR)
Domenica Fioredistella Iezzi (Univ. di Roma Tor Vergata, IT)
Margareta Kastberg (Univ de Franche Comté, FR)
Ludovic Lebart (CNRS / ENST, FR)
Jean-Marc Leblanc (Univ. de Créteil, FR)
Alain Lelu (Univ. de Franche Comté, FR)
Dominique Longrée (Univ. de Liège, BE)
Véronique Magri (Univ. Côte d'Azur / CNRS BCL, FR)
Pascal Marchand (Univ. de Toulouse, FR)
Damon Mayaffre (Univ. Côte d'Azur / CNRS BCL, FR)
Sylvie Mellet (CNRS, FR)
Michelangelo Misuraca (Univ. della Calabria, IT)
Denis Monière (Univ. de Montréal, CA)
Bénédicte Pincemin (CNRS, FR)
Céline Poudat (Univ. Côte d'Azur / CNRS BCL, FR)
Gérald Purnelle (Univ. de Liège, BE)
Elias Rizkallah (UQÀM, CA)
Pierre Ratinaud (Univ. de Toulouse, FR)
Max Reinert (CNRS / Univ. Paris-Saclay, FR)
André Salem (Univ. Sorbonne Nouvelle - Paris 3, FR)
Arjuna Tuzzi (Univ. di Padova, IT)
Mathieu Valette (Inalco, FR)
Jean-Marie Viprey (Univ de Franche Comté, FR)

Program Committee

Michelangelo Misuraca (Univ. della Calabria, IT) [CHAIR]
Enrica Amaturo (Univ. di Napoli Federico II, IT)
Laura Antonucci (Univ. di Foggia, IT)
Massimo Aria (Univ. di Napoli Federico II, IT)
Andrea Fronzetti Colladon (Univ. di Perugia, IT)
Fiorenza Deriu (Univ. di Roma La Sapienza, IT)
Giuseppe Giordano (Univ. di Salerno, IT)
Maria Gabriella Grassia (Univ. di Napoli Federico II, IT)
Francesca Greco (Univ. di Roma La Sapienza, IT)
Francesca Della Ratta (INAPP, IT)
Domenica Fioredistella Iezzi (Univ. di Roma Tor Vergata, IT)
Marina Marino (Univ. di Napoli Federico II, IT)
Mario Monteleone (Univ. di Salerno, IT)
Germana Scepi (Univ. di Napoli Federico II, IT)
Maria Spano (Univ. di Napoli Federico II, IT)
Arjuna Tuzzi (Univ. di Padova, IT)

Organising Committee

Maria Spano (Univ. di Napoli Federico II, IT) [CHAIR]
Alessandra Belfiore (Univ. della Campania Luigi Vanvitelli, IT)
Enrico Cafaro (Univ. di Napoli Federico II, IT)
Luigi Celardo (Univ. di Napoli Federico II, IT)
Luca D'Aniello (Univ. di Napoli Federico II, IT)
Nicola d'Alesio (Univ. della Campania Luigi Vanvitelli, IT)
Agostino Gnasso (Univ. di Napoli Federico II, IT)
Giorgio Infante (VADISTAT Per Simona Balbi, IT)
Rocco Mazza (Univ. di Napoli Federico II, IT)
Raffaele Mattera (Univ. di Napoli Federico II, IT)
Agostino Stavolo (Univ. di Napoli Federico II, IT)

TABLE OF CONTENTS

Introduction **pag. xvii**

Invited Speakers

Manuel J. COBO **1**
Bibliometric and science mapping analysis: models, software tools and future challenges

Stefano Maria IACUS **7**
Sentiment Analysis, Social Media and Subjective Well-Being

Max SILBERZTEIN **12**
Linguistic Resources for Corpus Processing: the ATISHS project

Contributions

Juan Abasolo, Naia Eguskiza, Aitor Iglesias **30**
Validación de léxico en lengua vasca mediante análisis lexicométrico de un corpus paralelo multilingüe

Juan Abasolo, Naia Eguskiza, Aitor Iglesias **37**
Construcción de un léxico en lengua vasca para su utilización con el software IRaMuTeQ

Suania Acampa **45**
A hybrid approach to opinion mining in Facebook's disinformative echo chambers

Touria Ait El Mekki, Bénédicte Grailles, Tsanta Randriatsitohaina **52**
Création d'une base de connaissances à partir des messageries spécialisées pour améliorer l'exploration et l'archivage des méls

Ramón Alvarez-Esteban, Mónica Bécue-Bertaut **60**
Order constrained clustering with local stopping rules. Application in textual analysis

Massimo Aria, Corrado Cuccurullo, Luca D'Aniello, Michelangelo Misuraca, Maria Spano **67**
Text Summarization of a scientific document: a comparison of Extractive unsupervised methods

Valerio Basile **74**
Towards Automatic Screening for Fibromyalgia in Italian Social Media Users

Federica Beghini **80**
Stylométrie, ADT et deep learning. Une étude de cas sur la prose romanesque de Milan Kundera

Alessandra Belfiore, Maria Spano, Katerina Mandenaki, Walter Giordano **88**
Exploring Companies' communication strategies in the Covid-19 era

Alessandra Belfiore, Agostino Gnasso, Corrado Cuccurullo, Massimo Aria **95**
AI and ML in accounting and finance: A bibliometric review
Loredana Bellantuono **102**
Complex networks and natural language processing to unfold the social enterprises environment
Pietro Belloni, Margherita Silan, Giulia Cuman **104**
Fake news spreading and sentiment of Italians during the first COVID-19 lockdown
Marion Bendinelli **111**
A Preparatory Corpus-Based Study of Swiss Touristic Brochures. Euphoric markers and Phraseology
Abdallah Benkadja, Ismaïl Biskri, Nadia Ghazzali **119**
Hybride CNN-SVM : vers une meilleure classification des données manuscrites
Valentina Biagioli, Francesca Greco, Giuseppina Spitaletta, Annachiara Liburdi, Rachele Mascolo, Orsola Gawronski, Riccardo Ricci, Emanuela Tiozzo, Ercole Vellone, Teresa Grimaldi Capitello, Michele Salata, Massimiliano Raponi, Immacolata Dall'Oglio **127**
Emotional Text-Mining in Pediatria: esplorazione del self-care nei bambini e nei giovani in condizioni croniche complesse
Magali Bigey **134**
Impact of the expression of feelings and emotions in fake news and their reception: Specific electronic dictionaries and automatic recognition with NooJ
Marco Bolpagni, Marco Broglio, Andrea Innocenzi, Tommaso Ulivieri **141**
EMOtivo: a classifier for emotion detection of Italian texts trained on a self-labelled corpus
Barbara Boschetto, Francesca Della Ratta, Federica Pintaldi, Maria Elena Pontecorvo, Alessia Sabbatini **148**
"Altro specificare". Effetto del Covid-19 sulla codifica dei campi "altro" nell'indagine Istat sulle forze di lavoro
Rosalie Bourdages, Roxane Gagnon, Denis Foucambert **156**
Évaluation et description de récits oraux spontanés d'élèves du primaire : pour l'utilisation des méthodes mixtes en sciences de l'éducation
Camille Bouzereau **164**
Complémentarité du quantitatif et du qualitatif pour l'analyse de données textuelles. Analyse du discours du Front National (2000-2017)
Giulia Bracco, Maria De Martino, Alessandro Laudanna **172**
Semantic and morphological properties of gender assignment in Italian: distributional data on common nouns

Andrea Briglia, Massimo Mucciardi, Giovanni Pirrotta **180**
The development of word frequency distribution in first language acquisition. An analysis on a spoken language corpus of French children
Joanna Byszuk, Quinn Dombrowski **188**
Stylometric investigations into translationese: The Baby-Sitters Club across languages
Stefania Capogna, Giulia Cecchini, Maria Chiara De Angelis, Francesca Greco, Emanuela Proietti **197**
Analisi comparativa dei rapporti nazionali: il caso del Progetto Erasmus+ ECOLHE
Rosanna Cataldo, Gabriella Punziano, Carmine Barricelli, Gabriele Luciano, Barbara Saracino, Ferdinando Iazzetta **204**
Three main dimensions in the construction of the expert during the Italian Covid-19 vaccination campaign: positional, reputational, and communicational spheres in comparison
Elena Catanese, Monica Scannapieco, Mauro Bruno, Luca Valentino **213**
The Italian Social Mood on Economy Index: recent methodological developments
Radek Čech, Miroslav Kubát, Ján Mačutek, Michaela Koščová **221**
Does an author leave a syntactic footprint ?
Roy Cerqueti, Valerio Ficcadenti, Gurjeet Dhesi, Marcel Ausloos **229**
A methodological strategy for assessing the Markovian stochastic structure of the text analysis-based rank-size laws
Marie Chandelier, Sascha Diwersy **234**
Analyse du profil sémantique des expressions biodiversité et diversité biologique dans la presse écrite française
Giacomo Chiara, Diego Romaioli, Alberta Contarello **243**
Rappresentazioni sociali della migrazione e dell'identità culturale in testi letterari di scrittrici e scrittori di origine africana
Barbara Cordella, Alessandro Gennaro **252**
Alla ricerca del benessere: le famiglie adottive
Rosario D'Agata, Domenico De Stefano, Francesco Santelli **259**
The 'words' of no green pass communities on twitter. A two-mode semantic network analysis
Angela Maria D'Uggento, Albino Biafora, Claudia Marin, Fabio Manca, Massimo Bilancia **266**
Sentiment analysis of emotions in social media during the COVID-19 pandemic
Giovanni De Gasperis, Pasquale Pavone, Sergio Bolasco **274**
La risorsa di Italiano Standard ad alta variabilità linguistica per misurare la peculiarità di un corpus
Alberto Maria De Mascellis, Maria Gabriella Grassia, Agostino Stavolo **282**
Topic modeling and sentiment analysis: an analysis of tweets about the game CyberPunk 2077

Francesca Della Ratta, Monya Ferritti **290**
Gli insegnanti dopo l'anno della didattica a distanza
Alessandro Delmonte, Matteo Farné **298**
Evaluating customer satisfaction through Amazon review analysis : the bluetooth earphones example
Vanina Deneux Le Barh **306**
Du discours sur le travail avec les chevaux aux valeurs communes des professions anthropoéquines
Francesco Paolo Di Candia, Nicoletta Roberto, Domenica Fiordistella Iezzi **314**
Italian Institutional communication in pandemic period: a chronological analysis of Prime Minister speeches
Sami Diaf, Florian Schütze **321**
Estimating Uncertainty in Narrative Economics Using Latent Semantic Scaling
Sami Diaf, Rachid Toumache **328**
Cross-lingual Macroeconomic Narratives: Different Views Of The Same Problem ?
Anne Dister, Hubert Naets, Patrick Watrin **334**
Présence des noms d'agents au féminin dans un corpus de presse. Étude comparée Québec-Belgique
Sascha Diwersy, Ivana Didirková, Christelle Dodane, Fabrice Hirsch, Giancarlo Luxardo **340**
Les temps de la crise sanitaire au prisme d'une série chronologique : une étude phonético-textométrique
Alexandre Dniestrowski, Louis Allard **348**
The "Augmented" Historian
Chuanming Dong, Philippe Gambette, Catherine Dominguès **354**
Extraction et caractérisation de noyaux d'événements liés à la pollution industrielle
Maximiliano Duran **361**
Automatic retrieving of derived Quechua verbs
Maciej Eder **369**
Improving the performance of word frequencies in authorship attribution
Louis Escouflaire, Antonin Descampe, Cédrick Fairon **377**
L'évolution de la subjectivité linguistique dans le journalisme web du XXIe siècle : analyse d'un corpus belge francophone d'articles de 2010 à 2021
Shira Fano, Gianluca Toschi **385**
The economy and the Web: RETI a regional economic twitter index
Manuel Favaro, Marco Biffi, Simonetta Montemagni **392**
Trattamento automatico del linguaggio e varietà storiche di italiano: la sfida della lemmatizzazione

Marco Ferracci, Germana Scepi, Maria Spano, Michelangelo Misuraca **400**
Nodi discorsivi e processi di progettazione condivisa: Milano e Torino
Rosa Ferri, Valentina Bua, Marzia Curia, Alessia D'Avack, Antonino Firetto, Federica Mauro, Francesca Greco **408**
Text Mining e Salute Organizzativa: il caso di Roma Capitale
Marzia Freo, Alessandra Luati **414**
Lasso-based variable selection methods in text regression: the case of short texts
Andrea Fronzetti Colladon **422**
Being known, being known for something and being good: A new textual measure of organizational reputation
Augusto Gamuzza, Francesca Greco, Anna Maria Leonora **424**
Text mining e analisi ermeneutica: il caso delle pratiche di volontariato internazionale
Franco M. T. Gatti **432**
Rehearsals of ideological placement on a corpus of parliamentary questions
Giuseppe Giordano, Maria Carmela Catone **441**
Islands and structural holes for mapping topics in scientific field. An application to Social Research Methodology
Maria Gabriella Grassia, Marina Marino, Rocco Mazza, Agostino Stavolo **446**
Analysis of the public debate on DDL Zan on Twitter: an application of the Structural Topic Model
Deborah Grbac **453**
From meta-data to research-data. How to use text analysis techniques to do Information retrieval as a propaedeutic practice to Data literacy
Francesca Greco, Raffaella Gallo, Alessandro Polli, Fiorenza Deriu **461**
A comparison between machine learning and non-supervised techniques using the GP Index
Francesca Greco, Gevisa La Rocca, Giovanni Boccia Artieri **468**
Sentiment analysis dei corpora multilingua: La percezione pubblica del Covid-19 in Europa
Francesca Greco, Alessandro Polli, Federico Siciliano **475**
Leveraging Deep Learning models to assess the temporal validity of Emotional Text Mining method
Massimo Guarino, Violetta Simonacci, Michele Gallo **482**
Linguistic complexity and engagement in social media communication for cultural heritage

Domenica Fioredistella Iezzi, Roberto Monte **489**
Sales forecast and electronic word of mouth: the power of feelings
Julie-Pier Jean, Sylvia Kasparian **495**
Cree Legends and Narratives au prisme d'IRaMuTeQ
Kristina Kocijan, Krešimir Šojat **503**
Who is Guilty and Who is Responsible in the Croatian Parliament: A Linguistic Approach
Tomasz Kruszewski, Joanna Michalak **511**
Dynamic and content of COVID-19 related tweets during early stage of pandemic
Daniela Laricchiuta, Andrea Termine, Carlo Fabrizio, Noemi Passarello, Francesca Greco, Fabrizio Piras, Eleonora Picerni, Debora Cutuli, Andrea Marini, Laura Mandolesi, Gianfranco Spalletta, Laura Petrosini **519**
The affecting meaning of words: multidimensional analysis of emotional communication
Sara Laurita, Rossella Nicoletti, Alfredo Fortunato, Carmelofrancesco Origlia **527**
Framing the presence of tourism policies in the Calabrian public agenda during Covid-19 crisis: analysis of online news from local newspapers using 3-wave datasets
Ludovic Lebart **536**
Poems and Songs. What can Textual Data Analysis do?
Virginie Lethier, Emilie Née, Hugo Dumoulin **544**
Caractériser les logiques disciplinaires et institutionnelles dans les rapports d'autoévaluation
Daniele Licari, Maria Francesca Romano, Giovanni Comandé **552**
Automatic Anonymization of Italian Legal Textual Documents using Deep Learning
Rosaria Lombardo, Matteo Pascoli, Nicola Grandi **560**
Text Mining and Variants of Correspondence Analysis for analysing written Italian of University students
Carlos Maciel, Damon Mayaffre, Laurent Vanni **568**
Corpus non alignés et ADT. Essai de comparaison entre les présidents français et brésiliens de l'ère contemporaine
Marina Marino, Rocco Mazza, Gabriele Luciano **576**
Online communication and political participation: a study dedicated to the referendum Cannabis Legale
Raffaele Mattera, Michelangelo Misuraca, Germana Scepi, Maria Spano **584**
Clustering of financial time series: a bibliometric analysis

Damon Mayaffre **591**
Contribution à l'histoire d'un concept : occcurrences et cooccurrences de 'souveraineté' dans les discours de Charles de Gaulle (1958-1969) et d'Emmanuel Macron (2017-2022)
Simona Mercurio **599**
What about corruption? A text analytics method for literature review
Silvia Monaco, Anna Cortellino, Francesca Greco, Michela Di Trani **607**
Using Emotional Text Mining to explore the Spanish culture of Organ Donation
Cyrielle Montrichard **613**
Questionner la posture énonciative d'un journal à visée humoristique: étude textométrique et contrastive
Jean Moscarola, Florence Laval, Caroline Mothe **621**
Recherche Qualitative : Analyse lexicale et sémantique de grands corpus d'entretiens ou de documents
Alejandro J. Napolitano Jawerbaum, Patrick Juola **629**
Co-authorship in Carmen Mola, a case study
Luisa Orrù, Marco Cuccarini, Christian Moro, Monia Paita, Davide Bassi, Giovanni Da San Martino, Nicolò Navarin, Gian Piero Turchi **636**
From fake news identification to news persuasion level study: the application of a machine learning model with the M.A.D.I.T. methodology
Luisa Orrù, Jessica Neri, Marco Cuccarini, Antonio Iudici, Nicolò Navarin, Gian Piero Turchi **644**
Biographical change and Machine Learning with MADIT: a predictive model proposal for data and intervention analysis
Aylin Pamuksaç **651**
Similarities and disparities among French modal verbs through co-occurrence correspondence analysis
Daniel Pélissier, Jérôme Bousquié **659**
Le temps dans le discours, expérimentation d'un protocole d'observation des caractéristiques temporelles d'un corpus d'avis de salariés
Manon Pengam, Agata Jackiewicz, Pascal Marchand **667**
Saisir par la textométrie la notion de radicalisation dans des corpus institutionnels hétérogènes
Marie Pérès, Jean-Marc Leblanc **675**
Réalité virtuelle: quels apports pour la visualisation en textométrie/lexicométrie ?
Loredana Piervisani, Mariachiara Figura, Paola Arcadi, Angela Durante, Ercole Vellone, Rosaria Alvaro **683**
Vocazione e professione: l'imperativo morale dell'infermiera italiana nel periodo fascista

Bénédicte Pincemin, Serge Heiden, Franck Mazuet **691**
The Textometric Concept of Active Corpus. Illustration by an Analysis Scenario based on Annotation then Projection
Sophie Piron **699**
Stylométrie des Elemens de la grammaire Françoise (1780) ou comment Lhomond a supplanté ses prédécesseurs
Lucia Ráčková **708**
Les erreurs interférentielles et le portfolio linguistique du locuteur, deux variables corrélées par le rho de Spearman
David Reymond, Manuel Durand-Barthez, Heman Khouilla, Sandrine Wolff **712**
Retrouver l'inventeur-auteur: la levée d'homonymies d'autorat entre les brevets et les publications scientifiques
Annabel Richeton, Virginie Lethier, Margareta Kastberg Sjöblom **720**
La stabilité du cadre discursif du Ministère des Affaires Etrangères à l'épreuve de la textométrie
Valentina Rizzoli, Andrea Azzoni, Giulia Rivellini, Arjuna Tuzzi **728**
Classificazione manuale vs automatica dei resoconti delle chiamate ricevute da Telefono Amico Italia (2016 -2020)
Valentina Rizzoli, Laura Soledad Norton, Alessandro Meneghini, Mauro Sarrica **736**
How far is the risk? Detecting socio-psychological processes from consciousness to denial in social media exchanges
Valentina Rizzoli, Matilde Trevisani, Arjuna Tuzzi **744**
Life cycle of ideas in the Journal of Personality and Social Psychology: A history of US social psychology
Minerva Rojas **751**
Analyse contrastive interlangue des marqueurs discursifs en Français Langue Étrangère et en Français Langue Maternelle
Maria Francesca Romano, Giovanni De Gasperis, Pasquale Pavone, Sergio Bolasco **758**
Potenzialità di TaLTaC nella anonimizzazione di sentenze della magistratura
Roberto Rondinelli, Stefano Marmani, Valerio Ficcadenti **765**
An analysis of the Jesus' community in the Gospel of Matthew, Mark, Luke, John, and Acts of Apostles networks
Corinne Rossari, Cyrielle Montrichard, Claudia Ricci, Linda Sanvido **772**
Subjectivity and connectors: study of the parameters of discourse genres and languages
Simona Ruggia, Minerva Rojas **780**
Les marqueurs discursifs : étude comparative de leur utilisation et fréquence chez les locuteurs francophones et dans les manuels de français langue étrangère

Camilla Salvatore, Silvia Biffignandi, Annamaria Bianchi **787**
Assessing Corporate Social Responsibility Communication on Twitter: A Composite Indicator Approach
Sara Santilli, Stefano Sbalchiero **794**
The use of IRaMuTeQ for longitudinal analysis. A study on postgraduate course participants' reflections..
Andrea Sciandra **801**
Dimensionality Reduction of Unstructured and Network Data for Stance Detection
Mariangela Sciandra, Alessandro Albano **809**
A two-stage LDA algorithm for ranking induced topic readability
Ludovica Segneri, Andrea Fronzetti Colladon, Claudia Fabiani, Anna Laura Pisello **815**
Text mining for socially responsible behavior: fostering the adoption of green technologies
Gerardo Sierra, Juan-Manuel Torres-Moreno, Mercè Lorente **817**
Exploration de contextes définitoires en français : cas d'étude sur la COVID-19
Andrea Simonetti, Nicoletta D'Angelo, Giada Adelfio **827**
Marked Hawkes processes for Twitter data
Valérie Thon, Laurent Vanni, Dominique Longrée **834**
Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison: à propos de l'étude stylistique des lettres de Pierre Damien
Alice Tontodimamma, Stefano Anzani, Valerio Basile, Marco Stranisci, Lara Fontanella **842**
Comparison of two annotation schemes to derive offensiveness scores in HurtLex
Gian Piero Turchi, Luisa Orrù, Annalisa Trovò, Christian Moro, Antonio Iudici, Nicolò Navarin **850**
Introducing Dialogic Process Analysis in Natural Language Processing
Laurent Vanni, Magali Guaresi, Véronique Magri **858**
Convolution et marqueurs multidimensionnels. Description des représentations genrées dans un corpus de films français
Pierre Wavresky **866**
Crise du Covid et ressenti des acteurs du secteur social et médico-social : une analyse textométrique de journaux de bord et d'entretiens
Emma Zavarrone, Alessia Forciniti, Martha Friel **874**
Exploratory textual network for Italian Third Mission
Emma Zavarrone, Alessia Forciniti, Marta Muscariello **879**
SustDict: a customized lexical-based dictionary for CSR
Lichao Zhu, Fabrice Issac, Touria Ait El Mekki, Olivier Hû **887**
Construction of an encyclopedic dictionary of tourism from scientific corpus

Leveraging Deep Learning models to assess the temporal validity of Emotional Text Mining method

Francesca Greco, Alessandro Polli, Federico Siciliano

Sapienza Università di Roma – francesca.greco@uniroma1.it
alessandro.polli@uniroma1.it
federico.siciliano@uniroma1.it

Abstract
The paper present the training of a surrogate model to mimic the functioning of a text mining social profiling method, Emotional Text Mining, applied on the Italian tweets containing #Covid-19 during the first lockdown. This surrogate model is based on state-of-the-art Deep Learning models in the field of Natural Language Processing (NLP). These models feature an architecture called Transformer, which adopts a mechanism of self-attention that weighs the significance of each part of the input data. The model was trained on the tweets and factors of an initial time period to predict the polarisation of a tweet according to each factor found using the ETM technique. The model was then applied to subsequent periods. Using appropriate metrics, it was then possible to quantify the need to re-run the technique from scratch on the following time periods, while also being able to assess which factors still remained relevant over time.

Keywords: Natural Language Processing, Emotional Text Mining, Neural Network, Machine Learning, Covid-19, Social Media

Riassunto
Il lavoro presenta l'addestramento di un modello surrogato per imitare il funzionamento di un metodo per la profilazione sociale, l'Emotional Text Mining, applicata sui tweet italiani che contenevano #Covid-19 prodotti durante il primo lockdown. Questo modello surrogato è basato su modelli di Deep Learning all'avanguardia nel campo del Natural Language Processing (NLP). Questi modelli sono caratterizzati da un'architettura chiamata Transformer, che adotta un meccanismo di auto-attenzione che pesa l'importanza di ogni parte dei dati di input.
Il modello è stato addestrato sui tweet e i fattori di un primo periodo di tempo per prevedere la polarizzazione di un tweet in base a ciascun fattore trovato con la tecnica ETM. Il modello è stato poi applicato ai periodi successivi. Utilizzando metriche appropriate, è stato poi possibile quantificare la necessità di rieseguire la tecnica da zero sui periodi di tempo successivi, potendo anche valutare quali fattori sono rimasti rilevanti nel tempo.

Parole chiave: Elaborazione del linguaggio naturale, Emotional Text Mining, Rete Neurale, Machine Learning, Covid-19, Social Media

1. Introduzione

L'ampia diffusione di Internet e la nascita dei Social media, ha permesso alle persone di cercare e condividere informazione su base giornaliera e rapidamente. Questa grande quantità di materiale generato dagli utenti può anche essere consultato facilmente e a basso costo; per esempio, il famoso social media Twitter fornisce API che permettono ai ricercatori di accedere al loro enorme archivio di dati. Il contenuto dei tweets può quindi essere elaborato utilizzando una metodologia per la profilazione sociale come l'Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2020) per identificare il sentiment degli utenti dei social media su un argomento. Poiché le opinioni, le emozioni ed i sentimenti mutano nel tempo, diventa necessario applicare il metodo a più riprese per diversi periodi di tempo. Sfortunatamente, però, questo processo è computazionalmente molto oneroso perché è richiesto applicarlo a grandi volumi di dati integralmente. In questo articolo, prendiamo spunto dalle recenti scoperte nel campo del Deep Learning per cercare di (i) stimare quanto sia valido il risultato della tecnica dell'ETM in periodi successivi a quello su cui è stata utilizzata e (ii) quando si rende necessario ripetere l'analisi per intero.

2. Lavori correlati

2.1. Emotional Text Mining

L'utilizzo di tecniche di text mining per misurare opinioni, sentimenti ed emozioni degli attori sociali è molto diffuso e si basa fondamentalmente su due approcci: quello semantico, in cui le parole vengono considerate solo per il loro significato esplicito, e quello semiotico, in cui le parole vengono considerate come simboli dei modi di pensare delle persone. Quest'ultimo offre il vantaggio non indifferente di connettere la comunicazione con il comportamento, effettuando di fatto una profilazione a partire dalla comunicazione. L'Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2020) adotta un approccio semiotico al testo e si basa su un modello socio-costruttivista ad orientamento psicodinamico che considera le emozioni non come delle sensazioni (correlato biologico) ma come l'espressione del funzionamento inconscio e sociale della mente (Carli, 1990; Greco, 2016). In tal senso l'ETM si configura come un metodo per la profilazione sociale poiché identifica le categorie simbolico-culturali che organizzano il comportamento, le sensazioni, le aspettative, gli atteggiamenti e la comunicazione degli attori sociali. Di conseguenza, l'ETM non guarda al contenuto esplicito della comunicazione ma alla sua modalità, vale a dire quali parole vengono selezionate e come vengono associate.

Mentre il funzionamento mentale procede dal livello semiotico a quello

semantico nel generare il testo, la procedura statistica simula il processo inverso del funzionamento mentale, dal livello semantico a quello semiotico. Per questo motivo, l'ETM esegue una sequenza di procedure di sintesi, dal pretrattamento del testo e la selezione dei termini, all'analisi multivariata, per identificare il livello semiotico, partendo da quello semantico (Greco and Polli, 2020). È una procedura automatica, veloce e relativamente semplice, che può essere utilizzata per estrarre informazioni in merito a quelle che potremmo definire le opinioni emotivamente guidate. A differenza di altre tecniche di sentiment analysis non solo misura il sentiment ma fornisce anche informazioni qualitative sullo stesso.

2.2. Natural Language Processing

L'elaborazione del linguaggio naturale (NLP) vede l'unione dei campi della linguistica e dell'informatica, focalizzandosi su come programmare i computer per analizzare grandi quantità di dati in linguaggio naturale. Sono vari i task specifici che ricadono all'interno di questo obiettivo più generale, come quello di categorizzare dei documenti, controllare se una notizia sia falsa o capire che sentimento l'autore stia esprimendo. Negli ultimi anni, il campo del NLP ha visto l'emergere di metodi di apprendimento basati su reti neurali, in particolare quelli appartenenti alla categoria del Deep Learning. Questi modelli, costituiti da architetture grandi e complesse, fanno uso di grosse quantità di dati e hanno dimostrato di riuscire a raggiungere risultati all'avanguardia in molti compiti del linguaggio naturale.

I primi modelli di rete neurale applicati a NLP erano basati su strutture di tipo ricorsivo (Recurrent Neural Network - RNN); queste ricevono i dati in maniera sequenziale, elaborando i dati ricevuti ad ogni istante e mantenendo un aggregato che riassume l'informazione passata. Nonostante vari miglioramenti, le RNNs sono state superate nell'elaborazione del linguaggio con i Transformers. Questi modelli adottano quello che viene definito Meccanismo di Attenzione (Vaswani et al., 2017), riescono cioè a guardare la frase in ingresso nella sua interezza ponderando il significato di ogni parola della sequenza. Questa caratteristica, oltre a far raggiungere risultati notevoli, permette una maggiore parallelizzazione rispetto alle RNN e riduce quindi i tempi di allenamento. Questo ha aiutato lo sviluppo di sistemi pre-addestrati come BERT (Devlin et al., 2019), allenati su grandi dataset linguistici, quali Wikipedia, e che possono essere messi a punto per compiti specifici.

3. Metodi

La tecnica di Emotional Text Mining permette di effettuare una profilazione degli utenti dei media attraverso i loro post. Tuttavia, questa suddivisione in diversi gruppi di profili è dipendente dal tempo, poiché l'esecuzione della procedura su un altro periodo di tempo può portare a risultati diversi. La procedura è però

computazionalmente costosa, e ciò non permette di eseguirla agilmente per diversi periodi di tempo. Per superare questo ostacolo, abbiamo addestrato un modello surrogato, basato su Transformer, per imitare il funzionamento della tecnica ETM. Abbiamo inoltre valutato e impiegato diverse metriche per quantificare la necessità di eseguire la tecnica da zero sui periodi di tempo successivi.

3.1. Modello

Sono stati raccolti tutti i tweet in lingua italiana contenenti la parola "coronavirus" successivamente a tre importanti conferenze stampa del Capo di Governo durante la prima fase di lockdown per un totale di 1.884.670 messaggi costituiti dall'81% di retweet. I 6 periodi temporali, nell'anno 2020, corrispondono a: 4-11 Marzo, 24-31 Marzo, 10-14 Aprile, 30 Aprile - 9 Maggio, 17-26 Maggio, 3-12 Giugno. Il primo sample dei dati (numero di tweets = 80.519) è stato analizzato con ETM (Boccia Artieri et al., 2021a; 2021b), per poi essere usato come input della rete neurale. In breve, l'ETM esegue un algoritmo di clustering sulla matrice termini-documenti e un'analisi delle corrispondenze sulla matrice cluster-termini, producendo un insieme di fattori, ognuno caratterizzato da un insieme di termini i quali ne danno il significato, e un insieme di cluster, posizionati nello spazio dei fattori. Per ogni tweet abbiamo perciò un cluster di appartenenza e perciò i fattori più significativi per esso. Maggiori informazioni possono essere trovate sul relativo articolo (Boccia Artieri et al., 2021a; 2021b).

Per strutturare la nostra rete neurale, siamo partiti da un modello pre-allenato basato su BERT, Neuraly[1], basato su una versione di BERT per l'italiano, Italian BERT[2], e messo a punto per effettuare sentiment analysis su un dataset di tweet in italiano, un task simile al nostro. La prima parte del nostro modello perciò genera la rappresentazione vettoriale della frase ricevuta in ingresso. Abbiamo poi aggiunto un'altra porzione di rete neurale specifica per il nostro task. L'output, viene poi passato a tre strutture parallele ognuna responsabile della predizione di un singolo fattore. Per ogni fattore i 3 neuroni di output ci danno una probabilità, rispettivamente quella dell'input di essere più vicino al polo negativo, al polo positivo o specificatamente a nessuno dei due poli di quel fattore.

3.2. Metriche

Per capire quanto il modello funzioni adeguatamente in periodi diversi da quello su cui è stato allenato, abbiamo selezionato delle metriche ad-hoc. Le prime 2 sono applicate separatamente per ogni singolo fattore, considerando le probabilità di essere polarizzati negativamente, neutralmente o positivamente.

La prima metrica, chiamata Max-Value, consiste nel prendere per ogni tweet il

[1] https://huggingface.co/neuraly/bert-base-italian-cased-sentiment
[2] https://huggingface.co/dbmdz/bert-base-italian-cased

massimo di questi 3 valori, per poi effettuare una media. In questa maniera, vogliamo valutare l'incertezza del modello: tanto più questa metrica tende al 100%, tanto più il modello sarà certo della sua predizione, tanto più tende al 33%, tanto più il modello sarà incerto tra i 3 output possibili.
La seconda metrica, chiamata Poles-Diff, calcola per ogni tweet la differenza, in valore assoluto, tra le probabilità di essere polarizzato positivamente e negativamente, per poi mediare i risultati. Vogliamo così valutare quanto il modello consideri i tweets polarizzati secondo ogni fattore. Più il valore tende al 100%, più l'opinione delle persone è polarizzata, mentre più tende allo 0% e meno quel fattore è in grado di esprimere l'emozione del pubblico.
L'ultima metrica, Four-Frequency, prende invece in considerazione la clusterizzazione dei tweets; per ogni tweet, viene usato il modello per predire una polarizzazione per ogni fattore, scegliendo quella con più alta probabilità. I tweet vengono quindi divisi in cluster in base ai valori ottenuti per ogni fattore, e selezioniamo i 4 cluster più presenti. Più i 4 valori si avvicinano al 25%, più l'opinione del pubblico è equamente descritta dai cluster trovati. Se uno dei 4 valori tende verso il 100%, più è probabile che quel cluster sia caratterizzato da un fattore per ora non considerato.

4. Risultati

Il modello di Deep Learning è stato allenato sul primo periodo di dati (4-11 Marzo 2020). Questi dati, sono stati divisi in set di training (72%), quello su cui è stato effettivamente allenato il modello, set di validation (18%), utilizzato per scegliere quando fermare l'allenamento della rete neurale, e set di test (10%), per valutare la performance del modello. Dopo aver allenato il modello, questo viene utilizzato per effettuare le predizioni per i tweets appartenenti ai periodi temporali seguenti, per poi calcolare le metriche da noi scelte.
Sul dataset di test, si raggiunge un'accuratezza del 93.7%. Questo ci suggerisce che il modello è in grado di replicare in maniera eccellente la tecnica di ETM per il primo periodo temporale.

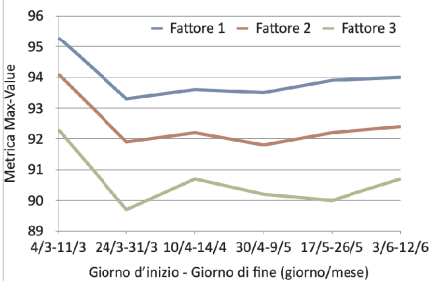

Figura 1. Valori metrica Max-Value per ogni fattore nei 6 periodi temporali considerati

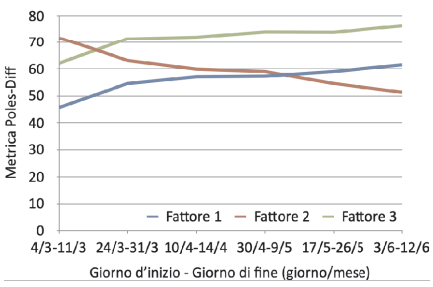

Figura 2. Valori metrica Poles-Diff per ogni fattore nei 6 periodi temporali considerati

La figura 1 mostra, per ogni fattore, il valore della prima metrica Max-Value. Possiamo constatare come i valori siano in ordine decrescente rispetto al numero del fattore; questo è da aspettarsi in quanto, in base alla tecnica di ETM, sappiamo il primo fattore è più importante del secondo, che è più importante del terzo, per spiegare la clusterizzazione. Notiamo inoltre una diminuzione netta nel secondo periodo, che potrebbe rappresentare la comparsa di un nuovo fattore, che non possiamo vedere se non riapplicando la tecnica di ETM.
La figura 2 mostra la metrica Poles-Diff calcolata per ogni fattore. Per il primo e il terzo fattore la metrica aumenta, indicando che i tweet risultano più polarizzati secondo essi. Per il secondo fattore vediamo invece come diventi meno importante per la categorizzazione; possiamo ipotizzare che un nuovo fattore stia comparendo e lo stia andando a sostituire.
La figura 3 mostra la frequenza percentuale per i 4 clusters più presenti nelle predizioni della rete. È interessante notare come, a partire dal secondo periodo, i primi 3 cluster mantengano all'incirca la stessa frequenza, decrescendo lentamente nel tempo, mentre il quarto cluster aumenti a loro discapito, andando a superare il 40%. Il quarto cluster rappresenta il sentimento riguardo il pericolo del Covid a livello nazionale. Ipotizziamo che l'aumento in frequenza sia collegato al fatto che dopo Marzo il pericolo è ormai nei confini nazionali.

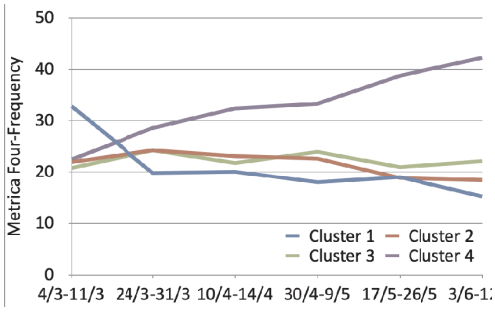

Fig 3. Valore metrics Four-Frequency per i 6 periodi temporali considerati

5. Conclusioni

Abbiamo mostrato come, con l'ausilio di un modello di Machine Learning, sia possibile imitare il funzionamento della tecnica di ETM. Grazie alle predizioni effettuate con tale modello e all'uso sapiente di metriche selezionate ad-hoc, siamo riusciti a valutare il cambiamento di opinione del pubblico in periodi temporali successivi a quello su cui è stata applicata la tecnica di ETM. Già nel secondo periodo, possiamo vedere un cambio repentino nell'espressione dei fattori, indicativo della capacità del primo e terzo fattore di rappresentare le dimensioni di senso presenti nella comunicazione e della decrescente rappresentatività del secondo fattore. Ciò suggerisce che a partire dalla fine di

marzo si venga a costituire un nuovo fattore peraltro confermato da analisi condotte in studi precedenti (Boccia Artieri et al., 2021a; 2021b). Infine, i risultati ottenuti suggeriscono come l'ETM potrebbe essere utilizzato come metodo per effettuare una prima classificazione automatica del testo per costituire il learning set. Infatti, la decodifica manuale del learning set può risultare onerosa in termini di tempi e costi nonché comportare problemi di classificazione in relazione alla grandezza dello stesso.

References

Boccia Artieri G., Greco F. and La Rocca, G. (2021). The flame of the Coronavirus in Italy. Looking for fear arousing appeal during the media hype period on Twitter. International Review of Sociology, 31(2), 287-309, https://doi.org/10.1080/03906701.2021.1947950

Boccia Artieri G., Greco F. and La Rocca, G. (2021). Lockdown and Breakdown in Italians' Reactions on Twitter during the First Phase of Covid-19. PACO, 14(1), 261-282. https://doi.org/10.1285/i20356609v14i1p261

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Giuliano L. and La Rocca G. (2010). *Analisi automatica e semi-automatica dei dati testuali: Strategie di ricerca e applicazioni* (vol. II). Led.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.

Greco F. and Polli A. (2020). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management,* 51: 101934, https://doi.org/10.1016/j.ijinfomgt.2019.04.007.

Lebart L. and Salem A. (1994). *Statistique Textuelle.* Dunod.

Steinbach M., Karypis G. and Kumar V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, vol. 400, pp. 525–526. Boston.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.