

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

PROPRIETÀ LETTERARIA RISERVATA

Copyright 2018 - UniversItalia - Roma

ISBN 978-88-3293-137-2

A norma della legge sul diritto d'autore e del codice civile è vietata la riproduzione di questo libro o di parte di esso con qualsiasi mezzo, elettronico, meccanico, per mezzo di fotocopie, microfilm, registratori o altro. Le fotocopie per uso personale del lettore possono tuttavia essere effettuate, ma solo nei limiti del 15% del volume e dietro pagamento alla SIAE del compenso previsto dall'art. 68, commi 4 e 5 della legge 22 aprile 1941 n. 633. Ogni riproduzione per finalità diverse da quelle per uso personale deve essere autorizzata specificatamente dagli autori o dall'editore.

Program Committee

Ramón Álvarez Esteban: Univ. of León, E
Valérie Beaudouin: Telecom ParisTech, F
Mónica Bécue: Poly. Univ. of Catalunya, E
Sergio Bolasco: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
François Daoust, UQÀM, Montreal, CDN
Anne Dister, FUSL, Bruxelles / UCL, Louvain, B
Jules Duchastel: UQÀM, Montreal, CDN
Serge Fleury: Univ. Paris 3, F
Cédric Fairon: UCL, Louvain, B
Luca Giuliano: Sapienza Univ. of Rome, I
Serge Heiden, ENS, Lyon, F
Domenica Fioredistella Iezzi, Univ. of Tor Vergata, I
Margareta Kastberg, Univ. of Franche Comté, F
Ludovic Lebart: CNRS / ENST, Paris, F
Jean-Marc Leblanc: Univ. of Créteil, F

Alain Lelu: Univ. of Franche Comté, F
Dominique Longrée, Univ. of Liège, B
Véronique Magri: Univ. of Nice Sophia-Antipolis, F
Pascal Marchand: Univ. of Toulouse, F
William Martinez: Univ. of Lisboa, P
Damon Mayaffre: CNRS, Nice, F
Sylvie Mellet: CNRS, Nice, F
Michelangelo Misuraca: Univ. of Calabria, I
Denis Monière: Univ. of Montréal, CDN
Bénédicte Pincemin: CNRS, Lyon, F
Céline Poudat: Univ. of Nice Sophia-Antipolis, F
Pierre Retinaud: Univ. of Toulouse, F
André Salem: Univ. Paris 3, F
Monique Slodzian: Inalco, F
Arjuna Tuzzi: Univ. of Padua, I
Mathieu Valette: Inalco, F

Organising Committee

Domenica Fioredistella Iezzi: Univ. of Tor Vergata, I
Sergio Bolasco: Sapienza Univ. of Rome, I
Livia Celardo: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
Francesca della Ratta: ISTAT, I
Fiorenza Deriu: Sapienza Univ. of Rome, I
Francesca Dolcetti: Sapienza Univ. of Rome, I

Andrea Fronzetti Colladon: Univ. of Tor Vergata, I
Francesca Greco: Sapienza Univ. of Rome, I
Isabella Mingò: Sapienza Univ. of Rome, I
Michelangelo Misuraca: Univ. of Calabria, I
Arjuna Tuzzi: Univ. of Padua, I
Maurizio Vichi: Sapienza Univ. of Rome, I
Francesco Zarelli: ISTAT, I

Local Organisation

Francesco Alò, Giulia Giacco,
Paolo Meoli, Vittorio Palermo, Viola Talucci

Table of contents

Introduction XVII

Acknowledgements XIX

Invited Speakers

GERMAN KRUSZEWSKI

Memorize or generalize? Searching for a compositional RNN in a haystack

Adam Liška XXIII

BING LIU

Scaling-up Sentiment Analysis through Continuous Learning XXIV

PASCAL MARCHAND

La textométrie comme outil d'expertise :

application à la négociation de crise. XXV

GEORGE K. MIKROS

Author Identification Combining Various Author Profiles. Towards a Blended

Authorship Attribution Methodology XXVI

ROBERTO NAVIGLI

From text to concepts and back: going multilingual

with BabelNet in a step or two XXVII

Contributors

MOTASEM ALRAHABI¹, CHIARA MAINARDI¹

Identification automatique de l'ironie et des formes apparentées dans un

corpus de controverses théâtrales 1

MOHAMMAD ALSADHAN, SASCHA DIWERSY,

AGATA JACKIEWICZ, GIANCARLO LUXARDO

Migrants et réfugiés : dynamique de la nomination de l'étranger 10

R. ALVAREZ-ESTEBAN, M. BÉCUE-BERTAUT, B. KOSTOV, F. HUSSON, J-A

SÁNCHEZ-ESPIGARES

Xplortext, a R package. Multidimensional statistics for textual data science. 19

ELENA, AMBROSETTI, ELEONORA MUSSINO, VALENTINA TALUCCI

L'evoluzione delle norme: analisi testuale delle politiche sull'immigrazione in

Italia 26

MASSIMO ARIA, CORRADO CUCCURULLO

A bibliometric meta-review of performance measurement, appraisal, management research 35

LAURA ASCONE

Textual Analysis of Extremist Propaganda and Counter-Narrative: a quantitative investigation 44

LAURA ASCONE, LUCIE GIANOLA

Analyse de données textuelles appliquée à des problématiques de sécurité et d'enquête judiciaire 52

SIMONA BALBI, MICHELANGELO MISURACA, MARIA SPANO

A two-step strategy for improving categorisation of short texts 60

CHRISTINE BARATS, ANNE DISTER, PHILIPPE GAMBETTE, JEAN-MARC LEBLANC, MARIE PERES

Appeler à signer une pétition en ligne : caractéristiques linguistiques des appels 68

MANUEL BARBERA, CARLA MARELLO

Newsgroup e lessicografia: dai NUNC al VoDIM 76

IGNAZIA BARTHOLINI

Techniques for detecting the normalized violence in the perception of refugee / asylum seekers between lexical analysis and factorial analysis 83

PATRIZIA BERTINI MALGARINI, MARCO BIFFI, UGO VIGNUZZI

Dal corpus al dizionario: prime riflessioni lessicografiche sul Vocabolario storico della cucina italiana postunitaria (VoSCIP) 90

MARCO BIFFI

Strumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano postunitario 99

ANNICK FARINA, RICCARDO BILLERO

Comparaison de corpus de langue « naturelle » et de langue « de traduction » : les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues 108

FELICE BISOGNI, STEFANO PIRROTTA

Il rapporto tra famiglie di anziani non autosufficienti e servizi territoriali: un'analisi dei dati esplorativa con l'Analisi Emozionale del Testo (AET).... 117

ANTONELLA BITETTO, LUIGI BOLLANI

Esperienza di analisi testuale di documentazione clinica e di flussi informativi sanitari, di utilità nella ricerca epidemiologica e per indagare la qualità dell'assistenza 126

GUIDO BONINO, DAVIDE PULIZZOTTO, PAOLO TRIPODI

Exploring the history of American philosophy in a computer-assisted framework 134

MARC-ANDRE BOUCHARD, SYLVIA KASPARIAN

La classification hiérarchique descendante pour l'analyse des représentations sociales dans une pétition antibilinguisme au Nouveau-Brunswick, Canada 142

LIVIA CELARDO, RITA VALLEROTONDA, DANIELE DE SANTIS, CLAUDIO SCARICI, ANTONIO LEVA

Analysing occupational safety culture through mass media monitoring..... 150

BARBARA CORDELLA, FRANCESCA GRECO, PAOLO MEOLI, VITTORIO PALERMO, MASSIMO GRASSO

Is the educational culture in Italian Universities effective? A case study..... 157

MICHELE A. CORTELAZZO, GEORGE K. MIKROS, ARJUNA TUZZI

Profiling Elena Ferrante: a Look Beyond Novels 165

FABRIZIO DE FAUSTI, MASSIMO DE CUBELLIS, DIEGO ZARDETTO¹

Word Embeddings: a Powerful Tool for Innovative Statistics at Istat 174

Gibbons A. (1985). *Algorithmic Graph Theory*. Cambridge University Press. . 182

VIVIANA DE GIORGI, CHIARA GNESI

Analisi di dati d'impresa disponibili online: un esempio di data science tratto dalla realtà economica dei siti di e-commerce 183

ALESSANDRO CAPEZZUOLI, FRANCESCA DELLA RATTA, STEFANIA MACCHIA, MANUELA MURGIA, MONICA SCANNAPIECO, DIEGO ZARDETTO

The use of textual sources in Istat: an overview..... 192

FRANCESCA DELLA RATTA, GABRIELLA FAZZI, MARIA ELENA PONTECORVO, CARLO VACCARI, ANTONINO VIRGILLITO

Twitter e la statistica ufficiale: il dibattito sul mercato del lavoro 200

SAMI DIAF

Gauging An Author's Mood Using Hidden Markov Chains 209

MARC DOUGUET

Les hémistiches répétés 215

FRANCESCA DRAGOTTO, SONIA MELCHIORRE

«Mangiata dall'orco e tradita dalle donne». Vecchi e nuovi media raccontano la vicenda di Asia Argento, tra storytelling e Speech Hate 223

CRISTIANO FELACO, ANNA PAROLA

Il *cosa* e il *come* del processo narrativo. L'uso combinato della Text Analysis e Network Text Analysis al servizio della precarietà lavorativa 233

ANA NORA FELDMAN

Hablando de crisis: las comunicaciones del Fondo Monetario Internacional 242

VALERIA FIASCO

Brexit in the Italian and the British press: a bilingual corpus-driven analysis 250

VIVIANA FINI, GIUSEPPE LUCIO GAETA, SERGIO SALVATORE

Textual analysis to promote innovation within public policy evaluation 259

ALESSIA FORCINITI, SIMONA BALBI

- A proposal for Cross-Language Analysis:
violence against women and the Web 268

BEATRICE FRACCHIOLLA, OLINKA SOLENE DE ROGER

- La verbalisation des émotions 276

**LUISA FRANCHINA, FRANCESCA GRECO, ANDREA LUCARIELLO,
ANGELO SOCIAL, LAURA TEODONNO**

- Improving Collection Process for Social Media Intelligence: A Case Study . 285

**ANDREA FRONZETTI COLLADON, JOHANNE SAINT-CHARLES, PIERRE
MONGEAU**

- The impact of language homophily and similarity of social position on
employees' digital communication 293

MATTEO GERLI

- Looking Through the Lens of Social Sciences: The European Union in the EU-
Funded Research Projects Reporting 300

LUCIE GIANOLA, MATHIEU VALETTE

- Spécialisation générique et discursive d'une unité lexical L'exemple de
joggeuse dans la presse quotidienne régionale 312

PETER A. GLOOR, JOAO MARCOS DE OLIVEIRA, DETLEF SCHODER

- The Transparency Engine – A Better Way to Deal with Fake News 319

FRANCESCA GRECO, LEONARDO ALAIMO, LIVIA CELARDO

- Brexit and Twitter: The voice of people..... 327

FRANCESCA GRECO, GIULIO DE FELICE, OMAR GELO

- A text mining on clinical transcripts of good and poor outcome
psychotherapies 335

FRANCESCA GRECO, DARIO MASCHIETTI, ALESSANDRO POLLI

- DOMINIO: A Modular and Scalable Tool for the Open Source Intelligence 343

LEONIE GRÖN, ANN BERTELS, KRIS HEYLEN

- Is training worth the trouble? A PoS tagging experiment with Dutch clinical
records..... 351

FRANCE GUERIN-PACE, ELODIE BARIL

- Les outils de la statistique textuelle pour analyser
les corpus de données d'enquêtes de la statistique publique 359

SERGE HEIDEN

- Annotation-based Digital Text Corpora Analysis within the TXM Platform 367

DANIEL HENKEL

- Quantifying Translation : an analysis of the conditional perfect in English-
French comparable-parallel corpus..... 375

DANIEL DEVATMAN HROMADA

- Extraction of lexical repetitive expressions from complete works of William
Shakespeare..... 384

OLIVIER KRAIF, JULIE SORBA

Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale) 392

CYRIL LABBE, DOMINIQUE LABBE

Les phrases de Marcel Proust 400

LUDOVICA LANINI, MARÍA CARLOTA NICOLÁS MARTÍNEZ

Verso un dizionario *corpus-based* del lessico dei beni culturali: procedure di estrazione del lemmario 411

DANIELA LARICCHIUTA, FRANCESCA GRECO, FABRIZIO PIRAS, BARBARA CORDELLA, DEBORA CUTULI, ELEONORA PICERNI, FRANCESCA ASSOGNA, CARLO LAI, GIANFRANCO SPALLETTA, LAURA PETROSINI

"The grief that doesn't speak": Text Mining and Brain Structure 419

GEVISA LA ROCCA, CIRUS RINALDI

Icone gay: tra processi di normalizzazione e di resistenza. Ricostruire la semantica degli hashtag..... 428

LUDOVIC LEBART

Looking for *topics*: a brief review..... 436

GAËL LEJEUNE, LICHAO ZHU

Analyse Diachronique de Corpus : le cas du poker..... 444

JULIEN LONGHI, ANDRE SALEM

Approche textométrique des variations du sens..... 452

LAURENT VANNI¹, DAMON MAYAFFRE, DOMINIQUE LONGREE

ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables 459

LUCIE LOUBERE

Déconstruction et reconstruction de corpus... À la recherche de la pertinence et du contexte 467

HEBA METWALLY

L'apport du *corpus-maquette* à la mise en évidence des niveaux descriptifs de la chronologie du sens. Essai sur une Série Textuelle Chronologique du *Monde diplomatique* (1990-2008). 474

JUN MIAO, ANDRE SALEM

Séries textuelles homogènes..... 491

SILVIO MIGLIORI, ANDREA QUINTILIANI, DANIELA ALDERUCCIO, FIORENZO AMBROSINO, ANTONIO COLAVINCENZO, MARIALUISA MONGELLI, SAMUELE PIERATTINI, GIOVANNI PONTI SERGIO BOLASCO, FRANCESCO BAIOCCHI, GIOVANNI DE GASPERIS

TaLTaC in ENEAGRID Infrastructure..... 501

ISABELLA MINGO, MARIELLA NOCENZI

The dimensions of Gender in the International Review of Sociology. A lexicometric approach to the analysis of the publications in the last twenty years 509

ADIEL MITTMANN, ALCKMAR LUIZ DOS SANTOS

The Rhythm of Epic Verse in Portuguese From the 16th to the 21st Century 514

DENIS MONIERE, DOMINIQUE LABBE

Le vocabulaire des campagnes électorales 522

CYRIELLE MONTRICHARD

Faire émerger les traces d'une pratique imitative dans la presse de tranchées à l'aide des outils textométriques 532

ALBERT MORALES MORENO

Evolución diacrónica de la terminología y la fraseología jurídico-administrativa en los Estatutos de autonomía de Catalunya de 1932, 1979 y 2006 541

CEDRIC MOREAU

Comment penser la recherche d'un signe pour une plateforme multilingue et multimodale français écrit / langue des signes française ? 556

JEAN MOSCAROLA, BORIS MOSCAROLA

Conclusion ADT et visualisation, pour une nouvelle lecture des corpus Les débats de 2ème tour des Présidentielles (1974-2017) 563

MAURIZIO NALDI

A conversation analysis of interactions in personal finance forums 571

STEFANO NOBILE

Analisi testuale, rumore semantico e peculiarità morfosintattiche: problemi e strategie di pretrattamento di corpora speciali 578

DANIEL PELISSIER

L'individu dans le(s) groupe(s) : focus group et partitionnement du corpus 586

BENEDICTE PINCEMIN, CELINE GUILLOT-BARBANCE, ALEXEI**LAURENTIEV**

Using the First Axis of a Correspondence Analysis as an Analytical Tool. Application to Establish and Define an Orality Gradient for Genres of Medieval French Texts 594

CELINE POUDAT

Explorer les désaccords dans les fils de discussion du Wikipédia francophone 602

MATTHIEU QUIGNARD, SERGE HEIDEN, FREDERIC LANDRAGIN,**MATTHIEU DECORDE**

Textometric Exploitation of Coreference-annotated Corpora with TXM: Methodological Choices and First Outcomes 610

PIERRE RATINAUD

Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ 616

LUISA REVELLI

- Il parametro della *frequenza* tra paradossi e antinomie:
il caso dell'*italiano scolastico* 626

PIERGIORGIO RICCI

- How Twitter emotional sentiments mirror on the Bitcoin
transaction network 635

CHANTAL RICHARD, SYLVIA KASPARIAN

- Analyse de contenu versus méthode Reinert : l'analyse comparée d'un corpus
bilingue de discours acadiens et loyalistes du N.-B., Canada 643

VALENTINA RIZZOLI, ARJUNA TUZZI

- Bridge over the ocean: Histories of social psychology in Europe and North
America. An analysis of chronological corpora 651

LOUIS ROMPRE, ISMAÏL BISKRI

- Les « itemsets fréquents » comme descripteurs de documents textuels 659

CORINNE ROSSARI, LJILJANA DOLAMIC, ANNALENA HÜTSCH, CLAUDIA RICCI, DENNIS WANDEL

- Discursive Functions of French Epistemic Adverbs: What can Correspondence
Analysis tell us about Genre and Diachronic Variation? 668

VANESSA RUSSO, MARA MARETTI, LARA FONTANELLA, ALICE TONTODIMAMMA

- Misleading information in online propaganda networks 676

ELIANA SANANDRES, CAMILO MADARIAGA, RAIMUNDO ABELLO

- Topic modeling of Twitter conversations 684

FRANCESCO SANTELLI, GIANCARLO RAGOZINI, MARCO MUSELLA

- What volunteers do? A textual analysis of voluntary activities in the Italian
context 692

S. SANTILLI, S. SBALCHIERO, L. NOTA, S. SORESI

- A longitudinal textual analysis of abstract presented at Italian Association for
Vocational guidance and Career Counseling'
Conferences from 2002 to 2017 700

JACQUES SAVOY

- A la poursuite d'Elena Ferrante 707

JACQUES SAVOY

- Regroupement d'auteurs dans la littérature du XIXe siècle 716

STEFANO SBALCHIERO, ARJUNA TUZZI

- What's Old and New? Discovering Topics in the American Journal of
Sociology 724

NILS SCHAETTI, JACQUES SAVOY

- Comparison of Neural Models for Gender Profiling 733

LIONEL SHEN

- Segments répétés appliqués à l'extraction de connaissances trilingues 740

SANDRO STANCAMPIANO	
Misurare, Monitorare e Governare le città con i Big Data	748
FADILA TALEB, MARYVONNE HOLZEM	
Exploration textométrique d'un corpus de motifs juridiques dans le droit international des transports	755
JAMES M. TEASDALE	
The Framing of the Migrant: Re-imagining a Fractured Methodology in the Context of the British Media.	763
MARJORIE TENDERO¹, CECILE BAZART	
Results from two complementary textual analysis software (Iramuteq and Tropes) to analyze social representation of contaminated brownfields	771
MATTEO TESTI, ANDREA MERCURI, FRANCESCO PUGLIESE	
Multilingual Sentiment Analysis.....	780
JUAN MARTÍNEZ TORVISCO	
A linguistic analysis of the image of immigrants' gender in Spanish newspapers.....	788
FRANCESCO URZÌ	
Lo strano caso delle frequenze zero nei testi legislativi euroistituzionali.....	796
SYLVIE VANDAELE	
Les traductions françaises de <i>The Origin of Species</i> : pistes lexicométriques .	805
PIERRE WAVRESKY, MATTHIEU DUBOYS DE LABARRE, JEAN-LOUP LECOEUR	
Circuits courts en agriculture : utilisation de la textométrie dans le traitement d'une enquête sur 2 marchés	814
MARIA ZIMINA, NICOLAS BALLIER	
On the phraseology of spoken French: initial salience, prominence and lexicogrammatical recurrence in a prosodic-syntactic treebank <i>Rhapsodie</i>	822

Abstracts

FILIPPO CHIARELLO, GUALTIERO FANTONI, ANDREA BONACCORSI, SILVIA FARERI	
What kind of contributions does research provides? Mapping issue based statements in research abstracts	833
FILIPPO CHIARELLO, GIACOMO OSSOLA, GUALTIERO FANTONI, ANDREA BONACCORSI, ANDREA CIMINO, FELICE DELL'ORLETTA	
Technical sentiment analysis: predicting the success of new products using social media.....	835

FIorenza DERIU, DOMENICA FIOREDISTELLA IEZZI
 Citizens and neighbourhood life: mapping population sentiment in Italian cities..... 837

FRANCESCA DI CARLO, ROSY INNARELLA, BRIZIO LEONARDO TOMMASI
 Vax network: profiling influential nodes with social network analysis on twitter..... 838

DAVIDE DONNA
 Alteryx 840

VALERIO FICCADENTI, ROY CERQUETI, MARCEL AUSLOOS
 Complexity of US President Speeches 841

PETER A. GLOOR
 Measuring the Dynamics of Social Networks with Condor 842

IOLANDA MAGGIO, DOMENICA FIOREDISTELLA IEZZI, MATTEO FATIGHENTI
 "BIG DATA" Words Trend Analysis using the multidimensional analysis of texts 844

MARIO MASTRANGELO
 Itinerari turistici, network analysis e text mining 845

MARIA FRANCESCA ROMANO, GUIDO REY, ANTONELLA BALDASSARINI PASQUALE PAVONE
 Text Mining per l'analisi qualitativa e quantitativa dei dati amministrativi utilizzati dalla Pubblica Amministrazione..... 847

ALESSANDRO CESARE ROSA
 Taglio cesareo e Vbac in Italia al tempo dei Big Data: una proposta di ulteriore contributo informativo..... 849

DOMINIO: A Modular and Scalable Tool for the Open Source Intelligence

Francesca Greco¹, Dario Maschietti², Alessandro Polli³

¹ La Sapienza University of Rome, Prisma S.r.l. – francesca.greco@uniroma1.it

² Prisma S.r.l – d.maschietti@prismaprogetti.it

³ La Sapienza University of Roma – alessandro.polli@uniroma1.it

Abstract

Prisma has developed an innovative technology for the Open Source Intelligence (OSINT) which aims to provide a solution for those processes of knowledge management, which require the intervention of a human operator, unaided by information technology (IT) support, in one or more stages of the procedure. Such intervention involves a considerable waste of time and resources that could be reduced through the use of an IT tool, partially or totally automating entire stages of the procedure. DOMINIO is a platform that implements tools for automatic online information aggregation, its analysis, the possible alignment with traditional databases and the representation through infographic and georeferencing tools, in order to generate a report. This paper describes the platform architecture, the main algorithms used in the analysis stage of the contents and possible directions of development.

Abstract

Prisma ha sviluppato una tecnologia innovativa finalizzata all'Open Source Intelligence (OSINT) che intende fornire risposta alle necessità di knowledge management, che richiedono l'intervento di un operatore umano, non assistito da supporti di information technology (IT), in una o più fasi della procedura. Tale intervento comporta un notevole dispendio di tempo e risorse che potrebbe essere ridotto attraverso l'utilizzo di uno strumento di IT, automatizzando parzialmente o totalmente intere fasi della procedura. DOMINIO è una piattaforma che implementa strumenti per l'aggregazione automatica di informazioni on line, la loro analisi, l'eventuale allineamento con banche dati di tipo tradizionale, la rappresentazione attraverso tool di infografica e georeferenziazione, allo scopo di generare una reportistica. Il presente contributo descrive l'architettura della piattaforma, i principali algoritmi adottati nella fase di analisi dei contenuti e le possibili direzioni di sviluppo.

Keywords: knowledge management, Open Source Intelligence tool, Information Technology,

1. Introduction

There is a close link between data management and knowledge on the one hand, and knowledge and innovation on the other. The growing mass of unstructured information from disparate channels (search engines, RSS feeds, social networks) and traditional databases entails the need to drastically simplify the preparation, analysis and reporting stages required to structure the information. In fact, only a structured information translates into knowledge. Knowledge, in turn, is a major driver of innovation and, properly managed, it translates into a competitive advantage. The idea at the basis of the tool OSINT (Open Source Intelligence) stems from the needs expressed by analysts – mainly involved in the field of sentiment analysis and opinion mining industry. However, this idea is enough comprehensive to encompass all those activities of knowledge management, similar to the former, which require intervention by a human operator, unaided by IT support (Information Technology), in one or more stages of the procedure, the intervention of which involves a great deal of time and resources. Although in high-end solutions machine learning systems are starting to spread, the available technology is still characterized by significant limitations, especially in the presence of unstructured information. In particular, with regard to supervised machine learning systems, intervention is required by an operator in the initial stages of the procedure and, in general, with reference to any automated system applied to the analysis of a text, it is still impossible to identify complex cognitive functions (for example, irony). Of course, these problems are immanent in many fields of OSINT, and they also affect the stage of reporting, which requires a direct involvement of the analyst, unaided by IT. So, the availability of an IT tool that minimizes human operator intervention – partially or totally automates entire stages of the procedure – would result in substantial advantages, like time savings, increased productivity and the resulting increased efficiency in the allocation of human and financial resources.

Prisma has developed an innovative technology of OSINT, which aims to fix the problems briefly described above. The platform implements tools for automatic aggregation of the online information, their analysis, the alignment with traditional databases, the representation through infographic and georeferencing tools, aimed to automate also the phase of elaboration of the final report.

This paper will describe the architecture of the platform, the main analysis modules and the possible directions of development.

2. Platform Architecture

DOMINIO is an OSINT (Open Source Intelligence) platform that automatically aggregates information from online and traditional databases, analyses it and generates reports on a user-defined subject. The platform collects information by querying several channels: search engines (Google, Yahoo, Bing), social networks (Facebook, Twitter, Google+), RSS feeds, blogs (Blogger, Wordpress, Tumblr), traditional databases. The goal of DOMINIO is to build a structured set of contents, as broad as possible, and to carry out a wide range of qualitative and quantitative analysis. DOMINIO stores these contents within a non-relational database (DB) (MongoDB, 2018; Morphia, 2018), classifying the various documents by channel of origin (Twitter, Facebook, RSS, etc.) to ensure the homogeneity of the collections.

Among the options, the DOMINIO user can make queries on-demand or in a continuous mode. The on-demand option carries out an asynchronous search, while the continuous mode option enables to aggregate periodically data and to track a subject over an extended time span. The DOMINIO's architecture allows the user to switch from one mode to another; the availability of two searching modes allows overcoming the trade-off between accuracy of analysis and speed of processing.

With regard to one or more subjects selected by the operator, DOMINIO performs synchronous or asynchronous research on a set of Internet channels, such as search engines (Google, Yahoo, Bing), social networks (Facebook, Twitter, Google+), RSS feeds, blogs (Blogger, Wordpress, Tumblr). The user can also extend the search to the Deep Web, through specific search engines, such as Torch or Grams.

Moreover, to meet specific information needs, DOMINION can match these search results with the information achievable from the traditional databases to support many types of analysis (brand reputation, country risk assessment, opinion polls, cyber security, etc.), considerably increasing the operability and flexibility of the tool.

Among the traditional databases already available, DOMINIO includes:

- IHS Jane's (2018), which provides updates on military and political situation, terrorist acts, civil wars, transportation system, for most of the countries in the world;
- Bureau Van Dijk (2018), which collects firms data on ratings, shareholdings, equity investments and M&A;
- MIG (a geographic information database drawn up by one of the authors).

In addition, for specific information purposes, DOMINIO is open for interfacing with Enterprise Resource Planning databases (like SAP, Oracle, etc.) through market tools (Business Object, Quick View).

The search results are recalled by the analyst, who operates from a CMS (Content Management System) application to manage the structured set of content and conduct a wide range of qualitative and quantitative analyses (from simple summary statistics to sophisticated multivariate analyses and text and opinion mining techniques).

The statistical methods implemented on DOMINIO are chosen by the Prisma research team according to a set of criteria that privileges the suitability of one algorithm to automate entire stages of the procedure, in accordance with the original design idea. Moreover, the modular architecture of DOMINIO, described briefly below, allows a quick integration of the latest analysis tools and innovative methodologies produced in the academic field.

Once the stage of content analysis is completed, the CMS application generates a micro-site containing the results (geo-referenced maps, summary statistics, multivariate analysis results, textual and semantic analysis of sentiment analysis, etc.). After selecting a graphic layout for the final report, the analyst has only to write notes and final remarks.

The possibility of including features generating automatic and/or auto-completion comments, customizable by the user, is also being studied. Once the last stage is completed, the report is ready for online publication or traditional diffusion in pdf format, or linked to external services.

From an architectural point of view, DOMINIO is designed following the most modern criteria of modular software design, with the parallel development of the platform's modules. In short, in order to ensure a greater fault tolerance and high safety standards, the system is divided into three independent logical units (cfr. Figure 1):

- DOMINIO Engine Unit (MEU), which implements the features of 1) scraping information from the sources mentioned above (web, social networks, RSS feeds, traditional databases); 2) storage of results on MEDB database; 3) qualitative and quantitative analysis;
- DOMINIO RESTurl Unit (MRESTU), which receives requests from the MCMS unit, verifies the consistency and forwards the request to the unit ME. Upon receiving the response, it implements the request by adding additional fields (username, token, etc.) and returns them to the MCMS client. The MRESTU unit contains the database (MRESTDB) for user profiling;
- DOMINIO Content Management System Unit (MCMSU), which manages the stage concerning the reporting and archiving of reports according to pre-logical criteria (organization by topic, chronologically, for templates, etc.).

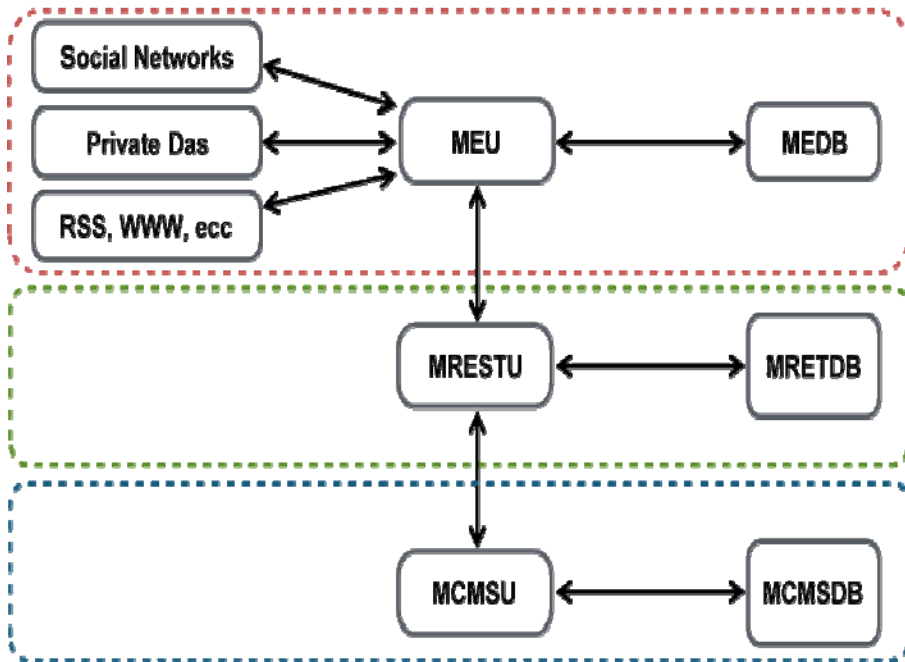


Figure 1 - DOMINIO General Overview

3. Main analysis modules

3.1. Country Threat Assessment

The Country Threat Assessment module supports the Company Intelligence and Security analyst in the country's risk assessment process. Through a responsive type interface, it aggregates information from major global industry databases (eg, IHS Jane's) giving an assessment of external and internal risk and that due to political and socio-economic factors and potential outbreaks or revolutionary movements for 192 different countries. Country Threat Assessment is integrated with intelligence information updated weekly on each country. Through an automatic report, data is aggregated into a single file by optimizing timing of risk assessment and providing a solid foundation for any further detailed analysis. DOMINIO offers the possibility of making a full or partial information download, and the generation of an automatic report, thus optimizing any drafting processes.

3.2. Due Diligence

The Due Diligence module supports the Economic Intelligence analyst in the process of business valuation in relation to suppliers, partners and customers. Among the sectors analysed in the module are included

assessments of profitability and financial performance as well as creditworthiness. Through a simple and intuitive interface, the module aggregates information from leading industry databases and returns an economic, financial and credit risk profile on hundreds of millions of businesses around the world. The Due Diligence Module also allows an assessment of individuals, through the analysis of individuals exposed politically, returning an automatic report that integrates the main aspects of each business and its economic risk analysis.

3.3. Open Source Intelligence

On completion of the aggregation of large amounts of data from major social networks (Facebook, Twitter, Youtube) and the main Italian newspapers based on predetermined keywords analyst, a statistical representation of the main trending topic is returned and an output of structured data for subsequent multivariate analysis is generated. Furthermore, the module allows the geo-referencing of content, highlighting even at geographic levels useful signs for the analyst. As for each of DOMINIO's modules, it is possible to generate automatic reporting.

3.4. Geographic Information Module

This is a module that analyses the information inferable from a dataset of basic statistical information and related indicators, with reference to a multitude of subjects, 9 of which are in a current stage of development. The basic statistical information, refers to the division of the Italian territory into provinces, covering a time period between 1995 and the latest available year, which for some subject areas is ongoing or, more frequently, the previous year to the current one. The dataset will be supportive to a wide range of applications - from forecasting and scenario analysis, counterfactual analysis to spatial analysis.

3.5. Text Mining Module

On completion of the automatic analysis of textual data using statistical methods (Lebart et Salem, 1994; Feldman et Sanger, 2006; Bolasco, 2013), in order to extract structured information, the main statistical methods of analysis of textual data implemented in DOMINIO are: factor analysis (correspondence analysis, multiple correspondence analysis); cluster analysis (k-mean, bisecting k-mean, fuzzy clustering, etc.); network analysis; Markov analysis; pattern recognition.

For example, during the French presidential campaign of 2017 we analysed the sentiment about migration, that was one of the most debated theme. We performed an Emotional Text Mining (Greco et al., 2017) in order to explore

the emotional content of the Twitter messages concerning migration written in French in the last two weeks before the first round of the presidential election in 2017. The aim was to analyse the opinions, feelings and shared comments, classifying the contents and the sentiments. We retrieved the messages from the Twitter repository collecting a sample of over one hundred thousand tweets. The large size corpus of 2.154.194 tokens (TTR = 0,01; Hapax percentage = 40,4) underwent a multivariate analysis based on a bisecting *k*-means algorithm (Savaresi et Boley, 2004) to classify the text, and a correspondence analysis (Lebart et Salem, 1994) to detect the latent dimensions setting the cluster per keywords matrix. The advantage connected with this approach is to interpret the factorial space according to words polarization, thus identifying the emotional categories that generate migration representations, and to facilitate the interpretation of clusters, exploring their relationship within the symbolic space (Greco, 2016).

The results interpretation allowed for the detection of seven representations of migrants that corresponded to three different sentiments: positive (42%), negative for the community (45%), and negative for migrants (13%). We considered as negative the representation of migrants as squatters, invaders, terrorists, trafficking slaves and migration victims, and positive the sport heroes and the EU solidarity target. Among the negative clusters, we distinguished negativity according to the direction of the action: squatters, terrorists and invaders are negative for the community and trafficking slaves and migration victims are negatives for migrants themselves (see Greco et al., 2017). Moreover, it was possible to highlight the connection between the real life events and the tweets production. While the terrorist attack three days before the first round of voting in the centre of Paris had slightly modified the production of messages, the candidates' interviews had a higher impact. This suggests that the medialization was more important than the terrorist attack in the production of messages (see Greco et al., 2017).

4. Conclusion

The innovative aspect that characterizes DOMINIO is the ability to aggregate data of different types and from different channels of information, automatically, simply and transparently. Moreover, its structure allows for the integration of the latest analytical tools and innovative methodologies produced in academia. By means of an automated reporting system, the analyst is supported in the assessment of risk and the collection of information in the geopolitical and economic field and from open sources. The set of modules allows the analyst to generate knowledge from an ever-growing amount of data by optimizing the processes of assessment and risk reduction.

References

- Bolasco S. (2013). *L'analisi automatica dei testi: Fare ricerca con il text mining*. Carocci.
- Bureau von Dijk (2018). *A Moody's Analytics Company*. Bureau von Dijk, <https://www.bvdinfo.com/it-it/home>
- Feldman R. and Sanger J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.
- Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Franco Angeli.
- Greco F., Maschietti D. and Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *RIEDS*, 71(2): 125:36.
- IHS Jane's (2018). *Jane's Information Group*. IHS Jane's, <http://www.janes.com>
- Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod
- MongoDB (2018). *MongoDB for GIANT ideas*. MongoDB, <https://www.mongodb.com>
- Morphia (2018). *The Java Object Document Mapper for MongoDB*. MongoDB, <https://mongodb.github.io/morphia/>
- Savaresi S.M. and Boley D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4): 345-362.