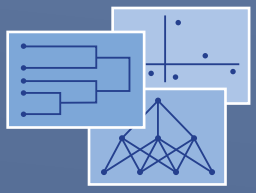
Studies in Classification, Data Analysis, and Knowledge Organization

Domenica Fioredistella Iezzi
Damon Mayaffre
Michelangelo Misuraca *Editors*

Text Analytics

Advances and Challenges



Springer

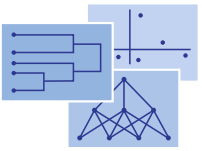
Studies in Classification, Data Analysis, and Knowledge Organization

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

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 inexpert 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 — Domenica Fioredistella Iezzi
Nice, France — Damon Mayaffre
Rende, Italy — Michelangelo Misuraca

Contents

Techniques, Methods and Models

Text Analytics: Present, Past and Future

Domenica Fioredistella Iezzi and Livia Celardo

Abstract Text analytics is a large umbrella under which it is possible to report countless techniques, models, methods for automatic and quantitative analysis of textual data. Its development can be traced back the introduction of the computer, but the prodromes date back, the importance of text analysis has grown over time and has been greatly enriched with the spread of the Internet and social media, which constitute an important flow of information also in support of official statistics. This paper aims to describe, through a timeline the past, the present and the possible future scenario of text analysis. Moreover, the main macro-steps for a practical study are illustrated.

1 Introduction

Data and statistics represent a key for decision-making processes. Nowadays, the amount of data produced by human activities generated the age called "data revolution" [1, 2]. Currently, unstructured data (documents, posts, tweets, video, photos, open questions in a survey, product complains and reviews) represent the most important part of the information present, but not available, to companies and decision-makers. As early as 1995, a Forrester Research report estimated that 80% of useful information in a company is hidden in unstructured documents and therefore not available for the same company [3]. Currently, the exponential growth in the use of social media and social networking sites such as Facebook, Twitter and Instagram have created a new information flow. According to Global Digital Report of "We Are Social and Hootsuite" (2020), about half of the world's population (3.8 billion

D. F. Iezzi (✉) · L. Celardo
Department of Enterprise Engineering, Tor Vergata University,
Via del Politecnico, 1, 00133 Rome, Italy
e-mail: stella.iezzi@uniroma2.it

L. Celardo
e-mail: livia.celardo@uniroma2.it

D. F. Iezzi et al. (eds.), *Text Analytics*, Studies in Classification,
Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-52680-1_1

people) regularly use social media, while 4.54 billion people are connected to the Internet, with almost 300 million users who had access to the Internet for the first time in 2019.[1]
Textual analytics is an automatic and quantitative approach for collecting, processing and interpreting text data. Generating new knowledge from a text analytics process is very complex. It is a chain of operations strong link to research goals, that can allow to have information in real time, and help to make a decision in uncertainty conditions.

This field of research represents an extraordinary challenge for the analysis of many research fields (artificial intelligence, social media, linguistics, statistics, mathematics, marketing, medicine, …). During the Coronavirus Disease (COVID-19) pandemic, e.g. Google users searched countless times the expressions "I can't smell". A sort of a "truth machine" on one's health. Indeed, anosmia, or loss of smell, has been proven to be a symptom of COVID-19, and some estimates suggest that between 30 and 60 per cent of people with the disease experience this symptom. In this way, it was possible to identify new outbreaks of pandemic propagation[2] [4].

This paper aims is to examine the main changes from the origin of text analytics to the present and with a look towards the future. A zoom on scientific works on these issues in the past 20 years is explored.

This article is structured as follows. In Sect. 2, the prodromes of the text analysis are described; in Sect. 3, origins and developments are discussed; in Sect. 4, the main steps of an text analysis are explored; in Sect. 5, the main applications developed in the last 20 years are explored, and conclusions are also drawn on possible future scenarios.

2 The Text Analysis Pioneers

The quantitative approach to the study of language originates from experimental psychology and the need to demarcate the differences between this new science and philosophy. In 1888, in a research on the expression of emotions through words, Benjamin Bourdon analysed the *Exodus* of the Bible and calculated the frequencies by rearranging and classifying them, eliminating the stop words [5].

A real interest in the quantitative analysis of the lexicon (Lexicometry) is due to stenographers such as Keading and Estoup. Keading coordinated a research in 1898 on the frequencies of graphemes, syllables and words in the German language. Estoup, one of the major exponents French shorthand, in a study published in 1907, defined the notion of rank as a position occupied by a word in a list of words sorted according to decreasing frequencies [6].

[1] See the website https://wearesocial.com/it/blog/2020/01/report-digital-2020-i-dati-global.

[2] Stephens-Davidowitz, in an article published in the New York Times (April 5, 2020), suggested that this methodology could be adopted to search for places where dissemination has escaped the reporting of official data as in the case of the state of Ecuador.

Table 1 Ranks and frequencies in Joyce's Ulysse

Rank	Frequency
10	2653
100	265
1,000	26
10,000	2

An acceleration to grow to text analysis comes from psycholinguistic studies on children's language. Bouseman (1925) demonstrated the existence of two distinct linguistic styles. Emotionally unstable children used verbs preponderantly [7].

In the 1940s, Boder obtained a grammatical measure called "adjective–verb ratio", using different literary genres and discovering that in the drama, the genre closest to spoken language, verbs are five times more than adjectives in 85% of cases [8].

Zipf's work, which takes up what Estoup had started, states "the principle of relative frequency", better known as "Zipf's Law". In a sufficiently long corpus, by classifying words by decreasing ranks, there is a fundamental relationship between the frequency (f) and the rank (r) of the words (Zipf's law): $fr = c$, where c is a constant [9, 10].

Zipf himself carries out an experiment on *Joyce's Ulysse* [27]. From the vocabulary of a corpus of 260,000 occurrences, Zipf noted stating from the highest rank, the frequency of words is inversely related and that the product of frequency (f) and rank (r) is approximately a constant (see Table 1).

Zipf's contribution to linguistics, albeit ingenious, is quite controversial and has given a strong impetus to other theoretical and empirical contributions [11, 26] that only with subsequent development of calculating have found wider fields of application, e.g. the frequency of access to web pages. From this debate, it arose that it is more appropriate to express the law as $fr^a = c$, which on a logarithmic scale expresses itself as $alog(r) + log(f) = c$. This formula can also be written as follows:

$$log(f) = c + alog(r) \tag{1}$$

This is the equation of a straight line, where a indicates the slope or *richness of the vocabulary*, which, in turn, depends on the size of the text [21].

Many indices on the richness of the vocabulary are based on the measure of the relationship between type (V) and token (N): V/N (type/token ratio). This measure allows to connect only texts of equal size. It is well known that the ratio between types and tokens (TTR) decreases systematically with increasing text length because speakers/writers have to repeat themselves. This makes it impossible to compare texts with different lengths.

The Yule index (K) is a measure to identify an author, assuming that it would differ for texts written by different authors. Let N be the total number of words in a text, $V(N)$ be the number of distinct words (dictionary of a corpus), $V(m, N)$ be the number of words appearing m times in the text, and $m_{\max}$ be the largest frequency of a word. Yule's K is then defined as follows, through the first and second moments

of the vocabulary population. Yule defines the two quantities S_1 and S_2 (two first moments about zero of the distribution of words), such that

$$S_1 = N = \sum_m mV(m, N) \tag{2}$$

$$S_2 = N = \sum_m m^2 V(m, N) \tag{3}$$

$$K = C\frac{S_2 - S_1}{S_1^2} = C[-\frac{1}{N} + \sum_m^{m_{max}} m^2 V(m, N)\frac{m^2}{N^2}] \tag{4}$$

where C is a constant enlarging of the value of K, defined by Yule as $C = 10^4$. The theme of word frequencies in a text, treated by Zipf and Yule, returns to the studies of Guiraud , who identifies the relationships between the length of the language and the extension of the vocabulary. Guiraud argues that there are very few words that cover more than half of the occurrences that make up the majority of words. Guiraud underlines that concentration of the most frequent words (thematic words) and the dispersion of the less frequent words represented a measure of the richness of the vocabulary (Guiraud, 1954):

$$G = \frac{V}{\sqrt{N}} \tag{5}$$

Guiraud showed empirically that his index is stable over texts between 1000 and 100,000 words [29]. The richness of the vocabulary finds a mathematical formulation with Herdan (1956), which proposes the use of a measure based on the exponential function [28].

3 Timeline Representation Framework

Computer science has a leading role in the development of text analysis. It presupposes, in fact, the automatic treatment of a text without reading by the researcher. The introduction of calculators from the mid-1940s and then the widespread use allowed their growth and application in all disciplinary areas.

Its fortune is originated from the fruitful encounter between computer science, linguistics, statistics and mathematics [23, 27].

Towards the end of the 1950s, the Besançon study centre began a mechanographic examination of Coraille's works, on which Muller's studies on lexicometric analyses were based (Muller, 1967, 1977). The perspectives change and the texts become a representative sample of the language. In which specific expressions are searched for Indices and specific vocabulary measures are born (Tounier, 1980; Lafon, 1984),

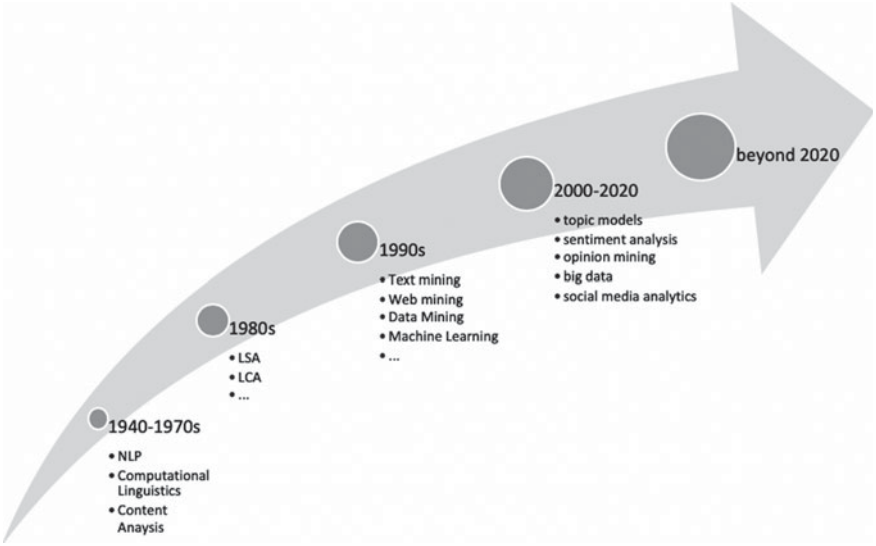


Fig. 1 Text analysis timeline

which soon become widely used by scholars thanks to the implementation of the HyperBase software (Brunet, 1993) [27].

Figure 1 presents a timeline of text analysis from 1940 to the present day. In the first phase (1940–1970's) text analysis studies focus on natural language processing (NPL), computational linguistics and content analysis. Among the most significant works are the first quantitative studies by Guiraud (1954, 1960) and Hedan (1964). The analysis of the content, which has its prodrome in the chiropractic culture of 4000 B.C., has known hybridizations with the computer [31] (Krippendorff, 2004).

The new objective is no longer just to describe the language of a specific work (e.g. The Divine Comedy) or an author (Corneille or Shakespeare), but to describe the written or spoken language through a representative sample of different textual genres (letters, newspaper articles, theatre,…). In the 1960s, "Corpus Linguistics" was born, a new method of investigating language which takes as its reference a finite set of linguistic executions that can be from spoken or written language. This approach is not universally shared by the scientific community, since the language is alive, therefore potentially infinite.

A corpus can only represent an incomplete way of all possible statements. According to Noam Chomsky, the object of linguistics is constituted by an inevitably partial set of performances but by the implicit and unconscious structures of natural language from which it is necessary to reconstruct the rules. If a corpus is a representative sample of the reference language, although it cannot claim to exhaust the complexity of the linguistic analysis, it can provide reliable, it is generalizable, and controllable information. In the 1960s, Jean-Paul Benzecrí introduced a new and innovative approach based on graphic forms. This approach of orientation is opposite

to that of Chomsky, with the creation of the French School. Ludovic Lebart and Alain Morineau (1985) develop the SPAD software (*Système Portable pour l'Analyse des Donneés*) which allows a wide application of these techniques to a large community of scholars from different fields of knowledge. Andrè Salem produced innovative textometry methods with the dissemination through the Lexico software proposal.

The French school develops (Benzecrí, 1977; Lebart 1982) which bases its originality on the analysis of graphic forms, repeated segments, analysis of the correspondence of lexical tables (LCA). Semantic space representation and latent semantic analysis (LSA) are the most relevant analysis techniques introduced in the 1980s [27]. In the 1990s, the widespread use of the Internet made available an enormous circulation of textual documents. Techniques for the extraction of relevant information are developed, such as data mining, text mining and web mining, which together with the techniques and machine learning models have allowed the dissemination and use of textual data in the following corporate and institutional areas. At the dawn of the development of the Internet, Etienne Brunet recognizes the role of the web as a reference for corpora linguistics: "*The Internet has become a huge storehouse of information and humanities researchers can find the statistics they need [...]. Statistics have always had ample space, large amounts of data and the law of large numbers. The Internet opens up his wealth, without control, without delay, without expenses and without limits"* [21].

At the beginning of 2000, there were many proposals for the identification of topics, such as the Latent Dirichlet allocation or analyses based on the mood of users as sentiment analysis [25]. Starting from 2014, big data and social media have been used to support official statistics, and the United Nations Global Working Group (GWG)[3] on big data for official statistics was created under the UN Statistical Commission. The GWG provides strategic vision, direction and the coordination of a global program on the use of new data sources and new technologies, which is essential for national statistical systems to remain relevant in a fast-moving data landscape. It is evident that from a methodological point of view, big data sources, the social data generation mechanism does not fall under the direct control of the statistician and is not known. This radically distinguishes this information from the usual statistical outputs derived both from traditional sample surveys, whose sampling design is designed and controlled by statistical institutes, and from administrative sources, whose data generation mechanism is usually known to survey designers. As a result, there is no rigorous methodology to date that guarantees the general validity of statistical information derived from textual social media data, customer reviews and posts. In particular, users cannot be considered as a representative sample of the population. It follows that these data can be explored, but they cannot guarantee the accuracy of any inferences on the population. In Italy, the Italian Institute of Statistics (Istat) proposed a statistical tool capable of assessing the mood of Italians on well-defined topics or aspects of life, such as the economic situation.[4]

[3] See the website: https://unstats.un.org/bigdata/.

[4] See the website https://www.istat.it/it/archivio/219585.

4 Text Analytics Process

The steps of a text analysis process, albeit strongly linked to the objectives of the analysis, elaborate a process of chain analysis that is expressed in different macro-phases (Fig. 2).

The first step is characterized by the definition of the research problem. The project, in fact, should be clear, well-focused and flexible. In this phase, it is necessary to specify the population under study, defined deadlines and goals. If an inductive approach is applied, it is appropriate to examine any external influences such as backgrounds and theories. In the deductive approach, the existence of data or information can be used to test hypotheses, benchmark and build models.

In the second step, the corpus is generated, collecting a set of documents. Therefore, many aspects related to the objectives and the sample collection will be clarified at this stage.

The third step is dedicated to pre-processing, which is extrinsic in a set of sub-phases dependent on the type of document and the aims of the analysis. A corpus that comes from a social survey with open questions, or a corpus that presented a collection of Tweets will get "the need" for a very different treatment. In a corpus from a survey, pretreatment requires a fairly traditional simple standardization, with the removal of punctuation, numbers and so on, and then it will be chosen whether to carry out a morphological normalization such as lemmatization [21]. Moreover, any meta information from survey can support eventually the building of a lexical table [27]. In the case of a corpus coming from a social media, there will be emojis and emoticons and many special characters that can be removed or incorporated the analysis [22].

In the previous step, the pre-processed text document is first transformed into an inverted index, that, in the fourth step, is transformed into a term-document matrix (TDM) or document-term matrix (DTM). In a TDM, the rows correspond to terms, and the columns correspond to documents. Alternatively, in a DTM, the

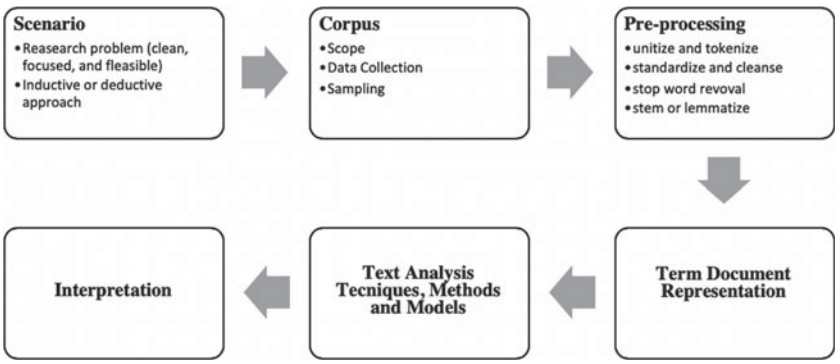


Fig. 2 Steps of a text analysis process

rows correspond to documents, and the columns correspond to terms. Local, global and combinatorial weighting can be applied to TDM or DTM.

In the fifth phase, methods, techniques and models based on goals are applied. The most used techniques in the exploration phase are the latent semantic analysis (LSA) and the correspondence analysis of lexical tables (LCA), and clustering techniques. For the research of the topics, a widespread probabilistic approach is the application of the probabilistic topic models. In recent years, machine learning classification techniques and sentiment analysis models have developed.

The analysis phase is followed by the interpretative phase, which allows to generate knowledge. This sixth phase of the supply chain can end the process or start another one if the results are not satisfactory.

5 Scientific Research and Collaborations

The world's scientific community has been publishing an enormous amount of papers and books in several scientific fields, in the way that it is essential to know which databases are efficient for literature searches. Many studies confirmed that today, the two most extensive databases are Web of Science (WoS) and Scopus [30]. Many articles are comparing Scopus and WoS databases, trying to find which one is the most representative of the existing literature [12, 13, 15, 16, 19, 20, 24]. These comparative studies concluded that even if these two databases are constantly improving, Scopus could be preferred due to its quality of outcomes, time savings and ease of use and because Scopus provides about 20% more coverage than WoS. For these reasons, we decided to implement the research about text analytics subject on Scopus. We selected all the articles in English, with no limits in time, including in their abstract, using the keyword "text analytics". The selected articles from the database were 580 (excluding the manuscripts published in 2020). In Fig. 3 we displayed, by year, the number of published articles.

We noticed that, using the keyword "text analytics" the first manuscript was published in 2003. Under this umbrella name, in 2019 the number of published articles was about 120 while until 2012 there were less than 25 manuscripts published by year. Moreover, from 2018 to 2019, there was a large jump in the number of published articles. After we had analysed the most active authors in the text analytics field, we investigated the co-authorship network [18] of the selected articles (Fig. 5), in order to describe the interactions between authors in this area. For a better readability of the graph, we constructed the network keeping only those nodes with a centrality degree score higher than 4. Among the authors, Liu S. has the highest centrality degree ($D = 19$), showing that it is the one who works more together with the other authors. On the other hand, Atasu K. and Polig R. have the highest number of links connecting each other. If we look to which subject area these articles belong, we can observe in Fig. 4 that most of the manuscripts are related to computer science area (41%), but also to many other areas, such as engineering (10.8%), social sciences (9.7%) and mathematics (8.6%). It means that, even if text analytics seems to be strongly related

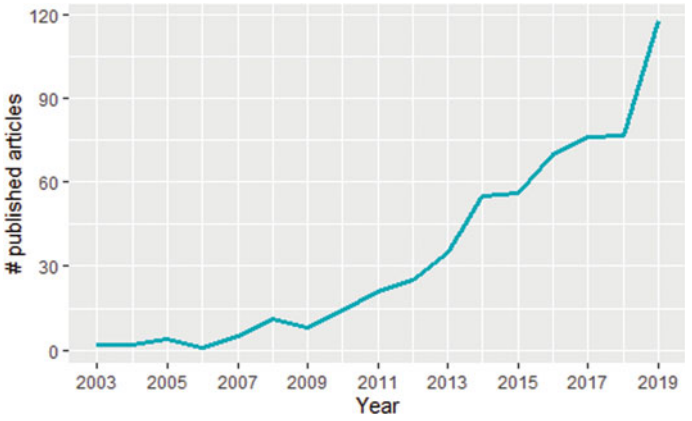


Fig. 3 Time plot of published articles related to text analytics subject from 2000 to 2020

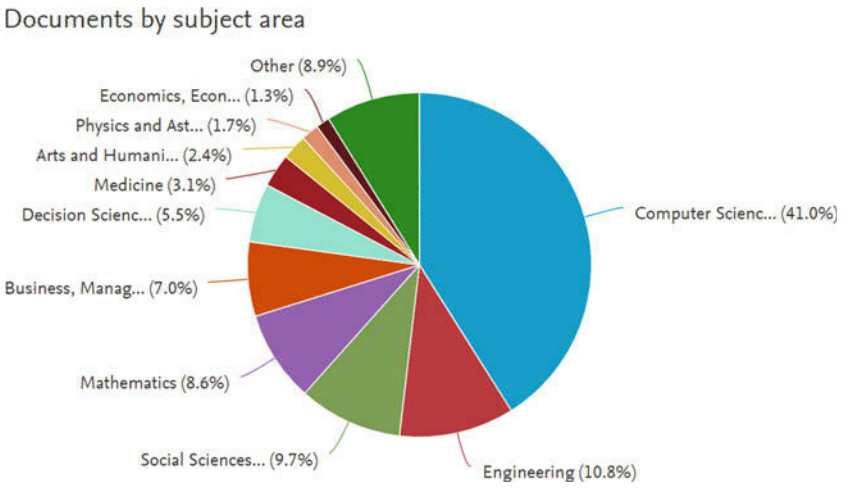


Fig. 4 Published articles related to text analytics subject, by subject area (2003–2019)

to the information technologies, it is a multidisciplinary topic, since it appeared in articles belonging to many areas (Fig. 6).

Table 2 shows the total number of published manuscripts listed by top eight authors. The most active authors in this field, from the Scopus query, are Raphael Polig (IBM Research), Kubilay Atasu (IBM Research) and Shixia Liu (Tsinghua University). The articles year of publication by these authors ranges from 2003 to 2019; 39 out of 41 (almost all) manuscripts belong to computer science area, confirming that top authors in the field of text analytics work in the information technology sector. The most productive researchers work in the USA (48%), India (13%) and Germany (5%).

Fig. 5 Co-authorship network

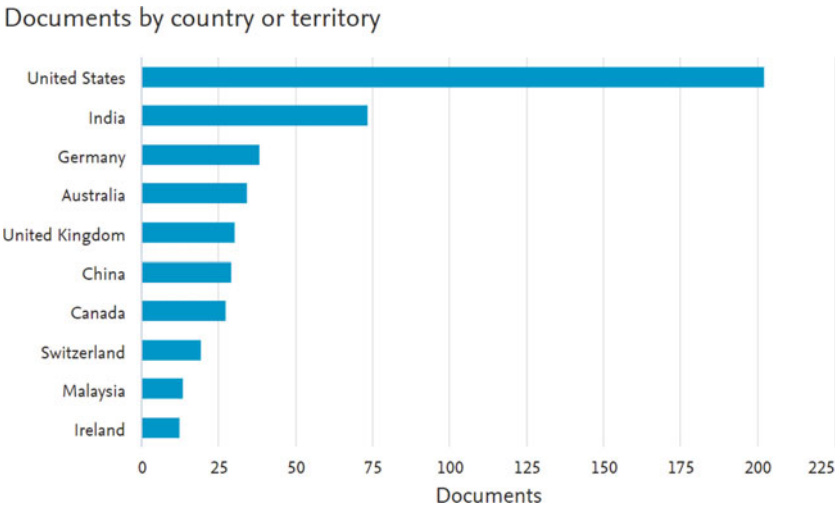


Fig. 6 Articles (no.) by Country

Table 2 Top eight authors, by number of published articles and country

Author	Article (no.)	Country
Polig R.	10	Switzerland
Atasu K.	8	Switzerland
Liu S.	7	China
Hagleitner C.	6	Switzerland
Chiticariu L.	5	USA
Dey L.	5	India
Palmer S.	5	Australia
Zhou Y.	5	USA

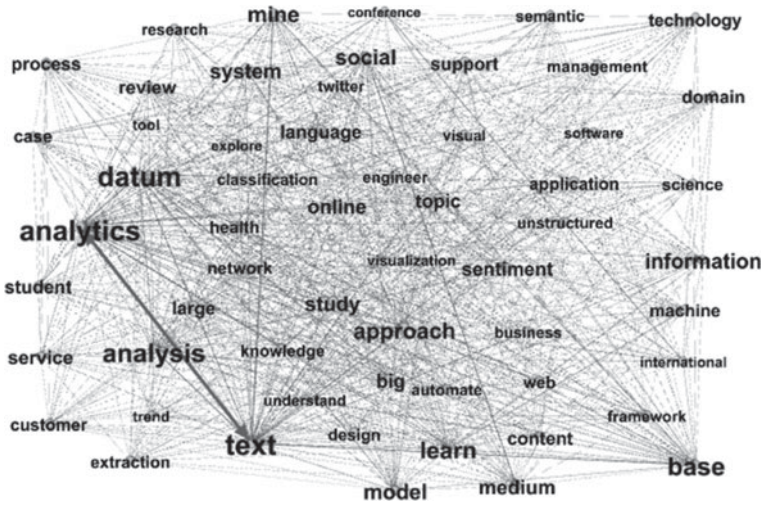


Fig. 7 Word co-occurrences network

In order to better understand the topics covered in the articles, we constructed a one-mode network of the words [14, 17] from the articles' titles. Figure 7 shows a graph, collecting all the titles of the selected manuscripts in order to construct a corpus composed of these short texts. This corpus had, before pre-processing, 6,149 occurrences and 1806 forms. We lemmatized the corpus, and we removed hapax forms and stop words (articles, prepositions, auxiliary verbs, conjunctions and pronouns), obtaining a term-document matrix of size (550×580). From the term-document matrix, we obtained the adjacent matrix of terms (550×550); for a better readability of the graph, we constructed the word network keeping only those terms with a centrality degree score higher than 100.

From the network emerged that text analytics subject is mostly connected to topics belonging to computer science area, e.g. text mining, data mining, social media but is also connected to more general terms which regard issues connected to many

different areas, e.g. information, system, student, customer, learn, model, approach, service, confirming the multidisciplinary nature of this topic. We have reviewed the research topics addressed in the past 20 years from work on this topic, and we have found that although the expression "text analysis" is relatively new (the first articles were published in 2003), it currently incorporates all the techniques, methods and models for automatic text analysis. Networks of scientific collaborations should be increased, also between different sectors. If the birth and development of the text analysis are due to the hybridization of knowledge, today more than collaborations between different research fields can ensure its future.

References

1. Giovannini E (2014) Scegliere il futuro, il Mulino., Bologna
2. United Nations (2014) Independent expert advisory group on a data revolution for sustainable development . A world that counts: mobilising the data revolution for sustainable development. Independent Advisory Group Secretariat. Available from http://www.undatarevolution.org/report/
3. Forrester Research (1995) Coping with complex data. The Forrest Report, April
4. Stephens-Davidowitz S, Pinker S (2017) Everybody lies : big data, new data, and what the Internet can tell us about who we really are, New York, NY
5. Bourdon B (1892) L'espression des émotions et des tendence dans le language. Alcan, Paris
6. Estoup JB (1916) Gammes sténografiques. Institut sténografiques de France, Paris
7. Busemann (1925) Die Sprache der Jugend als Ausdruck des Entwicklungsbrbythmuds. Fischer, Jena
8. The adjective-verb quotient: a contribution to the psychology of language. Psychol Rec 3:310–343
9. Zipf GK (1929) Relative frequency as a determinant of phonetic change. Harvard Stud Classical Philology 40: l–95
10. Zipf GK (1935) The psycho-biology of language. Houghton Mifflin, Boston, MA
11. Mandelbrot B (1954) Structure formelle des textes et communication. Word lO:l-27
12. Bakkalbasi N, Bauer K, Glover J, Wang L (2006) Three options for citation tracking: Google Scholar, Scopus and Web of science. Biomed Digital Libraries 3(1):7
13. Boyle F, Sherman D (2006) Scopus: the product and its development. The Serials Librarian 49(3)
14. Celardo L, Everett MG (2020) Network text analysis: a two-way classification approach. Int J Inf Manage 51:102009
15. Malagas ME, Pitsouni EI, Malietzis GA, Pappas G (2008) Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. FASEB J 22(2):338–342
16. Harzing AW, Alakangas S (2016) Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics 106(2):787–804
17. Iezzi DF (2012) Centrality measures for text clustering. Commun Statistics-Theor Methods 41(16–17):3179–3197
18. Liu X, Bollen J, Nelson ML, Van de Sompel H (2005) Co-authorship networks in the digital library research community. Information Process Manage 41(6):1462–1480
19. Mongeon P, Paul-Hus A (2016) The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics 106(1):213–228
20. Vieira ES, Gomes JANF (2009) A comparison of Scopus and Web of Science for a typical university. Scientometrics 81(2):587–600
21. Bolasco S (2013) L'analisi automatica dei testi: fare ricerca con il text mining. Carocci, Roma

22. Celardo L, Iezzi DF (2020) Combining words, emoticons and emojis to measure sentiment in Italian tweet speeches. In: JADT 2020 : 15th international conference on statistical analysis of textual data, 16–19 Jun 2020, TOULOUSE (France)
23. Bolasco S, Iezzi DF (2012) Advances in textual data analysis and text mining-special issue Statistica Applicata Italian. J Appl Statistics 21(1):9–21
24. Aria M, Misuraca M, Spano M (2020) Mapping the evolution of social research and data science on 30 years of social indicators research
25. Bing L (2012) Sentiment analysis and opinion mining. Synthesis Lect Human Language Technol 5(1):1–167
26. Witten IH, Bell TC (1991) The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. IEEE Trans Information Theor 37:1085–1094
27. Lebart L, Salem A (1988) Analyse statistique des données textuelle. Dunod, Paris
28. Herdan G (1964) Quantitative linguistics or generative grammar? Linguistics 4:56–65
29. Guiraud P (1954) Les caractheres statistiques du vocabulaire, Paris, P.U.F
30. Aghaei Chadegani A, Salehi H, Yunus M, Farhadi H, Fooladi M, Farhadi M, Ale Ebrahim N (2013) A comparison between two main academic literature collections: web of science and Scopus databases. Asian Soc Sci 9(5):18–26
31. Krippendorff K (2012) Content analysis: an introduction to its methodology, 3rd edn. Sage, Thousand Oaks, CA, p 441

Unsupervised Analytic Strategies to Explore Large Document Collections

Michelangelo Misuraca and Maria Spano

Abstract The technological revolution of the last years allowed to process different kinds of data to study several real-world phenomena. Together with the traditional source of data, textual data became more and more critical in many research domains, proposing new challenges to scholars working with documents written in natural language. In this paper, we explain how to prepare a set of documents for quantitative analyses and compare the different approaches widely used to extract information automatically, discussing their advantages and disadvantages.

1 Introduction

Reading a collection of documents to discover recurring meaning patterns is a time-consuming activity. Besides the words used to write the document content, it is necessary to consider high-level latent entities that express the different concepts underlying the text. As stated by cognitive psychologists, these concepts are "the glue that holds our mental world together" [39]. They are an abstraction of a particular set of instances retained in the human mind from experience, reasoning and/or imagination. Highlighting the primary topics running through a document can increase its understanding, but this task is challenging when there is not any other prior knowledge about the document itself. A topic can be represented as a set of meaningful words with a *syntagmatic* relatedness [47].

Identifying the topic of large document collections has become more and more important in several domains. The increased use of computer technologies allowed drawing up automatic or semi-automatic strategies to extract meaning from texts written in natural language. Because of the multifaceted nature of the linguistic

M. Misuraca (✉)
DiScAG—University of Calabria, Rende, Italy
e-mail: michelangelo.misuraca@unical.it

M. Spano
DiSeS—University of Naples Federico II, Naples, Italy
e-mail: maria.spano@unina.it

D. F. Iezzi et al. (eds.), *Text Analytics*, Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-52680-1_2

phenomenon, scholars belonging to diverse fields investigated this research issue founding their proposals on sometimes distant theoretical frameworks, with proper terminology and notation. Nevertheless, the mathematical (and statistical) base of quantitative natural language processing often led to very similar solutions.

Aim of this paper is briefly explaining how to prepare and organise a set of documents for quantitative analyses (Sect. 2), comparing the leading solutions proposed in the literature to extract a knowledge base as a set of topics (Sect. 3) and discussing advantages and disadvantages of the different approaches (Sect. 4).

2 From Unstructured to Structured Data

Let us consider a collection of documents written in natural language. Texts encode information in a form difficult to analyse from a quantitative viewpoint because their content does not follow a given data model. In the framework of an information mining process, it is then necessary to process the texts and obtain a set of structured data that can be handled with statistical techniques.

2.1 Data Acquisition and Pre-treatment

A document belonging to a given collection can be seen as a sequence of characters. Documents are parsed and tokenised, obtaining a set of distinct strings (*tokens*) separated by blanks, punctuation marks or—according to the particular analysed phenomenon—other kinds of special characters (e.g. hashtags, at-signs, ampersands). These tokens correspond to the words used in the documents. However, since the expression "word" is too generic and does not encompass other valuable combinations of words—e.g. collocations and multiword expressions—*term* is conventionally used instead of word. The particular scheme obtained with tokenisation is commonly known as *bag-of-words* (BoW), viewing each document as a multiset of its tokens, disregarding grammatical and syntactical roles but keeping multiplicity.

Once the documents have been atomised into their basic components, a pre-treatment is necessary to reduce language variability and avoid possible sources of noise, improving in this way the effectiveness of the next analytical steps [48, 55]. First of all, because of case sensitivity, it is often necessary to change all terms' characters to lower case. To keep track of the variety of language and proceed uniformly in the whole collection, it is then possible to perform other normalising operations such as fixing misspelt terms or deleting numbers. A more complex operation considers the inflexion of the terms, reducing variability at a morphological level. Two alternative solutions are commonly considered, *stemming* and *lemmatisation*. The main difference between the two approaches is that the first reduces inflected terms to their roots by removing their affixes, whereas the second uses terms' lemma transforming each inflected term in its canonical form. Even if stemmers are easier and faster to

implement, it is preferable to use lemmatisers to preserve the syntactic role of each term and improve the readability of the results (see [27, 28, 54]). When the texts have been pre-treated, it is possible to build the so-called *vocabulary* by stacking identical tokens/terms and counting the number of occurrences of each vocabulary entry (*type*) in the collection of documents.

Additional vocabulary pruning can be carried on to avoid non-informative terms, removing the so-called *stop words*, i.e. the most common terms used in the language and in the specific analysed domain. For the same reason, rare terms with a low number of occurrences are usually removed from the vocabulary. At the end of this process, the documents' content is ready to be processed as structured data.

2.2 Data Representation and Organisation

Vector space model [45] is commonly used to represent a document from an algebraic viewpoint. Let us consider a collection $\mathscr{D}$ containing n different documents. Each document d_i can be seen as a vector $(w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{ip})$ in a p-dimensional vector space $\mathscr{R}^p$ spanned by the terms belonging to the vocabulary. When a term j occurs in the document, its value w_{ij} in the vector is nonzero. This latter quantity can be seen as a weight representing the importance of the term j in d_i, i.e. how much the term contributes to explain the content of the document.

Several ways of computing these *term weights* have been proposed in the literature, mainly in information retrieval and text categorisation domains. According to Salton and Buckley [44], three components have to be taken into account in a weighting scheme:

- *term frequency*, used to express the relative importance of the terms in each document (local weight);
- *collection frequency*, used to capture the discrimination power of the terms with respect to all the documents (global weight);
- *normalisation*, used to avoid biases introduced by unequal document lengths, where the length is represented as the total number of tokens used in a document.

A primary distinction can be done between *unsupervised* and *supervised* term weighting schemes, depending on the use of available information on document membership. Unsupervised term weighting schemes include *binary* weights, *raw frequency* weights (*tf*), *normalised frequency* weights and *term frequency—inverse document frequency* weights (*tf-idf*). Several studies showed the influence of these schemes in a text mining domain (e.g. [2, 40]). Binary weights just consider the presence or the absence of a term j in a document d_i by assigning to w_{ij} a value of 1 or 0, respectively. Raw frequency weights are calculated as the number of occurrences of a term in a document and correspond to the absolute frequency. Normalised frequency weights incorporate $1/(\max tf_{ij})$ as normalisation factor of the raw frequencies, where $max\ tf_{ij}$ is the highest number of occurrences observed in the document d_i. In all

these schemes, the global weight and the length normalisation factor are set to 1. To include the discrimination power of the terms into the weighs, it is possible to consider a *tf-idf* scheme. The raw frequencies or the normalised frequencies are multiplied by a global weight $log(n/n_j)$, where n/n_j is the reciprocal of the fraction of documents containing the term j [50]. Often the *tf-idf* weights are normalised by the Euclidean vector length of the document, obtaining a so-called *best fully weight* [44]:

$$w_{ij} = \frac{tf_{ij} \cdot log(\frac{n}{n_i})}{\sqrt{\sum_{d_i} (tf_{ij} \cdot log(\frac{n}{n_i}))^2}}$$

Supervised term weighting schemes follow the same rationale of the *tf-idf*, using as global weights other measures developed for feature selection, such as χ^2, *gain ratio* or *information gain* [15, 17].

The different document vectors can be juxtaposed to form a *document-term* matrix $\mathbf{T}$ with n rows and p columns. This matrix is a particular contingency table whose marginal distributions provide different information depending on the chosen weighting scheme. The transposition of the matrix in a term-document matrix does not change data reading. In a binary document-term matrix $\mathbf{T_b}$, the marginal row totals state how many terms of the vocabulary are used in each document, whereas the marginal column totals state to the number of documents in which each term appears. In a raw frequency document-term matrix $\mathbf{T_{tf}}$, the marginal row totals state the lengths of the different documents, whereas the marginal column totals represent the term distribution in the overall collection, i.e. the vocabulary. When other weighting schemes are used, the marginal totals are not directly interpretable. As an example, the two marginal distributions of a tf-idf document-term matrix represent the total tf-idf per document and the total tf-idf per term, respectively.

When the research interest concerns the association between terms, it is possible to build a p-dimensional *term-term* square matrix $\mathbf{G}$. This matrix is commonly called *co-occurrence* matrix, because the generic element $g_{ii'}$ $(i \neq i')$ represents the number of times two terms co-occur together. The easiest way to obtain $\mathbf{G}$ is from the matrix product $\mathbf{T_b}^\top \mathbf{T_b}$, where $g_{jj'}$ $(j \neq j')$ is the number of documents in which the term i and the term i' co-occur. The diagonal elements g_{jj} correspond to the marginal column totals of $\mathbf{T_b}$. The co-occurrence matrix can be also obtained by splitting in sentences the document of the collection and building a binary sentence-term matrix $\mathbf{S_b}$. The generic element $\hat{g}_{jj'}$ of $\hat{\mathbf{G}} = \mathbf{S_b}^\top \mathbf{S_b}$ can be read as the number of sentences in which the term i and the term i' co-occur and more properly interpreted as the co-occurrence of the two terms in the collection. It is possible to obtain a different granularity of term-term co-occurrences changing the way documents are split, e.g. breaking documents in correspondence of stronger punctuation marks as full stops or splitting each document in clauses. Other measure can be used to express term association, such as *simple matching coefficient*, *Jaccard similarity*, *cosine similarity* [52, 53]. Following the same logic of term-term matrix, a n-dimensional *document-document* square matrix $\mathbf{H}$ can be derived to represent the pairwise document asso-

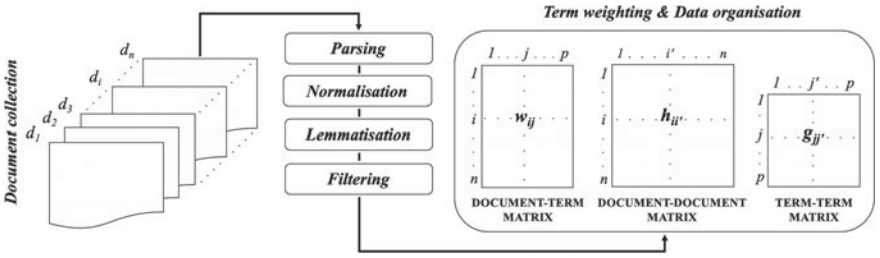


Fig. 1 From unstructured data to structured data: a workflow model

ciation in the collection. The document-document matrix is obtained from the matrix product $\mathbf{T_b T_b}^{\mathsf{T}}$, where $h_{ii'}$ is the number of terms shared by documents d_i and $d_{i'}$. The diagonal elements h_{ii} correspond to the marginal row totals of $\mathbf{T_b}$. As above for term-term matrix $\mathbf{G}$, different measure can be used to express pairwise association in a document-document matrix $\mathbf{H}$ [51].

A schematic drawing of the different steps above described is shown in Fig. 1.

3 Looking for Knowledge: A Comparison Between Text Analytics

Taking into account the different textual data representations and the distributional assumptions that can be made about the data themselves, it is possible to set up a general classification of the approaches mostly used in recent years to extract information from document collections automatically.

3.1 Factorial-Based Approach

Factorial techniques aim at minimising the number of terms required to describe a collection of documents, representing the most relevant information in a low-rank vector space. The dimensionality reduction is reached by considering linear combinations of the original terms expressing the content of the documents and visualising these latent association structures onto factorial maps.

Methods like *lexical correspondence analysis* (LCA:[5, 32]) and *latent semantic analysis* (LSA:[16]) are commonly used. LCA was developed in the framework of the French school of data analysis and is generally used to identify the association structure in the document-term matrix [4], using a raw frequency term weighting scheme. LSA became popular in the information retrieval domain to represent the semantic structures of a collection of documents, starting from a document-term matrix with a raw frequency or a tf-idf term weighting scheme. Although the two