SAPIENZA
UNIVERSITÀ DI ROMA

Department of Social Sciences and Economics

Ph.D. course in Applied Social Sciences

XXXI cycle

TEXT MINING FOR SOCIAL SCIENCES: NEW APPROACHES

(S.S.D. SECS-S/05)

Livia Celardo

Tutor: Prof.ssa D. Fioredistella Iezzi

Co-tutor: Prof. Luigi M. Solivetti

This research work is an edited and extended version of the papers:

Celardo L., Iezzi D.F., Vichi M. (2016). Multi-mode partitioning for text clustering to reduce dimensionality and noises. In (edited by) Mayaffe D., Poudat C., Vanni L., Magri V., Follette P. *JADT 2016: Statistical Analysis of Textual Data*, pp. 181-192. Les Press de Fac Imprimeur.

Celardo, L. (2017). Classifying textual data: a two-way approach. *Working papers series - PhD Course in Applied Social Sciences*, 6/2017.

Celardo, L. (2018). Opportunities of Using Big Data in Social Sciences: Work Injuries through Media Analysis. *Working papers series - PhD Course in Applied Social Sciences*, 9/2018.

Celardo L., Vallerotonda R., De Santis D., Scarici C., Leva A. (2018). Analysing occupational safety culture through mass media monitoring. In (edited by) Iezzi D.F., Celardo L., Misuraca M. *JADT 2018: Statistical Analysis of Textual Data*, pp.150-156. UniversItalia.

Greco F., Alaimo S.L., Celardo L. (2018). Brexit and Twitter: The voice of people. In (edited by) Iezzi D.F., Celardo L., Misuraca M. *JADT 2018: Statistical Analysis of Textual Data*, pp.327-334. UniversItalia.

Contents

Part I
Theory and problems

Part II
Real data implementations

Introduction

The rise of the Internet has determined an important change in the way we look at the world, and then the mode we measure it. In June 2018, more than 55% of the world's population has an Internet access[1]. It follows that, every day we are able to quantify what more than four billion people do, how and when they do it. This means data.

The availability of all these data raised more than one questions: *How to manage them? How to treat them? How to extract information from them?* Now, more than ever before, we need to think about new rules, new methods and new procedures for handling this huge amount of data, which are characterized by being unstructured, raw and messy.

One of the most interesting challenge in this field regards the implementation of processes for deriving information from textual sources; this process is also known as *Text Mining*. Born in the mid-90s, Text Mining represents a prolific field which has evolved – thanks to technology evolution – from the Automatic Text Analysis, a set of methods for the description and the analysis of documents.

Textual data, even if transformed into a structured format, present several criticisms as they are characterized by high dimensionality and noise. Moreover, online texts – like social media posts or blogs comments – are most of the time very short, and this means more sparseness of the matrices when the data are encoded. All these findings pose the problem of looking at new and advanced solutions for treating Web Data, that are able to overcome these criticisms and at the same time, return the information contained into these texts. The objective is to propose a fast and scalable method, able to deal with the findings of the online

[1] https://www.internetworldstats.com/stats.htm

texts, and then with big and sparse matrices. To do that, we propose a procedure that starts from the collection of texts to the interpretation of the results. The innovative parts of this procedure consist of the choice of the weighting scheme for the term-document matrix and the co-clustering approach for data classification. To verify the validity of the procedure, we test it through two real applications: one concerning the topic of the safety and health at work and another regarding the subject of the Brexit vote. It will be shown how the technique works on different types of texts, allowing us to obtain meaningful results.

For the reasons described above, in this research work we implement and test on real datasets a new procedure for content analysis of textual data, using a two-way approach in the Text Clustering field. As will be shown in the following pages, Text Clustering is a process of unsupervised classification that reproduces the internal structure of the data, by dividing the text into different groups on the basis of the lexical similarities. Text Clustering is mostly utilized for content analysis, and it might be applied for the classification of words, documents or both. In latter case we refer to two-way clustering, that is the specific approach we implemented within this research work for the treatment of the texts.

To better organize the research work, we divided it into two parts: a first part of theory and a second one of application. The first part contains a preliminary chapter of literature review on the field of the Automatic Text Analysis in the context of data revolution, and a second chapter where the new procedure for text co-clustering is proposed. The second part regards the application of the proposed techniques on two different set of texts, one composed of news and another one composed of tweets. The idea is to test the same procedure on different type of texts, in order to verify the validity and the robustness of the method.

Part I

Theory and problems

Chapter I

AUTOMATIC TEXT ANALYSIS IN THE WEB 4.0 ERA

In recent times, the diffusion of the Internet has generated a revolution in many areas of Social Sciences. Today, people increasingly share information online, creating a huge amount of available data; for that reason, the Web represents today the first source of the so-called “Big Data”.

As extensively reported in the literature, Big Data represent a huge and important source of information for social research, even if many open questions about the extraction and the utilization of them persist. Then, along with the growth of the Internet, the importance of text analysis applications is growing proportionally to the exponential growth of electronic text. In this sense, Big Data represents a significant opportunity for the development of text analysis methods and applications. Actually, with the Internet going inside the lives of more and more people, email, chat, newsgroups, blogs, etc. have become very popular and they generate a huge amount of text data every day (Agarwal, Godbole, Punjani and Roy, 2007). To have an idea, in 2016 global Internet networks carried more than 20,000 Gigabytes per second (CISCO, 2017); the 90% approximately of these data are unstructured (photos, texts, videos, etc.). In order to analyse them, specific methods designed for unstructured data are required.

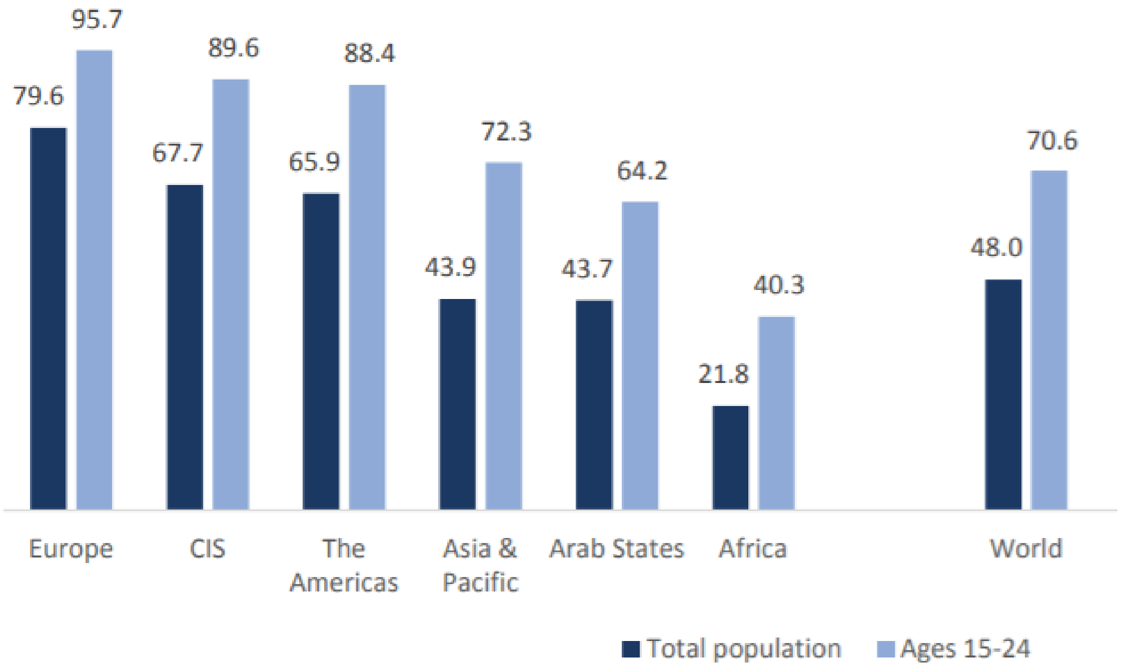
In this framework, Automatic Text Analysis is the set of methodologies for describing and analysing textual data; putting together linguistics, statistics and informatics, Automatic Text Analysis allows the extrapolation of structured information from large set of texts. Documents, although transformed into a structured data, are still characterized by a high dimensionality and noise; so, what matters is the method of the analysis and not only the data source or the amounts (Iacus 2014). Texts are different compared to structured data, so they require different methodologies, designed explicitly for their features. Moreover, online texts have specific findings – e.g. high dimensionality, noisy text, short structure of the sentences – that require new methods, able to manage and elaborate these data.

This chapter is structured as follow. Section 1.1 introduces the emergent phenomenon of the “Data Revolution”, that is the constant arise in the data availability, due to the Internet diffusion. Section 1.2 presents a brief review about the evolution and the main concepts of the Automatic Text Analysis. Section 1.3 shows one of the most important statistical method for Automatic Text Analysis, the Text Clustering.

1.1 Data revolution

Social Sciences are now experiencing an historic change coming from the availability of enormous quantity of highly informative data in every field (King, 2014). In the last half-century, the information base of social science research was primarily the survey; it consists of a systematic and standardized approach to the collection of information on individuals, households, or other entities through the questioning of systematically identified individuals, and it is coming into prominence as a research technique only in the last 50 years (Rossi, Wright and Anderson, 2013). The domain of that instrument in research finds motivation in many strong points, such as the low costs for large sample reaching, the low work for data collection and elaboration and the high level of standardization of the process. Thanks to technology evolution, in the course of time the development of new kind of survey data collection (telephone and computer assisted interviews) has been entailed to improve performances. Indeed, in the last years the growing diffusion of the Internet use has determined a crucial turning point for the research: the emergence of the *Web Data*, generated by the interaction between individuals and the cyberspace. With the rise of the online users, these data become ever much significant. The United Nations specialized agency for information and communication technologies (ITU) has estimated that the in 2017 the 48% of the total population used Internet (Figure 1.1), while, the proportion of households with Internet access at home is now equal to 53.6% (ITU, 2017).

Figure 1.1 – Proportion (%) of individuals using the Internet in various regions of the world, by age (2017).



Source: ITU (2017).

According to the Cisco's Visual Networking Index initiative, who estimates the communication capacity of the Internet measuring the traffic moving through it, in 2016 global IP traffic was 1.2 Zettabytes[2] per year or 96 Exabytes[3] per month (CISCO, 2017); the 90% of these data are unstructured and many types: photos, texts, likes, etc. They represent an important source of information, providing additional knowledge we may use to understand the characteristics of modern society more in-depth, especially regarding the reports

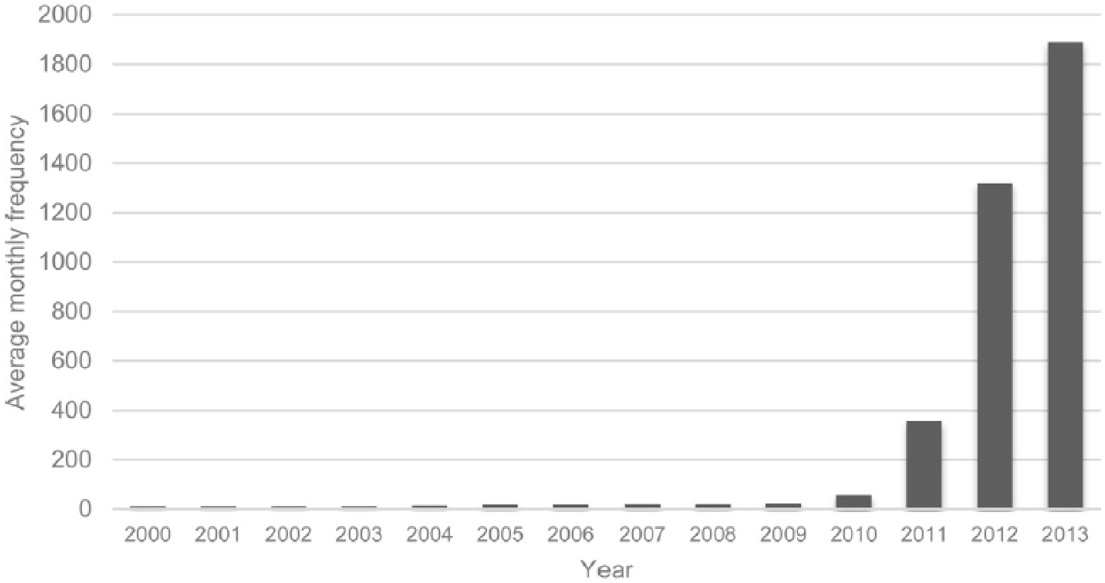
[2] The zettabyte (ZB) is a multiple of the unit byte for digital information. The prefix *zetta* indicates multiplication by the seventh power of 1000, so: 1 ZB = 1000^7 bytes.

[3] The exabyte (EB) is a multiple of the unit byte for digital information. The prefix *exa* indicates multiplication by the sixth power of 1000, so: 1 EB = 1000^6 bytes.

on events whose measurement is out of reach. In spite of this, the crucial point in treating these data is not just the dimensionality, but mainly the heterogeneity; then, it is necessary to use good and correct statistical methods in order to extract meaningful information from them.

More generally, Web Data belong to what we know as *Big Data*, defined as those data sets where traditional software of data management and data analysis are not capable to deal with them. As the industry has rapidly adopted Big Data, also academics and scientific journals have promptly covered the topic (Figure 1.2).

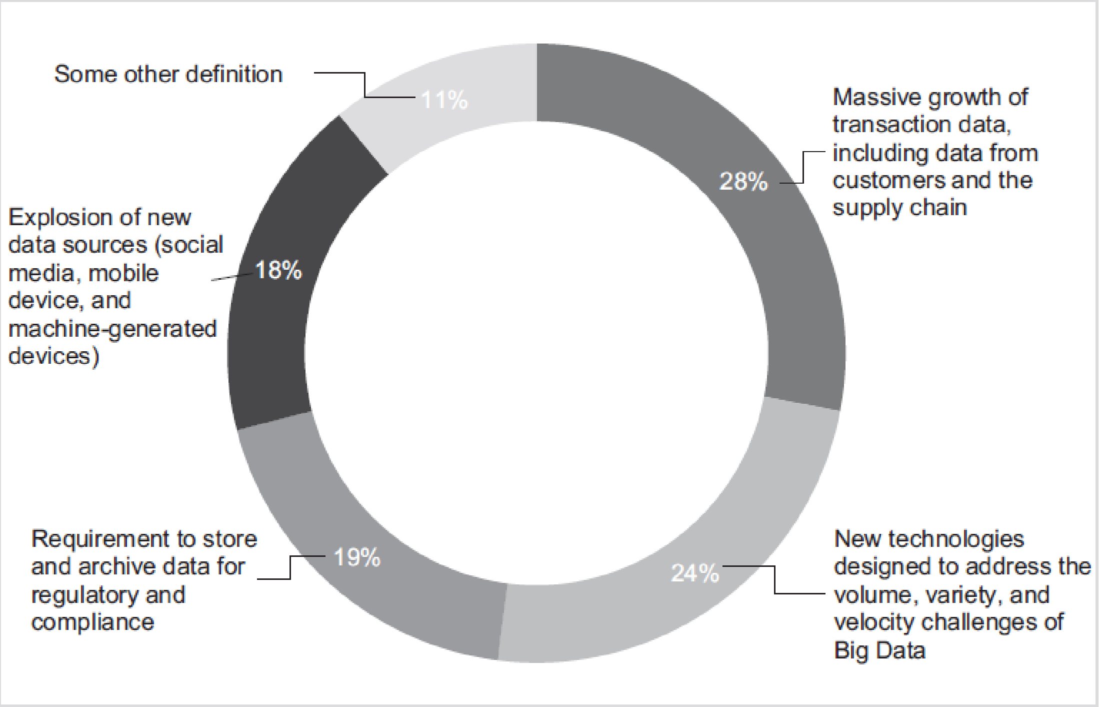
Figure1.2 – Frequency distribution of documents containing the term "Big Data" in ProQuest Research Library (2000-2013).



Source: Gandomi and Haider (2015).

In the last years, the definition of Big Data has evolved rapidly, also raising some confusion. This is evident from an online survey conducted by Harris Interactive in 2012, where executives tried to give their definition of Big Data (Figure 1.3).

Figure 1.3 – Definitions of Big Data, based on an online survey of 154 global executives in April 2012.



Source: Gandomi and Haider (2015).

Due to the need of a more precise explanation, at the beginning of the 21st century it has been proposed by Laney (2001) to define Big Data as that kind of information we can identify through three main properties, also known as the *three V's*:

1. *Volume*. Even if the definition of this aspect depends on many factors, generally Big Data exceed the capacity of traditional computing methods;
2. *Variety*. It refers to the structural heterogeneity of these data. Big data are mostly unstructured – like texts, images, audio tracks and videos – and raw, messy and not ready to be used by traditional data analysis software;
3. *Velocity*. It concerns to the rate or the frequency at which data are produced and the speed at which it should be analysed – due to the growing need for real-time analyses.

In addition to these dimensions, recently other factors have been identified as Big Data properties (Gandomi and Haider, 2015). The first one is the *veracity*: IBM defined the fourth V as the unreliability inherent in some sources of data. For example, sentiments in social media are uncertain in nature, so it is necessary to address these data using tools and analytics developed for management and mining of uncertain data. There is also *variability*: SAS describes this element as the constant variation in the data flow and in the Big Data velocity. *Complexity*, according to SAS, refers to the fact that these data are generated through several sources. The last one is *value*: Oracle introduced this attribute as the ratio of the data value on its volume. Based on the Oracle's definition, Big Data are often characterized by "low value density"; this means that the data received in the original form usually has a low value relative to its volume.

Starting from the source, we may identify three main types of Big Data (Iacus, 2014):

1. *Administrative data*. They are generated and stored by persons or organizations for regulatory or government activities;
2. *Transaction data*. They come from high frequency financial transactions, like credit/debit card operations;
3. *Social media/networking data*. They are generated by the voluntary interaction of people with the express purpose of sharing with others

(Couper, 2013; Iacus, 2014). These data are created on social media, which refers to online platforms that allow users to create and exchange content (Gandomi and Haider, 2015); they are mainly composed of digital texts – i.e. opinions and comments written by people who want to share their own thought with others – containing a lot of noise.

Social media data can be used also by institutions to investigate general opinions about specific themes; indeed, these information could be used to integrate and enhance official statistics, in order to have a better comprehension of the observed phenomena. For instance, the European Joint Research Centre has founded the Europe Media Monitor[4], a research project of daily monitoring of online news from different media, that aims to extract and synthetize information from the articles. Media news and newscasts are unstructured data produced by media channels; they are highly numerous and frequent, heterogeneous (texts, images and videos) and easily downloadable from the web.

Big Data diffusion is having a huge impact also on institutions where production of statistics and data analysis is the core dealing; specifically, we are talking about organizations responsible for official statistics production (National Statistical Institutes). The role of official statistics is fundamental for the society, so the way Big Data will be involved has consequences for everyone. The increasing need of the National Statistical Institutes (NSIs) in using Big Data and integrating different sources has been expressed through the realization of many projects. For instance, the Italian National Institute of Statistics (ISTAT) aims to move toward a new organisational structure – from a vertical to a horizontal approach, supporting a model where survey data, administrative data and data from new sources are integrated in a single data asset; it is the Integrated System of Statistical Registers, whose aim is to increase quality, efficiency, time saving

[4] http://emm.newsbrief.eu/NewsBrief/clusteredition/it/latest.html

and statistical outputs (Alleva, 2017). In this scenario, the ISTAT has also developed the ARCHIMEDE project (Integrated Archive of Economic and Demographic Micro Data) that brings together microdata[5] from administrative sources related to individuals and households living in Italy; starting from the integration of these sources, it intends to expand databases by collections of data that can be used for research, planning and public policy evaluation (Mazziotta, 2017). At the European level, 22 partners between NSIs and Statistical Authorities are carrying out the ESSnet Big Data project, in order to prepare the European Statistical System (ESS) for integration of Big Data sources into the production of official statistics (Struijs et al., 2017).

[5] Microdata is data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment (OECD).

1.2 Automatic Text Analysis

One of the most important ability for any researcher or practitioner is to manage the information contained within documents in order to extrapolate and interpret the content (Giuliano and La Rocca, 2008). Thanks to technology evolution and Internet diffusion, today most of these documents are digital; this allows the easy memorization, visualization and information extraction by means of specific software.

Nowadays, two main approaches are recognized in the international practise for the treatment of texts: the *Computer Assisted Qualitative Data Analysis* (Conrad and Reinarz, 1984) and the *Automatic Text Analysis*. The first approach is a semi-automatic procedure used to ease the reading of the texts by a priori questions, while the Automatic Text Analysis is a basket of methods for the description and the analysis of textual data (Iezzi, 2009), without any direct reading of them. Using the instruments of statistics, computer science and linguistics, Automatic Text Analysis allows a quantitative measurement of large texts (Feldman and Sanger, 2007) in a very short time (Bolasco and De Mauro, 2013). The general objectives of the Automatic Text Analysis are (Bolasco, 2012; Lebart, Salem and Berry, 1997):

- to study the meaning of what is expressed in natural language, with the aim of describing its lexical characteristics, extracting specific information from it, or define its ‘structural’ features;
- to analyse the content of a ‘context unit’ (entire documents or fragments of texts) and produce global knowledge by using multi-dimensional statistical methods (both mapping techniques and automatic classification techniques), with the aim of grasping the fundamental sense of the documents studied, beyond the simple words.

Starting from the objectives described above, it is possible to define the type of analysis in two different ways (Bolasco and De Mauro, 2013): in the first case we deal with the *lexical analysis*, where the aim is to investigate the vocabulary or how the discourse is produced, while in the other case we refer to the *textual analysis*, where the interest is in examining the content of the text. The lexical treatment of texts regards the study the terminology of the corpus, that is the set of keywords extracted from the text; the words represent the unit analysis, also named text unit. In fact, lexical analysis provides a study of the language, where the corpus is analysed (as a whole) from the point of view of language without considering fragmentations and/or sub-groups created according to some partition. The lexical analysis is thus achieved by navigating the vocabulary in which each word is associated, in addition to the number of occurrences, the length, the number of characters, the grammatical category and other characteristics. The vocabulary is therefore the reference domain for each lexical activity as a list composed of several thousand entries. On the other hand, in textual analysis the object of study is the corpus, as a set of occurrences that follow each other in sentences along the entire width of the text, according to the development of the discourse. The corpus is then studied as a collection of texts to be analyzed, compared and categorized. The unit of analysis is now the context unit or the fragment – and not yet the type, which can be both a sentence or the entire text.

1.2.1 From text statistics to text mining

Automatic Text Analysis has been developed since the 1950s, but only in the mid-1960s with the increasing use of computers it has seen the expansion of software with standardised procedures and the application of these techniques in the fields of Social Sciences and marketing. Over the years, the interest for quantitative studies on texts has shifted from a linguistic viewpoint (developed up to the 1960s) to a lexical type (around the 70s), up to textual or even better lexico-textual perspective since the 80s (Bolasco, 2004).

The first approaches to quantitative analysis of texts were made for linguistics purposes (Zipf, 1935; Zipf, 1949; Yule, 1944; Guiraud, 1954; Herdan, 1956; Herdan, 1964). In the 60s Benzécri started some pioneering studies (Benzécri, 1963), giving rise to what we know as the *analyse des données*[6] (Benzécri, 1973). The birth of the *analyse des données* has determined a remarkable leap in the field of text analysis, as the first methods for textual data has been proposed. During the 1970s and the 1980s there was in the field of Text Analysis a decisive turning point, since the concept of textual statistics based on the analysis of graphical forms[7] was introduced (Muller, 1973; Lafon, 1984). At the same time, in Italy Zampolli and De Mauro laid the foundation for the quantitative linguistics implementing the first linguistic resources (Bortolini and Zampolli, 1971; De Mauro, 1980). The interest in literacy gradually changed towards texts coming from other kind of sources – field investigations, analysis of short texts (abstracts, bibliographies, messages, etc.). At the end of the 80s, Lebart and Salem defined the boundaries of the textual statistics (Lebart and

[6] The *analyse des données* – i.e. "data analysis" – is a family of multidimensional statistical methods, characterized by a propensity to the study of large datasets and by the utilization of graphical representation.

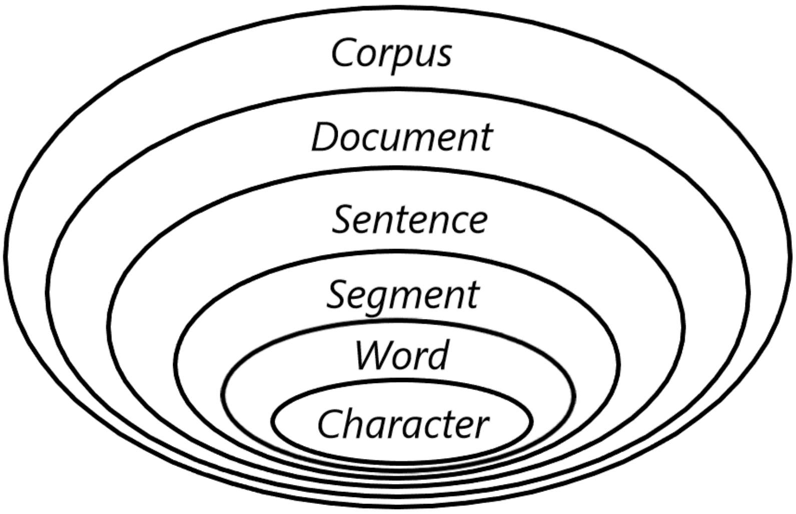
[7] A graphical form – i.e. word – is a sequence of characters, and it may represent the textual unit of a corpus.

Salem, 1988) – based on graphical forms instead of headwords, implementing new software for text analysis (Spad_T, Lexico). Simultaneously, quantitative linguists developed instruments for computational linguistics, creating new lexicons (De Mauro, 1993). In this context, the growing diffusion of information technology has improved the attention to the potentialities of text analysis procedures, up to what we know as Text Mining (Hearst, 1997; Sullivan, 2001), a set of techniques born in the mid-90s that allow the extraction of information from textual data using Information Retrieval and Information Extraction (Bolasco, 2005). Text Mining is an emerging area of Computer Science that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing (NPL), information retrieval and knowledge management (Feldman and Sanger, 2007); its objective is to synthetize, categorize, classify and select of texts for extracting valuable information (Bolasco and De Mauro, 2013).

1.2.2 Definitions and general concepts

Any Text Analysis starts collecting a set of similar documents (books, papers, news, abstracts, tweets, reviews, etc.) in order to structure a *corpus*, that is a collection of comparable texts. A text is a sequence of words; any word in the corpus – also counting words that occur more than once – is called *token*, while the different words in the corpus are the *types*. A corpus *vocabulary* is the list of the types together with their frequencies. Within a corpus, the words that occur just one time are called *hapax*. The number of tokens represents the size of the corpus in terms of occurrences, while the number of types is the size of the vocabulary in terms of different words.

Figure 1.4 – Corpus structure, different units of which the text is composed of.



Source: Our elaboration from Lebart, Salem and Berry (1997).

Within a corpus, is firstly necessary to identify the various units of which the text itself is composed of (Figure 1.4). The operation of subdividing the corpus into minimal units – that are not to be subdivided further – is called segmentation or tokenization. This phase of breaking the text up into distinct units is followed by a phase of identification, that is, grouping identical units together (Lebart, Salem and Berry, 1997). On the basis of the purposes of the research, different units are chosen and grouped together; for example, a researcher investigating a collection of news in the field of labour market might require that the words "worker" and "workers" be grouped into the same unit, so the presence or the absence of one or both terms is verified once.

A simple way to analyse a corpus is by choosing as units the types – also known as graphical forms. The graphical forms coming from the text segmentation may be further normalized, through *lemmatization* or reduction to their roots. In the first case each type is reduced to its base form, so for instance "am" and "is" are grouped into the unit "to be". In the other case, the procedure – called *stemming* – produces a stronger drop of the vocabulary, then for example the words "compression" and "compressed" are reduced to the unit "compress". The motivation in doing lemmatization or applying stemming to the corpus lies in the fact that they allow a reduction of the vocabulary; as mentioned before, one of the most crucial problem in text mining is the high dimensionality of the data and the consequent presence of high sparse matrices.

After the segmentation of the corpus is done, often is not necessary to keep in the analyses all the units previously identified; in fact, the different words are not equally important and distinctive (Bolasco, 2004), so it might be necessary to select some keywords in order to preserve the major of the information and simultaneously reduce the dimensionality. The process of selecting keywords from a corpus is done removing those words that belong to a stop list; these forms – i.e. *stop words* – are selected because for the purposes of the analysis they don’t contain any relevant information. A stop list is created by the researcher or by the user, time to time, on the basis of the corpus characteristics and the research objectives. For instance, in the case of content analysis function words – i.e. forms with little lexical meaning that express grammatical relationships – are often included in the stop list, while content words (nouns, verbs, adjectives and adverbs) are selected as keywords. The keyword selection is hence the process of selecting a sub-set of relevant types that contains the higher level of information content, and in this sense represents a feature selection method.

Because of the high dimensionality of textual data, it is strictly necessary to reduce it through specific methods. There are various ways for handling with high dimensionality; the main approaches regard the techniques for *feature selection* and *feature extraction* (Balbi, Misuraca and Spano, 2018). Feature extraction lets the reduction of the dimensionality through the projection of the objects in a sub-space. On the other hand, feature selection techniques use external information to discard terms that are not relevant for the analysis (Balbi, 2010) and they are implemented for several reasons (James, Witten, Hastie and Tibshirani, 2013):

1. the simplification of the data through the removing of the noise makes them easier for users to interpret;
2. it takes a shorter time to analyse the text;
3. it decreases the dimensionality;
4. it reduces the variance.

In a text analysis procedure, tokenization, normalization and keyword selection activities belong to what we known as the pre-processing phase. It is an essential part where the corpus is pre-processed and then prepared for the following phases. The pre-processing allows not only the identification of text units, but also the reduction of the data dimension through the elimination of the noise.

Noisy unstructured text is quite common in informal settings such as chat, SMS, email, newsgroups and blogs, automatically transcribed text from speech, and automatically recognized text from printed or handwritten material. In fact, documents produced under such circumstances are typically highly noisy containing spelling errors, abbreviations, nonstandard words, false starts, repetitions, missing punctuations, missing letter case information, pause-filling words (like “um” and “uh”), and other text and speech disfluencies (Agarwal, Godbole, Punjani and Roy, 2007).

1.2.3 Encoding texts: from unstructured to structure format

A quantitative approach to texts requires a preliminary transformation of them in structured data, through a system of codes (Iezzi, 2012). The most common way to do that is by using the Space Vector Model proposed by Salton, Wong and Yang (1975), also known as bag-of-words model, where each text is represented by a vector of weighed terms of the form:

$$\mathrm{d}_j = \left(w_{1,j}, w_{2,j}, \ldots, w_{i,j}, \ldots, w_{k,j}\right),$$

where $\mathrm{w}_{i,j}$ represents the weight for the term t_i attached to the document d_j. Then, by joining these vectors we obtain the term-document matrix – also known as lexical table, a mathematical matrix that describes the presence of terms that occur in the corpus:

$$\begin{pmatrix} w_{1,1} & \ldots & w_{1,j} & \ldots & w_{1,n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ w_{i,1} & \ldots & w_{i,j} & \ldots & w_{i,n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ w_{t,1} & \ldots & w_{t,j} & \ldots & w_{k,n} \end{pmatrix}.$$

The principal advantage of the term-document matrix is the low complexity of the tool, even if these matrices are very sparse – i.e. with a lot of zero entries.

After the text is transformed into a matrix, it is necessary to choose a scheme for weighting the terms, where a weight is a numerical value which is directly proportional to the importance of a word in a document. The most common weighting schemes are (Iezzi, 2012):

1. *Boolean scheme*: w_{ij} takes the value 1 if the term *i* is present in the document *j*, and value 0 otherwise. It expresses the presence or the absence of each word in the documents;

2. *Term Frequency (TF) scheme*: it shows how many times each term occurs in the texts. Therefore, w_{ij} is equal to the frequency of *i* in the document *j*:

   $$w_{ij} = n_{ij};$$

3. *Normalized Term Frequency (NTF) scheme*: it gives adjusted term frequencies on the basis of the documents length. Then, w_{ij} is equal to the frequency of the word *i* in the document *j*, divided by the maximum frequency of *i* in the corpus:

   $$w_{ij} = \frac{n_{ij}}{\max_{n_{ij}}},$$

   where n_{ij} is the frequency of the word *i* in the document *j* and $\max_{nij}$ is the maximum frequency of the term *i* in the corpus;

4. *Term Frequency - Inverse Document Frequency (TFIDF) scheme* (Salton and Buckley, 1988): it is a weighting scheme where the emphasis is given to those terms that occur with a high frequency in few documents, because rare words are more discriminative than general terms – occurring in many documents. Then, for this scheme:

   $$w_{ij} = \frac{n_{ij}}{\max_{n_{ij}}} \log \frac{N}{n_i},$$

   where n_{ij} is the frequency of the word *i* in the document *j*, $\max_{nij}$ is the maximum frequency of the term *i* in the corpus, *N* is the total number of documents and n_i is the number of documents in which the word *i* appears.

The choice of which scheme is better for weighting the terms depends on the research objectives; the TFIDF index, for instance, is widely used to identify in the text the significant language, composed of those terms that discriminate the different documents.

Another way to weight the matrix in order to discriminate the texts is by using the *specificities method* (Lafon, 1984); it calculates the over and under-using of each word in the different documents by a comparison of the different frequencies with the expected frequency of the terms in the corpus. Based on the observed frequency in one text, the specificities method quantifies for each term the discrepancy between the expected and the observed frequency: positive specificities have a higher than expected frequency while negative specificities appear less frequently that they should (Drouin, Francœur, Humbley and Picton, 2017). The measure of the specificity of a word in a document is given by the formula:

$$z = \frac{(x - x_t)}{\sigma_x},$$

where x is the word frequency in the document, x_t is the expected value of the word frequency in the document and σ_x is the standard deviation of x. Assuming the corpus as a population and each document as a sample, we used the hypergeometric law as probabilistic distribution model. Under specific conditions, this distribution is closely approximated by a normal distribution. The specificity is calculated using a statistical test, that under the assumption that terms are distributed evenly in the corpus and thus, in the documents, compares the observed (x) with the expected frequency (x_t); this test is made under the hypothesis of a normal distribution, so it is possible to assume the rules of a standardized variable z. Then, within a document, we define a term as “specific” if z is higher than |2|, while a value close to zero indicates that the word appears in the document as expected.

1.3 Text clustering

As mentioned before, in order to analyse the content of a corpus, statistical methods can be implemented on textual data. Statistical models for Automatic Text Analysis work for to globally representing a text, in terms of meaning. This level of study is related to the content, which is based on the study of co-occurrences. Based on the correlation or similarity between lexical profiles, it means discovering some models that give information on the system of meanings within the text.

The most used methods refer to two main categories: methods for *dimensionality reduction* – e.g. lexical correspondence analysis – and methods for *text classification*. The lexical correspondences analysis is a multivariate technique that allows to synthesize the information contained in a large matrix of textual data, showing on the factorial plane the association between the forms and looking for the best simultaneous representation of the row and column elements, in order to study the interdependence between characters (Lebart, Salem and Berry, 1997). On the other hand, the goal of text classification is to automatically classify the text documents into one or more defined categories. Algorithm for text classification can be supervised or unsupervised; in the latter case we also talk about Text Clustering.

Text Clustering is a process that allows to classify large sets of documents in groups based on their attributes, with the aim of reproducing the internal structure of the data (Iezzi 2010); the main objective is to split the corpus in different subgroups on the basis of words/documents similarities (Iezzi and Mastrangelo 2014). Text clustering is mostly utilized to analyse the content of the corpora, through the identification of words groups, each one representing a subject in the corpus. This technique corresponds to Cluster Analysis, and it doesn't require external information related to categories; in fact, clustering methodology is especially appropriate when no prior information is available about the data (Feldman and Sanger, 2007). However, Feldman and Sanger (2007) identified in text clustering some distinctive characteristics that structured data partitioning have not: first, the major complexity and richness of internal structure in text documents implies a more problematic phase of dimension reduction; in fact, most of the words in a corpus are irrelevant to the categorization and represent only a noise in the data. Secondly, the problem of finding meaningful and concise descriptions of the clusters, that are not merely the centroids; lastly, the measurement of algorithm quality, that in text mining undoubtedly requires the subjective human judgment.

1.3.1 One-way versus two-way approaches

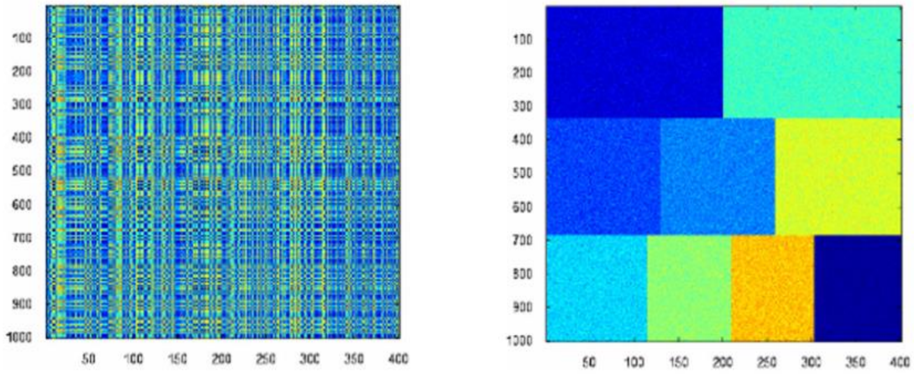
In text mining the most used partition algorithm for clustering is the *k*-means proposed by MacQueen (1967), because of its efficiency, even when it processes big sparse matrices (Iezzi, Mastrangelo and Sarlo, 2012; Iezzi, 2012b). Then, for general two-way data matrix (objects × variables), *k*-means algorithm is connected to one-way clustering, in which its objective is to classify objects. In text mining field another well-known algorithm for one-mode clustering is the Reinert's method (1983), a descendant hierarchical classification algorithm implemented within the IRAMUTEQ software that, using the words co-occurrences matrix, produces groups of similar lexical units. On the other hand, sometimes it could be necessary to identify syntheses both in the direction of objects and variables, or, in text mining framework, *texts* and *documents*. In fact, very useful is the co-clustering approach – also known as bi-clustering, subspace clustering, bi-dimensional clustering, simultaneous clustering, block clustering – that concerns simultaneous partitioning of rows and columns; the key idea is to identify submatrices of the observed data matrix, where each block specifies an object cluster and a variable cluster (Rocci and Vichi, 2008). In the literature several algorithms have been proposed, and Hartigan (1972) explains that the principal advantage of this approach is the direct interpretation of the clusters on the data.

In text mining contest, co-clustering is a very useful methodology (Balbi, 2012); the strength of this approach lies in finding clusters of documents characterized by groups of terms (Balbi, Miele and Scepi, 2010) with a high dimensionality reduction (Tjhi and Chen, 2006). Co-clustering methods allow overcoming some typical issues of textual data transformed into matrices that are large, sparse and non-negative. For its features, co-clustering is utilized in many studies in which it is involved in multiple attribute analysis; in text mining the study of co-clustering is proposed to deal with multipartition of texts and words in digital library, because this approach is very useful in the observation of the co-occurrence of terms and documents in the same corpus (Xu, Zhang and Li, 2010). Although this approach presents many advantages and it may also improve the interpretation of the clusters, there are still few proposals in this direction, while one-way partition is even now widely utilized for information retrieval.

1.3.2 *Text co-clustering*

Two-way approach for text clustering involves partitioning of rows and columns, so both words and texts are grouped; in this case, if rows and columns are treated asymmetrically, the result is a *sequential* clustering of terms and texts. Instead, when rows and columns are treated symmetrically, the result is the *simultaneous* clustering of words and documents – also known as “co-clustering” or “two-mode partition; the key idea is to identify sub-matrices of the observed data matrix (Figure 1.5), where each block specifies an object cluster and a variable cluster (Rocci and Vichi 2008).

Figure 1.5 - Data matrices (1000 × 400), before and after the rows and columns permutation, obtained with a two-mode multi-partitions model.



Source: Rocci and Vichi (2008).

In text co-clustering, two main approaches are well known:

1. *Information-theoretic co-clustering,* where the bag of words table represents an empirical joint probability distribution of two discrete random variables. In this approach, the optimal co-clustering maximizes the mutual information between the clustered random variables subject to constraints on the number of row and column clusters.
2. *Co-clustering with graph partitioning,* in which **X**=[w_{ij}] is a term-document matrix of dimensions ($k \times n$), where k is the word types, n the documents, and w_{ij} is the weight of each word in a document, that corresponding to normalized term-frequency. **X** can be represented as a bipartite graph $G = (V_1 \cup V_2, E)$, where V_1 and V_2 are the vertex sets in the two bipartite positions of the *G* graph, and I is the edge set (Dhillon, 2001; Iezzi, 2010). Each node in V_1 corresponds to one of the *k* terms, and each node in V_2 corresponds to one of the *n* documents. An undirected edge exists between $i \in V_1$ and node $j \in V_2$ if document *j* contains the term *i.*

Co-clustering is more robust to sparseness, noise and high-dimensional data, because the main aim is to exploit the “duality” between rows and columns, so at all stages the row clusters incorporate column clusters, and vice versa. Agrawal *et al.* (1999) underline that co-clustering is also related to the problem of sub-space clustering, in fact, the data is clustered by simultaneously associating it with a set of points and subspaces in multidimensional space. In this case, the data can be represented as sparse high dimensional matrices in which most of the entries are 0. Methods for subspace clustering can be adjusted to the co-clustering; for instance, Li *et al.* (2004) proposed an adaptive iterative subspace clustering for documents. Moreover, sub-space clustering can be considered a procedure of local feature selection, in which the words or repeated segments and/or documents selected are specific for each group. Principal Component Analysis (PCA) is a traditional way to select the features as linear combination of dimensions (Jolliffe, 1986). In text mining, traditional matrix approximations, such as Singular Value Decomposition or PCA, do not preserve non-negativity or sparsity, because it has the disadvantage that the components extracted by this method have exclusively dense expressions, therefore interpretation can be very difficult.

Text co-clustering allows to overcome some typical issues of textual data transformed into matrices, because they are large, sparse and non-negative. For its features, co-clustering is utilized in many studies where multiple attribute analysis is involved. Numerous algorithms and several applications are proposed in multiple domains, e.g. in bioinformatics, in marketing, in medical science, in business and economics, and in many other fields. Dhillon *et al*. (2003) present a co-clustering algorithm that monotonically increases the preserved mutual information by intertwining both the row and column clustering at all stages. Balbi *et al.* (2010) proposed to use Goodman and Kruskal index on BOW as a criterion to prediction.

In the next chapter, a new procedure for text co-clustering is proposed. The objective is to overcome the many complexities of textual data proposing a method for analysing the content of documents, that is able to process also very high dimensional, noisy and short texts.

Chapter II

NEW PROPOSAL FOR TEXT CLUSTERING: A TWO-WAY APPROACH

Recently, there has been a large and continuous proliferation of unstructured information, especially regarding online documents, e-books, journals, technical reports and digital libraries (Iezzi 2012). Indeed, people increasingly share information online, and that creates a huge volume of available textual data (King 2014) on the Web.

There are many open questions about these data – *are they representative? Which information do they contain? How to treat them? How to extract the content?* Among all, one of the most relevant issues concerns the methods that can be used to treat them, because techniques set on structured data are not able to perform satisfactory analyses on high dimensional textual data, so it is necessary to involve new methods to elaborate “words”. Textual data, especially those coming from the Web, are wide and unstructured, containing a great amount of noise. Noise in text is represented by all those words that are not relevant and significant for the analysis; for instance, in content analysis terms without informative contents – e.g. conjunctions, articles – are often eliminated from the vocabulary, in order to preserve only the terms that can be show the subjects treated in the text. Noisy text, high dimensionality and sparse matrices are the common findings in text analysis, so the new methods shall meet these problems proposing solutions able to effectively analyse texts.

For the reason presented above, in this chapter we present a new procedure we implemented (Celardo, Iezzi and Vichi, 2016) for classifying textual data on the basis of documents specificities, reducing the high dimensionality and removing most of noise. In text clustering the most used approach is *one-way partition*, because it is fast and efficient also when elaborating big matrices. On the other hand, *two-way partition* lets the classification of both rows and columns, so more information are extracted from data. Then, we propose a co-clustering approach where simultaneously both rows and columns are classified, so it is possible to individuate peculiarities in language with regard to the different documents. In addition, in this field a two-way approach is more efficient in overcoming some important issues of text clustering, such as high dimensionality and sparse matrices. To validate the procedure, we applied it on three famous corpora already used in several studies: Cranfield, Medline and Cisi collections. The results will show how the procedure gives and effective and efficient way for analysing big lexical matrices, extracting the content and together classifying the documents on the basis of the subjects treated.

Section 2.1 presents a brief review about the general issues in text clustering, motivating the need of a new method. Section 2.2 shows methodologically the new procedure of text co-clustering, starting from collecting the texts to the interpretation of the results. Section 2.3 exposes a real data analysis performed on text using the procedure of co-clustering proposed. Section 2.4 gives some final remarks on the procedure proposed.

2.1 Background: why a new procedure?

As mentioned in the previous chapter, text clustering is a method for analysing and synthetizing a collection of texts, on the basis of document similarities. The objective of the method is to extract from a collection of documents a set of meaningful information that reveals the content of these texts. Text clustering, through the classification of words/documents in groups of similar objects, allows to infer the topics of a corpus starting from the word co-occurrences. In this field, two main approaches are known: *one-way* and *two-way* clustering. The first approach regards the classification of words or documents, so it allows to produce groups of words that co-occur in more documents or groups of similar texts – i.e. that have more words in common. On the other hand, co-clustering lets the simultaneous classification of word and documents, so "blocks" of homogenous objects are identified. Each block specifies a set of words that frequently co-occur in a group of documents. In this way, it is possible to identify the topics of a corpus, and also in which texts these subjects are peculiar.

In text clustering, the main difficulties are related to high dimensionality and sparsity of the matrices. With the rising interest in Web Data, these two findings are more and more important; in fact, online texts – e.g. posts, tweets, comments, reviews, etc. – are characterized to be numerous, noisy and often "short", so each document is composed of just few terms. This means that when analysing online texts, the term-document matrices are very big and sparse – shorter the texts, less the occurrences – so clustering techniques need to be able in treating computationally these matrices, and also efficient in extracting meaningful information. The issue raised above motivates the need for new and more efficient way of analysing these data. The objective of this part of the research work is to propose a new procedure – from texts collection to words/documents partition – that may overcome these issues. The innovative parts in this approach we are proposing regard:

- The use of the specificities method as weighting scheme for the term-document matrix: it allows to reduce the sparseness of the matrix and to identify the important terms, also when the length of the documents is very short;
- The application of a co-clustering algorithm to the term-document matrix: this method, compared to one-way classification, is more robust to sparseness, noise and high-dimensional data and it consents to classify at the same time words and documents.

In the next paragraph, the procedure is exposed in detail.

2.2 The method

The procedure here presented (Celardo, Iezzi and Vichi, 2016) allows the text co-clustering through the simultaneous classification of both terms and documents of the lexical table; the method produces blocks of specific words connected to groups of documents. In this way, the content of a corpus is identified with regards to the “weight” of each subject within the corpus. In fact, each content – inferred by the list of terms – is connected to the entire corpus or to a specific cluster of documents, and then to a sub-part of the dataset. The connection between groups of words and texts is given by the centroids matrix, where for each block the average value is expressed; it follows that the interpretation of the centroids values depends of the weighting scheme of the term-document matrix.

For a better accessibility of the text, the terminology used in this paragraph is listed here:

k	number of terms to be classified
n	number of documents to be classified
$\mathbf{X} = [w_{ij}]$	$k \times n$ term-document matrix, weighted by the specificities method
$\mathrm{T} = \{t_1,...,t_k\}$	set of k terms to be classified
$\mathrm{D} = \{d_1,...,d_n\}$	set of n documents to be classified
$\mathrm{P} = \{P_1,...,P_c\}$	partition of T into c classes, where P_l is the l^{th} class of P
$\mathrm{Q} = \{Q_1,...,Q_m\}$	partition of D into m classes, where Q_p is the p^{th} class of Q
$\mathbf{U} = [u_{il}]$	$k \times c$ membership function matrix, assuming values $\{0,1\}$, specifying for each term t_i if it belongs to P_l

$\mathbf{V} = [v_{jp}]$ $n \times m$ membership function matrix, assuming values $\{0,1\}$, specifying for each document d_j if it belongs to Q_p

$\mathbf{C} = [c_{lp}]$ $c \times m$ centroid matrix specifying the mean of document d_j in the class Q_p

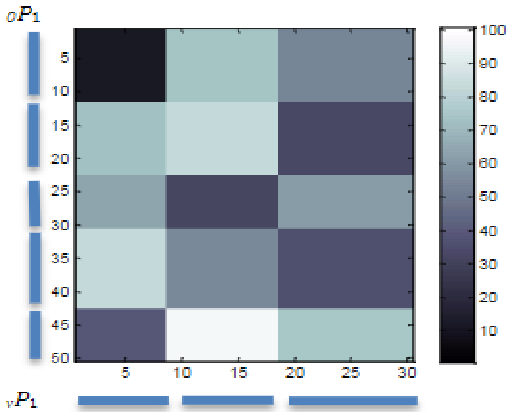
We start collecting texts about a certain phenomenon, so the first step consists of indexing documents, i.e. transforming them in structured data through a system of codes. Next, we pre-process data eliminating hapax forms and types without informative content – i.e. articles, numbers, auxiliary verbs, conjunctions, prepositions and pronouns – and reducing terms to the base form through lemmatization. Using the Space Vector Model, each text is then represented by a vector of weighted terms; we used as weighting scheme the specificities method for the reasons specified in the previous paragraph. By joining these vector, a ($k \times n$) terms-documents matrix **X** was obtained, where k is the total number of the terms or keywords and n is the total number the texts. This first phase is carried out with the IRaMuTeQ[8] software.

Therefore, to simultaneously classify rows and columns – using the MATLAB[9] software – the double *k*-means model proposed by Vichi (2001) was applied on $\mathbf{X}=[w_{ij}]$ for several partition combinations. It is a two-mode single-partition algorithm for splitting the data matrix into rectangular blocks of homogeneous values (Figure 2.1); the Euclidean distance was setting to calculate the distances matrix.

[8] *IRaMuTeQ* is a free software package using the R interface for multidimensional analysis of texts and questionnaires. It allows for statistical analysis of the corpus text and tables on people / characters. It is based on the software R and the Python language.

[9] *MATLAB* (matrix laboratory) is a multi-paradigm numerical computing environment; a proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms and creation of user interfaces.

Figure 2.1 – Two-mode single-partition of data matrix ***X*** *(50 × 30), with values 0-100, splits into (5 × 3) rectangular blocks of homogeneous values.*



Source: Vichi (in press).

The model of the double *k*-means is specified as follows:

$$\mathbf{X} = \mathbf{UCV'} + \mathrm{E}$$

where unknown partitions for rows and columns, specified by membership matrices **U** and **V**, are needed to be identified in order to best reconstruct the matrix **X**. The least-square assessment of the model leads to the formulation of the following quadratic optimization problem with respect to variables u_{il}, v_{jp} and c_{lp} :

$$\min \left\| \mathbf{X - UCV'} \right\|^2$$

subject to

$$u_{il} \in \{0, 1\} \qquad i=1,\ldots,I;\ l=1,\ldots,L;$$

$$\sum_{l=1}^{L} u_{il} = 1 \qquad i=1,\ldots,I;$$

$$v_{jp} \in \{0, 1\} \qquad j=1,\ldots,J;\ p=1,\ldots,P;$$

$$\sum_{p=1}^{P} v_{jp} = 1 \qquad j=1,\ldots,J.$$

We choose the optimal number of groups by calculating for both rows and columns the Calinski-Harabasz (CH) index (Calinski and Harabasz, 1974). The CH index is a well-known internal validation method (Iezzi, 2012b) that evaluates the cluster validity based on the best trade-off between the compactness and separation criteria:

$$\mathbf{CH} = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)},$$

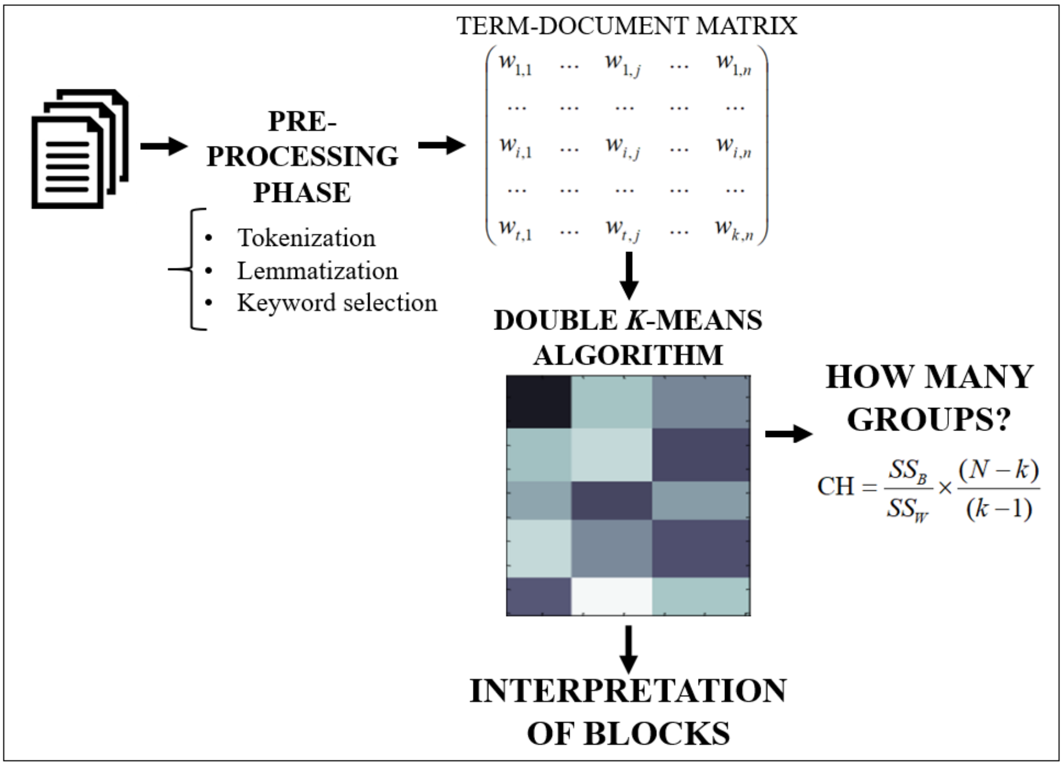
where SS_B is the overall between cluster variance, SS_W is the overall within cluster variance, k is the number of clusters and N is the number of observation. The optimal number of clusters corresponds to the highest level of the CH index.

After the optimal combination of words and documents clusters is identified, the interpretation of the result starts from the centroids matrix. In fact, the double *k*-means algorithm implemented on a term-document matrix produces:

- two membership matrices **U** and **V**, for the rows and the columns respectively;
- a centroid matrix **C** identifying the average specificity of the terms in connection with the documents.

Within the centroid matrix, the higher values correspond to language specificities connected to certain documents. In this way, it is possible to identify the contents of the corpus, and where – i.e. in which texts – they are specific or not. The overall procedure is graphically synthetized in the Figure 2.2.

Figure 2.2 – Text co-clustering procedure steps, from the collection of documents to the interpretation of groups.


TERM-DOCUMENT MATRIX
$\begin{pmatrix} w_{1,1} & \cdots & w_{1,j} & \cdots & w_{1,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{i,1} & \cdots & w_{i,j} & \cdots & w_{i,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{t,1} & \cdots & w_{t,j} & \cdots & w_{k,n} \end{pmatrix}$
PRE-PROCESSING PHASE
Tokenization
Lemmatization
Keyword selection
DOUBLE *K*-MEANS ALGORITHM
HOW MANY GROUPS?
$\mathrm{CH} = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$
INTERPRETATION OF BLOCKS

Source: Our elaboration.

In order to test the validity of this procedure, we implemented a real data analysis using three famous corpora: Cranfield, Medline and Cisi. The test is shown in the next paragraph.

2.3 Procedure validation through data analysis

For the real data analysis, we tested the procedure previously presented on three famous collections already used for *Information Retrieval* (Dhillon, 2001; Salton and Buckley, 1988; Lee, Chuang and Seamons, 1997; Fagan, 1989; Dumais, 1991; Acid, De Campos, Fernández-Luna and Huete, 2003; Sandhya, Lalitha, Sowmya, Anuradha and Govardhan, 2011; Fagan, 2017; Balbi, Miele and Scepi, 2010) and *Latent Semantic Indexing* (Kontostathis and Pottenger, 2006; Hofmann, 2017; Gao and Zhang, 2003; Efron, 2005; Ding, 2005; Kontostathis, Pottenger and Davison, 2005):

- *Cranfield* (CRAN): 1,398 abstracts in aeronautics and related areas originally used for tests at the Cranfield Institute of Technology in Bedford, England;
- *Medline* (MED): 1,033 abstracts in biomedicine received from the National Library of Medicine;
- *Cisi* (CISI): 1,454 abstracts in library science and related areas published between 1969 and 1977 and extracted from Social Science Citation Index by the Institute for Scientific Information.

These datasets can be freely downloaded from *ftp://ftp.s.ornell.edu/pub/smart*. In order to test the capability of the method in classifying the documents and also identifying the contents, we created a mixture of the three corpora, obtaining a large corpus – since now called CMS corpus – of 3,885 documents and over than 500,000 occurrences (Table 2.1).

Table 2.1 – Main lexical indexes for the CMS corpus.

Documents	***Occurrences***	***Types***	***Hapax***
3,885	562,583	21,713	8,907

With the pre-processing phase – lemmatization, stop words and hapax removal – we obtained a term-document matrix of dimension (9,288 × 3,885), with 57% reduction of the vocabulary. The matrix was then weighted using the specificities method, so we performed the double *k*-means algorithm on several combinations for rows and columns – from (2×2) to (9×9). On these combinations we calculated the Calinski-Harabasz index for identifying the optimal combination of clusters. For documents, the solution of three clusters was stable for all the combination, while for the terms we found that the CH index suggested both the solutions of four and six clusters (Table 2.2). Then, we analysed the centroid matrix of the two combinations – (3×4) and (3×6) – finding that the first one was the most interpretable (Table 2.3).

Table 2.2 – Row clusters evaluation for the CMS corpus, Calinski-Harabasz index.

No. clusters	2	3	4	5	6	7	8	9
CH index	21.25	28.55	**30.18**	26.36	**34.94**	23.36	22.64	22.83

Table 2.3 – Centroid matrix (Terms × Documents).

	Cluster 1 (38%)	***Cluster 2 (33%)***	***Cluster 3 (29%)***
Cluster 1	0,0055	0,0056	0,0096
Cluster 2	**0,1294**	0,0040	0,0049
Cluster 3	-0,0005	-0,0049	**0,1275**
Cluster 4	-0,0157	**0,2378**	-0,0110

The centroid matrix exposes the relationship between row and column clusters; the entries interpretation depends on the weighting scheme chosen to construct the term-document matrix. In this case, because we chose the specificities method to weight the matrix, a value close to zero indicates that the words of the selected cluster are not specific of the documents in the connected column cluster. Otherwise, a value far from zero shows a connection between the two clusters, in terms of specificity.

The centroid matrix calculated on the CMS corpus (Table 2.3) shows that the first cluster of words is not related to any specificity with the document clusters; it means that this group contains the general terms – i.e. those words that are present in all the documents (Table 2.4). In fact, in this cluster we found those terms that are connected to the scientific language – e.g. relation, investigate, variation, involve, support, etc. – used to describe the subjects. The second cluster of terms presents a connection with the first cluster of documents; the words characterising this group are related to information retrieval topic (library, information, term, book, etc.) and the connected column cluster contains mostly Cisi abstracts (Table 2.5). The third row cluster shows an interconnection with the third cluster of documents; the terms belonging to this group – cell, patient, child, treatment, blood, rat, etc. – regards a medical subject and the document cluster covers most of the Med abstract. The fourth cluster of words indicates an association with the second column cluster; the terms of this group concern an aerospace subject and the documents connected to these words are mainly Cran abstracts.

Table 2.4 – Word clusters, first 25 words listed by frequency of occurrence.

Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***
Experiment	Library	Case	Flow
Require	Information	Find	Result
Relation	Study	Increase	Pressure
Derive	System	Cell	Method
Reduce	Datum	Patient	Effect
Similar	Base	Type	Present
Change	Time	Normal	Numb
Influence	Analysis	Rate	Boundary
Vary	Form	Group	Layer
Difference	Term	Growth	Theory
Investigate	Paper	Level	Problem
Involve	Large	Control	Body
Variation	Discuss	Child	Show
Limit	Field	Occur	Solution
Comparison	Index	Suggest	Obtain
Property	Develop	Treatment	Heat
Pattern	Subject	Produce	Test
Frequency	Research	Activity	Wing
Exist	Include	Rat	Match
Maximum	Book	Total	High
Long	Work	Blood	Equation
Series	General	Due	Distribution
Support	Year	Decrease	Shock
Good	Development	Day	Surface
Represent	Process	Specific	Value

As shown, the co-clustering procedure has clearly identified the specific language connected to the different collection of abstract; through the centroid matrix, it has been possible to recognize not only the specific topics connected to the three datasets but also the common language shared by all the texts. The procedure has allowed to identify the different subjects of the corpus, distinguishing for each topic where it is specific – i.e. in which documents. Moreover, the procedure has correctly classified the three set of abstracts in groups – with a correct classification rate of 90%, 98% and 97%, respectively – as displayed in the confusion matrix (Table 2.5).

Table 2.5 – Confusion matrix (Corpora × Document clusters).

	Cluster 1	***Cluster 2***	***Cluster 3***
Cran	25	**1264** *(90%)*	109
Med	19	0	**1014** *(98%)*
Cisi	**1416** *(97%)*	1	37

2.4 Final remarks

Information extraction from documents could be a challenging task, even more if data are noisy and high dimensional. Today, new methods and new procedure are necessary to treat not only conventional texts – e.g. papers, news, books, etc. – but also online data, coming from the interaction between people. Then, a useful technique should allow the user analysing all the types of texts, moving from unnoisy long documents to short communications.

In this chapter, a new procedure of text co-clustering has been proposed; we constructed it thinking about versatility and scalability. Therefore, in this chapter the procedure has been tested on three famous corpora, composed of abstracts. In the next two chapters the procedure will be implemented on other two different corpora: the first one is related to a medium size collection of news, while the second one is about a set of thousand tweets related Brexit subject. The results will show that this method can be successfully implemented on different type and size of texts, obtaining an effective analysis of the content.

Part II

Real data implementations

Chapter III

THE INAIL PROJECT: MEDIA MONITORING OF WORK INJURIES

The first application of this research work regards a pilot project of media monitoring recently developed in collaboration with the National Institute for Insurance against Accidents at Work (INAIL). Its objective is to find out how the press deal with the culture of safety and health at work. To monitor mass media, the Institute has created a relational database of news concerning occupational injuries and diseases, that was filled with information obtained from the newspaper articles about work-related accidents and incidents, including the text itself of the articles. This database, called *News Repository*, was not created to replace the official statistics based on accidents reports or to provide quantitative data on job-related injuries, but it should be considered as a tool for observing how mass media deal with safety and health at work. However, this should not preclude the possibility of carrying out comparisons between data from the newspapers and other existing information systems, so the repository has been structured allowing comparisons with information from other sources.

In keeping with that, the ultimate objective is to identify the major lines for awareness-raising actions on safety and health at work. Then, 1,858 news articles regarding 580 different accidents were collected (within a predetermined period of time); for each injury, not only the news texts but also several variables were identified. The hypothesis is that, for different kind of accidents, a different language is used by journalists to narrate the events. To verify it, news have been pre-processed and several analyses were implemented on these data, using

Automatic Text Analysis techniques. The identification of various ways in reporting the events, in fact, could provide new elements to describe safety knowledge, also establishing collaborations with journalists in order to enhance the communication and raise people attention toward workers' safety. At the end, the results obtained have confirmed the starting hypothesis, and they have also provided useful tools to understand the media communication related to safety in workplaces.

3.1 The project

The study described here grew out of the collaboration between the *Department of Social Sciences and Economics* of Sapienza University of Rome and the Headquarters for Research of INAIL (Italian National Institute for Insurance against Accidents at Work) where, since 2012 a team of researchers has developed the idea of monitoring the mass media in view of prevention against accidents at work (INAIL, 2015). The idea behind this project, raised within the Institute and developed through this collaboration, is based on the hypothesis that there is an information asymmetry between real events and media news, in terms of both qualitative and quantitative aspects. In the interest of accident prevention, the Institute decided to start a mass media monitoring in order to identify and correct the critical issues present in media communication.

In the field of Health and Safety at Work, the archiving and reasoned organization of news from the press or web can make a contribution to the knowledge of the phenomenon, as in other domains[10]. For this reason, operators, societies and private associations have already undertaken experiences of news archives. These archives, often limited to specific territories or sectors, have so far been predominantly used for the purpose of quantification of the phenomenon. However, it is important to know how the topic is treated and, consequently, the level of safety culture within the descriptions. With this in mind, the Institute has realized a database of events related to health and safety at

[10] Actually, scholars have already used newspaper texts in order to analyse ideologies and opinions. Pollak et al. (2011) have used news reports about the 2007 Kenyan election and post-election crisis in order to provide ideological differences between local and international press coverage. Balahur and Steinberger (2009) explored sentiment analysis in newspaper articles, establishing guidelines for positive/negative opinions. Fortuna et al. (2009) investigated patterns in media through the observation of terms choices in different topics.

work, filled with information from the mass media. The database, called *News Repository on Occupational Safety and Health* (NeRO), does not replace the official sources and does not have the purpose of providing quantitative data on events, but it should be considered as an instrument in the sense just specified. This must not preclude the possibility of carrying out checks and comparisons with other data and other information systems already existing and, for this, the repository has been structured to facilitate comparisons between information of different sources.

3.2 The database

News Repository on Occupational Safety and Health (NeRO) is a tool created to allow analysis of contents and texts of news related to injuries at work and occupational diseases by mass media. It arose from a previous collaboration with the University of Bologna, concerning the implementation of a database about agricultural accidents, in a prevention perspective. Actually, for agriculture the phenomenon is not entirely covered by official databases, which do not deal with events occurred to hobbyists, assistants and family helpers. We then decided to extend the information gathering to all economic sectors and create a new database release containing additional technical specifications, guessing that NeRO utility doesn't lie so much in identifying accidents not recorded by official databases, as rather in detecting how (and if) news is presented. In fact, the strategic objective is to increase public awareness and safety culture, through a different approach, which will also be based on the study of news items and their composition and communication dynamics. So, the first operational purpose is understanding:

- what kind of terms are used for a news item about an accident at work or occupational diseases;
- what inspires a title;
- how the same news is dealt with by different sources;
- how it can give rise to different interpretations depending on who communicates the news itself;
- whether some aspects of the event are considered.

Our study plans to analyse the cultural characteristics of mass media communication regarding occupational safety and health, observing the attitude of the mass media (and journalists) towards the subject and the way to perceive news depending on words used. We have kept in touch with reporters dealing with occupational safety and health and, when this investigation is over, we expect to review with them the results and cooperate in finding the path to a greater awareness of the issue, whereas among news readers and viewers there are small employers, employees and workers.

3.2.1 Portrayal of the tool

News Repository on Occupational Safety and Health is an ad hoc relational database, centred on newspaper articles gathering about accidents at work, but it is potentially arranged to also gather news on near misses, occupational diseases and incidents from all kind of sources (press, television or radio). It involves several digital interconnected tables related to different entities (Table 3.1), which contain structured (i.e. based on appropriate classifications) and unstructured (i.e. textual) information.

Table 3.1 – Entities of the database NeRO, description of the information available for each entity.

Entity	***Description***
Event	Specific accident or occupational disease drawn from the news. It constitutes the central entity of the database, as well as the unit of analysis
News	Newspaper article related to one or more *Events*
Person	Person involved in a work accident or occupational disease
Injury	Type of injury
Disease	Type of disease
Legal personality	Company where an "Injury" took place, or a “Disease” was developed
Activity	Economic activity of the company
Location	Geographical location of the “Event”

From the conceptual scheme of entities to be entered, attributes and relations, a first release was produced in ANSI C++ language. It ran on all the major platforms (Windows, OS X and Linux), but it didn’t provide some features that would have been useful for data-entry and analyses. We achieved an Access release, thereby overcoming those problems. Information retrieval relates to occurrences happened in Italy and is performed online, exploiting Google Alert Service (using some suitable keywords). The reference unit is the event (right now, we are restricting events to accidents) and different aspects and information are linked to it: one or more articles about it, one or more workers injured, and so on. The data-entry interface consists of a series of thematic screens, starting from the opening one, which covers the list of already recorded events. These screens allow to enter the following data, step by step:

- [Screen “Event”]
  Text containing event report (written down taking worthwhile information from all the articles about that event), date of the event, venue, company where accident occurred (if appropriate), economic sector (according to ATECO 2007 classification);
- [Screens “News”]
  Texts of each article about the event, its source (newspaper name or press affiliation), article title, web url, date of the article to be compared to the event one;
- [Screens “Worker” and Sub-screens “Accident” and “Harms, disorders or diseases”]
  Injured worker’s biographical data, information about accident (using all the variables of ESAW classification, European Statistics on Accidents at Work), type of injury, physical implication or resulting disease (according ICD 10 classification).

3.3 Data analysis

The data collection resulted from a pilot phase of the project, and it was created reporting articles about occupational accidents occurred in five Italian regions (Apulia, Lazio, Campania, Sicily and Emilia-Romagna), in order to verify through a smaller dataset the good performance of the procedure. The news have been selected from the websites of the national and local newspapers using the *Google Alert* tool with the following keywords (both singular and plural forms): *work accident*; *work injury*; *white death*; *on-the-job fatality*.

At the end of the data collection, the repository was composed of 1,858 articles regarding 580 different work accidents. From the repository we extracted and merged the tables related to events and news, creating a single dataset composed of structured and unstructured information (Table 3.2).

On articles and events data, we involved different analyses. Initially, we analysed some event characteristics – economic activity, fatality, ongoing and road accident variables – comparing their distributions with official data, in order to examine if mass media give the same relevance to all the accidents. After that, we focused our attention on the news, analysing them through several text mining methods (Bolasco and De Mauro, 2013; Feldman and Sanger, 2007; Lebart, Salem and Berry, 1997). In order to analyse the communication features of mass media concerning our topic, we explored by way of textual analysis many aspects of the language used by the journalists to narrate the events. We started calculating some lexical indexes, in order to have an overall view on the corpus from a linguistic perspective. Then, we compared the corpus extracted from the NeRO database with other two corpora – InforMo and the Uniform Code on Occupational Safety and Health (81/2008) – to explore the differences between news and institutional language. After that, we focused our attention on the bi-grams, applying on the corpus a Sequence Analysis (also known as

Markovian Analysis), in order to investigate sequences of words. We did it using the T-LAB software (T-LAB, 2017). This method allows to calculate for all the predecessors and successors of each lexical unit the transition probabilities between pairs of words (also known as Markov chains). Next, we implemented a content analysis for discovering the subjects treated in news using the Reinert's method (Reinert, 1983); it is a descendant hierarchical classification algorithm implemented within the IRAMUTEQ software that, using the words co-occurrences matrix, produces groups of similar lexical units. So, we implemented on the term-document matrix the co-clustering procedure presented in the previous chapter; the aim was to identify if there are language specificities connected to some event characteristics. The results will show how the language used by journalists depends on some event characteristics, so to better understand the differences between news texts, we analysed the specificities related to the modalities of one specific variable, regarding the fatality of the accidents.

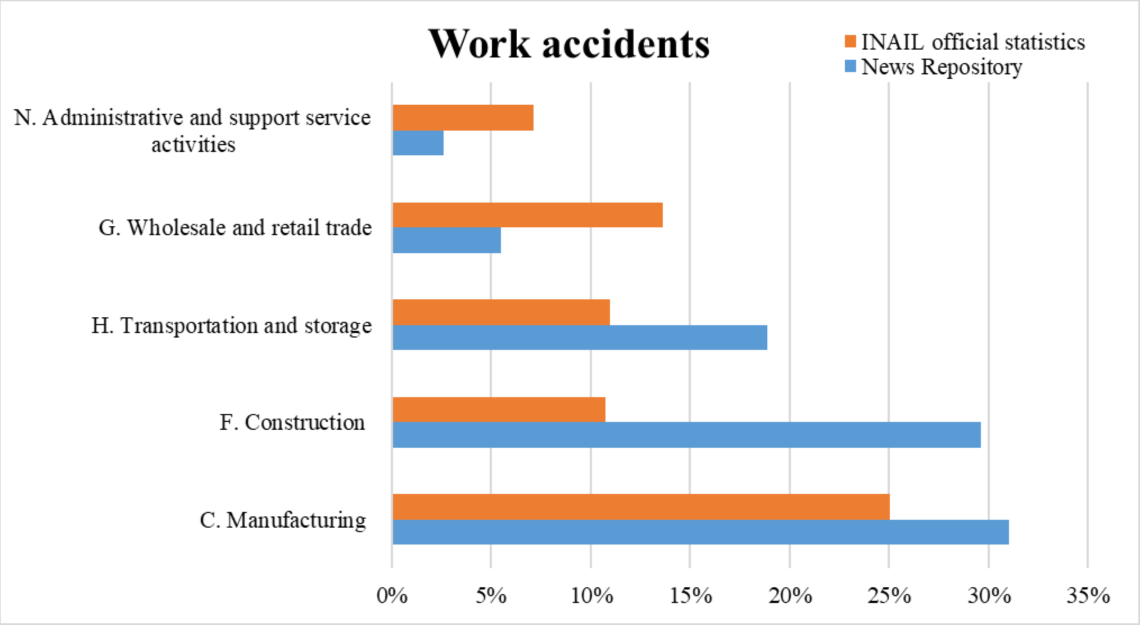
Table 3.2 – Dataset variables, extracted from the NeRO repository.

Labels	***Meaning***
Year	When the accident happened
Newspaper	In which journal the article was published
Article	Text of the news
Event	Accidents listed by sequential numeration
Area	Area of worker’s residence
Economic activity	Economic activity where the worker is employed by the NACE Rev. 2 classification
Place	Where the injury took place
Gender	Worker’s gender
Nationality	Worker’s nationality
Fatality	Fatal accident or not
Ongoing	If the accident occurred ongoing
Road	Road accident or not
Motivation	Cause of the injury

3.3.1 Results

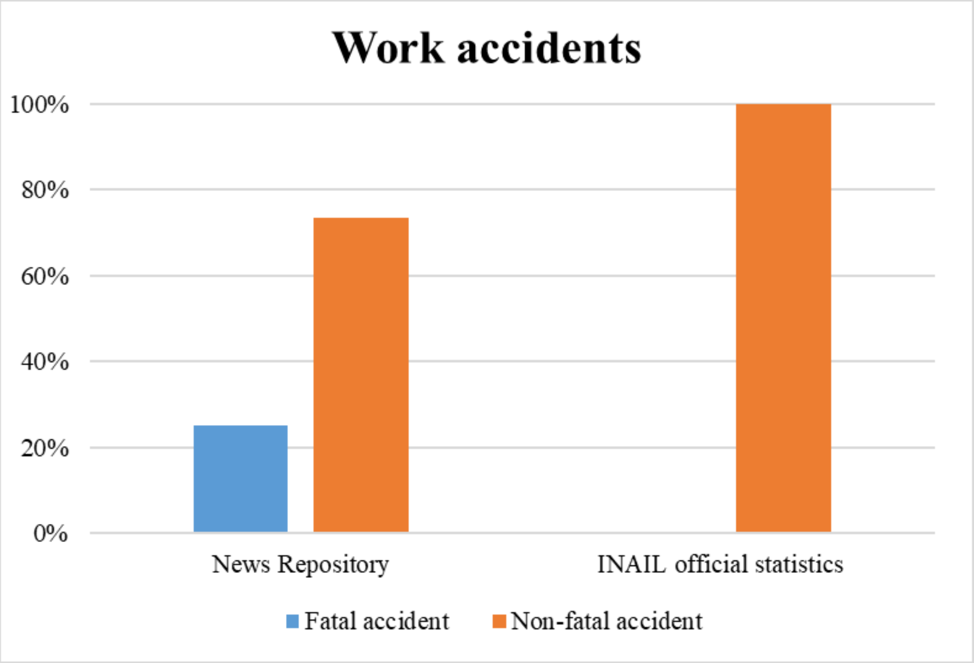
As previously mentioned, in the first part of the analyses we considered the events connected to the news. We focused on four variables, comparing the data from the repository with the INAIL official statistics of accidents reports (Figure 3.1-3.3).

Figure 3.1 – Work accidents by economic activity, first five sectors (see Appendix for further details – Table A.1).



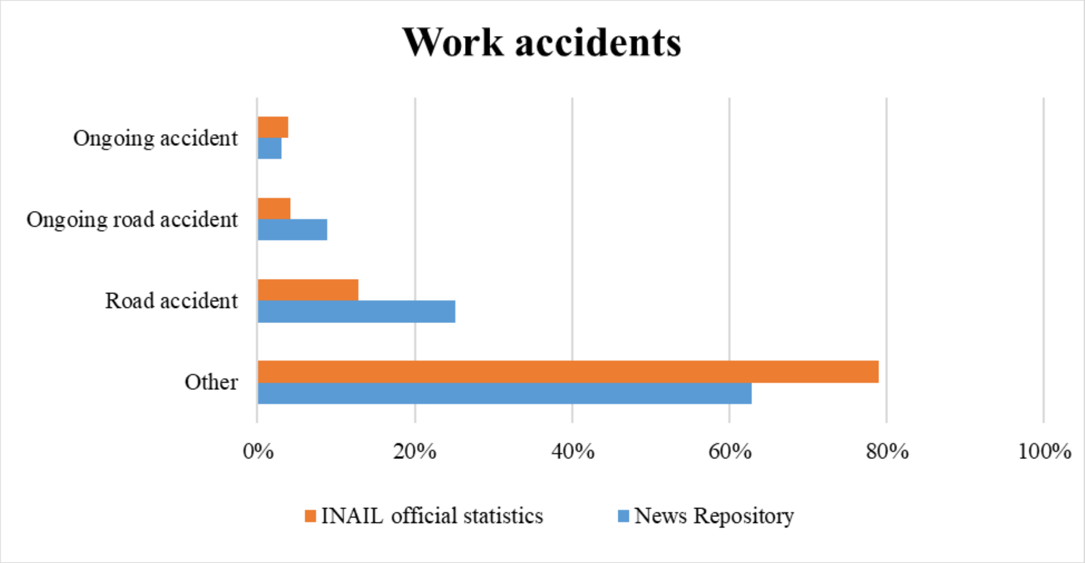
Source: Our elaboration.

Figure 3.2 – Work accidents by mortality (see Appendix for further details – Table A.2).



Source: Our elaboration.

Figure 3.3 – Work accidents by situation (see Appendix for further details – Table A.3).



Source: Our elaboration.

The Figures 3.1, 3.2 and 3.3 show that there is an asymmetry between accidents reports and the events described by mass media. Through the comparison, we observed that newspapers tend to focus on certain accidents rather than others: road accidents, on-the-job fatalities and those economic activities perceived more dangerous for workers – like construction or manufacturing – are associated with a higher level of journalists' attention.

In the second part of our analysis, we worked on the news texts; from the collection of the 1,858 articles, we obtained a corpus – since now called NeRo corpus – of medium dimension, composed of 326,938 occurrences and 14,057 types (Table 3.3). In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the Type-Token Ratio[11] (TTR: 4.3%) and the Hapax Percentage[12] (HP: 33.89%). According to the large size of the corpus, both lexical indicators highlighted its richness and indicated the possibility to proceed with the analysis.

Table 3.3 – Main lexical indexes for the NeRO corpus.

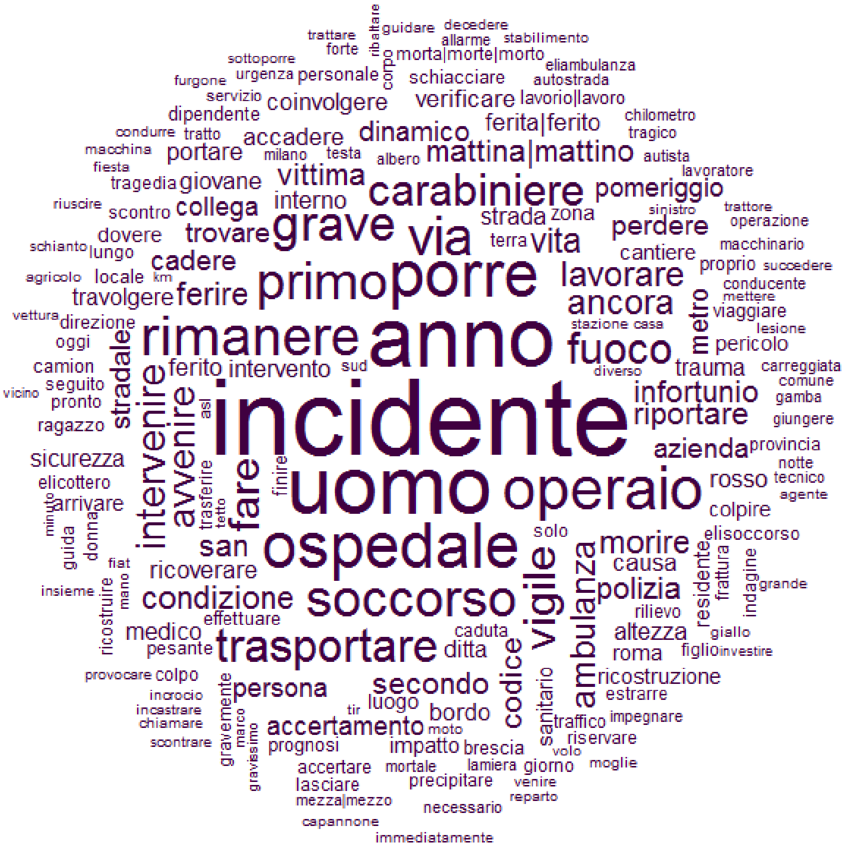
Documents	***Occurrences***	***Types***	***No. of Hapax***
1,858	326,938	14,057	4,763

[11] The Type-Token Ratio (TTR) is the ratio obtained by dividing the types of a corpus by its tokens; a higher TTR indicates more variation in the lexicon, and then a more richness language.

[12] The Hapax Percentage (HP) expresses the ratio between the number of hapax and the number of tokens in a corpus.

After lemmatization, hapax and stop-words removal (articles, numbers, auxiliary verbs, conjunctions, prepositions and pronouns), the number of types has fallen to 7,074. Then, we removed also the words not recognized by the dictionary, obtaining a set of 4,183 keywords. In the figure, it is possible to see the word cloud of the terms with high frequency, implemented with the Iramuteq software (Figure 3.4). In the Appendix, the list of terms is showed (Table A.4). This tool is a graphical instrument that allows to display the most frequent terms of a corpus; bigger is the word in the cloud, more time it occurs within the text. From the Figure, it is possible to see that the words most used in the news are those connected to the description of the injury: accident, man, worker, hospital, year, fire, serious, rescue, etc.

Figure 3.4 – Word cloud of the first 200 terms, most frequent words.



Source: Our elaboration on Iramuteq software.

To better understand the linguistic differences between news and technical reports regarding safety and health at work we compared the most frequent terms resulting from the NeRO database, InforMo and the Uniform Code on Occupational Safety and Health (Table 3.3).

Table 3.3 – Comparing different languages, most frequent words.

NeRo database	***InforMo database***	***Uniform Code***
Accident	To injure	Article
Year	Worker	Security
Man	Internal	Provision
To put	Ground	Worker
Hospital	Operation	Clause
Worker	To find	Legislative
To remain	Subway	Decree
Way	Car	Risk
Aid	To fall	Health
First	To do	Activity

From the table it is possible to observe how the news language is far from the technical and the normative ones. Even if the role of the journalists is to narrate the accidents, the comparison shows that news texts contain the emotive aspects of the injuries rather than the description of them.

The next step was to apply a Sequence Analysis to the corpus. We selected several words in order to analyse their predecessors and successors; we reported only the chains with transition probability higher than 0.1 (Table 3.5). In the results it is not important to focus on predecessors rather than successors (and vice versa), but the objective of this analysis is the association – i.e. the closeness – between pairs of words, independently from the order.

Table 3.5 – Sequence analysis, Markov chains.

Terms	***Predecessor***	***Successor***	***Probability***
Fatality	Tragic		0.40
Accidentally	Equipment		0.20
Accidentally	To slip		0.20
Distraction	Instant		0.13
Tragic		Event	0.15
Tragic		Destiny	0.30
Terrible	Misfortune		0.17

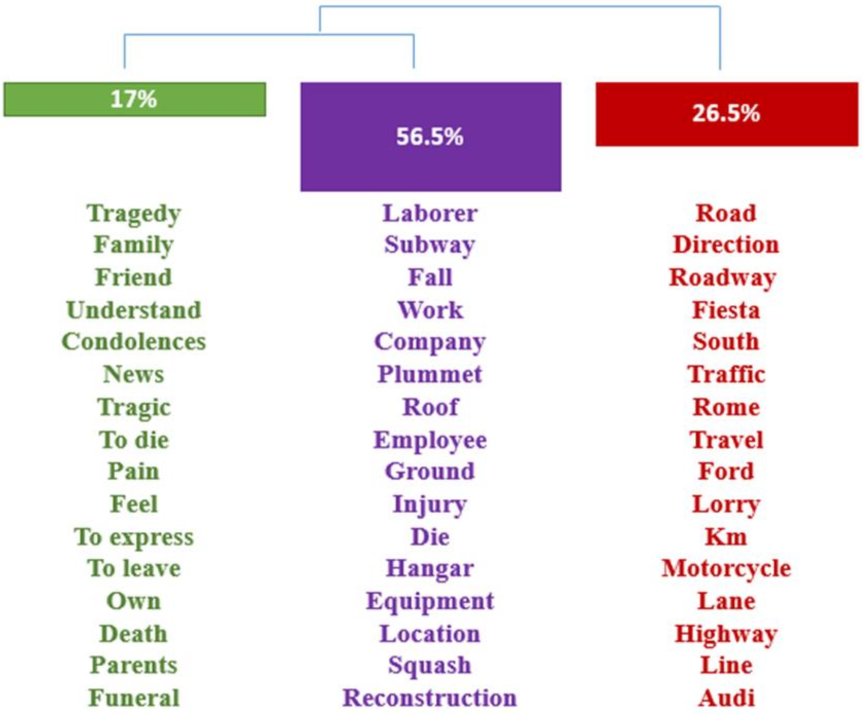
With this technique we investigated connections between some selected words and the others. Specifically, we concentrated on terms like fatality, tragic, accidental, to find if accidents have been communicated by media like something that is predictable, or not. As shown in the table, there are many significant chains; in fact, the association of specific words in the texts – e.g. tragic fatality, tragic destiny, an instant of distraction – transmits the idea that work injuries are random, fortuitous, and most important, something we cannot prevent.

To identify the main subjects of the corpus we performed a content analysis using the Reinert’s method for descending hierarchical classification. We implemented this method using as unit analysis both the segments and the texts. The implementation on the segments produced 4 overlapped clusters, while in the other case we identified three well-separated groups. Then, using as unit analysis the texts the algorithm detected three clusters of words (in order, the red, the blue and the green ones), identifying distinct contents within the corpus (Figure 3.4):

1. EVENT (56.5%). In this group, the largest one, we found all those terms that are connected to the accidents description;

2. ROAD (26.5%). This cluster identifies the focus given in the media news to road accidents by journalists. This group is very interesting, because it confirms the relevance given to injuries that took place on the road;
3. RITE (17%). In this group are contained those terms related to the ritual aspects connected to events like accidents, where the emotive part is predominant.

Figure 3.4 – Result of cluster analysis, Reinert's method.



Source: Our elaboration on Iramuteq software.

As previously mentioned, with the pre-processing phase we obtained a term-document matrix of dimension (4,183 × 1,858). The matrix was then weighted using the specificities method, so we performed the double *k*-means algorithm on several combinations for rows and columns – from (2×2) to (9×9). On these combinations we calculated the Calinski-Harabasz index for identifying the optimal combination of clusters. For both documents and terms, the CH index suggested the solutions of two and four clusters (Table 3.6). Then, we analysed the centroid matrix of the two combinations – (2×2) and (4×4) – finding that the second one was the most interpretable (Table 3.7).

Table 3.6 – Row clusters evaluation for the NeRO corpus, Calinski-Harabasz index.

N. clusters	2	3	4	5	6	7	8	9
CH index	**32.11**	23.42	**26.83**	21.49	20.97	20.26	13.58	11.10

Table 3.7 – Centroid matrix (Terms × Documents).

	Cluster 1 ***(43%)***	***Cluster 2*** ***(36%)***	***Cluster 3*** ***(17%)***	***Cluster 4*** ***(4%)***
Cluster 1	0.0163	0.0062	0.0058	0.0073
Cluster 2	0.0235	0.0002	**0.2406**	0.0821
Cluster 3	0.0414	**0.1533**	0.0196	-0.0469
Cluster 4	0.0166	-0.0034	0.0116	**0.6378**

The centroid matrix calculated on the NeRO corpus (Table 3.7) shows that the first cluster of words is not related to any specificity with the document clusters; it means that this group contains the general terms – i.e. those words that are present in all the documents (Table 3.8). In fact, in this cluster we found those terms that are connected to the description of the events – e.g. place, tragedy, worker, factory, to happen, installation, rescuer. The second cluster of terms presents a relationship with the third cluster of documents; the words characterising this group are related to the description of road events (road, kilometre, onboard, motorway, crossroad, etc.) and the connected column cluster contains mostly news of non-fatal road accidents (Table 3.8). The third row cluster shows an interconnection with the second cluster of documents; the terms belonging to this group (hospital, assistance, ambulance, doctor, trauma, to intervene, etc.) regards the narration of the rescue and the first aid, and the document cluster covers most of the news about non-fatal accidents. The fourth cluster of words indicates an association with the fourth column cluster; the terms of this group concern mostly the death of the worker (to leave, mortal, to kill, homicide, etc.) and the documents connected to these words are mainly fatal road accidents. It is interesting to note that in this last group of words are reported those terms connected to the worker’s family.

Table 3.8 – Word clusters, first 25 words listed by frequency of occurrence.

Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***
Resident	Accident	Man	Victim
Place	Firefighter	Year	To crush
Death	To occur	Hospital	Son/Daughter
Tragedy	Police	Labourer	Mortal
Province	Again	Assistance	Carriageway
To arrive	Road	To transport	Wife
Body	Kilometre	To intervene	To run over
Tragic	Wound	Ambulance	To kill
Operation	Wounded	Life	Unintentional
Strong	Young	Condition	Homicide
Worker	To involve	To injure	Biker
Service	On board	Injury	To arrest
House	Impact	Code	To track down
To cause	Motorway	Company	Automobile
Factory	Crossroad	Colleague	Omission
Serious	Direction	Verification	Minibus
Warehouse	Crash	To admit to hospital	Pirate
Minute	Lorry	Intervention	To stop up
To happen	Car	Cause	Escape
Civil	To extract	Trauma	Concert
Passing	Traffic	Safety	Scooter
Installation	Fiat	Doctor	Musical
Military	Driver	Construction site	Band
Family	Metal sheet	To reconstruct	Anterior
Rescuer	Chauffeur	To hit	To jump

From the cluster analysis we have seen that there are some interesting characterizations of the language used in newspapers. Some variables, like the fatality of the injury and the accident site, present a strong lexical differentiation among the modalities; this means that who is narrating the event – i.e. the journalist – uses a specific language to describe the accident, on the basis of these characteristics.

Finally, to better understand the lexical differences connected to the fatality of the event, we analysed the specificities (Bolasco and De Mauro, 2013; Lafon, 1980; Lebart, Salem and Berry, 1997) for this particular variable (Table 3.9).

Table 3.9 – Analysis of the specificities for the variable "accident mortality".

Fatal accident - No	***z = test-value***	***Fatal accident - Yes***	***z = test-value***
Hospital	59.17	Tragedy	35.68
Serious	58.84	Family	27.17
To transfer	54.90	Useless	23.62
Dangerous	28.38	To leave	19.84
Rescue	24.13	Victim	18.68
Ambulance	24.09	Tragic	17.71
Leg	23.12	Friend	14.95
Injury	22.06	Band	14.89
Trauma	20.55	Condolence	12.65
Hand	18.84	Province	12.15
Fracture	16.70	Son	11.49
Helicopter	13.70	Wife	11.48
Bus	12.23	Escape	10.63
Crossroad	10.20	Mayor	9.11

Starting from the results showed in table, we can observe and confirm that there is a significant difference in the language utilized when the accident is fatal or not. The terms used in the case of a non-fatal event are related to the description of the injury, while in the case of a mortal accident the situation is completely different: the words utilized refer to the emotional sphere of the event, so concepts like the family or the unpredictability are very often used to describe what was happened.

3.4 *Final remarks*

The project here presented showed how News Repository on occupational safety and health can contribute to analyse occupational safety and health, although in some institutions there are already databases dedicated to newspaper articles dealing with this subject. Actually, in addition to news texts, NeRO database provides several systematized information, enabling to filter news according to various search criteria and, above all, to carry out a number of studies and organized analysis on textual data, too.

The analyses conducted in this first phase gave us many causes for reflection. What we found shows that if official statistics provide the quantification of the phenomenon, the mass media monitoring offers a different point of view connected to the construction of prevention culture in the occupational environment. The analyses implemented in this chapter showed that there are several asymmetries between accidents reports and media news, in terms of both quantitative and qualitative aspects. This knowledge acquisition has relevant implications, firstly because potential asymmetries could be reduced in order to improve prevention and safety in workplaces, maybe shrinking job-related accidents. If mass media have the power to create public opinion on specific issues, the importance of monitoring the information and encouraging truthful communication is very clear. This goal could be achieved by means of a strict collaboration with institutions and journalists, through the improvement of knowledge, capabilities and awareness. The National Institute for Insurance against Accidents at Work is committed to improving safety knowledge and accidents prevention through the monitoring of the information.

Until now, a preliminary phase was carried on in order to identify the potentiality of the project; in the future, the monitoring will be systematic, and more advanced techniques will be implemented in order to have a continuous and constant update on mass media production.

Chapter IV

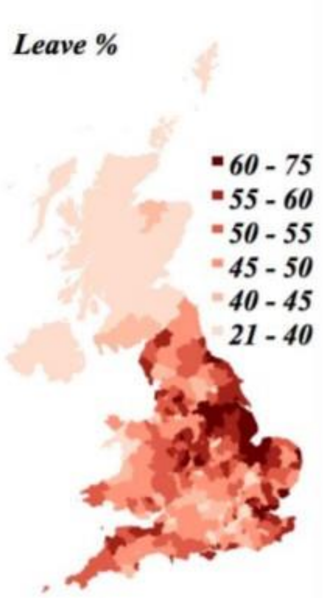
BREXIT AND TWITTER: THE VOICE OF PEOPLE

The second application of this research work regards the investigation of online discourses connected to the exit of the United Kingdom from the European Union (EU). The motivation in analysing this theme lies in the fact that there is an increase in Euroscepticism among EU citizens nowadays, as shown by the development of the ultra-nationalist parties among the European states. Regarding the European Union membership, public opinion is divided in two. British referendum in 2016, where citizens chose to "exit" shaking the public opinion, and the following general election in June 2017, where the British Europeanist parties won the election, are clear examples of this fracture. There are still few studies concerning the investigation of Brexit discourses within the social media and most of them focus on the 2016 British referendum. Due to that, this exploratory research aims to identify how Brexit and the EU are discussed on Twitter time after the vote, through a text mining approach. We collected all the tweets written in English language containing both the terms "Brexit" and "EU", for a period of 10 days. Data collection has been performed with R software, resulting in a large corpus to which we applied multivariate techniques in order to identify the contents behind the shared comments. Specifically, we implemented on the data a co-clustering procedure in order to identify linguistic specificities connected to the online discourses.

4.1 The Brexit vote

The 2016 British referendum (also known as "Brexit"), where 52% of citizens chose to “Leave” (Figure 4.1), took place on Thursday 23rd June 2016 and it was probably the most important political event in recent British history and a central theme in the political agenda because of the possible implications and consequences of citizens' choices. On the question: "*Should the United Kingdom remain a member of the European Union or leave the European Union?*", more than 17 million of people belonging to the electorate voted to leave the EU (Alaimo, 2018; Alaimo and Solivetti, in press).

Figure 4.1 – British referendum, percentage of leave by local government district.



Source: (Alaimo, 2018).

For further evidence of this division, the following general election of June 2017 saw the affirmation of the main Europeanist parties (especially the Labour Party) and the results led to a *hung Parliament*. Brexit has shaken the European public opinion as it revealed the relevance of the anti-Europeanist trend. During the 60th Anniversary of the Treaties of Rome in 2017, millions of citizens expressed their support to the EU participating in Europeanist demonstrations in many European cities.

One useful starting point for explaining the results of Brexit is to focus on the electoral issue: the relationship between the UK and Europe. This has always been a central and rather controversial issue in the British public debate. The media, public opinion and the political class have always been deeply critical and sceptical about the European integration. This position influences citizens' attitudes towards the Union, which is not only considered distant and inadequate to resolve everyday issues (immigration, unemployment, and so on), but it is often perceived as their major cause, by limiting the political and economic power of United Kingdom. The electoral outcome created disbelief all over the world. *Britain is the home of the term Euroscepticism* (Spiering 2004, p.127). But, while it is clear that a large proportion of UK residents are sceptical about Europe, it is not clear enough that this position coincides with the wish to leave the EU. However, Euroscepticism should not be confused with this wish. Szczerbiak and Taggart (2008) have distinguished two different types of Euroscepticism: the Hard Euroscepticism that is a principled opposition to the EU and European integration and Soft Euroscepticism that concerns on one (or a number) of policy areas lead to the expression of qualified opposition to the EU.

4.2 Data analysis

Although there are several studies exploring British Euroscepticism, only few of them investigate the Brexit discourses within the social media. Moreover, to our knowledge there are no study analysing opinions on Brexit after more than one year from the vote. Due to that, we decided to perform a quantitative study, where the online discourses regarding both Brexit and European Union subjects are analysed using a co-clustering approach. The aim is to explore the contents shared by users on Twitter on these themes. For this paper, we used as data source one of the most important and known blog tools, Twitter. It is an online platform for sharing real-time, character limited communication with people partaking of similar interests that, in 2017, counted over than 300 million users and an average of about 500 million of tweets sent per day.

The corpus was collected in order to explore the contents on Brexit and EU in twitter communications; so, for ten days – from September 22nd to October 2nd, 2017 – we scraped all the messages in English language containing together the words Brexit and EU. The data extraction was carried out with the TwitteR package of R Statistics (Gentry, 2016). After that, the corpus was pre-processed through lemmatization, hapax and stop words removal. From the corpus we removed also the keywords "Brexit" and "EU", used for the data search on twitter. The resulting term-document matrix was then weighted by the specificities method, using the Iramuteq software. On this matrix, the co-clustering approach presented in the second chapter was applied, in order to find content specificities in the text.

4.2.1 Results

From Twitter we extracted 221,069 messages, from which we excluded all the retweets, resulting a large corpus – since now called Brexit corpus – of 37,318 tweets and 618,255 tokens (Table 4.1).

Table 4.1 – Main lexical indexes for the Brexit corpus.

Documents	***Occurrences***	***Types***	***No. of Hapax***
37,318	618,255	27,409	14,338

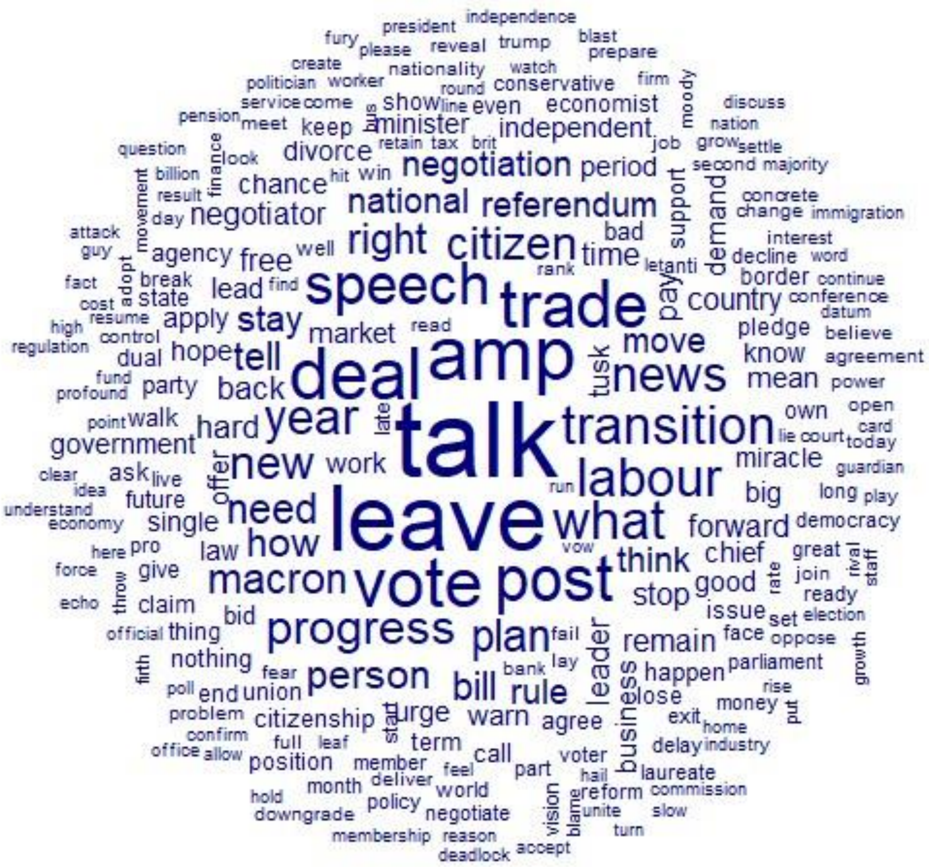
In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the Type-Token Ratio and the Hapax Percentage (Table 4.2). According to the large size of the corpus, both lexical indicators highlighted its richness and indicated the possibility to proceed with the analysis.

Table 4.2 – Lexical indicators, calculated on the Brexit corpus.

Type-Token Ratio	***Hapax Percentage***
0.04	52.31%

On the corpus we implemented a strong pre-processing, due to the raw nature of the data. In fact, data from social media are characterized by being really noisy, therefore they require a more in-depth phase of treatment. In this phase we lemmatized the corpus, we removed stop-words, hapax and terms used to select the corpus – i.e. “Brexit” and “EU” words. In the stop list we included also those terms not recognized by the dictionary, in order to do a stronger keywords selection. So, the pre-processing phase allowed us to identify a set of 1,957 keywords, representing 97% of the tweets. In the Figure 4.2 we represented graphically the first 200 terms of the corpus, by using the word cloud tool of Iramuteq software, where it is possible to see the most used words in the Brexit debate – e.g. talk, leave, deal, vote, post, transition, trade, speech, labour, etc. In the Appendix, the list of terms is showed (Table A.5).

Figure 4.2 – Word cloud of the first 200 terms, most frequent words.



Source: Our elaboration on Iramuteq software.

So, on the term-document matrix of dimension (1,957 × 36,383) weighted by the specificities method, we calculated the Calinski-Harabasz Index in order to define the number of clusters for rows and columns. After calculating the index values for several partition combinations (from two to ten for each dimension), the Index suggested to classify the words in three groups, while for the columns several partitions were acceptable. Then, we analysed the different combinations – (3 × 2), (3 × 5), (3 × 7) – finding that the second one was the most interpretable (Table 4.3).

Table 4.3 – Column clusters evaluation for the Brexit corpus, Calinski-Harabasz index.

N. clusters	2	3	4	5	6	7	8	9
CH index	**91.06**	63.29	65.01	**65.91**	54.12	**62.44**	59.40	53.88

Found the best combination, we implemented the double *k*-means algorithm (3 × 5) on the term-document matrix, obtaining the centroid matrix exposed in the Table 4.4.

Table 4.4 – Centroid matrix (Terms × Documents).

	Cluster 1 (55%)	***Cluster 2 (20%)***	***Cluster 3 (12%)***	***Cluster 4 (11%)***	***Cluster 5 (2%)***
Cluster 1	0,005	0,003	0,004	0,000	0,000
Cluster 2	0,002	**0,063**	0,003	**0,149**	0,012
Cluster 3	-0,002	0,000	**0,090**	-0,003	**0,309**

As shown in the table, the algorithm has identified four blocks of specificities. In fact, the centroid matrix shows that:

- The first row cluster does not contain specific terms. This means that this group displays the common language, used in all the messages;
- The second group of terms is connected to the column clusters no. 2 and 4, so in that group we found words that are specific for the tweets belonging to the second and the fourth clusters;

- The third and last cluster of words is connected to the groups no. 3 and 5, so it is specific of the third and the fifth groups of tweets.

In the Table 4.5, the groups of words are presented.

Table 4.5 – Words clusters, first 25 words listed by frequency of occurrence.

Cluster 1	***Cluster 2***	***Cluster 3***
Negotiation	***Economic Transformation***	***British Identity***
Stay	Leave	Home
Junker	Move	Sound
Ambassador	Transition	Cake
Cry	Late	Plan
Track	Deal	Datum
Surge	Trade	Live
Peer	Retain	Finish
Shape	Post	Id
Turmoil	Macron	Idea
Survive	Urge	National
Interview	Month	Citizen
Ups	Reform	Country
Minority	Chief	Citizenship
Politician	Resume	Bank
Minister	Lay	Card
Cheer	Profound	Eat
Original	Divorce	Ditch
Rejoin	Success	Alarm
Military	Frustration	Confirm
Operate	Tusk	Right
Half	Dynamic	Tier
Stray	New	Law
Beware	Chance	Office
Silly	Tell	Need
Stumble	Period	Share

The results of the two analyses showed a strong relationship between the terms “Brexit” and “EU”. The first group of words (*Negotiation*) is related to the need of defining new rules and settlements within the negotiation and it has no specificities related to the different documents, so it contains general terms, used more or less equally in all the texts. On the other hand, for the other two groups of words, there are more effective specificities; the second cluster of words (*Economic transformation*) is about the definition of new economic agreements, and it is connected to the 31% of the tweets, while the third one (*British identity*), related to the requirement in specifying a new identity after Brexit, is representative of the 14% of the corpus documents. Then, starting from the analysis of the contents we found that the Twitter communications on Brexit focuses primarily on the concept of negotiation. The remaining part of the messages take into account both the Brexit economic features and the need of the national identity redefinition.

4.3 Final remarks

As shown in the previous paragraph, the co-clustering procedure applied on Twitter data regarding the Brexit vote has identified the language specificities connected to the different messages. Even if social media data, like tweets, are raw, really noisy and short – in terms of single document length, the procedure proposed in this research work and then applied to Web Data was able to process and analyse data, extracting information from them.

The two-way classification of the data collected with the aim of investigating discourses about Brexit has shown that the dominant theme is about the terms of negotiation for the exit of the nation from the European Union. However, from this analysis two further important aspects emerged: first, the economic transformation deriving from the vote outcome, that the United Kingdom has to manage; second, the issue of the "identity" for Britain citizens, now that the nation is not anymore part of the Europe. So, two years from the vote, British people seem to focus their attention on three issues: the new asset, the economic consequences, and the national identity.

Conclusions

Social Sciences are undergoing a complex transformation, which has been called "Big Data revolution" (Ceron, Curini and Iacus, 2016b); even if they are categorised as "Big", they are firstly "Data", therefore good statistical techniques are required in order to extract meaningful results from these sources.

Today Big Data, especially those ones coming from the Web, play a crucial role in information production; they create numerous occasions to improve official statistics and they also influence the system where measurements are produced. Then, the emergence of new data sources and the availability of these data require new tools and methodologies. The answer to these challenges lies in the integration of sources, methods, and skills (Alleva, 2017). Big Data, such as structured data, have strengths and weaknesses. Limitations of Web Data make it difficult for them to replace traditional survey data collection; however, these new data are increasingly more accessible, and they can be easily managed. Moreover, Web Data contain different and additional information, such as opinions or sentiments, that can be quickly collected and analysed. So, the point is not if Big Data will replace the survey, but in which way we could integrate the two approaches, in order to expand the range of research methods. Statistical organizations are more and more committed to the integration of data sources; the requirement is to expand the knowledge and increase the quality of the information provided. In fact, if the survey remains a crucial tool for the quantification of the phenomenon, new data are now available and they could be used in order to improve the understanding of the society.

We started this research work presenting the potentialities versus the soft points – well-known in the scientific literature – in the use of textual data, underline the need in looking for new and advanced solutions for treating texts.

As widely explained, Web Data are characterized by being freely accessible, but also hardly treatable from a statistical point of view because of their unstructured nature. In this framework, textual data represent one of the most interesting source of information, so many methods have been proposed to analyse them. Starting from the weaknesses of textual data, in the second chapter of this research work, we have described our methodological proposal in terms of a new procedure of co-clustering for treating textual data; our objective was to implement a technique for content analysis able to deal with the issues of texts we have previously identified. To test the method, real applications of the procedure– on different data – have been shown. To verify the robustness and the scalability of the technique, we chose to apply the procedure to two different corpora: one was composed of news and another was a combination of posts from Twitter. The dissimilarity between datasets mainly lied in the dimension of the corpora and in the length of the single documents; so, the aim was to demonstrate that the procedure is able to work with every type of text.

The corpus of the first application focused on a mass media monitoring, with the aim of analysing the general opinion concerning safety and health at work. The data collection was carried on within a project recently launched by the National Institute for Insurance against Accidents at Work; the idea was that storage and organization of news might improve the knowledge of the phenomenon, in terms of how mass media deal with safety and health at work. In this perspective, the Institute has created a relational database of news concerning work-related injuries and diseases, filled with information obtained from the newspapers about occupational accidents and incidents. Until now a preliminary phase was carried on – where also the co-clustering approach has been used – and first results have been presented in the third chapter. The second application presented in the last chapter of this research work regarded the analysis of the tweets concerning the Brexit vote; the objective was to understand the content of the comments on this topic within a social media, namely Twitter. Again, the co-clustering procedure has been tested on these data, and the results presented.

The results obtained in the two applications have shown the capability of the procedure in treating different kind of corpora; in fact, both the analyses have produced meaningful outcomes, allowing us to extract the contents of the two collections of text.

References

Acid S., De Campos L.M., Fernández-Luna J.M. and Huete J.F. (2003). An information retrieval model based on simple Bayesian networks. *International Journal of Intelligent Systems*, 18(2), 251-265.

Agrawal R., Gehrke J., Raghavan P. and Gunopulos D. (1999). Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Conference*.

Agarwal S., Godbole S., Punjani D. and Roy S. (2007). How much noise is too much: A study in automatic text classification. *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference,* 3-12.

Alaimo L.S. (2018). Demographic and socio-economic factors influencing the Brexit vote. *Rivista Italiana di Economia Demografia e Statistica*, 72(1).

Alaimo L.S. and Solivetti L.M. (in press). Territorial determinants of the Brexit vote. *Social Indicators Research*.

Alleva G. (2017). Emerging challenges in official statistics: New sources, methods and skills. In (edited by) Petrucci A. and Verde R. *SIS 2017. Statistics and Data Science: new challenges, new generations*, 43-44. Firenze University Press.

Balahur A. and Steinberger R. (2009). Rethinking sentiment analysis in the news: From theory to practice and back. *Proceeding of WOMSA*.

Balbi S. (2012). Beyond the curse of multidimensionality: high dimensional clustering in text mining. In (edited by) Bolasco S. and Iezzi D.F. *Advances in Textual Data Analysis and Text Mining - Special Issue Italian Journal of Applied Statistics,* 22(1).

Balbi S., Miele R. and Scepi G. (2010). Clustering of documents from a two-way viewpoint. In *10th International Conference on Statistical Analysis of Textual Data*.

Benzécri J. P. (1963). *Cours de linguistique mathématique*. Université de Rennes.

Benzécri J. P. (1973). *L'analyse des données*. Dunod.

Bolasco S. (2004). L’analisi statistica dei dati testuali: intrecci problematici e prospettive. In *Applicazioni di analisi statistica dei dati testuali*, 9-26. Casa Editrice Università La Sapienza.

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. *Quaderni di Statistica*, (7).

Bolasco S. (2012). Introduction to the automatic analysis of textual data via a case study. In (edited by) Bolasco S. and Iezzi D.F. *Advances in Textual Data Analysis and Text Mining - Special Issue Italian Journal of Applied Statistics*, 22(1), 5-19.

Bolasco S. and De Mauro T. (2013). *L'analisi automatica dei testi: fare ricerca con il text mining*. Carocci Editore.

Bortolini U. and Zampolli A. (1971). Lessico di frequenza della lingua italiana contemporanea: prospettive metodologiche. In *Atti del Convegno Internazionale di Studi “L'insegnamento dell'italiano in Italia e all'estero”*, (2), 639-648.

Caliński T. and Harabasz J. (1974). A dendrite method for cluster analysis. In *Communications in Statistics - Theory and Methods*, 3(1), 1-27.

Celardo L., Iezzi D.F. and Vichi M. (2016). Multi-mode partitioning for text clustering to reduce dimensionality and noises. In (edited by) Mayaffe D., Poudat C., Vanni L., Magri V., Follette P. *JADT 2016: Statistical Analysis of Textual Data*, 181-192. Les Press de Fac Imprimeur.

Celardo L. (2017). Classifying textual data: a two-way approach. *Working papers series - PhD course in Applied Social Sciences*, 6/2017.

Celardo L. (2018). Opportunities of using big data in social sciences: Work injuries through media analysis. *Working papers series - PhD Course in Applied Social Sciences*, 9/2018.

Celardo L., Vallerotonda R., De Santis D., Scarici C. and Leva A. (2018). Analysing occupational safety culture through mass media monitoring. In (edited by) Iezzi D.F., Celardo L. and Misuraca M. *JADT 2018: Statistical Analysis of Textual Data*, 150-156. UniversItalia.

Ceron A., Curini L. and Iacus S.M. (2016a). First-and second-level agenda setting in the twittersphere: An application to the italian political debate. *Journal of Information Technology & Politics*, 13(2), 159-174.

Ceron A., Curini L. and Iacus S.M. (2016b). *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Taylor & Francis.

CISCO. (2017). *The Zettabyte Era: Trends and Analysis*. Retrieved on June 10, 2018 (https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf).

Conrad P. and Reinarz S. (1984): Qualitative computing: Approaches and issues. *Qualitative Sociology*.

Couper M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods,* 7(3), 145-156.

De Mauro T. (1980). *Guida all'uso delle parole*. Ed. Riuniti.

De Mauro T. (1993). *Lessico di frequenza dell'italiano parlato*. Etas.

Dhillon I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *ACM KDD Conference*.

Dhillon I., Mallela S. and Modha, D.S. (2003). Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining,* 89-98.

Ding C.H. (2005). A probabilistic model for latent semantic indexing. *Journal of the American Society for Information Science and Technology*, 56(6), 597-608.

Drouin P., Francœur A., Humbley J. and Picton A. (2017). *Multiple perspectives on terminological variation* (Vol. 18). John Benjamins Publishing Company.

Dumais S.T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2), 229-236.

Efron M. (2005). Eigenvalue-based model selection during latent semantic indexing. *Journal of the American Society for Information Science and Technology*, 56(9), 969-988.

Fagan J.L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2), 115-132.

Fagan J.L. (2017). Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *ACM SIGIR Forum*, 51(2), 51-61.

Feldman R. and Sanger J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fortuna B., Galleguillos C. and Cristianini N. (2009). Detection of bias in media outlets with statistical learning methods. In *Text Mining*, 57-80. Chapman and Hall/CRC.

Gandomi A. and Haider M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137-144.

Gao J. and Zhang, J. (2003). Sparsification strategies in latent semantic indexing. In *Proceedings of the 2003 Text Mining Workshop*, 93-103.

Giuliano L. and La Rocca G. (2008). *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*. LED Edizioni Universitarie.

Greco F., Alaimo S.L. and Celardo L. (2018). Brexit and Twitter: The voice of people. In (edited by) Iezzi D.F., Celardo L. and Misuraca M. *JADT 2018: Statistical Analysis of Textual Data*, 327-334. UniversItalia.

Guiraud P. (1954). *Les caractères statistiques du vocabulaire*. Presses universitaires de France.

Hartigan J.A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337), 123-129.

Hearst M.A. (1997). Text data mining: Issues, techniques, and the relationship to information access. In *Presentation notes for UW/MS workshop on data mining*, pp. 112-117.

Herdan G. (1956). *Language as chance and choice*. Noordhoff.

Herdan G. (1964). *Quantitative linguistics*. Butterworth & Co. Publishers.

Hofmann T. (2017). Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, 51(2), 211-218.

Iacus S.M. (2014). Big data or big fail? The good, the bad and the ugly and the missing role of statistics. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, 5(1), 4-11.

Iezzi D.F. (2009). *Statistica per le Scienze Sociali. Dalla progettazione dell'indagine all'analisi dei dati*. Carocci Editore.

Iezzi D.F. (2010). Topic connections and clustering in text mining: an analysis of the JADT network. In *Statistical Analysis of Textual Data*, 2(29), 719-730.

Iezzi D.F. (2012a). Centrality measures for text clustering. *Communications in Statistics - Theory and Methods*, 41(16-17), 3179-3197.

Iezzi D.F. (2012b). A new method for adapting the *k*-means algorithm to text mining. *Italian Journal of Applied Statistics*, 22(1), 69-80.

Iezzi D.F., Mastrangelo M. and Sarlo S. (2012). Text clustering based on centrality measures: an application on job advertisements. *11es Journées Internationales d'analyse statistique des données textuelles*, pp. 515-524.

Iezzi D.F. and Mastrangelo M. (2014). The IEMA fuzzy *c*-means algorithm for text clustering. *12es Journées Internationales d'analyse statistique des données textuelles,* 239-248.

Iezzi D.F. and Zarelli F. (2015). What tourists say about the Italian national parks: a web mining analysis. *Rivista Italiana di Economia, Demografia e Statistica*, 69, 73-82.

INAIL. (2015). *Il monitoraggio dei mass media in materia di salute e sicurezza: Strumenti per la raccolta e l'analisi delle informazioni*. Retrieved on June 2, 2018 (https://www.inail.it/cs/internet/docs/allegato_monitoraggio_mass_media.pdf).

ITU. (2017). *ICT Facts & Figures: The World in 2017*. Paris: ITU. Retrieved September 19, 2017 (https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICT FactsFigures2017.pdf).

James G., Witten D., Hastie T. and Tibshirani R. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer.

Jolliffe I.T. (1986). *Principal Component Analysis*. Springer.

King G. (2014). Restructuring the social sciences: Reflections from Harvard's institute for quantitative social science. *PS: Political Science & Politics*, 47(1), 165-172.

Kontostathis A., Pottenger W.M. and Davison B.D. (2005). Identification of critical values in latent semantic indexing. In *Foundations of Data Mining and Knowledge Discovery*, 333-346. Springer, Berlin, Heidelberg.

Kontostathis A. and Pottenger W.M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, 42(1), 56-73.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1(1), 127-165.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie* (Vol. 24). Slatkine.

Laney D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research* Note, 6(70).

Lebart L. and Salem A. (1988). *Analyse statistique des données textuelles: questions ouvertes et lexicométrie*. Dunod.

Lebart L., Salem A. and Berry L. (1997). *Exploring Textual Data* (Vol. 4). Springer Science & Business Media.

Lee D. L., Chuang H. and Seamons K. (1997). Document ranking and the vector-space model. *IEEE Software*, 14(2), 67-75.

Li T., Ma S. and Ogihara M. (2004). Document clustering via adaptive subspace iteration. *ACM SIGIR Conference*.

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281-297.

Mazziotta M. (2017). Well-being composite indicators for Italian municipalities: Case study of basilicata. *Working Papers Series – PhD Course in Applied Social Sciences*, 1/2017.

Muller C. (1973). *Initiation au méthodes de la statistique linguistique*. Classiques Hachette.

Nardi P.M. (2015). *Doing Survey Research*. Routledge.

Pollak S., Coesemans R., Daelemans W. and Lavrač N. (2011). Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. *Quarterly Publication of the International Pragmatics Association*, 21(4), 647-683.

Reinert M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse des Données*, 8(2), 187-198.

Rocci R. and Vichi M. (2008). Two-mode multi-partitioning. *Computational Statistics & Data Analysis*, 52(4), 1984-2003.

Rossi P.H., Wright J.D. and Anderson A.B. (2013). *Handbook of Survey Research*. Academic Press.

Salton G., Wong A. and Yang C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620.

Salton G. and McGill M.J. (1983). *Introduction to Modern Retrieval*. McGraw-Hill Book Company.

Salton G. and Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

Sandhya N., Lalitha Y.S., Sowmya V., Anuradha K. and Govardhan A. (2011). Analysis of stemming algorithm for text clustering. *International Journal of Computer Science Issues (IJCSI)*, 8(5).

Struijs P., Consten A., Daas P., Debusschere M., Ilves M., Nikic B., Nowicka A., Salgado D., Scannapieco M. and Swier N. (2017). The ESSnet Big Data: Experimental Results. In (edited by) Petrucci A. and Verde R. *SIS 2017. Statistics and Data Science: New Challenges, New Generations*, 969-976. Firenze University Press.

Sullivan D. (2001). *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. John Wiley & Sons, Inc.

Tjhi W.C. and Chen L. (2006). A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recognition Letters*, 27(3), 151-159.

T-LAB. (2017). *User's Manual T-LAB Plus 2017*. T-LAB. Retrieved July 14, 2017 (http://tlab.it/en/download.php).

Vichi M. (2001). Double *k*-means clustering for simultaneous classification of objects and variables. *Advances in Classification and Data Analysis*, 43-52. Springer Berlin Heidelberg.

Vichi M. (2013). Robust Two-mode clustering. *Proceedings of the 59th World Statistics Congress of the International Statistical Institute.*

Vichi M. (in press). Two-mode Clustering.

Wanta W. and Ghanem S. (2007). Effects of agenda setting. *Mass Media Effects Research: Advances through Meta-Analysis*, 37-51.

Xu G., Zhang Y. and Li L. (2010). *Web Mining and Social Networking: Techniques and Applications* (Vol. 6). Springer Science & Business Media.

Yule G.U. (1944). *A Statistical Study of Vocabulary*. Cambridge University Press.

Zipf G.K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin.

Zipf G.K. (1949). *Human Behaviour and the Principle of Least-Effort*. Addison.

Appendix

Table A.1 – Work accidents by economic activity, comparison between News repository and INAIL official data.

Economic activity (NACE Rev. 2 classification)	***News Repository (%)***	***INAIL official statistics (%)***
C. Manufacturing	31.03	25.05
F. Construction	29.59	10.75
H. Transportation and storage	18.85	10.95
G. Wholesale and retail trade	5.49	13.64
N. Administrative and support service activities	2.63	7.15
E. Water supply	2.15	2.76
Q. Human health and social work activities	2.15	11.24
B. Mining and quarrying	1.67	0.17
I. Accommodation and food service activities	1.67	6.86
S. Other service activities	1.19	1.97
D. Electricity, gas, steam and air conditioning supply	0.95	0.38
J. Information and communication	0.95	1.48
M. Professional, scientific and technical activities	0.72	2.58
R. Arts, entertainment and recreation	0.48	1.18
K. Financial and insurance activities	0.24	1.43
P. Education	0.24	1.46
L. Real estate activities	0.00	0.91
T. Activities of households as employers	0.00	0.02
U. Activities of extraterritorial organizations	0.00	0.03

Table A.2 – Work accidents by mortality, comparison between News repository and INAIL official data.

Fatality	***News Repository (%)***	***INAIL official statistics (%)***
Fatal accident	25.00	0.20
Non-fatal accident	73.45	99.80
Not classified	1.55	-

Table A.3 – Work accidents by situation, comparison between News repository and INAIL official data.

Ongoing and/or road accident	***News Repository (%)***	***INAIL official statistics (%)***
Ongoing accident	3.10	3.94
Ongoing road accident	8.89	4.20
Road accident	25.11	12.84
Other	62.90	79.02

Table A.4 – First 200 terms used for the construction of the word cloud, NeRO corpus.

No.	**Term**	**Frequency**	**No.**	**Term**	**Frequency**
1	Incidente	2398	101	Pronto	226
2	Anno	1839	102	Tragedia	225
3	Uomo	1789	103	Estrarre	225
4	Porre	1372	104	Solo	222
5	Ospedale	1342	105	Terra	221
6	Operaio	1272	106	Frattura	221
7	Rimanere	1177	107	Colpo	220
8	Via	1128	108	Traffico	220
9	Soccorso	1103	109	Precipitare	219
10	Primo	1100	110	Lasciare	217
11	Grave	1057	111	Provincia	214
12	Trasportare	951	112	Trasferire	212
13	Vigile	922	113	Giorno	211
14	Carabiniere	914	114	Guida	211
15	Fuoco	878	115	Oggi	211
16	Fare	822	116	Elicottero	209
17	Intervenire	763	117	Riservare	205
18	Avvenire	742	118	Figlio	205
19	Ancora	663	119	Seguito	204
20	Lavorare	662	120	Caduta	203
21	Ambulanza	661	121	Indagine	201
22	Vita	628	122	Albero	199
23	Condizione	585	123	Impegnare	199
24	Morire	581	124	Conducente	199
25	Ferire	572	125	Giungere	199
26	Mattino	563	126	Milano	197
27	Secondo	561	127	Mortale	197
28	Vittima	559	128	Notte	195
29	Riportare	559	129	Carreggiata	192
30	Infortunio	556	130	Sud	191

31	Codice	555	131	Lamiera	189
32	San	544	132	Gamba	189
33	Polizia	536	133	Corpo	187
34	Azienda	487	134	Macchinario	187
35	Trovare	476	135	Mezzo	184
36	Perdere	469	136	Testa	184
37	Cadere	464	137	Comune	182
38	Stradale	461	138	Autista	180
39	Dinamico	459	139	Tratto	177
40	Metro	432	140	Necessario	176
41	Collega	427	141	Asl	173
42	Persona	427	142	Tragico	172
43	Accertamento	421	143	Operazione	172
44	Pomeriggio	418	144	Autostrada	172
45	Ferito	416	145	Allarme	169
46	Accadere	414	146	Chilometro	168
47	Ricoverare	414	147	Forte	168
48	Altezza	412	148	Lavoratore	167
49	Strada	408	149	Eliambulanza	167
50	Intervento	406	150	Urgenza	166
51	Verificare	405	151	Servizio	165
52	Ferito	405	152	Tecnico	164
53	Rosso	401	153	Stazione	163
54	Interno	400	154	Diverso	162
55	Giovane	391	155	Mano	161
56	Coinvolgere	390	156	Guidare	160
57	Bordo	388	157	Sottoporre	159
58	Causa	385	158	Tir	158
59	Portare	382	159	Casa	158
60	Ditta	374	160	Insieme	157
61	Trauma	372	161	Provocare	156
62	Travolgere	372	162	Lesione	156
63	Sicurezza	348	163	Incrocio	155

64	Medico	346	164	Trattare	155
65	Sanitario	334	165	Venire	155
66	Cantiere	322	166	Decedere	154
67	Impatto	316	167	Macchina	154
68	Residente	312	168	Fiesta	154
69	Roma	312	169	Chiamare	151
70	Ricostruzione	309	170	Fiat	151
71	Dovere	305	171	Sinistro	148
72	Colpire	294	172	Moglie	148
73	Arrivare	293	173	Agente	146
74	Zona	291	174	Stabilimento	145
75	Pericolo	274	175	Tetto	145
76	Luogo	274	176	Riuscire	145
77	Schiacciare	270	177	Furgone	144
78	Viaggiare	270	178	Mettere	144
79	Finire	268	179	Agricolo	143
80	Direzione	262	180	Gravissimo	142
81	Proprio	261	181	Scontrare	140
82	Personale	260	182	Grande	140
83	Lavorio|lavoro	257	183	Vettura	138
84	Brescia	253	184	Reparto	138
85	Prognosi	252	185	Capannone	138
86	Scontro	248	186	Trattore	138
87	Locale	248	187	Moto	137
88	Pesante	247	188	Condurre	137
89	Camion	245	189	Minuto	136
90	Ragazzo	241	190	Succedere	135
91	Dipendente	240	191	Immediatamente	134
92	Elisoccorso	239	192	Incastrare	134
93	Donna	238	193	Volo	134
94	Effettuare	237	194	Ribaltare	133
95	Ricostruire	235	195	Marco	130
96	Rilievo	233	196	Giallo	130

97	Gravemente	231	197	Schianto	129
98	Accertare	230	198	Km	129
99	Lungo	230	199	Vicino	127
100	Morto	228	200	Investire	127

Table A.5 – First 200 terms used for the construction of the word cloud, Brexit corpus.

No.	***Term***	***Frequency***	***No.***	***Term***	***Frequency***
1	Talk	3131	101	Policy	355
2	Leave	2687	102	Negotiate	354
3	Amp	2155	103	Live	353
4	Deal	1992	104	Vision	348
5	Vote	1950	105	Delay	348
6	Trade	1888	106	Reform	345
7	Post	1752	107	Start	343
8	Speech	1561	108	Pro	339
9	What	1427	109	Anti	335
10	Transition	1402	110	Member	332
11	Labour	1361	111	Laureate	331
12	Year	1299	112	Win	331
13	News	1297	113	World	329
14	Progress	1247	114	Money	325
15	New	1199	115	Parliament	324
16	Citizen	1165	116	Part	324
17	Plan	1109	117	Late	321
18	Right	1051	118	Voter	320
19	Need	1042	119	Decline	308
20	Macron	1039	120	Exit	308
21	Person	998	121	Conservative	307
22	How	979	122	Set	305
23	Tell	926	123	Ready	304
24	Bill	913	124	Great	301
25	Stay	850	125	Democracy	299
26	Think	810	126	Long	298
27	Rule	807	127	Face	298
28	Referendum	773	128	Break	297
29	Move	758	129	Join	293
30	National	755	130	Change	292

31	Negotiation	752	131	Job	287
32	Hard	718	132	Today	285
33	Leader	696	133	Nationality	284
34	Back	691	134	Come	283
35	Remain	673	135	Conference	280
36	Time	672	136	Lay	278
37	Pay	667	137	Rank	276
38	Stop	643	138	Bank	275
39	Forward	637	139	Oppose	275
40	Country	637	140	Rate	273
41	Negotiator	623	141	Open	268
42	Free	623	142	Rival	266
43	Work	613	143	Power	266
44	Offer	607	144	Believe	262
45	Market	601	145	Court	262
46	Lead	591	146	Control	261
47	Good	581	147	Second	260
48	Urge	577	148	Read	259
49	Business	571	149	Grow	256
50	Big	571	150	Tax	255
51	Tusk	566	151	Fear	254
52	Warn	566	152	Interest	253
53	Demand	565	153	Problem	252
54	Chief	560	154	Look	251
55	Mean	557	155	Worker	250
56	Miracle	547	156	Moody	247
57	Know	534	157	Full	246
58	Independent	530	158	Meet	246
59	Period	528	159	Deliver	244
60	Law	527	160	Agreement	239
61	Hope	520	161	Prepare	239
62	Apply	518	162	Trump	238
63	Minister	514	163	Downgrade	236

64	Divorce	512	164	Finance	234
65	Chance	511	165	Concrete	234
66	Single	507	166	Adopt	234
67	Government	497	167	Fail	232
68	Citizenship	476	168	Reveal	232
69	Party	465	169	Day	229
70	Agree	453	170	Profound	228
71	Support	449	171	Round	226
72	Bid	449	172	Point	226
73	Term	444	173	Let	225
74	Happen	443	174	Home	224
75	Call	441	175	Resume	223
76	Walk	437	176	Lie	221
77	Issue	436	177	Rise	221
78	Nothing	434	178	Result	220
79	Agency	432	179	Guardian	220
80	Bad	429	180	Independence	219
81	Ask	427	181	Please	219
82	Give	421	182	Retain	218
83	Even	421	183	Economy	216
84	Claim	420	184	Blame	216
85	State	418	185	Here	216
86	Future	416	186	Run	214
87	Pledge	416	187	Confirm	213
88	Union	390	188	Feel	212
89	Border	389	189	Politician	210
90	Economist	383	190	Election	210
91	Dual	379	191	Official	208
92	Thing	377	192	President	207
93	Show	373	193	Hit	206
94	Keep	372	194	Create	205
95	Own	363	195	Fund	203
96	Lose	361	196	Datum	203

97	End	361	197	Continue	203
98	Position	360	198	Membership	203
99	Month	356	199	Movement	203
100	Well	356	200	Service	202