Using Text Mining to Analyse Digital Transformation Impact on People

Florinda Matos[1], Valter Vairinhos[2, 3] and Ana Josefa Matos[2]
[1]DINÂMIA'CET-IUL - ISCTE-IUL, Lisbon, Portugal
[2]ICLab - ICAA - Intellectual Capital Association, Santarém, Portugal
[3]CINAV - Naval Research Centre - Escola Naval, Almada, Portugal
florinda.matos@iscte-iul.pt
valter.vairinhos@icaa.pt
anajosefa.matos@icaa.pt

DOI: 10.34190/EAIR.20.041

Abstract: Digital transformation is changing people's lives in many ways, creating competition between people and machines. All aspects of people's lives are being influenced with global impacts for society. In this context, many problems have emerged for which there is still no clear ideas of their effects on people's lives. To study these problems, new tools and methodologies are needed in order to compare large volumes of data. The analysis of texts, using Text Mining, has been gaining prominence, among researchers, as one of the most relevant methodologies. However, methodologies using Text Mining are not robust enough to allow researchers to compare data from different sources, such as report data and text data. The main objective of this paper is to propose an innovative Text Mining methodology that allows to compare different texts. This study is exploratory, and it is supported by quantitative methodologies. Using Text Mining to explore ECIAIR 2019 proceedings and other European reputed reports about digital transformation, and comparing the opinions expressed by researchers with those manifested by other people, it is intended to understand if there are coincidences in the language used by researchers and on the reports in what concerns what people feel about the impacts of digital transformation on their lives. This paper belongs to an ongoing research aiming to develop text mining tools that consider corpora as variables with specific values, treating those variables as statistic variables, contributing to the enrichment of the statistical methodologies used to study digital transformation impacts. The results show that there is a gap between the language of the investigators and the one used on the reports. At the same time, there are also overlaps in some topics analysed in the documents. These results indicate that there are topics that concern both the scientific community and the international organisations responsible for the preparation of public policy guiding reports.

Keywords: People, Robots, Digital Transformation, Text Mining, Cluster Analysis

1. Introduction and Contextualisation

Digital transformation and the large-scale application of artificial intelligence (AI), robotics and analytics is a process each day with more importance, leading to unavoidable sociological and ethical issues (Russel and Norvig, 2010*;* Russel, 2020). Russel (2020) presents a logic-mathematical attempt to redefine a concept of a robot, from the point of view of its provability benefits for man. According to this author, machines are good when there is no doubt of whom is the control. The decision to build machines with behaviour indistinguishable from humans is not only a matter of technical feasibility and economic interest. This is just one example of the pertinent questions currently being asked to researchers. Effectively, digital transformation has already begun to demonstrate that researchers will be vital in the development of a society that puts people's well-being at the centre of decisions.

The Covid-19 pandemic, with all of the impacts it has been having on the world in general, but particularly on people's lives and the economy, has brought to the public discussion the impacts on the new business and work models, with emphasis on the teleworking.

Understanding these phenomena, studying them, and drawing valid conclusions that allow decision-makers to make better decisions has become imperative. However, the scientific investigation of tools, that facilitate these processes, does not have the same velocity as the implementation of these processes. Thus, researchers often continue to analyse new and highly complex phenomena with the same tools they used in the past century. Namely, in terms of mathematical models, which facilitate the investigations in these areas, their evolution has been scarce, and new studies are still needed.

This paper presents an attempt to synthesise what the investigators, assembled in the last ECIAIR 2019 conference, are feeling, and saying about the impacts of digital transformation on important human activities.

These findings will be related with similar ideas and conclusions, expressed at the country level, in the texts of reports and questionnaires from the Eurobarometer 469, Special Eurobarometer 503 (2019), Impact of digitalisation, Eurobarometer 480, Internet Security, Special Eurobarometer 499 and Europeans Attitudes towards cyber-security.

The paper is structured in four sections in addition to the Introduction and Contextualization. The four sections are: section 1 – Empirical Research, dedicated to objectives and investigation hypothesis, data sources and its organisation as corpora and data analysis and software employed; section 2 is dedicated to a synthesis of the statistical methodology employed; section 3 is dedicated to data analysis and the interpretation of results; and lastly, section 4 is used to present the conclusions, their discussion and future work.

2. Empirical Research

2.1 Objectives and Investigation Hypothesis

The original motivation for this research was to relate the content of the papers presented in the 2019 edition of ECIAIR about the impact of artificial intelligence on people's lives, with the Eurobarometer reports content covering the study of results of public polls about similar topics (Griffiths and Kabir, 2019; EU 2017, 2018, 2019).

The main objective of this work is to study and relate, using a suitable methodology, the languages used by researchers when they produce knowledge about digital transformation and by the technicians when they design, analyse and synthesise the results of public opinion polls about such themes.

Since this problem turnout to be a specific instance of a general problem of global comparison of two corpuses, the development of a specific text mining methodology for that kind of problem is also an objective of this study.

The researchers express their results about knowledge management, digital transformation, robotics and artificial intelligence in scientific papers, or other publications and presentations in scientific meetings. The technicians express their results when they elaborate the specifications and results from Eurobarometer polls involving those themes. Both groups see the same themes from two distinct points of view: the researcher from the point of view of knowledge creation and the technicians from the point of view of discovering what people think about that.

The subjacent idea is that it would be natural to expect evidence of some noticeable overlapping between the two languages, since it is natural to expect that the results of a good scientific research influence the formulation, design, results, and synthesis of those studies. On the contrary, in the case of irrelevant scientific literature, some noticeable gaps are expected. More specifically, what people know – the terms, words meanings, sentences and mental models they use to generate the documents they produce (scientific papers or specifications and reports about results of opinion public polls) express the nature, quality or power of their scientific ideas or the excellence of their formation at the respective texts production times. If there exists a considerable gap between those languages, it is expected that those gaps are manifested in the documents produced by the two sets of writers and captured by the analysis instruments employed. On the other hand, if there exist relevant overlaps, these should also be expressed in some way in the texts produced and, hopefully, detected by the relevant methodology.

The meaning of those overlaps or gaps must be examined, explained, and valued in the specific context that is being considered, since they can mean distinct things to different observers– positive, negative or neither, in a specific context. For example, if there are strong detectable overlaps, that can imply that the ideas generated in the universities, where technicians studied, were not only valuable and fecund from the point of view of applications, but also mean that there existed a good knowledge diffusion and transmission processes.

The general idea is the perception that if research, expressed through papers and scientific books is successfully in some specific scientific domain, that should translate in the creation of new, useful and influential knowledge, that conditions the technicians' formation responsible for policies specification or the evaluation of its results. The documents (e.g. reports, studies, evaluation studies, etc.) are expressed in a language that reflects, in some detectable way, the scientific formation they received and, consequently, the language in which it is expressed. This means that it makes sense to study the texts involved in these two people groups to detect those influences, if there are any.

The presence, or absence, of such influences, expressed in documents generated along the time can explain the eventual emergence of trends in the elaboration of technical or normative documentation. Some domains where this kind of inquiry could interest are, just to mention a few, intellectual capital and intangibles, knowledge management research and its consequence in governance and legislation or organisation of scientific systems.

2.2 Data, Materials and Software.

As mentioned before, the two data sources used for this research were the proceedings of the European Conference on the Impact of Artificial Intelligence (ECIAIR 2019) and the reports about European public opinion regarding topics related with digital transformations and internet use, namely:

- Special Eurobarometer 460: Report. Attitudes towards the impact of digitalisation and automation on digital life.
- Flash Eurobarometer 469: Report.
- Special Eurobarometer 499: Report. European's attitudes towards cyber security.

The research papers in ECIAIR 2019 formed one of the two corpora used in the data analysis that follows. The contents structural parts of the three reports for Eurobarometer (EU, 2017, 2018, 2019) were used to form the second corpus (reports).

For Text Mining tasks (descriptive statistics of texts) *R language* was used, namely the package *quanteda* (Benoit et al. 2018a,2018b) was used to perform the basic tasks of counting, tokenising and stemming and the preparation of data structures to be used by other packages. Those packages were also used for modelling text data, namely in the implementation of *LDA - Latent Dirichlet Analysis* (see Grün *et a.l*, 2011, 2020).

In addition, graphics for Biplots (Gabiel, 1971, Galindo, 1985 and Nenadié *et al*., 2007) were produced by the *program Biplots PMP* (Vairinhos, 2003).

3. Methodology

The main statistical methodologies employed in this paper, in addition to the usual counting and graphical tasks, were Cluster analysis *using R package stats and biplots*. For inferential tasks, the *Probabilistic Topic Models (*Blei, 2012), was employed (Blei *et al.,* 2003). According to this probabilistic model, documents are assumed to be generated by finite sets of topics (latent - non directedly observable variables) modelled by a random probabilistic process. This means that each text in a set of documents (*corpus*) can be seen, theoretically, as being generated by a mixture of such topics, each one representing a specific subject. Given such a corpus, the topics are estimated using the whole set of words contained in the corpus. Such an estimation assumes the form of lists of words with specific probabilities of belonging to the topic. The estimation method used was the method LSA explained in Blei, et al. (2003) as implemented by *R package* (Grün, Kepler and Hornik 2011; Grün, et al. 2020). In this perspective, it is assumed that current ideas, concepts and perceptions of researchers - at the time the papers were written - about the subjects of an international meeting such as European Conference on the Impact of Artificial Intelligence and Robotics are well represented in the textual content of the resulting proceedings. It is also assumed that those papers represent a good sample of what was then known about the conference thematic contents, in this specific case, about what the main European research groups know, assume or feel about the impact of robotics and artificial intelligence on society, the topics being estimated by LDA using the ECIAIR 2019 papers.

On the other hand, the technical documents and reports that present to public opinion the study's results about Eurobarometer polls commanded by EU (2017, 2018, 2019) at the national level, are assumed to have been generated by another set of latent topics that explain the specific texts observed. This is the case of final reports (containing conclusions, discussion of results and synthesis) from Eurobarometer studies.

In this research, biplots were used to relate sets of documents and terms obtained through tokenization and counting of terms occurrence in texts. See Gabriel (1971), Galindo (1087), Nenadié *et al.* (2007). The methodology of Cluster analysis employed the *hclust* procedure from R stats package.

4. Data Analysis and Interpretation.

The documents representing papers contained in the Proceedings of ECIAIR19, (Griffiths, *et al.,* 2019), are numbered according to the sequence of pages *i-iv* of that publication. In the present research, those papers are numbered from 1 to 60, because some original papers were split in two documents. The reports are numbered from 61 to 87.

The main objective of this data analysis is to detect overlaps or gaps of scientific documents and reports. From now on, it will be considered the following notation P for papers and R for reports. The first issue addressed is to specify a criterion giving the meaning of the overlap between the languages employed by the two sets of documents – or, otherwise, given the meaning of the gap.

The methodologies of the data analysis used in this research (biplots and Cluster analysis) imply the Clustering of documents based in the language employed in its texts. The criteria was defined as it follows: "given two corpus, whenever in a clustering of those documents expressed by its proximities in function of words occurrence, in each Cluster the probability of occurrence of texts originated in both corpus is the same, there is a perfect overlap between the two sets of documents".

When documents from both corpora are mixed in a compound corpus and the common set of words in this compound is used to obtain homogeneous clusters of documents, the resulting clusters correspond to homogeneous sets of documents with similar meaning and content but covering distinct subjects or topics.

The simultaneous occurrence of two documents in the same Cluster mean that those documents use words with similar meanings. When those documents come from both corpus, this means that this Cluster manifests some overlap degree (measured by the performance measure), meaning that the associated documents have similar words and, possibly, similar contents. If this event is replicated for all Clusters, that would correspond to strong evidence supporting the idea that both groups use the same language for distinct topics associated with distinct clusters. For a cluster containing only documents from one of the two corpora, the overlap is null, which is reflected by the value 0 for the criterium implemented in the expression below. In synthesis, the overlap is maximum when the incertitude about the origin of the documents is maximum. In practice, those coincidences of documents from different *corpora* can manifest for some variable proportion of Clusters, allowing to examine the reasons for the overlaps and for the gaps. The reasons for the overlaps or for the gaps can be obtained from the contents of documents belonging to each Cluster.

The following expression assumes values in the interval *[0,1]* and can be used to measure overlap between two *corpora*:

$$ov = \left(\frac{4}{K}\right) * \sum_{i=1}^{i=K} pi * (1 - pi)$$

Where:

- *pi i=1..K* is the observed proportions of documents from *corpus 1* for Cluster *I;*
- *K* is the number of Clusters in the clustering.

This expression assumes values between 0 - all Clusters are formed by documents from just one *corpus* - to 1 - both corpora are equally represented in all Clusters (Vairinhos, 2020).

Figure 1 shows a biplot produced by the program *BiplotsPmd*, (Vairinhos, 2003), using the results of tokenization, stemming, and counting of terms in each one of the 87 documents occurring in the two *corpora* involved: scientific papers of ECIAIR 2019 (Griffiths, *et al.*, 2019) and reports from Eurobarometer (EU,2017,2018,2019). The blue points represent scientific papers from ECIAIR 2019 proceedings and technical chapters of reports generated when studying the results of the three Eurobarometer polls.

The Biplot in Figure 1 shows also, in red, the 150 more frequent terms occurring in those two sets of texts. The figure, also, shows that, from the three text Clusters detected by the Cluster analysis performed using the texts and words coordinates on biplot, the proximities among documents mean that those documents use similar terms in similar proportions. The proximity of a Cluster of words to a Cluster of documents indicates that the words of the Cluster of words can explain the proximities of the texts in that specific Cluster of texts. For example, in Figure 1, the words in Cluster of terms 5 help to explain the meaning of Cluster documents 2, all Reports, in the set {63,64,65,67,69}.

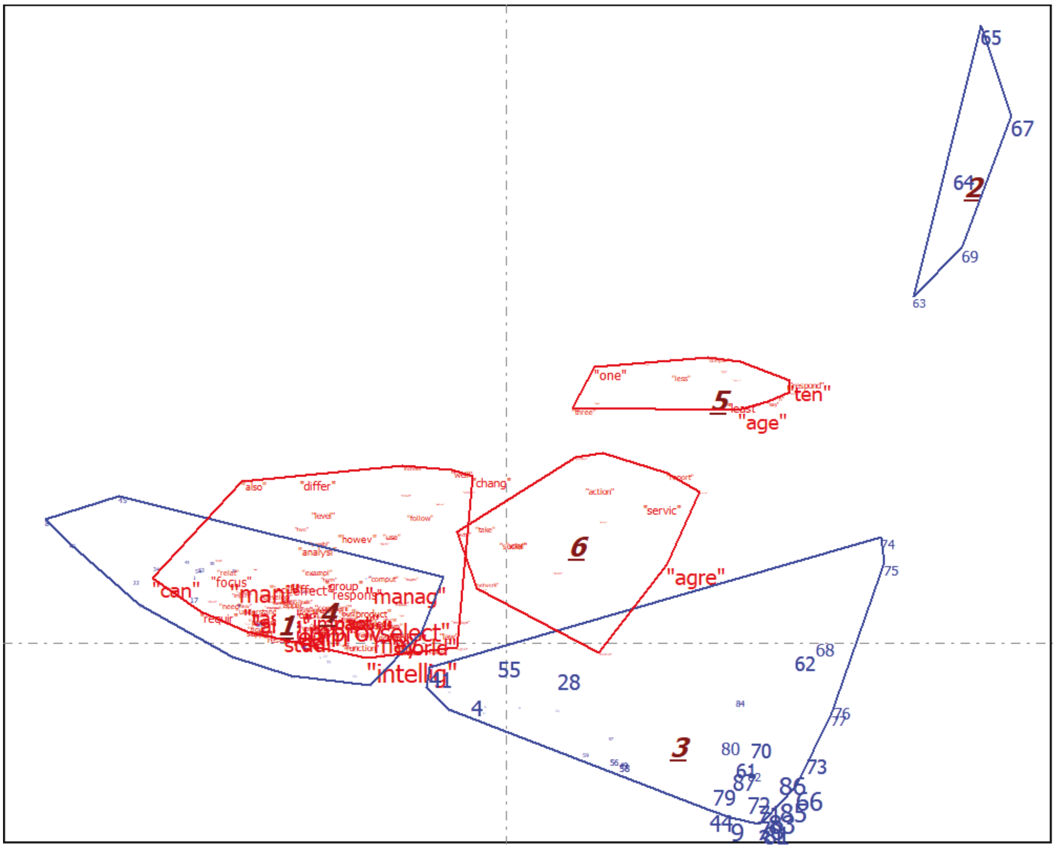

Figure 1: Biplot obtained from a table of 87 rows (documents) by 150 columns(terms) with frequencies of occurrence of the 150 more frequent terms in the 87 texts. Texts are represented by numeric identifiers in blue. Terms are represented in red. Cluster 3 represents a mixture of Papers and Reports. Cluster 1 is formed only by scientific Papers. Cluster 2 is formed only by Reports.

To get a more explicit and precise definition and enumeration of documents Clusters, a Cluster analysis was performed using now the program hclust (R package stats) where each document was represented by a vector of 150 frequencies. The resulting dendrogram can be seen in Figure 2. That figure clearly shows the presence of three Clusters of documents whose composition can be read at the bottom of Figure 2. The composition of those Clusters is, from top to bottom, considering only the numeric part of labels:

Cluster 1 = {1, 2,..8, 10,…48, 51, 52,53,54,55,60} - 48 Scientific Papers
Cluster 2 = {63,64,65,67,69} - 5 Reports
Cluster 3 = {9,28,31,33,44,49,50,56,57, 58,59,61,62,66,68,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87} - 33 Documents: 11 Scientific Papers + 22 Reports

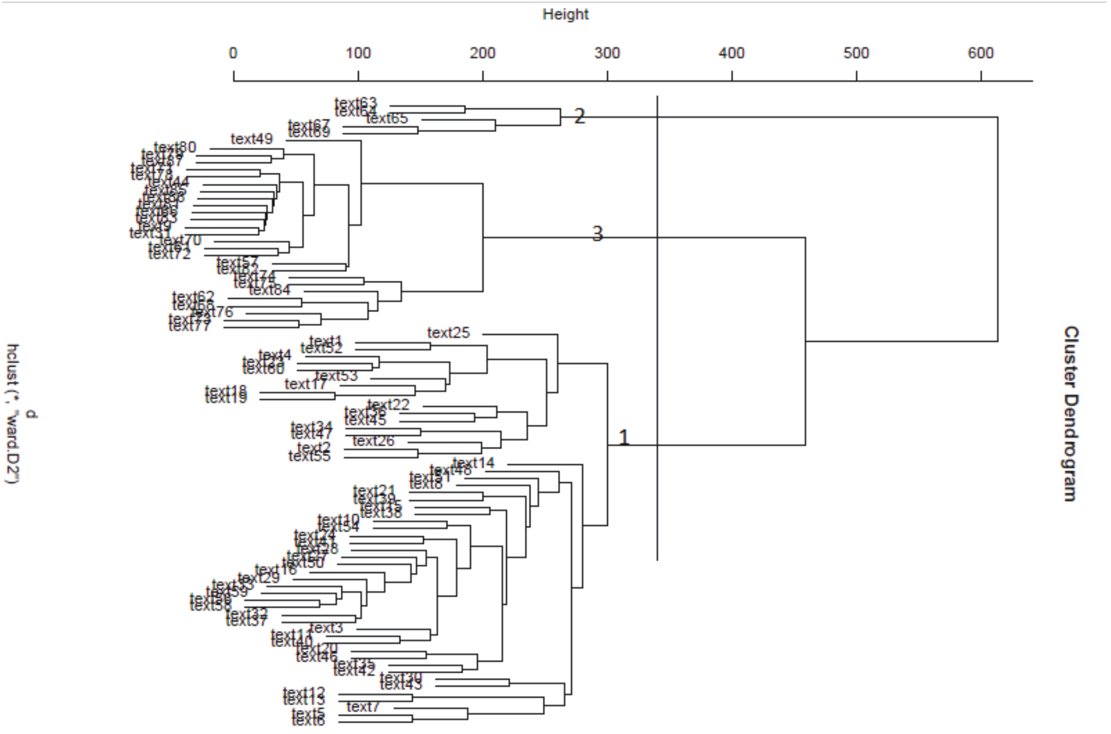

Figure 2: Cluster analysis results obtained by procedure hclust from R package stats (CRAN)

These Clusters are consistent with those shown in biplot of Figure 1 – both were obtained with the same documents aggregation method (Ward).

As can be confirmed with the results expressed in Figure 1, the Clusters 1 and 2 are "pure" in the sense that are formed just by one type of document (scientific papers for Cluster 1, reports for Cluster 2 and a mixture of Scientific papers and Reports for Cluster 3. Applying the criterium specified, this analysis shows that there is not a perfect overlap of languages used by scientific papers and reports, as Clusters 1 and 2 contain only one kind of document and in each one of those Clusters there is a maximal gap between texts. This is not the case with Cluster 3 where an overlap is detected since in this Cluster coexist documents of distinct nature but with similar language and, consequently, similar meaning. The overlap obtained with previous expression is *ov = 0.23 (low)*.

To synthetize the meaning of those three Clusters, LDA - Latent Dirichlet Analysis (Grün and Hornick,2011, Blei et al. , 2003; Blei 2012, Chen et al. , 2019), from the *R package topicmodels,* was used. In each one of those Clusters, the 3 more likely latent topics were estimated by their combination.

For Cluster 1, the estimated topics (given by its more probable term) are:
Cluster 1:
Topic1 = {artificial, intelligence, education, learn, higher, law, robot, digita, approach, hybrid}.
Topic2 = {analysis, intelligence, artificial, system, ethics, use, ai, robot, transform, technology}
Topic3 = {ai, impact, development, machine, dynamic, towards, human, emerge, work, future}

Cluster 2:
Topic1 = {cybercrime, perception, response, provide, assist, citizen, differ, type}.
Topic2 = {cybercrime, aware, experience, fight, use, internet, eb_499_i, eb_499_II, eb}
Topic3 = {internet, eb_499_i, use, eb_499_ii, concern, interact, cybercrime, fight, eb_499_iii}

Cluster 3:
Topic1 = {ai, promote, use, response, hard, interdisciplinary, ethic, design, system, possibility}.
Topic2 = {artificial, intelligence, find, digital, use, impact, boss, machine, human, choice}
Topic3 = {content, online, experience, cybercrime, continuous, specific, host, service, attitude, toward}

5. Discussion of Results and Future work

The main result of this research is the formulation of the problem of relating the documents of two corpora as an abstraction for a family of practical problems of comparison of the languages used by two groups of people when they address the same class of problems but under two point of view.

A methodology to deal with this problem was illustrated and a criterion for its performance specified. An instance of this problem - the comparison of languages used in scientific papers of ECIAIR 2019 and reports from Eurobarometer about polls on the same subjects - was analysed and the results show that the methodology makes sense and justifies future developments such as the use of statistic inference.

In synthesis, for the topics characterising Cluster 1 (papers from ECIAIR Proceedings) and Cluster 2 (reports – EU - 2017, 2018, 2019), it seems that there is no overlap in the language used in the documents that belong to those clusters: scientists and technicians use their specific languages. This is not the case for Cluster 3, where scientists and technicians seem to use similar languages to discuss specific topics.

Globally, as shown, the overlap obtained is only 0.23 in a scale [0,1].

Despite the limitations of this exploratory research, which relies on a small volume of textual data, it can serve as an alert to the need for greater coordination between scientific studies and reports that, as is known, are fundamental for public policies in country governance.

The suggested method of text mining is easily replicable and implementable using freely available software (free R packages identified in the paper) and a numeric performance criterion, also supplied.

References

Benoit, K., Watamabe, K., Wang, H., Nulty, P., Obeng, A., Müller, P., and Matsuo, A. (2018a). *quanteda*: An R Package for the Quantitative Analysis of Textual Data. *Journal of Open Source Software*, 3(30), 774.

Benoit, K., Watamabe, K., Wang, H., Nulty, P., Obeng, A., Müller, P., and Matsuo, (2018b). An R Quantitative Analysis of Textual Data. Repository CRAN. Available at: https://cran.r-project.org/web/packages/quanteda/index.html

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55, 77-84 [568]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022 [567, 580].

Chen, Z. And Doss, H. (2019). Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modeming. *Journal of Computational and Graphical Statistics*, 29(3), 567-585. Doi:10.1080/10618600.2019.1558063

EU - European Commission. Directorate-General for Communication (2017). Special Eurobarometer 460: Reports. Attitudes towards the Impact of Digitalization and Automation on Digital Life. Catalogue number KK-02-17-455-EN-N, ISBN: 978-92-79-68474-6. European Union, 2017. Doi: 10.2759/25616

EU - European Commission. Directorate-General for Communication (2018). Flash Eurobarometer 469: Report. Catalogue number KK-01-18-910-EN-N, ISBN: 978-92-79-93003-4, Doi: 10.2759/780040. European Union, 2018. Available at: http://ec.europa.eu/commfrontoffice/publicopinion.

EU - European Commission. Directorate-General for Communication (2019). Special Eurobarometer 499: Report. European's Attitudes towards Cyber Security. Catalogue number DR-02-20-011-EN-N, ISBN: 978-92-76-14938-5, European Union, 2019. Doi: 10.2837/672023. Available at : http://ec.europa.eu/commfrontoffice/publicopinion.

Galindo, M. P. (1985) *Contribuiciones a la Representación Simultánea de Datos Multidimensionales*, Tesis Doctoral, Universidad de Salamanca.

Gabriel, K. R. (1971) *The Biplot Graphic Display of Matrices with Application.*

Principal Component Analysis, Biometrika, 58 (3), pp. 453-467.

Griffiths, P., and Kabir, M. (2019). Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR 2019). *Academic Conference and Publishing International Limited*. E-Book ISBN: 978-1-912764-44-0. Book Version ISBN: 978-1-912764-45-7. Available at: http://academic-bookshop.com

Grün, B., and Hornick, K. (2011). *topicmodels*: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13), 1-30. Available at: http://www.jstatsoft.org/v40/i13/

Grün, B., Hornick, K., Blei, D. M., Lufferty, J., Phan, X.-H., Matsumoto, M., and Nishimura, T. (2020). Package "Topic Models". Repository CRAN.

Nenadié, O., and Greenacre, M. (2007). Correspondence Analysis in R, With Two - and Three - Dimensional Graphics: The ca Package. *Journal of Statistical Software,* May 2007, Vol 20, Issue 3. Published by American Statistical Association. Available at: http://www.amstat.org/

Russel, P., & Norvig, P. (2010). *Artificial Intelligence - A Modern Approach*. Third Edition. Prentice Hall.

Russell, S. (2020). *Provably Beneficial AI. The Turning Lecture*. The Alan Turning Institute,

Selivanov, D., Bickel, M., and Wang, Q. (2020). Package *text2Vec*. Modern Text Mining Framework for R. Repository CRAN. Available at: http://text2vec.org

Vairinhos, V. M. (2003). *Desarrollo de un Sistema de Minería de Datos Basado en los Métodos de Biplot*, Tesis Doctoral, Universidad de Salamanca.

Vairinhos,V.M. (2020). Comparing *corpus* for Text Mining. To appear.