SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF THE SILESIAN UNIVERSITY OF TECHNOLOGY 2019
ORGANISATION AND MANAGEMENT SERIES NO. 137

BUSINESS INTELLIGENCE – PUBLICATION ANALYSIS USING THE R LANGUAGE

Marcin WYSKWARSKI

Department of Economy and Informatics, Faculty of Organization and Management of Silesian University of Technology; marcin.wyskwarski@polsl.pl, ORCID: 0000-0003-2004-330X

Abstract: The purpose of the work was to analyse publications in the area of Business Intelligence. Only bibliometric data was used in the analysis. The analysis was performed using the R programming language. An attempt was made to determine whether by analysing bibliometric data, it is possible to obtain information on Business Intelligence systems. Aiming at achieving the adopted goal, in the second point of the work, selected information on Business Intelligence systems was presented. The third point presents the manner of collecting data. Further stages of the analysis were also presented. The fourth point contains the results of the conducted research. Among others, the number of publications in individual years and the most common words in titles, abstracts and keywords were presented. Using two topic modelling algorithms, topics were generated that can also be used to identify information related to Business Intelligence systems.

Keywords: Business Intelligence, bibliometrics, text mining, R language.

1. Introduction

The growing amount of data in the economic environment means that the demand for solutions that support decision-making processes by providing necessary information is constantly increasing (Shahid et al., 2016). This enables, among others, the concept of Business Intelligence focused on extracting information and discovering knowledge from data. Providing useful information and knowledge allows employees to make better decisions (Diaz, & Caralt, 2011).

The number of academic publications is also growing rapidly, which makes it increasingly difficult to read all publications on a given topic. This is due to, among others, the pressure to constantly provide new research results, which further contributes to conducting a lot of fragmented research (Briner, & Denyer, 2012).

http://dx.doi.org/10.29119/1641-3466.2019.137.17

One method of assessing the current state of knowledge in a given area is bibliometrics. This is a technique for the quantitative analysis of literature using mathematical and statistical methods. It enables the assessment of scientific creativity and the dissemination of knowledge on a given topic (Araujo, 2006). Bibliometrics is a statistical and quantitative analysis of academic results. It may include, among others, quantitative analysis of the content of articles, citations, impact factor, place of publication, keywords and cooperation between countries, authors and institutions (Ellegaard, & Wallin, 2015). Bibliometric analysis can be performed using various tools. One of them is the Bibliometrix package (Aria, & Cuccurullo, 2017). It is a package for the R language that enables the import of bibliographic data from Scopus and Web of Science. This package is widely used to evaluate scientific literature in the areas of healthcare, public administration, business and others (Addor, & Melsen, 2019; Aria, Cuccurullo, & Sarto, 2015; Cuccurullo, Aria, & Sarto, 2016).

The purpose of the work was to analyse the selected bibliometric data of publications from the area of Business Intelligence using the R programming language. The performed analysis was to provide answers to the following questions:

- How many publications were published in each year?
- Which issues were most often raised?
- Can text mining analysis of selected bibliometric data provide new information in the area of "Business Intelligence".

2. Business Intelligence

Literature on the topic lacks one universal definition of the term "Business Intelligence". This term was first proposed by Luhn in 1958 to describe an automated system designed to disseminate information and support decision making (Luhn, 1958).

Business Intelligence (BI) consists of systems combining data collection and storage, as well as knowledge management (Negash, & Gray, 2008). It is a broad concept covering the processes of collecting, integrating, analysing and visualising data performed to support and improve the decision-making process in the organisation (Fink, Yogev, & Even, 2017). Hannula and Pirttimäki define BI as organised and systematic processes used by enterprises to acquire, analyse and disseminate information relevant to their business operations. The information and knowledge obtained is used to support operational and strategic decisions (Hannula, & Pirttimäki, 2003).

Wixom and Watson write that BI "is a broad category of technologies, applications and processes for gathering, storing, accessing and analysing data to help its users make better decisions" (Wixom, & Watson, 2010). Yoon and others treat BI as "innovative tools for data analysis, query and reporting that [...] enables interactive access and manipulation of data in order to gain valuable insights and to support the management decision-making process across a broad range of business activities" (Yoon, Ghosh, & Jeong, 2014).

BI is usually not a single application but consists of different components. According to Kimball and Ross, the overall architecture of Business Intelligence systems includes the following elements (Kimball, & Ross, 2013):

- Data sources: Internal and external data repositories supplying data warehouses. Data may originate from, among others, databases, Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM), Supply Chain Management (SCM) and Online Transaction Processing (OLTP), as well as from text files, spreadsheets, sensors, etc.
- Extraction-Transformation-Load (ETL) process: Before data can be loaded into the data warehouse, it must first be extracted from the repositories and then processed, cleaned, filtered and redefined.
- Presentation Area: A place where organised data is stored and made available for user queries, applications and analytical tools. Usually it is a data warehouse, namely, a repository with properties such as stability, reliability, consistency and storage of historical information. It provides a global, shared and integrated view of the data, regardless of how the data will be used later.
- BI applications: These are solutions enabling access, analysis and sharing of information that is collected in a data warehouse. The processing uses, among others, OLAP tools and data mining tools.

Turban and others state that BI includes issues such as a data warehouse, data acquisition, data mining, business analytics and visualisation (Turban, Aronson, & Liang, 2005). According to the definition provided by the Gartner Group in 1996, "Business Intelligence is a series of technology or application systems which consist of a data warehouse (or data mart), reporting, data analysis, data mining, data backup and recovery components and which contribute to a better business decision and finally can help enterprises to keep a leading position in the competitive market" (Cheng, & Cheng, 2011).

3. Data source and publication analysis process

The study data was downloaded from the Clarivate Analytics Web of Science (WoS) database on 05.09.2019. From this database, publications were selected that had the phrase "Business Intelligence" in at least one of the following fields: title, abstract, keywords. Publication data was exported to files with the "bib" extension. The analysis included the following steps:

- Downloading data from the Clarivate Analytics Web of Science (WoS) database.
- Data import into the integrated programming environment of the R language (RStudio v.1.2.1335).
- Determining the number of documents in individual years, taking into account the type of documents.
- Data cleansing – among others, removing the so-called stopwords, bringing words to their basic form, removing all punctuation and digits, etc.
- Creation of keywords based on n-grams.
- Generating charts showing the most frequently used keywords and the most used words in abstracts and titles.
- Generating abstract topics using topic modelling algorithms.

The R programming language was used in the analysis. Among others, the "Bibliometrix" package (Aria, & Cuccurullo, 2017), "tm" package (Feinerer, 2008b, 2008a; Feinerer, Hornik, & Meyer, 2008) and the "topicmodels" package were used (Hornik, & Grün, 2011).

Thanks to the "convert2df" function, the "Bibliometrix" package ensured the conversion of data from files with the "bib" extension[1] to a data frame consisting of 42 columns. Each row of the data frame contained data on one publication. The number of publications was 4,024[2]. This amount of publications had the phrase "Business Intelligence" in at least one of the fields, such as title, abstract, keywords. After removing rows that had no value in one of these fields, the number of publications was limited to 3,087[3]. From these publications, a corpus of documents was created, which was identified by the code "3087_doc". Using the "biblioAnalysis" function (Bibliometrix package), an object of the "bibliometrix" class was created from the data frame. This object, together with the data frame, constituted the data source for further analysis.

The "tm" package enabled text mining analysis, while the "topicmodels" package enabled Topic Modelling algorithms.

4. The results of the performed analysis

Table No. 1. presents the number of documents with the phrase "Business Intelligence" in individual fields (i.e. title, abstract, keywords). It can be read that, among the analysed documents, 50 of them had the phrase "Business Intelligence" only in the title, and 95 in both the title and the abstract. The group of documents that had the phrase "Business Intelligence"

[1] The WoS database allows you to export a maximum of 500 publications at one time.
[2] The data frame was 4,024x42 (4,024 rows, 42 columns – 42 variables).
[3] The data frame was 3,087x42.

in all three fields (i.e. in the title, abstract and keywords) amounted to 833 documents. A second corpus was created from these documents, which was identified by the code “833_doc”.

During the research, other corpora were also analysed, e.g. a corpus made of 500 most frequently cited documents, a corpus made of documents published in 2016-2018, a corpus only containing the phrase “Business Intelligence” in the field of keywords, a corpus made of “proceedings paper” documents, etc. Due to the limited amount of workspace, the results are only shown for the “3087_doc” and “833_doc” corpora.

Table 1.
Number of documents with the phrase “Business Intelligence” in specific fields

Title	Abstract	Keywords	Number of documents
yes	yes	yes	833
no	no	yes	585
no	yes	no	982
yes	no	no	50
no	yes	yes	459
yes	yes	no	95
yes	no	yes	83
		Sum	**3,087**

Sources: own elaboration

Table 2 presents the number of publications analysed by document type. The most common document type for both the “3087_doc” and “833_doc” corpora was the so-called “proceedings paper”.

Table 2.
Number and type of documents analysed

Document type	**3087_doc**	**833_doc**
article	1045	270
article, book chapter	39	10
article, early access	1	0
article, proceedings paper	38	5
editorial material	6	0
editorial material, book chapter	2	0
proceedings paper	1,922	537
reprint	1	0
review	32	11
review, book chapter	1	0
Sum	**3,087**	**833**

Sources: own elaboration

Figure 1 and Table 3 show the number of publications in individual years. Taking into account the “3087_doc” corpus, most documents appeared in 2016 – 392 documents. In 2017, the number of publications dropped to 363, and in 2018, down to 292.

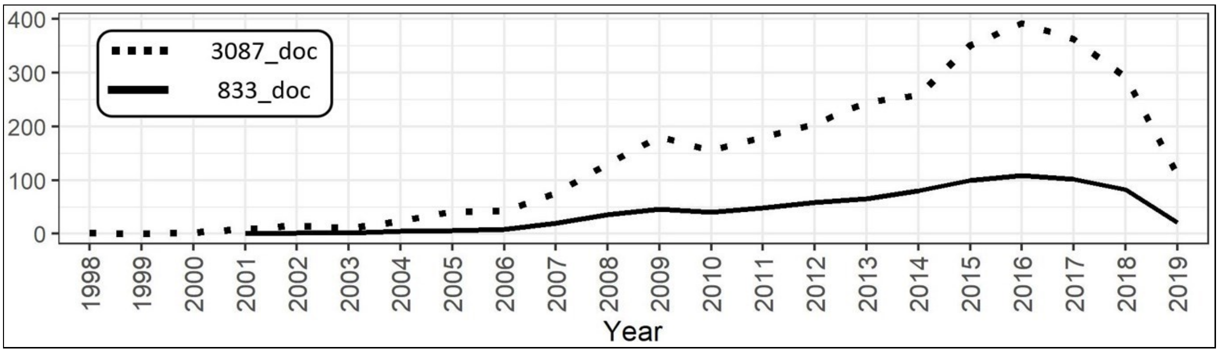

Figure 1. Number of documents analysed in individual years. Sources: own elaboration.

Table 3.
Number of documents analysed in individual years

Year	3087_doc	833_doc
1999	1	0
2000	3	0
2001	9	1
2002	15	2
2003	11	2
2004	23	5
2005	41	6
2006	43	8
2007	76	20
2008	130	36
2009	180	46
2010	155	40
2011	181	48
2012	204	59
2013	245	65
2014	258	80
2015	352	100
2016	392	109
2017	363	102
2018	292	83
2019	110	21
Sum	**3,084**	**833**

Sources: own elaboration.

Figure 2 shows the most common keywords. The keyword most often used by the authors was the phrase “Business Intelligence”. This was intentionally removed from the charts presented in Figure 2. In the “3087_doc” corpus, the phrase “Business Intelligence” appeared 1,842 times, and in the “833_doc” corpus, 763 times. Looking through the keywords, it can be seen that among them appear the names of solutions used to build BI systems, e.g. CRM, data warehouse, OLAP, ETL. Among them, there are also areas in which BI systems are used, e.g. knowledge management, decision support, decision making. There are also technologies and solutions that have been trending in recent years, e.g. cloud computing, big data analytics, machine learning, social media.

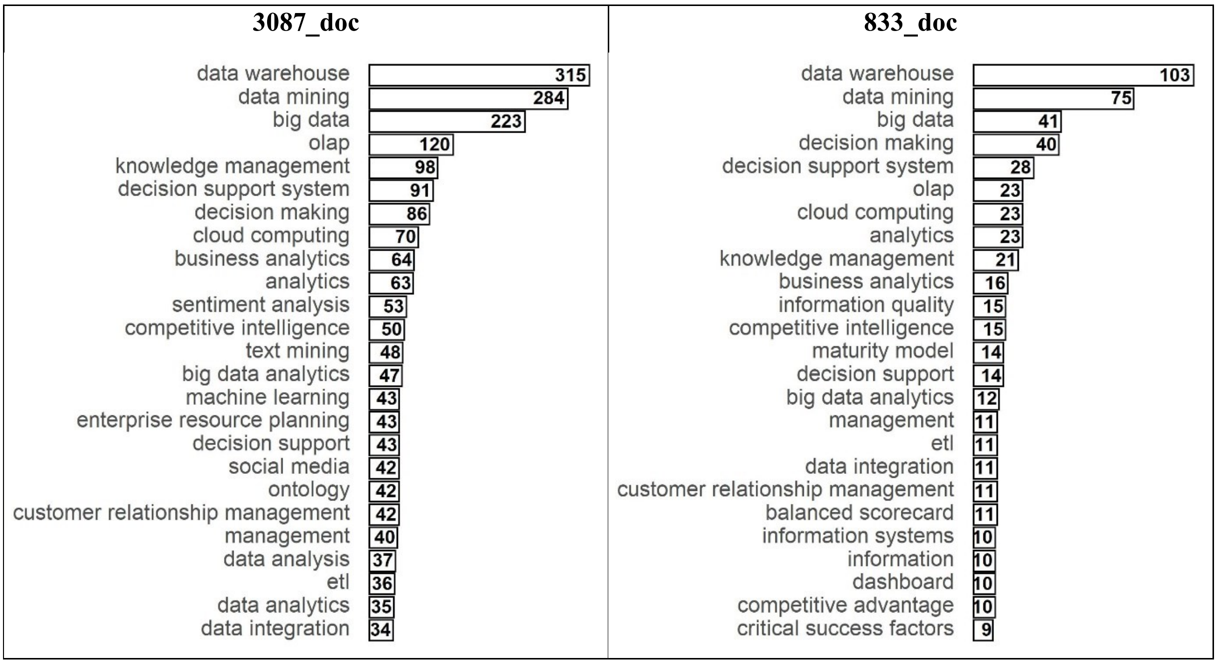

Figure 2. Most common keywords. Sources: own elaboration.

Analysing keywords, it was noticed that they also include phrases composed of several words, e.g. “data mining”, “big data”, “data warehouse”, “enterprise resource planning”, i.e. the so-called n-grams. By performing text mining analysis of abstracts and publication titles, it was decided to create n-grams based on keywords. In the charts representing the number of words most often used by the authors (fig. 3, fig. 4), one can notice the occurrence of n-grams. N-grams are also found in the word clouds presented in Figure 5. For example, 2-gram data_mining created from the words “data” and “mining”, visible in Figure 3, appeared in the titles of the analysed documents 108 times among documents from the “3087_doc” corpus, and 24 times for documents from the “833_doc” corpus.

The creation of n-grams allowed us to obtain additional information. Thanks to this, it was possible to calculate the occurrence of a specific phrase without breaking it into separate words. Failure to create n-grams would result in obtaining different results. For example, in the chart to the right of Figure 3, the word “data” would be a different value. The value of 26 would be increased by, among others, the value of 24 – the number of times “data” appeared in the 2-gram “data_mining”.

Figure 3 shows the most common words in the document titles. The most common phrase “Business Intelligence” was intentionally removed from the charts. Among the documents from the “3087_doc” corpus, the phrase “Business Intelligence” appeared in titles 917 times. For documents from the “833_doc” corpus, this phrase occurred 718 times.

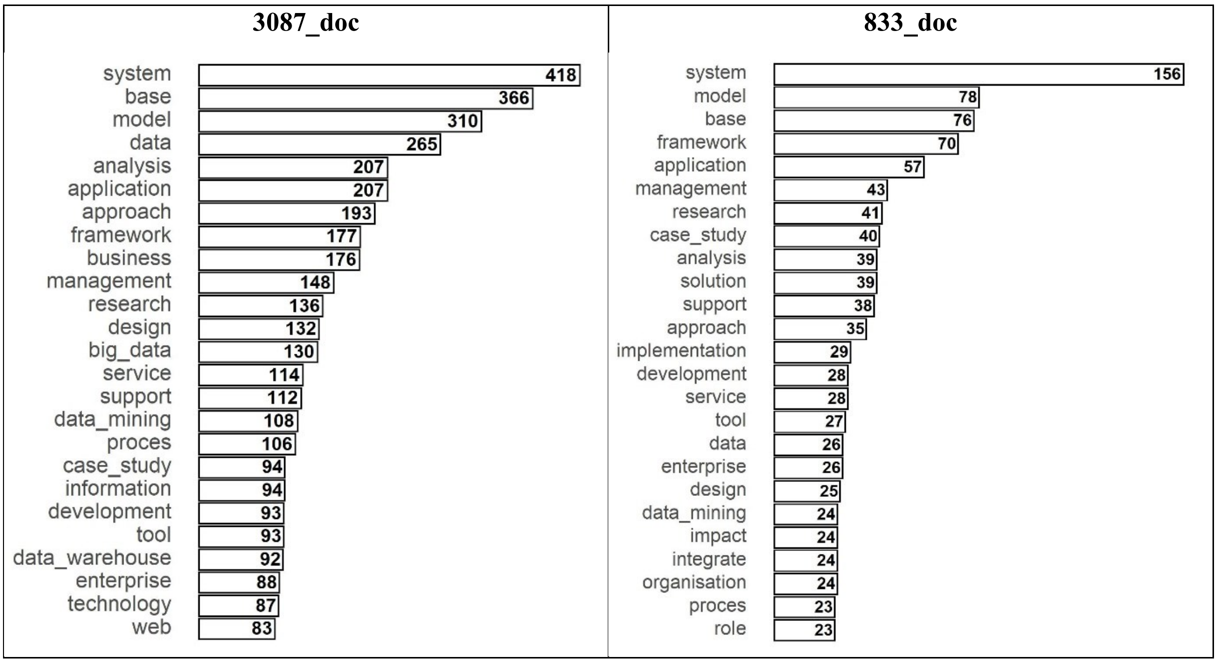

Figure 3. Most common words occurring in document titles. Sources: own elaboration.

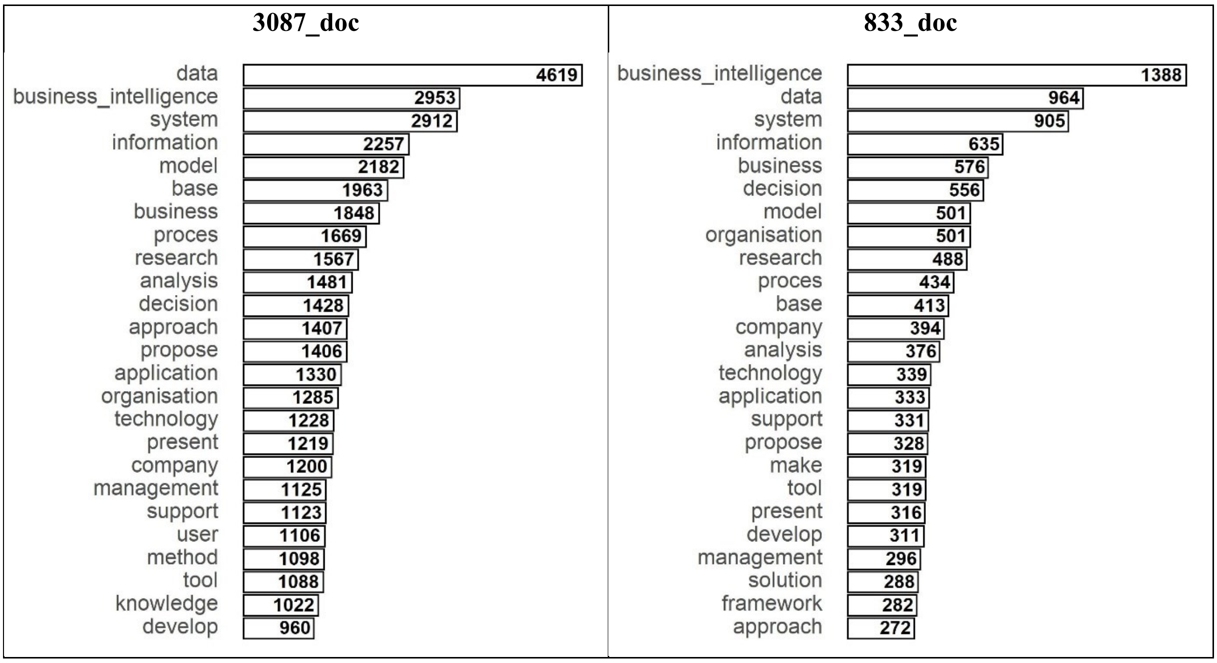

Figure 4. Most common words occurring in document abstracts. Sources: own elaboration.

Analysis of the frequency of words in the titles of the analysed documents also provides information about the research interests of the authors and about the BI systems themselves. For example, the words "model", "case study", "implementation", "impact" and "role" appear. On their basis, one can try to determine what the authors did, e.g. they examined the impact and role of BI systems and analysed selected case studies.

Figure 4 presents the most common words occurring in document abstracts in the form of charts. By analysing these words, one can also try to determine what is an important element of BI systems, as well as in what area and for what purpose are they used.

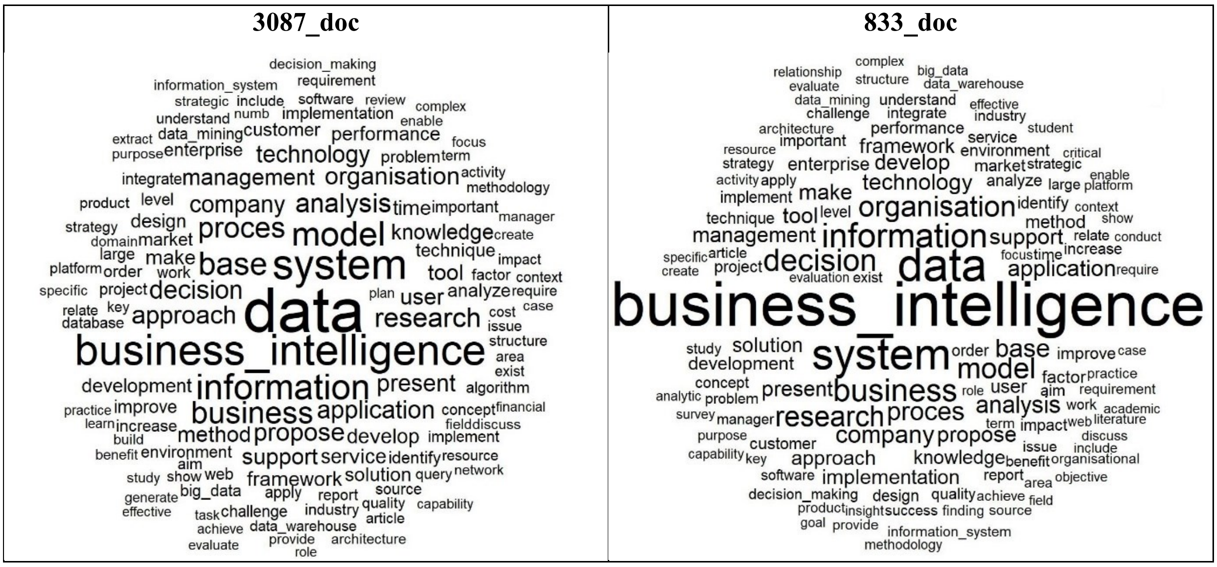

Figure 5. Word cloud of the most common words in document abstracts. Sources: own elaboration.

The most common words in abstracts are also presented in the form of a word cloud (Figure 5). In each of them, there were 125 words most frequently used by the authors. As can be seen, the most commonly used word in the abstracts of documents from the "3087_doc" corpus was the word "data", and for the "833_doc" corpus, it was the phrase "Business Intelligence".

Table 4.
CTM Model

Item	Topics
1	information business business_intelligence method application
2	data information analysis tool decision
3	business_intelligence system research process information
4	data business_intelligence system information user
5	business_intelligence organisation information framework research
6	model approach base business analysis
7	system business analysis propose organisation
8	data system base business_intelligence decision
9	model base research company method
10	data system process business application

Sources: own elaboration.

Using two topic modelling algorithms, abstract topics were generated describing the abstracts of the analysed documents. Topic models try to discover the hidden semantic structures represented by abstract topics within a collection of documents. The topic modelling algorithm assumes that each document is represented by a topic breakdown and that each topic is represented as a word breakdown. To generate the topics, the LDA algorithm (Latent Dirichlet Allocation) (Blei, Ng, & Jordan, 2003) and the CTM algorithm (Correlated Topic Model) were used (Blei, & Lafferty, 2005). Due to the volume of the work, only topics generated for documents from the "3087_doc" corpus were presented. The number of topics was predetermined at 10 and the number of words at 5. The topics may include n-grams treated as one word. Identified topics can also be used to identify information about Business

Intelligence systems. For example, from the words presented in the fourth topic (Table 5) generated by the LDA model, one can form the following sentence: "The analysed data from the database constitutes the source for the data warehouse".

Table 5.
LDA Model

Item	Topic
1	system management process tool implementation
2	model approach framework design propose
3	knowledge research development activity field
4	data database source analysis data_warehouse
5	research business_intelligence factor impact quality
6	time performance problem improve solution
7	information company customer market analyse
8	application service technology enterprise base
9	business decision organisation business_intelligence make
10	user method base technique web

Sources: own elaboration.

5. Conclusion

Analysing the number of publications issued in individual years, related to "Business Intelligence" systems, it can be concluded that this is still a current and trending topic.

Another conclusion arises after analysing the data from Table 1 showing the number of publications that had/did not have the phrase "Business Intelligence" in the fields: title, keywords, abstract. Before constructing a query to a database storing bibliometric data, it is worth considering whether we are interested in publications that have a given phrase in only one specific field, in any one of them or maybe in all of them at the same time.

Determining the most frequently used keywords and the most frequently used words in abstracts and article titles can be used to identify areas of research interest of authors and to determine the technologies and solutions that accompany Business Intelligence systems.

References

1. Addor, N., & Melsen, L.A. (2019). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. *Water Resources Research*, *55(1),* 378-390. https://doi.org/10.1029/2018WR022958.
2. Araujo, C. (2006). Bibliometria: Evolução histórica e questões atuais. *Em Questão, 12*, 11-32.

3. Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11(4),* 959-975. https://doi.org/10.1016/j.joi.2017.08.007
4. Aria, M., Cuccurullo, C., & Sarto, F. (2015). Exploring healthcare governance literature: Systematic review and paths for future research. *MECOSAN*, *23*, 61-80. https://doi.org/10.3280/MESA2014-091004.
5. Blei, D.M., & Lafferty, J.D. (2005). Correlated Topic Models. Proceedings of the 18th International Conference on Neural Information Processing Systems, 147-154. Retrieved from http://dl.acm.org/citation.cfm?id=2976248.2976267.
6. Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research, 3.* Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.
7. Briner, R.B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In D. Rousseau (Ed.), *The Oxford Handbook of Evidence-Based Management* (pp. 112-129). United Kingdom: Oxford University Press.
8. Cheng, L., & Cheng, P. (2011). *Integration: Knowledge Management and Business Intelligence,* 307-310. https://doi.org/10.1109/BIFE.2011.172.
9. Cuccurullo, C., Aria, M., & Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains. *Scientometrics*, *108(2),* 595-611. https://doi.org/10.1007/s11192-016-1948-8.
10. Ellegaard, O., & Wallin, J.A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105(3),* 1809-1831. https://doi.org/10.1007/s11192-015-1645-z.
11. Feinerer, I. (2008a). A Text Mining Framework in {R} and Its Applications (Department of Statistics and Mathematics, Vienna University of Economics and Business Administration). Retrieved from http://epub.wu-wien.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_e09.
12. Feinerer, I. (2008b). An Introduction to Text Mining. *R News, 8(2),* 19-22. Retrieved from http://cran.r-project.org/doc/Rnews/.
13. Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure. *Journal of Statistical Software, 25(5),* 1-54. Retrieved from http://www.jstatsoft.org/v25/i05.
14. Fink, L., Yogev, N., & Even, A. (2017). Business intelligence and organizational learning: An empirical investigation of value creation processes. *Information and Management*, *54(1),* 38-56. https://doi.org/10.1016/j.im.2016.03.009.
15. Hannula, M., & Pirttimäki, V. (2003). Business Intelligence Empirical Study on the top 50 Finnish Companies. *The Journal of American Academy of Business, Cambridge, 2(2),* 593-599.

16. Hornik, K., & Grün, B. (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*. https://doi.org/10.18637/jss.v040.i13.
17. Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: the definitive guide to dimensional modeling.* Wiley.
18. Luhn, H.P. (1958). A Business Intelligence System. *IBM J. Res. Dev., 2(4),* 314-319. https://doi.org/10.1147/rd.24.0314.
19. Negash, S., & Gray, P. (2008). Business Intelligence. *Handbook on Decision Support Systems, 2,* 175-193. https://doi.org/10.1007/978-3-540-48716-6_9.
20. Turban, E., Aronson, J.E., & Liang, T.-P. (2005). *Decision support systems and intelligent systems*. Pearson/Prentice Hall.
21. Wixom, B., & Watson, H. (2010). The BI-Based Organization. *International Journal of Business Intelligence Research*, *1(1),* 13-28. https://doi.org/10.4018/jbir.2010071702.
22. Yoon, T., Ghosh, B., & Jeong, B. (2014). User Acceptance of Business Intelligence (BI) Application: Technology, Individual Difference, Social Influence, and Situational Constraints. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 3758-3766. https://doi.org/10.1109/HICSS.2014.467.