

Semantic-based Clustering for Education-Science-Business Interaction Bibliometric Analysis

Oleksii Gorokhovatskyi¹, Nataliya Vnukova^{2,1}, Viktoriia Ostapenko¹ and Viktoriia Tyschenko¹

¹ Simon Kuznets Kharkiv National University of Economics, Nauki pr. 9-A, Kharkiv, 61064, Ukraine

² Scientific and Research Institute of Providing Legal Framework for the Innovative Development of the National Academy of Law Sciences of Ukraine, Chernyshevska st., 80, 61000, Kharkiv, Ukraine

Abstract

This paper presents the analysis of scientific publications on the interaction of education, science and business in the innovation economy on the basis of bibliometric software, sources from the Scopus scientometric database, supplemented by data visualization and descriptive analysis. The usage of clustering based on the word semantical similarity as well as clustering quality evaluation has been proposed to extend the data analysis opportunities in the scope of research topic evaluation. Different pretrained word embedding models were tested: GloVe, Word2Vec and transformers models. This allows us to evaluate the effective clustering quantity and extend the topic analysis using both the representation of our methods and known software (VOSViewer, Biblioshiny). It is shown also that performing the dimensionality reduction for this research is more effective before K-Means clustering than after it.

Keywords

Bibliometric software tools, Scopus, VOSviewer, Biblioshiny, innovative economy, education-science-business interaction, K-Means, word embeddings, pretrained models, clustering, clustering quality

1. Introduction

Despite the numerous studies on specific aspects of education, science and business etc., only a few have been published on the conceptual provisions of education-science-business interaction analysis in the innovation economy. To remedy this shortcoming, it is advisable to analyze the structure of publications on this topic using bibliometric analysis. Bibliometric analysis (bibliometrics) is the use of quantitative methods to study information resources. The disciplines related to bibliometrics are scientometrics and citation analysis, which deal with the quantitative analysis of all scientific achievements and scientific citations. Bibliometric analysis is a software for assessing the effectiveness of researchers, journals and institutions.

The relevance of bibliometric software is due to the fact that the modern wave of computer and information progress is transforming society and shaping the Internet generation, which makes it possible to highlight the possibility of solving any issue. To investigate the latest trends in research, it is advisable to conduct a bibliometric analysis, which is a quantitative statistical assessment of publications that is objective, rigorous, transparent and repeatable. Bibliometric research allows you to develop a unique perspective based on a fairly extensive analysis. Bibliometric technologies allow categorizing and analyzing large amounts of historical data obtained as a result of research conducted over a certain period in order to retrieve information from a repository. Bibliometric analyses rely on quantitative methods and therefore can avoid or mitigate bias, unlike systematic literature reviews, which usually rely on qualitative methods,

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

✉ oleksii.gorokhovatskyi@gmail.com (O. Gorokhovatskyi); vnn@hneu.net (N. Vnukova); viktoria.ostapenko@hneu.net (V. Ostapenko); vf_hneu@ukr.net (V. Tyshchenko)

ORCID 0000-0003-3477-2132 (O. Gorokhovatskyi); 0000-0001-9124-9511 (N. Vnukova); 0000-0002-4077-5738 (V. Ostapenko); 0000-0002-2530-185X (V. Tyshchenko)



© 2024 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

which can be tainted by interpretation bias from researchers with different academic backgrounds.

The paper is structured as follows. Section 2 shows the results of brief analysis of the related papers, the goals of the research and the contribution are presented at the end of the section. The analysis of typical software tools used in the bibliometric data processing is presented in Section 3. Section 4 contains the description of the dataset (including the details of its gathering), basic visualization results including known software tools, and the description of extended analysis of the topic using custom natural language processing and machine learning methods. Interpretation of the obtained results is presented in section 5. Finally, Section 6 contains the conclusions.

2. Related Works

This research is interdisciplinary and covers economic and computer sciences. There are many softwares for analyzing economic processes. Some of them include statistical data analysis software, market forecasting, economic scenario modelling software, and many others. They help economists and business analysts make better decisions based on data and analysis.

Scholars' consideration of education-science-business interaction analysis in the innovation economy has recently been updated through the use of bibliometric analysis. A systematic analysis of the scientific literature was conducted in order to identify progress in this area, the most fruitful contributions and promising trends for further research,

Study [1] conducted a bibliographic analysis using software that made it possible to identify such key factors as overall publication activity, the most productive and influential authors, journals, institutions and countries in the relevant field, and citation of publications through the analysis of joint citation, bibliographic linkage and common words. The analyzed approaches concerned various aspects of bibliometric analysis of large data sets from publications of powerful scientometric databases over many years, which allowed the authors to determine the list of methods that can be used to achieve the research goal [2].

The paper [3] presents the analysis of “innovation”, “ecosystem” and “development” keywords. According to the terms co-occurrence analysis the authors revealed 5 directions of future research: innovations in general, entrepreneurship and economic development, digital innovations and digitalization, sustainable development, smart environment.

The research productivity, impact, intellectual structure of global sustainable development goals has been analyzed in [4]. Authors investigated publications with “sustainable development goals”, “sustainable development”, “agenda”, and “United Nations” and found the inconsistency between visualization of author-provided keywords and keywords obtained from text analysis. Authors [5] also defined the current tendencies of research on social and environmental innovation as interdisciplinary that can be represented by specific stakeholders. Complex problems and creative solutions are addressed through the existence of collaborative networks between researchers and highlighted by the analysis of co-authorship networks that facilitate knowledge sharing, cross-diffusion of concepts and the creation of comprehensive solutions [6]. The scientometric productivity of countries, institutions, journals, and researchers in the field of stakeholder management research (Stakeholder” and “Management” terms have been analyzed) has been proposed also in [7].

The study [8] analyzed the list of more than 2500 journals of the Australian Council of Business Deans as a basis for business research and included an analysis of social media, which provided significant insights into the policy of aligning business research with development goals, especially in terms of formulating a system of interaction between them.

Investigation of sustainability and education has been proposed in [9] with the analysis of author' keywords in the publications according to the request “education for sustainable development” OR “education for sustainability”. Seven separate clusters were found that include “sustainable development”, “transformative learning”, “sustainable consumption”, “environment”, “case study”, “global citizenship education”, and “transformative education”

research topics. The research [10] defines the importance of education in sustainable development based on the use of bibliometric analysis, which reveals the characteristics of growth, research areas and methods, and also conducts a statistical analysis of the contributing forces of countries, institutions and authors, which proves the predominance of developed countries in the creation of scientific publications, on the basis of which promising topics for further research are formed.

The researchers [11] examined education as a competitive advantage of business based on content and bibliometric analysis, for which they conducted a bibliometric analysis of scientific publications indexed by the Scopus database using the Bibliometrix and VosViewer software products and the R Studio programming language. Following the results of the bibliographic analysis, four research clusters were formed, covering 10,914 keywords and 95,636 links were identified. The research [12] proposed the analysis of triple helix model and education as well as the usage of co-word analysis to predict the shape of the future research agenda in these topics. Two fields about helix and education have been analyzed by clustering of authors, citation and keywords, four clusters were identified for the latter. Future research topics have been proposed in the scope of each found cluster.

The authors [13] offer a bibliometric and visual analysis of 1747 scientific articles registered in the Scopus database using Vos Viewer and Biblioshiny, which identifies the current state of research, the most cited articles and authors and analysis of co-authorship, repetition and citation.

It is expected that the results of the reviewed studies will benefit researchers [14] by offering them insight into the current research landscape and serving as a valuable source for future research.

The majority of existing publications in the field of bibliometric analysis use known software tools to create network maps and data clustering. Our idea is to extend the pure visualization with other known technological ideas from natural language processing (NLP) [15] and machine learning (ML) [16] methods.

The goals of the research include:

- to investigate and visualize the education-science-business interaction in the innovation economy topic with known bibliometric software;
- to apply additional tools to get better insights about visualizations with clustering quality evaluation and performing the semantic word comparison instead of terms co-occurrences.

The contribution of the paper includes the application of known NLP and ML methods including semantic clustering of keywords and clustering quality evaluation for the extension of bibliometric visualizations.

3. Related Software

There are several main softwares that are predominantly used for analysing bibliometric analysis, each with its own strengths and weaknesses.

The sample of data for the study was formed from different databases: the Web of Science (WoS)[17], Elsevier Scopus [18], Dimensions, Cochrane Library, Lens i PubMed, etc. Each of them has unique properties and functions. Web of Science and Scopus are currently the most widely used international networks and integrated databases that allow scholars to study and evaluate publications, patents, reviews, and analytical documents. The selected publications [19] are analyzed by year of publication, country, title, publisher, open access level, funding agency, etc. it is advisable to analyze the thematic and citation statistics to conduct a comprehensive analysis of literature sources, according to the scientometric database Scopus, as it can provide a significant number of documents and offers more citation-rich data.

Bibliometric analysis can be performed using modern software. It provides a set of functions for searching, cleaning and analysing data, including bibliometric indicators, citation and common word networks, co-authorship analysis and journal impact factors. Futher Microsoft Excel, VOSviewer, Bibliometrix/Biblioshiny and NLP were used to analyse the publications.

Bibliometrix software (programming language R Studio) [21], VOSViewer [22] and another softwares [23] were based on the analysis of performance indicators. Technical information on VOSviewer and on the VOS mapping and clustering techniques is provided in the publications on the official cite [24]. The structure and characteristics of Biblioshiny, the possibilities of its use are presented in [25]. There is comparison of bibliometric software for education-science-business interaction analyze in the innovation economy (Table 1).

Table 1
Comparison of bibliometric software for education-science-business interaction analyze in the innovation economy

Criteria	Biblioshiny	VOSViewer
Scopus	The reference elements mentioned are not standardized, so they must be combined. In Dimensions, the algorithm that classifies search areas is not efficient	The data have to be exported in a CSV file and all data elements was included
Database		It is recommended to use the classic version of Web of Science, which offers more extensive possibilities for exporting data than the new version that was introduced recently
Web of Science	Plain text format is preferable in terms of data quality	
Workflow	Four levels of analysis (Source, Authors, Documents, Clustering). Three structures of knowledge (Conceptual, Intellectual, Social)	3 kinds of Visualization (Zooming and scrolling, Density and overlay visualizations, Screenshots) Techniques (Advanced layout and clustering techniques, Creating bibliometric networks)

Thus, Table 1 shows that Bibliometrix provides numerous tools that allow researchers to perform in-depth bibliometric analyses. One of these tools is a web-based application. Biblioshiny allows users without programming skills to perform bibliometric analysis using a graphical interface. The statistical software Biblioshiny should be used for semantic analysis of data to determine the frequency of simultaneous occurrence of keywords in scientific articles to simplify complex network relationships between keywords. VOSviewer is popular open-source software that are useful for creating visual maps and network diagrams of bibliometric data, while online platforms such as Google Scholar Metrics, Scopus Metrics and SciVal provide bibliometric analysis services. Researchers should choose the software that best suits their needs and research questions. Visualisation of scientific mapping, research clusters were formed and deep structures in the categorical data set were identified on thematic maps (niche topics; developing topics; declining topics; most common topics and main topics).

4. Application results of bibliometric software tool for education-science-business interaction analysis in the innovation economy

4.1. Dataset

At the first stage of the research, the search for Scopus-indexed scientific publications of different types was performed. We searched for such sources that contain terms "Education",

"Science", "Business", "Innovation" at the same time in the title, keywords, or abstract. As a result of this query, a set containing 1572 publications was obtained, which became the basis for further semantic analysis on the relevant subjects. The detailed description of the query and search results are shown in Table 2. The queries were further limited to original articles, conference abstracts, and books and book chapters. It is advised to exclude all those studies that included commentaries, editorials and letters, as well as articles or reviews that were published on preprint websites. Also, during the search the subject areas were limited to "Business, Management and Accounting", and "Economics, Econometrics and Finance". The data was uploaded on 11 February 2024.

Table 2
Structure and results of the bibliographic analysis of education-science-business interaction in the innovation economy

Criteria	Limitation	Result
Subject area, scope and coverage of publications with the eligibility criteria	Subject area Scope and coverage Search by parameters	«Education», «Science», «Business», «Innovation» Scientometric database: Scopus. Time period: 1984-2024 Search within: Article title, Abstract, Keywords
Information that has been received		All relevant information by articles n = 1572 Subject area:
Checking and selecting publications	Identify and verify records with filtering	Business, Management and Accounting, Economics, Econometrics and Finance Document type: Article, Conference paper, Book chapter, Book
For bibliometric research	Publications Citation	n = 405 n = 10705

According to the data received, the dynamics of publications for 1984-2024 was built (Fig. 1). As shown in Fig. 1, we can notice an upward trend in both the number of publications and citations. The research shows that over the past 30 years, the popularity of publications on this subject has increased, but a steady increase in its popularity began only in 2006 (Fig. 1), which is due to the relevant political and economic events in the country. This indicates a positive trend, growing interest in scientific community, and the relevance of the selected area for further research.

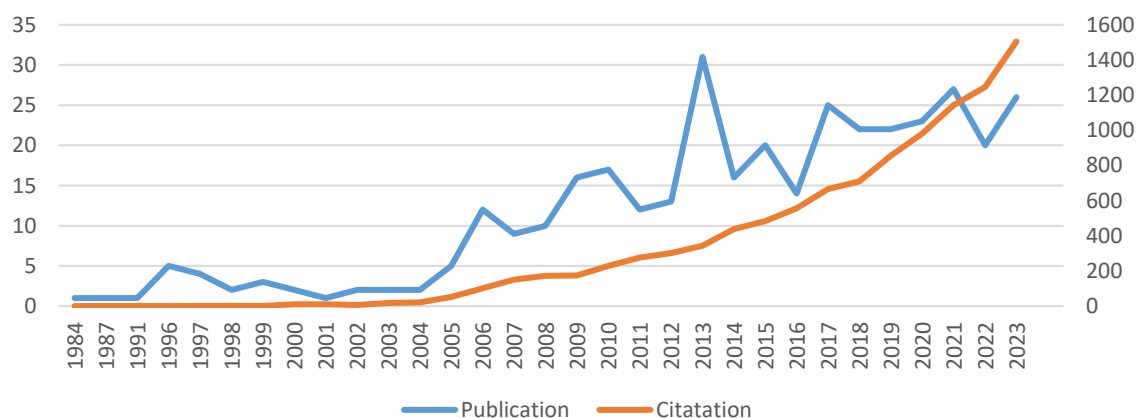


Figure 1: Trends in publications and citations in the Scopus database by search criteria and restrictions on education-science-business interaction in the innovation economy

According to the results of the research (Fig. 1), there is a strong increase in the number of publications and citations per year, with a peak in productivity in 2013. Applying concentration indices, we can further investigate that there is a corresponding concentration in production and citations in a particular area by region, institution and journal. In general, the upward trend in the number of citations allows us to assume their growth in the future and demonstrates the growing popularity of education, science and business interaction in the innovation economy for the scientific community.

However, it should be noted that an increase of citations is often associated with connections in scientific communities and the authority of certain countries and institutions in scientific research. The correlation between the growth of citations and the number of publications is not dependent, as the citation of a particular research paper can be extended over time and occur much later than the year of publication. Thus, the analysis of publications on a particular research topic proves its relevance, but does not provide extended information on the prospects for further research. In this context, it is recommended to conduct a more detailed bibliographic analysis using computer software.

4.2. Biblioshiny

A bibliographic database is a database of bibliographic records, an organized digital collection of references to published scientific literature, including journal articles, conference proceedings, patents, books, etc. They generally contain very rich subject descriptions in the form of keywords, subject classification terms, or abstracts. Information related to a bibliographic record are named bibliographic meta-dat [25]. The full set of bibliographic data was loaded from the Scopus database according to the previous stage of the study (Table 3).

Table 3
The main bibliometric data generated by the application Biblioshiny for education-science-business interaction in the innovation economy

Description	Results
Timespan	1984:2024
Authors	949
Author's Keywords	1244
Sources (Journals, Books, etc)	244
Authors of single-authored docs	120
References	14455
Documents	405
International co-authorships , %	17,04
Document Average Age	9,75
Co-Authors per Doc	2,43
Average citations per doc	28,86

Table 3 shows the main bibliometric data generated by bibliometric software Biblioshiny. In total, over the period 1984-2024 405 documents were found, among which 1244 keywords were selected. The average citation rate of the articles is 28.86, which confirms the significant attention that scientists around the world devote to the research issues. The considered studies are presented in 244 sources of various types. We can also mention the tendency to cooperate in conducting research in this area, as less than 30% of publications were written by a single author. On average, 3 authors were involved in each study, and international co-authorships accounted for 17%. According to the main bibliometric data obtained, it is advisable to analyse the structure of the most used keywords, words in titles and geographical concentration of research authors.

4.2.1. Tripod chart (Sankey chart)

Biblioshiny makes it easy to see the data flow by creating Sankey diagrams. Fig. 2 shows the tripod chart (Sankey chart) by country (AU_CO), keywords (DE) and terms in titles of publications (TI_TM) to reflect the proportion of research areas on education, science and business interaction in innovation economy for each country and the dynamics of publications. First 20 terms, countries and article titles were used for plotting.

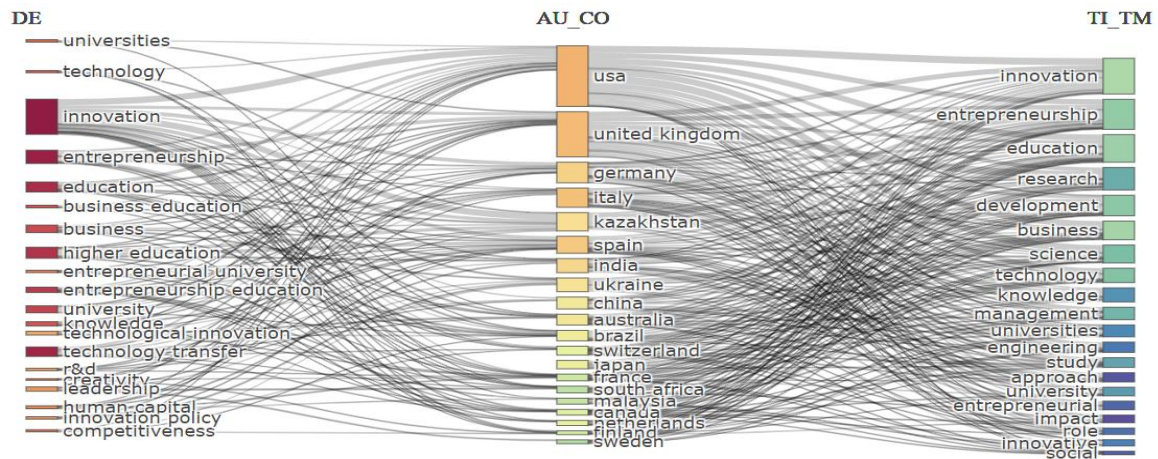


Figure 2: The relationship between article titles, keywords and countries of publications revealed in researches related to education-science-business interaction in the innovation economy

As one can see from Fig. 2, there is a concentration of publications on education-science-business interaction in innovation economy in the USA (78 items), the UK (46 items), Germany (21 items), Italy (17 items), Spain (16 items), Ukraine (9 items). It is essential to add that these countries form the vast majority of international co-authorships, as noted in Table 3. Ukraine ranks 8th in most publications in the research area. As for the keywords, the most commonly used are innovation, entrepreneurship, business, education, high education, university, knowledge and technology (transfer), and the titles of publications focus on science, research, development, management, etc. The analysis reflects the objective results of the search according to the query, i.e. it finds scientific publications by synonymous series of the requested terms, which also does not reveal certain policy areas and prospects for further research in this area.

4.2.2. Factor analysis

The basic idea behind factorial approaches is to reduce the dimensionality of data and represent it in a low-dimensionality space. Three alternative methodologies: Correspondence Analysis (CA), Multiple Correspondence Analysis (MCA), Multidimensional Scaling (MDS). The proximity between words corresponds to shared-substance: keywords are close to each other because a large proportion of articles treat them together; they are distant from each other when only a small fraction of articles uses these words together. The origin of the map represents the average position of all column profiles and therefore represents the center of the research field (meaning common and large shared topics) [25]. Extraction and presentation the most relevant information in the data set, using the Factorial Map tool in R show in Fig. 3.

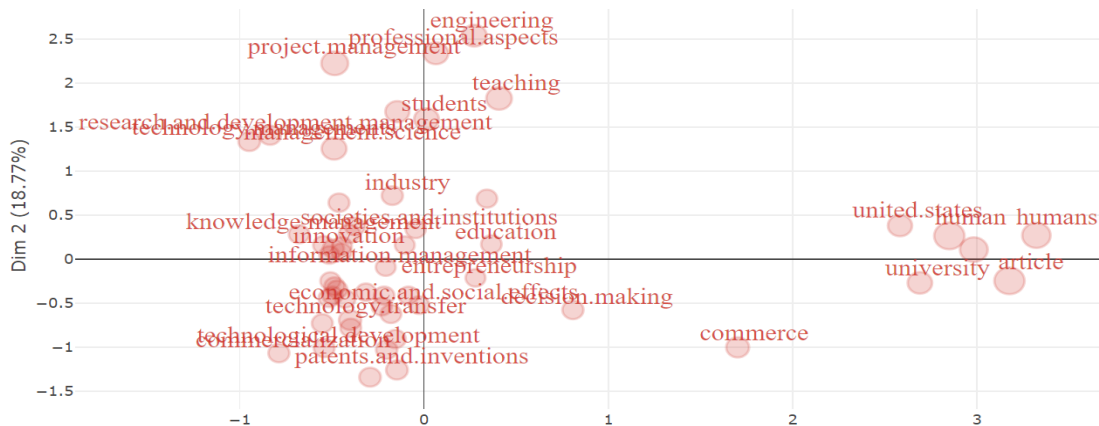


Figure 3: Conceptual structural map of factor analysis using multidimensional scaling, which integrates and correlates keywords of research on education-science-business interaction in the innovation economy

To interpret the results, the relative position and distribution of points along the measurements is used. The closer the words shown in Fig. 3, the more comparable their distribution is. The terms that are located closer to the center of the map are more common in this study and have received more attention during the analyzed period [25]. And those terms that are more evenly distributed are associated with less discussed research subjects. Thus, based on the results of the conceptual structural map, it can be concluded that such keywords as education-science-business interaction in the innovation economy are located close to the center, which indicates the most discussed topics of publications for the period under study and proves the relevance of the chosen research area [25]. Thus, according to the results of the conceptual structural map, it can be concluded that such keywords as education, research, economics and innovation are located close to the centre, which indicates their general use, but does not allow for a qualitative analysis of a large number of bibliographic sources, which is an objective necessity in today's realities.

Therefore, in order to deepen the quantitative characteristics of the bibliographic analysis and interpret it with qualitative conclusions, it is advisable to conduct a cluster analysis with the possibility of forming certain groups according to the subject of the research and focusing the attention of scientists on certain aspects of education-science-business interaction in the innovation economy.

4.3. VOSViewer

More detailed semantic analysis and visualization of the key areas of multidisciplinary research on the education-science-business interaction in the innovation economy can be achieved by using the VOSviewer tool and building visualization maps based on the results of search queries. Science maps use knowledge structures and describe the structural and dynamic elements of a research field. In this study, they were used to provide a comprehensive overview of significant trends and research findings, in the form of conceptual structures that identified major themes, directions and intellectual structures that classified how the author's work has influenced this research community (Fig. 4).

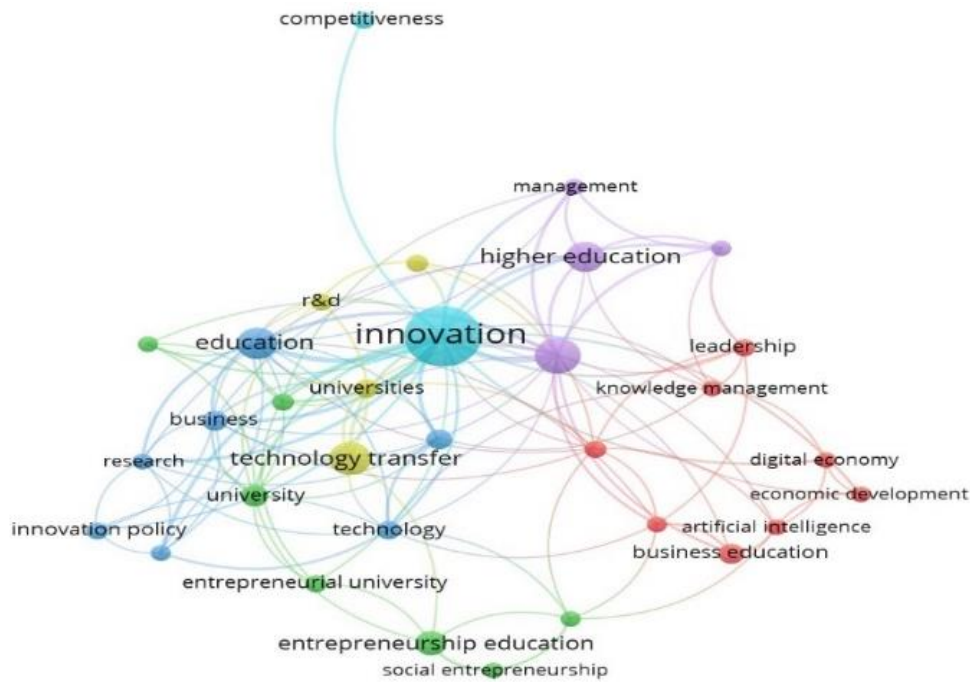


Figure 4: Visualization map of keywords in current research on education-science-business interaction in the innovation economy (different colors indicate different clusters, the size of the label indicates how often the keyword occurs, created with VOSViewer)

The building of the maps is based on the co-occurrences of the keywords that can be used for the analysis of relations between clusters and separate terms in clusters [26]. As one can see in the Fig. 4 all 32 terms which occurred at least 5 times in the keywords were grouped into 6 clusters with the most frequent term “innovation” that relates to items in other clusters. Some other interesting insights from this map include: the terms “business school” and “business” are located in different clusters and don’t have direct relation.

This methodology provides researchers with a framework for each subject cluster that can be used to limit the research related to a particular issue. The main goal was to recognise and identify relevant subjects. Themes are groups of keywords whose density and centrality can be used to organise them into a single circle and map them as a two-dimensional image. Emerging or disappearing issues are located in the lower left quadrant (green), and highly specialised/niche issues are located in the upper left quadrant (blue) (Table 4).

Table 4
Lists of clustered keywords for VOSviewer

Software	Cluster ID and keywords
VOSViewer	1: artificial intelligent, business education, COVID-19, digital economy, economic development, knowledge management, leadership, technological innovation
	2: business school, creativity, entrepreneurial university, entrepreneurship education, social entrepreneurship, social education, university
	3: business, education, innovation policy, knowledge, research, science, technology
	4: human capital, R&D, technology transfer, university
	5: entrepreneurship, high education, management
	6: competitiveness, innovation

From the obtained data presented in Table 4, we can offer some separate aspects deep research on education-science-business interaction in the innovation economy.

The keywords of Cluster 1 describe the of education-science-business interaction as a prerequisite and priority area for increasing the level of scientific and technological development and transformation of Ukraine's innovation economy. Cluster 2 defines an entrepreneurial or innovation-active university as a tool education-science-business interaction in the innovation economy. Cluster 3 focuses on the harmonisation of Ukraine's institutional framework with global trends and the implementation of innovative principles, approaches, and practices of education-science-business interaction in the innovation economy. The keywords of Cluster 4 define the university as a centre of human capital and a framework for R&D and technology transfer. The positioning, structuring and provision, as well as the identification of areas for managing education-science-business interaction in the innovation economy are considered by the researchers of Cluster 5. Representatives of Cluster 6 substantiate the impact of education-science-business interaction as the basis of the innovation economy on increasing competitiveness and sustainable development. It is possible to see a general idea of the direction in which further research in this field will develop and to form a road map based on these trends.

However, we will conduct further research to compare the results of using different computer software, in particular NLP.

4.4. Custom semantic clustering

Despite charts from different bibliometric software are bright and handy most of them still can show only the entire picture without significant details for the topic being investigated and flexible tuning of the parameters being used to create charts. For instance, it is interesting whether the quantity of clusters in Fig. 4 is successful enough. The usage of some explicit natural language processing and machine learning methods could be applied in order to get more flexible results with required detailing level.

We use some methods to process the Abstract field of the dataset in order to understand the landscape of keywords without taking into account the direct relation between them in terms of co-occurrence but considering the semantic similarity only.

The traditional NLP text processing routines include:

- text cleaning and removing unnecessary symbols;
- split of text to tokens (depending on separation symbols in text and the task being solved);
- stemming/lemmatization that allows to normalize the structure of the token;
- building of token embeddings;
- processing of embeddings accordingly to the problem (comparison, classification, training of neural networks, etc.).

4.4.1. Preparation

The content of abstract for each source in the dataset was tokenized to words, stop words from the NLTK English list [27] were removed. After that all terms were merged into a single list and count of occurrences for each term was calculated. The additional filtration according to the frequency of words was applied, so only words with the quantity bigger than 100 were used. There are 58560 words in the dictionary (8503 unique ones), 60 terms repeat with the quantity over 100.

4.4.2. Embeddings

The method to select or build embeddings is often the most important step in NLP pipeline as all calculations and comparisons are based on the quality of numerical representations for words or tokens. In this paper, we used such pretrained models, that allow to put the term as input and receive its numerical representation immediately:

- Gensim GloVe models [28, 29] pretrained on Wikipedia 2014 dump and Gigaword 5 datasets (6B tokens in total) named “glove-wiki-gigaword-50” and “glove-wiki-gigaword-300” [30, 31];
- Gensim Word2Vec models trained on “text8” dataset (first 100 million bytes of plain text from Wikipedia [30]) to produce word embeddings having 100 and 200 numbers;
- sentence transformers “all-mpnet-base-v2” [32] and “all-MiniLM-L6-v2” that represents sentence as vector having 768 and 384 numbers respectively [33, 34].

4.4.3. Clustering and dimensionality reduction

The problem we are trying to solve here is the analysis of clusters of keywords. There are a lot of clustering methods and we chose one of the simplest – K-means, that requires the effective quantity of clusters to be known beforehand. If the quantity of clusters is unknown it could be evaluated with elbow method [35, 36] or other clustering quality index like silhouette [37, 38] or Davies-Bouldin index [39].

Data visualization is difficult in multidimensional spaces, and vector embeddings are representatives of these spaces. So, the dimensionality reduction is required to show clusters of definitions we are researching. We used Principal Component Analysis (PCA) as a well-known method to reduce the dimension of embedding vectors.

The interesting question about the dimensionality reduction is whether to apply it before the clustering, or perform the clustering over full embedding vectors using all their representation power firstly and apply reduction only after that, e.g., for visualization purposes only.

The results of full word embeddings (for “glove-wiki-gigaword-50” model) clustering for different quantity of clusters and corresponding quality indices are shown in the left part of the Fig. 5. As one can see, elbow method is smooth and the effective quantity of clusters is unclear, the same is true for the curve built for Davies-Bouldin scores. Plot of silhouette indices seems to be the most interesting but the maximum value is only 0.1 for ten clusters that means that clustering is bad. We refer to these results in evaluation of the effective quantity of clusters as uncertain, because the maximum value for silhouette index corresponds to ten clusters, while the minimum value for Davies-Bouldin scores refers to 13 or 14 clusters, and there is no joint decision between these approaches.

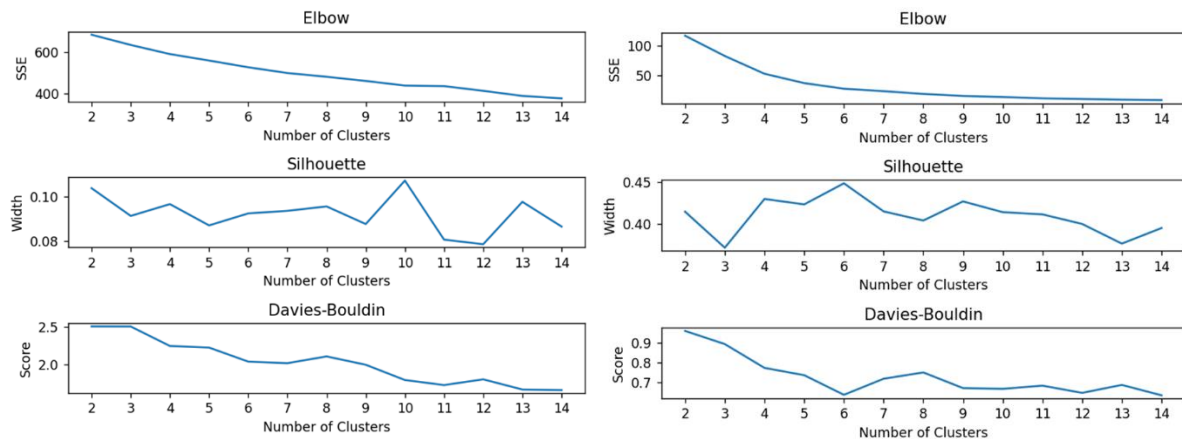


Figure 5: Visualization of full word embeddings clustering quality indices for different quantity of clusters (left) and visualization of clustering quality indices for different quantity of clusters after reducing dimensionality of embeddings to 2D (right) (both for “glove-wiki-gigaword-50” model)

The results after reducing the dimensionality of word embeddings to two-dimensional are shown in the right part of the Fig. 5. Elbow method is still smooth but both Davies-Bouldin and silhouette indices show the effective quantity of clusters to be 6 and maximum silhouette value

is about 0.45 but still is not very good though. Similar situation occurs for the second “all-mpnet-base-v2” model we tested clustering quality indexes for.

The results about the quality of clustering and evaluation of proper quantity of clusters for different word embedding models are shown in Table 5 with the best values highlighted in bold: maximal value for silhouette index and the minimal one for Davies-Bouldin score.

Table 5
Quantity of clusters and clustering indices for different embedding models

Model name	Quantity of clusters	Silhouette index	Davies-Bouldin score
all-mpnet-base-v2	3	0.51	0.62
all-MiniLM-L6-v2	3	0.52	0.62
glove-wiki-gigaword-50	6	0.45	0.64
Text-8 (length of embeddings is 100)	uncertain	N/A	N/A
Text-8 (length of embeddings is 200)	9	0.45	0.58
glove-wiki-gigaword-300	4 (uncertain)	0.53	0.64

The quantity of clusters was defined as a result of same decision for curves built on both indices, e.g., if the quantity of clusters is three – it means that silhouette curve reached maximum at this quantity and Davies-Bouldin score reached minimum at the same time.

As one can see, the most powerful and recent word embedding models based on transformers (“all-mpnet-base-v2” and “all-MiniLM-L6-v2”) both found three clusters with pretty the same clustering indices. The analysis of silhouette and Davied-Bouldin values allowed us to highlight from 4 to 9 clusters for other models. The curves based on the Word2Vec model for “text-8” dataset (the case with 100 values in embedding) shows inconsistent effective quantity of clusters, and somewhat partially consistent for “glove-wiki-gigaword-300” model.

Clustering results for transformers-based models (with cluster centers marked with black crosses) are shown in Fig. 6, both views contain three clusters. The text results of keywords clustering for all models are shown in Table 5.

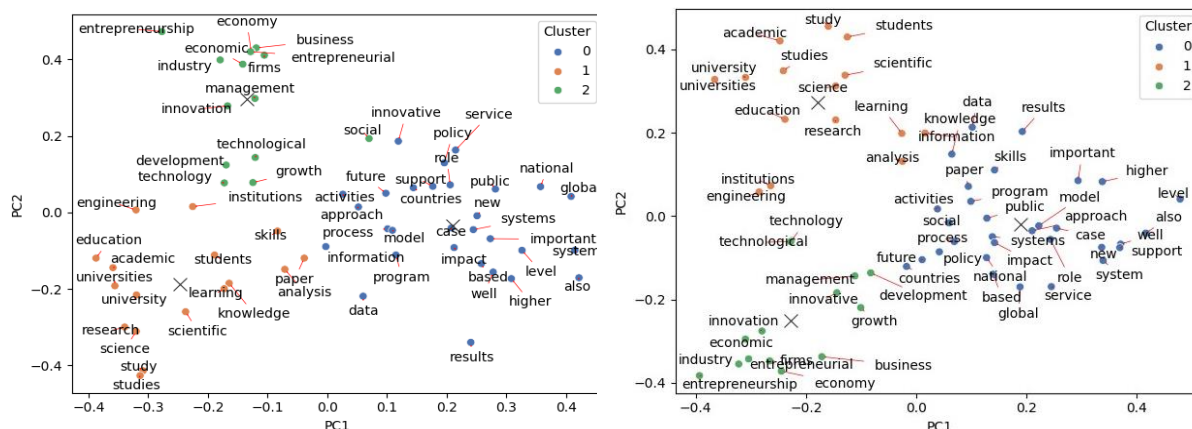


Figure 6: Clustering results for “all-mpnet-base-v2” model (left) and “all-MiniLM-L6-v2” (right), visualization improved with [40]

Fig. 6 shows clustering results for “all-mpnet-base-v2” model (left) and “all-MiniLM-L6-v2” (right), which allowed us to build 3 clusters each. The content of the clusters is similar in terms of keywords and proves certain interconnections and subtopics of research within the topic. Lists of clustered keywords for different embedding models are presented in Table 6.

Table 6
Lists of clustered keywords for different embedding models

Model name	Cluster ID and keywords
all-mpnet-base-v2	<p>0: level, new, systems, activities, role, also, future, innovative, impact, higher, system, based, data, results, national, important, global, approach, public, well, countries, policy, support, case, process, information, model, program, service</p> <p>1: research, paper, study, university, analysis, education, science, studies, students, knowledge, learning, engineering, skills, universities, academic, scientific, institutions</p> <p>2: business, social, economy, entrepreneurship, entrepreneurial, innovation, technology, development, economic, industry, management, technological, firms, growth</p>
all-MiniLM-L6-v2	<p>0: paper, level, new, social, systems, activities, role, also, knowledge, future, skills, impact, higher, system, based, data, results, national, important, global, approach, public, well, countries, policy, support, case, process, model, program, service</p> <p>1: research, study, university, analysis, education, science, studies, students, learning, engineering, universities, academic, scientific, institutions, information</p> <p>2: business, economy, entrepreneurship, entrepreneurial, innovation, technology, development, innovative, economic, industry, management, technological, firms, growth</p>
glove-wiki-gigaword-50	<p>0: business, social, systems, activities, role, technology, development, important, approach, management, institutions, model, program</p> <p>1: research, study, university, education, science, studies, students, knowledge, learning, engineering, universities, academic, scientific</p> <p>2: entrepreneurship, entrepreneurial, innovation, skills, innovative, technological</p> <p>3: economy, economic, global, growth</p> <p>4: paper, level, new, analysis, also, system, based, data, results, national, public, well, case, information, service</p> <p>5: future, impact, higher, industry, countries, policy, support, firms, process</p>
Text-8 (length of embeddings is 200)	<p>0: systems, role, impact, system, case, process, model</p> <p>1: national, public, institutions</p> <p>2: research, study, science, studies, learning, engineering, academic, scientific</p> <p>3: level, also, entrepreneurship, entrepreneurial, innovation, technology, future, development, skills, innovative, higher, important, global, well, management, technological, program</p> <p>4: level, also, entrepreneurship, entrepreneurial, innovation, technology, future, development, skills, innovative, higher, important, global, well, management, technological, program</p> <p>5: paper, analysis, knowledge, based, data, results, approach, information</p> <p>6: economy, economic, countries</p> <p>7: education, students, universities</p> <p>8: new, business, social, activities, industry, policy, support, firms, growth, service</p>
glove-wiki-gigaword-300	<p>0: paper, level, new, social, systems, activities, analysis, role, also, future, impact, higher, system, based, data, results, national, important, approach, public, well, management, countries, institutions, policy, support, case, process, information, model, program, service</p> <p>1: research, study, university, education, science, studies, students, knowledge, learning, engineering, universities, academic, scientific</p> <p>2: entrepreneurship, entrepreneurial, innovation, technology, skills, innovative, technological</p> <p>3: business, economy, development, economic, global, industry, firms, growth</p>

The analysis of semantic similarity of words from abstracts compared to the analysis of word co-occurrence (Table 4) allows us to propose some additional ideas. Clustering obtained from

word co-occurrence now seems to be somewhat too detailed as it contains both clusters from 2-3 words and clusters containing 7-8 terms with dense visualization of them (Figure 4). We can see that "university", "research", "higher education", "business education" are located in different clusters, which confirms the results obtained in section 4.2, but only provides a quantitative assessment of the bibliographic analysis of interaction in the innovation economy. As one can see from Fig. 6, there are separate educational cluster, economics/business/innovation cluster, and cluster with common words. In this case, the clusters are formed according to the keywords of the primary query and do not allow for a deep qualitative semantic analysis of the topic. Additionally, we have some numerical measurement of the quality of such clustering. Probably, the best choice is to clear common words and create only two clusters from this data.

5. Discussions

Bibliometric analysis is increasingly being used to assess quantitative and qualitative aspects of research trends and findings in a particular field, as well as to identify future research directions for scholars, policy makers, institutions and funding agencies.

The results of using bibliometric software tool of publications in the international environment are obtaining of education-science-business interaction in the innovation economy. This research is focused on identifying trends in research areas, countries, authors, and citations of publications in the Scopus scientometric database. Generalization of patterns in research related to the identified topics and the findings will provide valuable information on future research paths in a rapidly developing field, focusing on opportunities for future research.

The main bibliometric software, used in the research were Microsoft Excel, VOSviewer, Bibliometrix/Biblioshiny and NLP, which were used to create visualisation maps based on keywords and additional information from the Scopus scientometric database. The presented keyword-based visualisations integrate and correlate the knowledge of current research on the education-science-business interaction in the innovation economy.

Additionally, we performed and implemented the semantical clustering of keywords using different pretrained word embedding GloVe, Word2Vec and transformers models. This allows us to evaluate the effective clustering quantity and extend the topic analysis using both the representation of our methods and known software (VOSViewer, Biblioshiny).

It is shown that performing the dimensionality reduction for this research is more effective before clustering than after it.

The analysis of visualizations allowed us to form some insights about our topic of interest, e.g.:

- our research, as well as previous ones, shows that the bibliometric methodology and different databases can help researchers overcome the problems of managing large amounts of bibliometric data and implement retrospective and prospective analysis on a particular research area. The formed framework will allow to select a subset of features from a huge data set, and its results will allow to make grounded decisions in accordance with the request with bibliometric software;
- bibliometric data from scientific databases such as Scopus and Web of Science are not created exclusively for bibliometric analysis, and therefore may contain errors that affect the results of the analysis. In addition, the developers of the considered software pointed out the disadvantages of using certain databases, which should be taken into account in the research process. All this necessitates the critical formation of bibliometric data that will be used in a more detailed study. Multiple authors in their studies compare the bibliographic analysis of different databases, including Scopus and Web of Science. The developers of bibliometric software indicate that they recommend using, for example, Web of Science for VOSviewer, which offers more extensive possibilities for exporting data than Scopus, which exports the data in a CSV file and that is why it is necessary to make sure that all data elements should be included. But we believe that this is an advantage, as it allows you to make a selection according to certain criteria and restrictions that meet your needs;

- qualitative statements of bibliometrics are based on quantitative methods of bibliometric analysis, and the relationship between quantitative and qualitative results is often unclear and can be quite subjective. However, unlike our study, other researchers did not consider the possibility of conducting an analysis with different bibliometric software on the basis of the generated database. Also, most researchers aimed to offer a forecast of development in the research area based on a qualitative bibliometric result. Our study, unlike the previous ones, compares clustering within bibliographic analysis and provides recommendations on the possibility of conducting qualitative analysis.

6. Conclusions

The visualization, clusters for different embedding models and other results of the bibliometric analysis using a wide range of methodological and software made it possible to see the existing interrelationships of interdisciplinary research, their intersection points and development alternatives. Based on the results presented here, it is possible to develop further analysis on the chosen issue in regard to individual aspects, using software for bibliographic, citation and co-author analysis, which complement the meta-analysis and qualitative structuring in the original research.

The following results of the bibliometric analysis for education-science-business interaction in the innovation economy should be noted:

- The primary query was formed using the keywords "Education", "Science", "Business", "Innovation". The study was limited to open access articles published in English in the two fields of Business, Management and Accounting and Economics, Econometrics and Finance in the Scopus database, by a certain type of publication (article, conference presentation, book chapter, book). As a result, 405 publications for the period 1984-2024 with more than 10 thousand citations were compiled. A steady increase in the number of scientific articles since 2016 has been revealed, which indicates a growing interest of researchers in the topic of education-science-business interaction in the innovation economy.
- The research was conducted using various software. Quantitative information allows us to identify current trends and characteristics of research within the research topic. The quantitative analysis identified 949 authors, including 120 single-authored papers, while other studies were conducted in collaboration with more than two authors. International co-authorships made up 17%, and are represented by authors from such countries as the USA, the UK, Germany, Italy, Spain, and Ukraine. The geographical features of influence in the study of education-science-business interaction in the innovation economy are determined. These publications are presented in 244 Sources (Journals, Books, etc.), and the average citations is almost 29 per doc.
- Qualitative analysis, based on quantitative indicators, a systematic literature review, allows to reveal interrelationships and promising trends in the development of research. However, additional methods and approaches are needed for analysis of a large database to identify specific areas of research on interaction in the innovation economy. Using one sample of publications, each software forms special relationships between keywords and builds a unique clustering (Biblioshiny, VOSviewer and NLP).
- Clustering with Biblioshiny and NLP forms mainly 1-3 clusters, in which keywords are formed mainly by primary queries and reflect synonymous or associative keyword series. This proves that the vast majority of publications are on separate topics of education, science and business, rather than on their interaction in the innovation economy. VOSviewer has formed 6 clusters, the keywords of which can be used to identify promising areas for further research on education-science-business interaction in the innovation economy.

References

- [1] C. Yang, Q. Xiu, Bibliometric Review of Education for Sustainable Development, 1992–2022. *Sustainability* 15(14) (2023) 10823. doi:10.3390/su151410823.
- [2] Z.-H. Sun, T.-Y. Zuo, D. Liang, X. Ming, Z. Chen, S. Qiu, GPHC: A heuristic clustering method to customer segmentation. *Applied Soft Computing* 111 (2021). doi:10.1016/j.asoc.2021.107677.
- [3] L. Pilelienė, G. Jucevičius, Decade of Innovation Ecosystem Development: Bibliometric Review of Scopus Database. *Sustainability* 15(23) (2023) 16386. doi:10.3390/su152316386.
- [4] N. Alfirević, P. L. Malešević, K. M. Mihaljević, Productivity and Impact of Sustainable Development Goals (SDGs)-Related Academic Research: A Bibliometric Analysis. *Sustainability* 15(9) (2023) 7434. doi:10.3390/su15097434.
- [5] R. Raman, H. Lathabhai, S. Mandal, C. Kumar, P. Nedungadi, Contribution of business research to sustainable development goals: bibliometrics and science mapping analysis. *Sustainability* 15(17) (2023) 12982. doi:10.3390/su151712982.
- [6] S. A. Alanazi, M. Alruwaili, F. Ahmad, A. Alaerjan, N. Alshammari, Estimation of organizational competitiveness by a hybrid of one-dimensional convolutional neural networks and self-organizing maps using physiological signals for emotional analysis of employees. *Sensors* 21(11) (2021) 3760. doi:10.3390/s21113760.
- [7] S. Esparza-Rodríguez, T. G. Garcia, C. I. Rivas, Stakeholder management: a bibliometric analysis to understand the evolution of the research field. *Biblios Journal of Librarianship and Information Science* 84 (2022) 32–59. doi:10.5195/biblios.2022.1026.
- [8] D. Palupiningtyas, M. G. Sono, R. Pahrijal, Bibliometric Analysis of Social and Environmental Innovation Research Developments: Trend Identification, Key Concepts, and Collaboration in the Scientific Literature. *West Science Business and Management* 1(04) (2023) 245–254. doi:10.58812/wsbm.v1i04.247.
- [9] A. T. Gorski, E. D. Ranf, D. Badea, E. E. Halmaghi, H. Gorski, Education for sustainability - Some bibliometric insights. *Sustainability* 15(20) (2023) 14916. doi:10.3390/su152014916.
- [10] M. Aria, C. Cuccurullo, Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics* 11(4) (2017) 959–975. doi:10.1016/j.joi.2017.08.007.
- [11] D. Kharchenko, Content and Bibliometric Analysis of Education as a Competitive Advantage of Business. *Business Ethics and Leadership*, 7(2) (2023) 99–108. doi:10.21272/bel.7(2).99-108.2023.
- [12] H. Zakaria, D. Kamarudin, M. A. Fauzi, W. Wider, Mapping the helix model of innovation influence on education: A bibliometric review. *Frontiers in Education* 8 (2023). doi:10.3389/feduc.2023.1142502.
- [13] M. Sodom, N. Zulaikha, M. Yusoff, Unveiling the landscape of social media marketing in social science studies: A bibliometric analysis using VosViewer and Biblioshiny. *International Journal of Innovation and Business Strategy* (2023). doi:10.11113/ijibs.v18.146.
- [14] R. Tomaszewski, Visibility, impact, and applications of bibliometric software tools through citation analysis. *Scientometrics* 128 (2023) 4007–4028. doi:10.1007/s11192-023-04725-2.
- [15] S. Hosseini, H. Baziyad, R. Norouzi, Mapping the intellectual structure of GIS-T field (2008–2019): a dynamic co-word analysis. *Scientometrics* 126 (2021) 2667–2688. doi:10.1007/s11192-020-03840-8.
- [16] J. Baas, M. Schotten, A. Plume, G. Côté, R. Karimi, Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies *Quant. Sci. Stud.* 1 (2020) 377–386. doi:10.1162/qss_a_00019.
- [17] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W.M. Lim, How to conduct a bibliometric analysis: An overview and guidelines. *J. Bus. Res.* 133 (2021) 285–296. doi:10.1016/j.jbusres.2021.04.070.
- [18] G. D. Dzhunushalieva, R. Teuber, A bibliometric analysis of trends in the relationship between innovation and food. *British Food Journal. Sustainability* 15(23) (2023) 16386. doi:10.3390/su152316386.

- [19] L. Farinha, J.R. Sebastião, C. Sampaio, J. Lopes, Social innovation and social entrepreneurship: Discovering origins, exploring current and future trends. *Int. Rev. Public Nonprofit Mark.* 17 (2020) 77–96. doi:10.1007/s12208-020-00243-6.
- [20] A. T. Rosário, R. Raimundo, Sustainable Entrepreneurship Education: A Systematic Bibliometric Literature Review. *Sustainability* 16(2) (2024) 784. doi: 10.3390/su16020784.
- [21] G. Aparicio, T. Iturralde, A. Maseda, Conceptual structure and perspectives on entrepreneurship education research: A bibliometric review. *European Research on Management and Business Economics* 25(3) (2019) 105–113. doi:10.1016/j.iemeen.2019.04.003.
- [22] N. Koutsoupias, K. Mikelis, Text, Content and Data Analysis of Journal Articles: The Field of International Relations, in: *Proceedings of the Conference of the International Federation of Classification Societies*, Thessaloniki, Greece, 26–29 August 2019, pp. 113–120. doi:10.1007/978-3-030-60104-1_13.
- [23] N.C. Nelson, K. Ichikawa, J. Chung, M.M. Malik, Mapping the discursive dimensions of the reproducibility crisis: A mixed methods analysis. *PLoS ONE* 16(7) (2021). doi:10.1371/journal.pone.0254090.
- [24] VOSviewer, 2024. URL: <https://www.vosviewer.com/publications>
- [25] Bibliometrix, 2024. URL: <https://bibliometrix.org/biblioshiny/biblioshiny1.html>
- [26] M. Dominko, K. Primc, R. Slabe-Erkeret, A bibliometric analysis of circular economy in the fields of business and economics: towards more action-oriented research. *Environ Dev Sustain* 25 (2023) 5797–5830. doi:10.1007/s10668-022-02347-x.
- [27] List All English Stop Words in NLTK – NLTK Tutorial, 2019. URL: <https://www.tutorialexample.com/list-all-english-stop-words-in-nltk-nltk-tutorial/>
- [28] Gensim – Topic Modelling in Python, 2022. URL: <https://github.com/RaRe-Technologies/gensim>
- [29] Q. Le, T. Mikolov, Distributed Representations of Sentences and Documents, in: *Proceedings of the 31st International Conference on Machine Learning* (2014), pp. 1188–1196.
- [30] What is Gensim-data for? 2024. URL: <https://github.com/piskvorky/gensim-data>
- [31] J. Pennington, R. Socher, C. Manning, A. Moschitti, B. Pang, W. Daelemans, GloVe: Global Vectors for Word Representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.
- [32] all-mpnet-base-v2, 2024. URL: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>
- [33] Pretrained models, 2024. URL: https://www.sbert.net/docs/pretrained_models.html
- [34] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019. doi:10.18653/v1/D19-1410.
- [35] M. Syakur, B. Khotimah, E. Rochman, B. Satoto, Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster, in: *The 2nd International Conference on Vocational Education and Electrical Engineering (ICVEE)*, Surabaya, Indonesia, 2017, vol. 336. doi:10.1088/1757-899X/336/1/012017.
- [36] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, J. Liu, A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm, *J. Wireless Com Network* 31 (2021). doi:10.1186/s13638-021-01910-w.
- [37] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987) 53–65.
- [38] M. Halkidi, Y. Batistakis, M. Vazirgiannis, M., On clustering validation techniques. *Journal of intelligent information systems* 17 (2001) 107–145.
- [39] D. L. Davies, D. W. Bouldin, A cluster separation measure, in: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979) 224–227.
- [40] AdjustText - automatic label placement for matplotlib, 2024. URL: <https://github.com/Phlya/adjustText>. doi:10.5281/zenodo.10499815.