ANALYZING HIGHER EDUCATION RESEARCH TRENDS WITH LATENT DIRICHLET ALLOCATION: A TEXT MINING APPROACH

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

ABDULAZIZ ALRAYASHI

A MASTER OF SCIENCE THESIS
IN
THE DEPARTMENT OF COMPUTER ENGINEERING

APRIL 2023

ANALYZING HIGHER EDUCATION RESEARCH TRENDS WITH LATENT DIRICHLET ALLOCATION: A TEXT MINING APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

BY

ABDULAZIZ ALRAYASHI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
THE DEPARTMENT OF COMPUTER ENGINEERING

APRIL 2023

Approval of the Graduate School of Natural and Applied Sciences, Atilim University.

_____

Prof. Dr. Ender KESKINKILIC
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of **Master of Science in Computer Engineering Department, Atılım University**.

_____

Assoc. Prof. Dr. Gökhan ŞENGÜL
Head of Department

This is to certify that we have read the thesis ANALYZING HIGHER EDUCATION RESEARCH TRENDS WITH LATENT DIRICHLET ALLOCATION: A TEXT MINING APPROACH submitted by ABDULAZIZ ALRAYASHI and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Cansu Çiğdem EKİN
Supervisor

Examining Committee Members:

Asst. Prof. Dr. Damla TOPALLI
Computer Eng. Department, Atılım University _____

Asst. Prof. Dr. Cansu Çiğdem EKİN
Computer Eng. Department, Atılım University _____

Assoc. Prof. Dr. Elif POLAT HOPCAN
Department of Computer Education and
Instructional Technologies, Istanbul University _____

Date: *13 April 2023*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

ABDULAZIZ ALRAYASHI

Signature:

ABSTRACT

ANALYZING HIGHER EDUCATION RESEARCH TRENDS WITH LATENT DIRICHLET ALLOCATION: A TEXT MINING APPROACH

AlRayashi, Abdulaziz

Master of Science, Computer Engineering

Supervisor: Asst.Prof. Dr. Cansu Çiğdem EKİN

April 2023, #133 pages

This thesis investigates the research trends in higher education by analyzing a large corpus of over 69,000 academic publications indexed in the Web of Science (WOS). The study employs a combination of bibliometric analysis and topic modeling techniques, specifically Latent Dirichlet Allocation (LDA), to gain a deeper understanding of the higher education research landscape. The bibliometric analysis offers a comprehensive examination of the statistical distributions of WOS publications related to higher education. This includes aspects such as publication trends, document types, languages, and research areas, providing a solid foundation for understanding the research context in higher education. In parallel, the topic modelling analysis using LDA uncovers the main research trends, subjects, and frequently addressed topics in the field, shedding light on the evolution of research themes over time. This study provides a thorough picture of the current status of research on higher education by incorporating both bibliometric and topic modelling methodologies. The findings underscore the field's complexity and interdependence, emphasizing the need for a multidisciplinary strategy to address the possibilities and problems that exist in higher education. The results add to the body of knowledge in higher education research and are a useful tool for academics, decision-makers, and practitioners who want to remain up to date on the most recent advancements and trends in this field.

Keywords: Higher Education Research, Latent Dirichlet Allocation (LDA), Text Mining, Research Trends, Topic Modelling.

ÖZ

YÜKSEKÖĞRETİM ARAŞTIRMA EĞİLİMLERİNİN GİZLİ DİRİCHLET TAHSİSİ İLE ANALİZ EDİLMESİ: BİR METİN MADENCİLİK YAKLAŞIMI

AlRayashi, Abdulaziz
Y. Lisans, Bilgisayar Mühendisliği Bölümü
Danışman: Dr Öğr. Üyesi Cansu Çiğdem EKİN

Nisan 2023, #133 sayfa

Bu tez, Web of Science'ta (WOS) endekslenen 69.000'den fazla akademik yayından oluşan geniş bir külliyatı analiz ederek yüksek öğretimdeki araştırma eğilimlerini araştırmaktadır. Çalışma, yüksek öğretim araştırma ortamına ilişkin daha derin bir anlayış kazanmak için bibliyometrik analiz ve konu modelleme tekniklerinin, özellikle Gizli Dirichlet Tahsisi'nin (LDA) bir kombinasyonunu kullanır. Bibliyometrik analiz, yüksek öğretimle ilgili WOS yayınlarının istatistiksel dağılımlarının kapsamlı bir incelemesini sunar. Bu, yüksek öğretimdeki araştırma bağlamını anlamak için sağlam bir temel sağlayan yayın eğilimleri, belge türleri, diller ve araştırma alanları gibi unsurları içerir. Buna paralel olarak, LDA kullanan konu modelleme analizi, alandaki ana araştırma eğilimlerini, konuları ve sıklıkla ele alınan konuları ortaya çıkararak araştırma temalarının zaman içindeki gelişimine ışık tutar. Bu çalışma, hem bibliyometrik hem de konu modelleme metodolojilerini birleştirerek yüksek öğretim araştırmalarının mevcut durumunun kapsamlı bir resmini sunmaktadır. Bulgular, alanın karmaşıklığının ve karşılıklı bağımlılığının altını çizerek, yüksek öğretimde var olan olasılıkları ve sorunları ele almak için çok disiplinli bir stratejiye duyulan ihtiyacı vurgulamaktadır. Sonuçlar, yüksek öğretim araştırmalarındaki bilgi birikimine katkıda bulunur ve bu alandaki en son gelişmeler ve eğilimler hakkında güncel kalmak isteyen akademisyenler, karar vericiler ve uygulayıcılar için yararlı bir araçtır.

Anahtar Kelimeler: Yüksek Öğrenim Araştırması, Gizli Dirichlet Tahsisi (LDA), Metin Madenciliği, Araştırma Eğilimleri, Konu Modelleme

DEDICATION

To my loving family, inspiring mentors, and supportive friends, who have shaped my life's journey.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Asst. Prof. Dr. Cansu Çiğdem EKİN for their unwavering support, encouragement, and invaluable guidance throughout my thesis journey. Their expertise, patience, and constructive feedback were instrumental in shaping this work, and I am truly grateful for their mentorship.

I am also incredibly thankful to my family for their constant love, understanding, and encouragement during my academic pursuits. Their unwavering belief in my abilities and aspirations has been a great source of strength and motivation. I would like to extend special thanks to my parents, the engineer Mohammed AlRayashi and my lovely mother Nasra Ali for their unconditional support and for instilling in me the importance of hard work and perseverance.

In addition, I would like to express my appreciation to my friends who have been there for me during this challenging journey. Their camaraderie, advice, and shared experiences have been invaluable in helping me maintain balance and perspective throughout the process.

Lastly, I would like to acknowledge the academic community, my fellow students, and the faculty at Atilim University for providing a stimulating and supportive environment for my research. Their insights, suggestions, and critiques have significantly contributed to the development of this thesis.

Finally, I am deeply grateful to everyone who has played a part in my academic journey, and I hope that the knowledge and experience gained will serve me well in future endeavor

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

Acronym	Expansion
LDA	Latent Dirichlet Allocation
DM	Data Mining
KDD	Knowledge Discovery from Databases
CSV	Comma Separated Value file format
ML	Machine Learning
AI	Artificial Intelligence
ANN	Artificial Neural Network
KDT	Knowledge Discovery from Text databases
IR	Information Retrieval
IE	Information Extraction
POS	Part-of-Speech
BoW	Bag-of-Words model
VSM	Vector Space Model
TDM	Term Document Matrix
TF-IDF	Term Frequency-Inverse Document Frequency
LSI/LSA	Latent Semantic Indexing/Latent Semantic Analysis
SVD	Singular Value Decomposition
PLSI	Probabilistic Latent Semantic Indexing
PLSA	Probabilistic Latent Semantic Analysis

CHAPTER 1

INTRODUCTION

In the past few years, numerous institutions have expanded the volume of scientific articles, documents, and literature accessible on the internet. In the era of information, internet content is more readily available than ever. This has caused a sharp rise in the amount of data being kept and made available digitally each day. Unstructured data can be even more beneficial for researchers, businesspeople, and decision-makers than structured data, which is still crucial. When thoroughly studied, they offer useful information to many enterprises. Apart from rich-data formats like photos and videos and text documents like books, essays, Web pages, and social media material, the bulk of unstructured data is saved in text documents [1], [2]. The evaluation and analysis of literature to discern thematic patterns and progression in academic research fields have become increasingly challenging, if not unattainable, due to advancements in digital information publication and internet accessibility [3]–[5]. Automated Text Mining Methods, a discipline within Data Mining, can be employed to uncover latent semantic topics within a corpus of documents, effectively summarizing vast amounts of text while simultaneously revealing shifts in research themes and trends over time.

Nonetheless, the history of document classification extends far back, originating from the inception of libraries and forming a subsection within the realm of text mining. For instance, within the vast confines of the Library of Alexandria, Callimachus systematically arranged literary works into "Pinakes" tables. These tables encompassed divisions devoted to rhetoric, law, epic, tragedy, comedy, lyrical poetry, historical accounts, medical knowledge, mathematical principles, natural sciences, and an assortment of other subjects [6] Classifications often face challenges shortly after their establishment, a practice purportedly initiated by Aristophanes' "pugnacious"

critique of Callimachus' Pinakes [7]. The objective of classification is to group relevant elements into a single category, with documents potentially sharing various attribute such as language field and others. Research definitions inherently incorporate numerous aspects of knowledge within their terminology. Classifications are necessary at multiple levels to categorize research activities and publications, including journal level, article level, article topic, and so forth.

Concurrently, different types of courses can be pursued. For instance, scholarly disciplines, departments, and the Scopus All Science Journal Classification organize based on their specific domains of expertise. Conversely, the Medical Subject Headings Thesaurus represents an endeavour to encompass an extensive range of subjects and research methodologies.

The ongoing generation of new knowledge through research often defies neat categorization within pre-established frameworks [8] Despite counterclaims by some authors, these typical barriers to activity identification persist at all levels, particularly at the journal level [9], [10]. Owing to the fractal nature of classification difficulties, scholarly publications and articles are often segmented into various categories, leading to disputes regarding the employment of mutually exclusive multiple-class allocations and ontologically more dependable classification methods that streamline and elucidate reporting [11].

With the growing volume of information and the concurrent development of autonomous computer systems [12], [13]. data mining and data categorization, in particular text classification, are viewed as vital [14]. A common problem in the disciplines of machine learning and natural language processing (NLP) is classifying incoming texts into existing clusters by mining their similarity to the group [15] .Text classification is a classic supervised learning method (equivalent to automated metadata extraction) [16]. The process of classifying new texts entails labelling them in accordance with how similar they are to previously classified texts in the training set [14], as detailed by reference [16].

Automated text classification offers the benefits of accelerated data storage and retrieval. Moreover, manually sifting through large quantities of text is time-consuming and labour-intensive, with the possibility of subjective categorization errors. Text classification serves numerous functions, such as filtering spam emails,

categorizing news by topic in online newspapers, managing knowledge, and supporting internet search tools [13].

To examine vast amounts of textual data, academic research in higher education has progressively employed text mining methods [17]. simplifying the identification of patterns and trends within the data. A primary advantage of utilizing text mining in higher education research lies in its ability to assess unstructured data, including student evaluations, scholarly research, and additional text-based data sources [16].

Text mining has been extensively employed in higher education research to examine factors influencing academic achievement and identify trends in student satisfaction [17]. Other research has applied text mining to assess course and curriculum content to enhance student learning outcomes more effectively. Additionally, text mining has been utilized to analyze faculty members' research output and identify characteristics of prolific scholars.

In summary, text mining has substantially enhanced the precision and rigour of higher education research by offering a potent tool for scrutinizing vast volumes of intricate data and yielding innovative insights into the conduct and performance of students, faculty, and institutions [18].

By harnessing the capabilities of text mining, researchers can delve deeper into the intricacies of higher education, identifying trends, uncovering hidden patterns, and ultimately informing strategies and policies that foster improved academic outcomes and institutional effectiveness [19]. As the field of text mining continues to evolve, its applications and potential benefits for higher education research are expected to expand, further enhancing our understanding of the complex landscape of academia [18]

1.1 Problem Statement

The following restrictions/limitations have been discovered after reviewing significant literature on text mining in higher education:

• Traditional quantitative approaches have been used extensively in research, sometimes in conjunction with time-consuming manual content analysis. Furthermore, the scope of analysis has been restricted to a relatively small number of articles.
• Although certain studies have utilized basic machine learning algorithms for article classification, the precision and reliability of these systems might be limited.
• As higher education generates increasing volumes of text-based data over extended periods, employing conventional analytic approaches for analysis becomes progressively more challenging.

A significant portion of research has predominantly concentrated on the technical facets of text mining, such as categorizing articles according to their titles or document types, rather than offering a more exhaustive comprehension of the domain.

To carry out swifter, more accurate, and increasingly automated content analysis and classification within higher education research, it is crucial to adopt more efficient research methods, such as text mining. Although text-mining approaches have been employed in certain studies, it is crucial to apply more advanced text-mining techniques to uncover significant information that may enhance comprehension and guide changes in a variety of text-oriented data features in higher education.

1.2 Purpose of the Study

This study employs dimensionality reduction and linear models for classification techniques to examine academic publications on the Web of Science (WOS) with the aim of obtaining vital information and insights pertaining to this subject. The following is a summary of the investigation's primary research goals:

• Implementing Bibliometric Analysis Tools to evaluate the statistical distributions of WOS publications through an extensive literature review, encompassing aspects such as publication titles, abstracts, and document types from the Social Sciences Citation Index (SSCI).
• Utilizing Text Mining Instruments for the Recognition of Research Topics and Trends developments within the realm of higher education.

1.3 Importance of the Research

Some of the importance of conducting this investigation of text mining in higher education include the following:

• By outlining the state of the field and highlighting knowledge gaps regarding the use of text mining technologies in this setting, this study offers the potential to improve the literature review of academic articles in higher education.

• The results of this study add to the body of knowledge by giving a thorough understanding of the research issues and trends in the field.

1.4 Research Questions

The research, which is based on an analytical methodology, aims to answer the following questions:

1. What is the status of the publications in Higher Education from its existence to today?
2. Which are the most productive countries?
3. Which countries/regions and institutions were the major contributors?
4. What were the scientific collaborations among major contributors like?
5. In which journals were Higher Education studies mainly published?
6. What topics were commonly discussed/researched in Higher Education?
7. Is the number of articles concerning these topics increasing or decreasing?
8. How did research topics evolve over time?

1.5 Thesis Outline

- Chapter 1: Introductory.
- Chapter 2: Literature Analysis on Text Classification Algorithms.
- Chapter 3: The study's background is covered in this chapter.
- Chapter 4: Methodology.
- Chapter 5: Results.
- Chapter 6: Discussion and Conclusion.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In academic research, literature reviews are essential for evaluating the body of existing knowledge and defining the state of the problem being studied [20], [21] Various well-established approaches exist for examining and evaluating the existing literature, contingent upon the research subject or the issues the researcher seeks to address. A summary of these methodologies is provided. Literature reviews are an essential aspect of academic research for assessing the existing knowledge on a subject [20], [21]. Researchers have different methods at their disposal for examining and assessing the body of literature, depending on the research question and study subject. Among these methods, systematic reviews utilize a research-based, repeatable, and accessible process for examining relevant literature, analyzing the reviewer's methodology and conclusions, and addressing a specific research question, thereby reducing systemic errors and biases [22].Bibliometric analysis uses quantitative data analytics to examine and summarize publications and spot distinctive trends in research-based literature. Scientists use this method to assess the importance of academic publications and contrast the contributions of individuals, organizations, and nations [23] [24]. Meta-analysis is another research method that statistically analyzes the findings of multiple studies on the same topic, even when they use different reporting measures that may have some level of error. The main benefit of this technique is that it allows for the collection of data that provides statistically more dependable and accurate estimates than what is possible in a single investigation [25]. In qualitative research, content analysis is a methodical method for assessing topics,

writings, articles, or ideas. The relationships between certain words, thoughts, or ideas may be examined and categorized by researchers using content analysis [26].

Last but not least, trend analysis entails gathering data to spot trends within a certain academic topic. This strategy tries to present a qualitative, well-organized, and easily understandable explanation of the results of academic papers and scientific research by using techniques like percentage form and repetition [27]. [28].

The analysis of existing literature in this review is segmented into a pair of components. In the initial portion, we will delve into the key outcomes from numerous significant investigations, the majority of which employed conventional research methodologies. While text mining is the approach used in this thesis, the second section, which gives more weight to the findings of two important review papers, will concentrate on research techniques based on machine learning technology.

Scholars have utilized various techniques over time to assess the literature on text classification and data mining learning. Despite the progress made, there is still much work to be done in this field.

Since academics started employing statistical techniques for language-related problems in the 1960s, text classification has been an element of natural language processing [29] This innovation makes it feasible for computers to analyses and categorize enormous volumes of textual material rapidly and correctly, which was before not possible.

In the late 1990s, text classification initially employed Machine Learning methodologies, which demonstrated superior efficacy compared to traditional strategies. Specifically, Support Vector Machines (SVMs) were established as highly effective in this domain [30]. Supervised learning models such as SVMs have proven to be more effective than conventional linear models by creating hyperplanes between classes or clusters, resulting in improved identification of these classes or clusters. Maximum Entropy Models (MEMs)[31]are yet another effective approach for text classification. MEM is a probabilistic graphical modelling tool that finds patterns in data by ascribing probability depending on how often certain patterns appear in a particular dataset.

There has been significant progress over the last several decades in the development of increasingly effective algorithms and methodologies for automated text

classification tasks including sentiment analysis and document classification. These developments have made it feasible to understand the meaning of different sorts of writings in addition to being able to recognize them. As a result, cutting-edge software has been created, including chatbots and automated customer support systems that may run constantly without human interaction [31].

In the study [32] researchers Alper Kursat Uysal and Serkan Gunal carried out a pioneering investigation into text classification by thoroughly assessing various preprocessing approaches, including stemming, stop word elimination, tokenization, and conversion to lowercase, in two distinct domains (news and emails) for both English and Turkish languages. Utilizing Support Vector Machines (SVMs) with feature dimensions varying between 10 and 2,000, the authors attained notable accuracy levels. This research emphasizes the efficacy of the offered methodologies in the field of text classification for a range of topics and languages.

Their findings revealed that the Turkish news domain had an accuracy of 97.3% with the lowest feature size of 10, while English emails had the greatest accuracy at 98.8% with a feature size of 500. No matter the domain or language, the authors underline the necessity of preprocessing tasks for better results and insist that all potential combinations should be attempted to get better results. This demonstrates the critical function of preprocessing in effective machine-learning models and the need to take it into account when creating any model. Researchers may make sure they are getting the greatest outcomes from their models and using every benefit by making sure all potential combinations are examined.

In their work, the authors underlined that in order to get better results, all conceivable arrangements of the preprocessing jobs should be looked at. Preprocessing is an essential part of successful machine learning models; hence it is essential to keep this in mind while creating any model and the nation's finest, including the best of the best. This acts as a critical reminder for the development of any machine learning model.

To get the best performance in text classification, this research underlines the significance of choosing the right preprocessing tasks. Instead, then either activating or removing all preprocessing approaches, the authors advise evaluating every conceivable combination. The accuracy of text classification models may be considerably improved using this method. The study also offers insightful information

on the best preparation methods for various language domains. This knowledge will be helpful to academics and professionals who plan to use similar techniques for text analysis in future projects.

In their study, Howard and Ruder (2018) [33] describe transfer learning for NLP tasks and provide a thorough description of several transfer learning techniques and their efficacy on six datasets. The research emphasizes the significance of directionality and fine-tuning when employing a classifier, which may boost model performance by 0.5 to 0.7 times in comparison to techniques without fine-tuning. In an academic context, the writers contend that the Universal Language Model Fine-tuning (ULMFiT) technique proves to be an effective transfer learning approach for NLP tasks, as it surpasses competing methods concerning accuracy, precision, and recall metrics throughout all the datasets examined in this research.

Moreover, ULMFiT offers a variety of approaches that users may adopt depending on the particular demands or requirements for each activity, allowing them greater flexibility over how they apply their models and enhancing their accuracy. These techniques include discriminative fine-tuning and progressive unfreezing.

The work mentioned above offers academics the chance to investigate cutting-edge methods for text classification, to sum up. Researchers may improve accuracy, precision, and recall scores compared to conventional methods by combining Transfer Learning with ULMFiT's innovative fine-tuning procedures. The research emphasizes the importance of directionality and fine-tuning when employing a classifier since these factors may significantly improve model performance. In order to provide users more flexibility over how they utilize their models, it also provides several strategies that may be tailored to particular needs, such as discriminative fine-tuning or progressive unfreezing.

2.2. An Overview of Text Classification Techniques

It is helpful to look at two papers in order to better understand the various text classification techniques looked at in earlier study. The first is the study described in[34] , which created a classifier that assigns named entity tags to German Wikipedia entries using English Wikipedia as a reference. The project aims to minimize the requirement for human annotation in languages other than English while

simultaneously producing sizable, annotated corpora to evaluate named entity recognition in the German language. The developed classifier demonstrated strong recall (87%) and accuracy (97%) on diverse entity classifications, demonstrating its potential use in this study area.

Graph mining is a method for classifying the content of English news articles, and it was also suggested in [35] as a way for classifying texts. The approach relies on the supposition that every piece of text can be depicted through a graph illustrating the relationships among words, where each term is allocated, a weight corresponding to its importance within the written content. For those looking to speed up and improve text classification, this novel approach has been shown to provide more precise results than existing methods.

In [36] ,the authors introduce the innovative Trusted Platform Module (TPM) algorithm for text classification, which combines machine learning principles with technical specifications and real-world use cases, resulting in enhanced precision and effectiveness. For text classification tasks like telling spam from legitimate emails, this method was developed particularly for natural language processing. The suggested technique demonstrated an amazing accuracy rate of over 95% when tested on multiple datasets. This work demonstrates the possibility for combining machine learning with technical requirements to develop cutting-edge text classification techniques that may improve efficiency and accuracy.

The Relevant-Based Feature Ranking (RBFR) method, developed in [37] , uses machine learning to find and choose the most crucial characteristics from a feature space. The RBFR approach was compared to four well-known feature selection strategies, including the balanced accuracy measure, Gini index, odds ratio, and information gain, using three distinct datasets (20 newsgroups, Reuters, and WAP) and five machine learning models, namely KNN, RF, SVM, NB, and LR. As compared to the other examined strategies, the findings showed that using the RBFR algorithm significantly increased accuracy by 25.4305%. These findings show how effective this innovative approach is when used on large datasets with many parameters.

The results of a recent study by [38] were predicted using text classification. To forecast the outcome of court battles, the authors combined transformer-based classification algorithms with other data sources. The results demonstrated that the suggested models reliably surpassed benchmarks and even outdid human estimations in predicting the triumph of claims. By incorporating these models into a currently utilized claim administration system, the researchers effectively managed the case lifecycle and dealt with a multitude of operations reaching tens of thousands every month. This study is noteworthy because it offers insights into possible real-world applications of text classification outside of academia.

2.3. Text Mining in Higher Education - A Literature Review

This section of the literature review will concentrate on the use of text-mining techniques in the area of higher education. Throughout this portion, we will explore research that delved into the utilization of text mining for knowledge administration, assessment, and student engagement in tertiary education settings. By examining the current level of research in this field, we may get a better understanding of the potential advantages of text mining in higher education and pinpoint areas that need more study and development. J. S. Alejandro-Cruz describes how sustainability is included in international higher education studies using text-mining tools [39].

It was discovered by looking at papers from 1991 to 2018 that the USA, China, the UK, and Australia were the top contributors to sustainability research in higher education. The study also identified new trends, themes, and patterns in global sustainability studies, highlighting societal awareness and sustainable planning. In order to promote ecological consciousness and direct the achievement of sustainability goals within communities, organizations, and governmental bodies, the authors in [40] concluded that advanced education is essential. Examining how sustainability is included in studies of international higher education was the goal of this study. The researchers gathered 6,724 papers from the Web of Science and Scopus databases using text mining methods, covering global publications between 1991 and 2018. The results showed that 40.58% of all records for sustainability research in higher education were from the USA, China, the UK, and Australia combined. Furthermore, the analysis highlighted emerging trends and subjects, emphasizing social awareness

and sustainable strategies. According to the study's results, higher education institutions play a significant role in raising environmental awareness and assisting societal and governmental efforts to achieve sustainability objectives.

The analysis's primary objective was to identify the current research trends in technology-augmented learning (TEL) in higher education for 2,154 TEL-related articles relating to higher education were collected [41], The researchers used a hybrid bibliometric approach that includes textual analytics and direct citation network evaluation to evaluate TEL research papers from the Web of Science database. They employed cluster analysis, latent semantic analysis, and visual analytics to comprehend the growth of TEL in higher education. The research identified five main growth paths for TEL, including podcasting, social media, adoption, and criticism. The researchers provided a thorough summary of the gathered data by highlighting the important subcategories within each strand. The proposed approach presents an in-depth means of comprehending the complete progression of TEL research in tertiary education. In the study [42] a comprehensive examination of the theoretical advancements in e-learning within higher education settings are provided, specifically focusing on publications during the COVID-19 outbreak. By reviewing e-learning research and conducting bibliometric evaluations on 602 articles found in the WOS database, the authors were able to cover a publication period spanning from 2020 to 2021. The research revealed significant findings in the realm of e-learning, encompassing prominent concepts and areas like blended learning, interactive learning, and distance learning. The article also emphasized current advancements in e-learning research, such as the growing use of artificial intelligence and machine learning, which has been linked to the pandemic. The analysis concentrated on instructional approaches and concurrently proposed novel directions for exploration.

The work in [43]presents the core principles of text mining, covering essential concepts, prevalent methodologies, and popular software applications. The authors show the importance of text mining by the exploration of two promising practices and the presentation of two comprehensive examples. These promising practices involve (1) employing text analytics to understand and minimize course withdrawals, and (2) assessing student understanding and depth of learning in STEM fields, with a focus on physics. The in-depth examples include (1) refining questionnaire items for the

National Survey of Student Engagement (NSSE), and (2) implementing a learning analytics system at a community college (City University of New York [CUNY]: the Stella and Charles Guttman Community College, or CUNY Guttman). The study's results feature the identification of additional item options for the survey and the discovery of a connection between e-portfolio content and academic achievement.

In [44] a case study is presented, this case study demonstrates the use of text mining in analyzing open-ended student survey responses to improve student experience management (SEM), a concept derived from customer experience management (CEM). By applying text mining to a campus-wide survey at Arizona State University, researchers gained valuable insights into students' experiences with instructional technology. Open-ended questions encouraged the organic emergence of themes, revealing students' desire for an interconnected learning environment. Key findings include the importance of complementing online access with laptop usage and accessible course materials, as well as balancing virtual classes with human interactions. The work in [45] explores the emotions experienced by university students in Ecuador during the COVID-19 pandemic. Using quantitative, cross-sectional research methods, 55 unstructured anonymous interviews were conducted with students from 16 higher education institutions. Sentiment analysis techniques like Latent Dirichlet Allocation (LDA), MATLAB, and NVIVO were applied to 500 phrases from the interviews. Results indicated 64% negative, 11% neutral, and 25% positive emotions. LDA revealed two unobserved groups associated with feelings such as stress, tiredness, and effort. Future research could investigate the specific emotions and their causes in greater depth.

2.4. Contribution

The literature review that conducted in the previous sections has shown a lack in the papers that show the general views of the higher education literature and also the limitation in sample sizes of higher education research papers that considered in the researches. Utilizing text mining methods, one can analyze the vast amounts of unstructured data present, providing essential knowledge concerning prevailing

tendencies and concerns within higher education, thereby gaining a more profound comprehension of research patterns in this field. This research aims to bridge this void and contribute to a comprehensive understanding of higher education research trends.

2.5. Summary

To enhance the precision of article classification, text classification has been more and more popular in recent years. Machine learning algorithms provide a more effective substitute for labour and time-intensive classical approaches like bibliometric analysis and content analysis. A survey of the literature reveals that many methodologies have been used in research papers and publications, with supervised machine learning algorithms ,which need a labelled training dataset, being one of the most common.

The system is trained using the labelled training dataset, allowing it to identify newly undiscovered text using previously discovered patterns. Unsupervised machine learning methods provide a different strategy that does not need labelled training data. These techniques may be used for tasks like anomaly detection and clustering and are useful for revealing hidden patterns within data.

The research stresses the use of text mining methods in tertiary education, specifically the use of text categorization to improve the accuracy of article classification. Machine learning algorithms provide a more effective substitute for older approaches like bibliometric analysis and content analysis, which may be labour and time-intensive. Based on the research in the literature, numerous tactics have been employed in this domain, encompassing supervised machine-learning techniques requiring annotated training information. This data serves to educate the algorithm, allowing it to discern new, unfamiliar text by applying patterns acquired from the training information. Unsupervised machine learning approaches, on the other hand, do not need labelled training data and are used to find underlying patterns in data for tasks like grouping or spotting abnormalities. Depending on the job at hand, both supervised and unsupervised procedures have benefits; supervised methodologies provide exact results, while unsupervised methodologies, in the lack of previous knowledge, discover hidden links within vast datasets

CHAPTER 3

BACKGROUND OF THE STUDY

3.1. Understanding Bibliometrics: Definition, Concept, and Key Metrics

In today's highly competitive and rapidly evolving research landscape, it has become increasingly important to evaluate the impact and significance of scientific contributions. Bibliometrics is a powerful tool used to quantitatively analyze research impact and identify trends in various scientific fields [46]. In this comprehensive guide, we will explore the definition and concept of bibliometrics, along with the key terms and metrics used in this field, such as citation analysis, h-index, and impact factor.

3.1.1. Definition and Concept of Bibliometrics

Bibliometrics originates from the Greek terms 'biblion,' signifying book, and 'metron,' denoting measure. This subfield of information science utilizes quantitative examination and statistical methods to appraise the significance, influence, and patterns of disseminated research in a particular discipline. Frequently, bibliometrics is employed in gauging the accomplishments of researchers, organizations, and academic journals, thereby facilitating informed decisions in matters such as resource distribution, hiring, and research planning [47].

The concept of bibliometrics is rooted in the idea that scientific knowledge can be represented by publications, and the impact or value of these publications can be measured through various indicators. These metrics are based on variables including the quantity and frequency of citations, as well as the standing of the authors, institutions, and publications involved. By analyzing these indicators, bibliometrics

provides a quantitative means of understanding the relationships between research outputs and their impact on the scientific community.

A major component of bibliometrics in research assessment is determining the significance of scientific work. The Social Sciences Citation Index (SSCI) and Emerging Sources Citation Index (ESCI) are provided by the Web of Science (WOS) as indications of research impact across various scientific fields [48]. In the field of academia, higher learning establishments frequently employ such indices to assess the significance and calibre of scholarly inquiry, as well as pinpoint the institutions and investigators that accumulate the greatest number of citations. Education, sociology, and psychology are only a few of the social science fields that the SSCI covers. On the other side, the ESCI includes recent scientific fields and upcoming research topics. For higher education institutions and funding organizations, the SSCI and ESCI are useful instruments for assessing the significance and calibre of research in many domains.

In spite of bibliometrics' use in evaluating research, it's crucial to remember that it has its limits. Numerous factors, including the quality of published information, the influence of diverse scientific disciplines, and the magnitude of the scientific community, can potentially affect bibliometric outcome [48]. Moreover, bibliometric metrics must be used with other evaluation techniques, such as expert analysis, in order to give a thorough evaluation of research achievements. In conclusion, bibliometrics is often used to assess the caliber and significance of studies across a range of scientific fields and is crucial in determining the value of research in higher education [49]. Publication names, themes, abstracts, document kinds, language use, and Web of Science indexes are all indicators of how well research is done in higher education. It is essential to combine bibliometrics with other evaluation techniques in order to give a thorough evaluation of research achievements.

3.1.2. Application of Bibliometric in Higher Education Research

For informing and influencing higher education policies, practices, and advances, the results of this research are crucial. The need to recognize and comprehend the major currents, significant figures, institutions, and publications that influence the discourse grows as the body of study in this area expands. Higher education research outputs

may be statistically analyzed using bibliometrics, which also gives useful insights into the advancement of the discipline [50]. We will discuss bibliometrics in higher education research in this section and examine how they might be used to identify important actors and emerging trends.

3.1.2.1 Analyzing Research Trends

Bibliometrics may be used to investigate research trends in higher education by examining publication patterns and citation networks within the field. For instance, researchers may use bibliometric methods to identify the most often cited books, authors, organizations, and journals, as well as the most well-liked research topics and themes. This information can help to reveal the key areas of focus and interest within the higher education research community, as well as to uncover any gaps or underrepresented areas in the literature[50] .

3.1.2.2 Identifying Influential Authors, Institutions, and Publications

One of the key applications of bibliometrics in higher education research is the identification of the most well-known authors, organizations, and publications in the field. By looking at citation score data researchers may learn more about the individuals, organizations, and journals that have contributed the most to higher education research[51] . This information can be invaluable for decision-makers in various contexts, such as recruitment, funding allocation, and research collaboration, as well as for researchers seeking to situate their work within the broader scholarly landscape.

3.1.2.3 Limitations and Challenges

While bibliometrics may provide light on the state of research in higher education, it is important to recognize the constraints and difficulties that come with its use. While the area often incorporates multidisciplinary work that may be more likely to be referenced outside of its core discipline, citation statistics, for instance, may not accurately reflect the effect and relevance of research in higher education [52].

Moreover, citation styles might differ across various fields of study, which could introduce biases in the analysis.

Additionally, bibliometric indicators like the h-index and impact factor may not always provide a thorough evaluation of an author's or institution's contributions to higher education research because these metrics may be influenced by variables like self-citation, co-authorship patterns, and researcher age. To address these limitations, it is crucial to employ a combination of bibliometric indicators and complementary methods, such as expert opinion and qualitative analysis, when evaluating research performance in higher education.

3.1.3 Limitations of Bibliometrics in Evaluating Research Performance

Bibliometrics has become an essential tool in evaluating research performance, offering valuable insights into the impact, significance, and trends of scientific contributions across various fields. However, the use of bibliometric indicators is not without limitations and criticisms [53]. In this section, we will outline the key challenges and criticisms associated with the use of bibliometrics in evaluating research performance, highlighting potential biases, the overemphasis on citation counts, and the neglect of other factors that contribute to research quality [53].

3.1.3.1 Potential Biases

Bibliometric indicators can be influenced by various biases that may affect the accuracy and fairness of research evaluation. Some of these biases include [53]:

1-Self-citation bias: Researchers may cite their work to increase their citation counts, potentially skewing bibliometric indicators in their favour.
2-Citation bias: Certain research topics, disciplines, or methodologies may be more prone to citation than others, leading to an over- or underestimation of their impact.

3-Language bias: Research published in languages other than English may be less likely to be cited, potentially undervaluing the contributions of non-English-speaking researchers.

4-Age bias: older researchers may have higher citation counts due to their longer publishing history, which may not necessarily reflect their current research impact.

Overemphasis on Citation Counts

The use of citation counts as a stand-in for the significance and quality of research is one of the primary critiques levelled towards bibliometrics. Citations may provide light on a publication's impact, but they may not adequately represent the overall influence of research on society, policy, and practice [54]. Moreover, the emphasis on citation counts may inadvertently encourage researchers to choose widely cited study topics and publication venues over those that may be more original, riskier, or pertinent to their particular areas of interest and competence.

3.1.3.2 Neglect of Other Factors Contributing to Research Quality

Bibliometric indicators often fail to account for other factors that contribute to research quality, such as [54]:

1-Methodological rigour: Bibliometric indicators do not directly measure the methodological quality of research, which is a critical aspect of research performance and impact.

2-Reproducibility: The replicability of research findings is an essential component of research quality, but it is not captured by bibliometric indicators.

3-Interdisciplinary research: Bibliometric indicators may not adequately reflect the impact and value of interdisciplinary research, which often spans multiple fields and citation practices.

4-Research collaboration: The extent and nature of research collaboration, which can play a crucial role in advancing scientific knowledge, are not directly measured by bibliometric indicators.

3.2 Unveiling Research Trends in Higher Education through Knowledge Discovery from Databases

Knowledge discovery from databases (KDD) involves mining massive volumes of data for meaningful patterns, trends, and insightful information. The KDD process is particularly relevant in the context of higher education research, where huge amounts of data are created via publishing, citation, and other research activities. The three primary phases of the KDD process, data preparation, data mining, and knowledge interpretation will be briefly described in this section. We will also discuss the significance of data mining as a crucial component of KDD and how it relates to the evaluation of research trends in higher education [55].

3.2.1 Main Steps of the KDD Process

1-Data Preprocessing: The first step in the KDD process is to prepare the data for analysis. This process often includes data integration (i.e., combining data from several sources), data transformation (i.e., normalizing or aggregating), and data reduction (e.g., feature selection or dimensionality reduction). Data pretreatment is essential to guarantee the reliability and correctness of the next data mining and knowledge interpretation processes [56]

2-Data Mining: Data mining is the primary phase in the KDD process, and it focuses on applying various algorithms and methodologies to uncover patterns, trends, and correlations from preprocessed data. Supervised learning (such as classification and regression), unsupervised learning (such as clustering and association rule mining), and semi-supervised learning are the three main categories into which data mining methods may be categorized.
The study topic, the kind of data, and the intended analytical results all influence the choice of data mining techniques [57]

3-Knowledge Interpretation: Interpreting and confirming the outcomes of data mining is the last phase in the KDD process. This stage typically includes the evaluation of the mined patterns for their relevance, novelty, and potential impact, as

well as the presentation of the findings in a user-friendly and understandable format. Knowledge interpretation is crucial for transforming the raw output of data mining into actionable insights that can inform decision-making and drive innovation in higher education research [58]

3.2.2 Role of Data Mining in KDD and its Relevance to Higher Education Research

Data mining, a key phase in the KDD process, enables researchers to uncover hidden patterns, trends, and linkages within large and complex datasets. In the context of higher education research, data mining techniques can be used to examine various types of data, such as publication records, citation networks, and collaboration patterns, in order to identify the most significant authors, institutions, and research topics, as well as new trends and research frontiers[59].

By combining data mining methods with bibliometrics and other complementing tools, researchers may get crucial insights into the dynamics of higher education research and its impacts on policy, practice, and society. Moreover, the use of data mining to research in higher education may facilitate cooperative research, resource allocation, and strategic planning while also enhancing industry knowledge and innovation.

3.2.3. Data Mining Techniques for Analyzing Research Trends in Higher Education

Data mining techniques are crucial for extracting patterns, trends, and insights from massive, complex datasets in the area of higher education research. By using these techniques, researchers may better understand the research environment, identify emerging trends, and promote evidence-based decision-making[60] .We will discuss a number of data mining techniques that may be utilized to analyze trends in higher education research, such as clustering, classification, association rule mining, and sequential pattern mining. We will also go through each approach's benefits and drawbacks as well as potential applications in the context of research analysis in higher education.

- **Clustering**

Clustering is a method of unsupervised learning that groups data objects that are similar based on their traits or attributes. In the context of higher education research analysis, clustering can be used to identify research areas with similar characteristics, such as citation patterns or collaboration networks[61]. This can help reveal emerging research topics, sub-disciplines, and interdisciplinary areas of study.

- **Strengths**

Unsupervised learning allows for the identification of previously unknown patterns or trends. Can be used to explore large and complex datasets without predefined categories or labels [62]

Limitations:

- Results can be sensitive to the choice of similarity measure and clustering algorithm.
- Requires domain knowledge for interpreting and validating the identified clusters.

- **Classification**

Classification is a supervised learning technique that assigns data points to predefined categories or labels based on their features. In higher education research analysis, classification can be used to predict the research area, impact, or potential of a given publication, author, or institution based on their attributes and historical data [63].

- **Strengths**

Can provide accurate predictions when trained on large and representative datasets.
Offers a wide range of algorithms and techniques to choose from, depending on the specific research question and data characteristics [64]

- **Limitations:**

- Requires labelled data for training and validation, which can be time-consuming and labour-intensive to obtain.
- Predictive performance can be affected by class imbalance, noisy data, or the choice of the classification algorithm.

- **Association Rule Mining**

This technique makes it easier to identify connections between study topics, authors, academic institutions, or other factors in the context of higher education research analysis. Association rule mining is a method for finding connections or recurring patterns among variables in a dataset. Association rule mining may be very helpful in the study of higher education research by revealing prospective areas for cooperation, interdisciplinary research, or resource allocation [65].

- **Strengths**
- Can uncover hidden relationships and patterns in large and complex datasets [66].
- Provides easily interpretable rules that describe the associations between variables [66].
- **Limitations**

Requires the definition of appropriate support and confidence thresholds, which can affect the number of generated rules [66].

Can produce a large number of rules, necessitating further filtering or post-processing to identify the most relevant and interesting associations.

- **Sequential Pattern Mining**

In temporal or ordered data, sequential pattern mining serves as a method for detecting reiterative sequences or patterns. This technique is particularly adept at identifying recurring trends within data that follow a specific order or chronology. In higher education research analysis, sequential pattern mining can help uncover the evolution of research topics, collaboration patterns, or citation networks over time, providing insights into the development and progression of research fields [67]

- **Strengths**

  Can reveal trends and patterns in temporal or ordered data that are not apparent in static or unordered datasets.Offers a range of algorithms and techniques to handle various types of sequence data and research questions.

- **Limitations**

  Can be computationally intensive, especially for large and complex datasets. Requires the definition of appropriate support and gap constraints, which can affect the number of generated patterns [67].

Table 3.1 summarized the main three datamining techniques with their strengths and limitations.

Table 3-1:Strengths and Limitations of Data Mining Techniques

Data Mining Technique	Strengths	Limitations
Clustering	- Unsupervised learning allows for the identification of previously unknown patterns or trends.	- Results can be sensitive to the choice of similarity measure and clustering algorithm.
	- Can be used to explore large and complex datasets without predefined categories or labels.	- Requires domain knowledge for interpreting and validating the identified clusters.
Classification	- Can provide accurate predictions when trained on large and representative datasets.	- Requires labelled data for training and validation, which can be time-consuming and labour-intensive to obtain.
	- Offers a wide range of algorithms and techniques to choose from, depending on the specific research question and data characteristics.	- Predictive performance can be affected by class imbalance, noisy data, or the choice of the classification algorithm.
Association Rule Mining	- Can uncover hidden relationships and patterns in large and complex datasets.	- Requires the definition of appropriate support and confidence thresholds, which can affect the number of generated rules.
	- Provides easily interpretable rules that describe the associations between variables.	- Can produce a large number of rules, necessitating further filtering or post-processing to identify the most relevant and interesting associations.
Sequential Pattern Mining	- Can reveal trends and patterns in temporal or ordered data that are not apparent in static or unordered datasets.	- Can be computationally intensive, especially for large and complex datasets.
	- Offers a range of algorithms and techniques to handle various types of sequence data and research questions.	- Requires the definition of appropriate support and gap constraints, which can affect the number of generated patterns.

3.2.4. Challenges and Opportunities in Applying Data Mining to Higher Education Research

Employing data mining instruments to examine research tendencies in higher education holds significant potential for unveiling perceptive knowledge and promoting ingenuity. This approach can greatly contribute to the advancement and development of academic research. However, there are also challenges and opportunities associated with this approach, including issues related to data quality, data privacy, ethical considerations, interdisciplinary collaboration, and the development of novel methodologies [68].We will go through these possibilities and problems in further depth in this section, emphasizing the consequences for academics and practitioners involved in higher education research analysis.

- **Challenges**

1-Data Quality: The validity and reliability of data mining results are significantly impacted by the correctness of the input data [68]problems like missing, wrong, or inconsistent data may have a substantial influence on the performance of data mining algorithms and the overall results drawn from the research. Ensuring data quality through careful preprocessing and data cleaning is essential for obtaining meaningful insights from data mining techniques.

2-Data Privacy: Higher education research often involves sensitive data, such as personal information about researchers, students, or institutions. Ensuring data privacy and complying with relevant regulations, such as GDPR, is crucial when applying data mining techniques to higher education research. Techniques such as anonymization, aggregation, or differential privacy can be employed to protect individual privacy while still allowing for meaningful analysis.

3-Ethical Considerations: Data mining techniques can potentially lead to unintended consequences, such as biased or discriminatory outcomes, if not applied responsibly. Researchers and practitioners should be mindful of potential ethical concerns when

applying data mining techniques to higher education research and take appropriate steps to ensure fairness, transparency, and accountability in their analyses.

- **Opportunities**

1-Interdisciplinary Collaboration: Data mining techniques offer a unique opportunity to foster interdisciplinary collaboration in higher education research [68]. Researchers may create unique strategies to assess study patterns and unearth previously undiscovered insights by merging skills from many domains, such as computer science, statistics, and domain-specific knowledge.

2-Novel Methodologies: New opportunities for methodological innovation are created by the introduction of data mining tools to higher education research[69] . By developing and altering algorithms and techniques specifically tailored to the needs and challenges of higher education research, researchers may increase the robustness and usefulness of their studies and contribute to the body of knowledge in the field.

3-Integration with other research evaluation approaches: Integrating data mining tools with other research evaluation methodologies, such as bibliometrics, altmetrics, expert opinion, and qualitative evaluations, may give a more thorough and nuanced knowledge of research trends and effects in higher education [70].This integration can lead to more informed decision-making, resource allocation, and research collaboration, ultimately contributing to the development of evidence-based policies and practices that shape the future of higher education.

Despite the challenges associated with applying data mining techniques to higher education research, there are also significant opportunities for innovation and collaboration. By addressing issues related to data quality, data privacy, and ethical considerations, and capitalizing on the potential for interdisciplinary collaboration and methodological development, researchers and practitioners can harness the power of data mining techniques to drive advancements in higher education research [71]. The integration of data mining with other research evaluation approaches can further enrich our understanding of research trends and impact, paving the way for a more inclusive, diverse, and innovative higher education research landscape.

3.3. Text Mining

Text mining is a method that is becoming more and more popular for gaining important knowledge and insights from massive volumes of unstructured textual data. Text mining provides a potent method for analyzing and comprehending research trends in this field since the number of research papers, reports, and other text-based resources in higher education keeps increasing[72]. This section will present the concept of text mining, its core concepts, methods, and tools, as well as how they relate to the analysis of research trends in higher education.

3.3.1. Text Mining Definition and Concept

Frequently denoted as text analytics or natural language processing (NLP), text mining constitutes an automated technique for extracting, processing, and analyzing unstructured textual data to uncover intriguing patterns, tendencies, and insights. This method allows for efficient examination and understanding of vast textual information[73]. Text mining combines techniques from various fields, including computer science, linguistics, and statistics, to transform raw text data into structured, machine-readable formats that can be further analyzed and interpreted [74]

3.3.1.1 Key Concepts and Methods in Text Mining

1-Tokenization: As a crucial preliminary phase in text mining, tokenization facilitates the investigation of individual tokens and their interrelations within the given textual content. Tokenization encompasses the process of segmenting a text into smaller components, commonly referred to as tokens, which may include words, phrases, or sentences[75] .

2-Stopword Removal: Common words like "and," "the," and "in" have little meaning and might add noise to text mining analysis. Stop words should be eliminated in order to concentrate on more significant keywords and phrases and decrease the dimensionality of the text data[76] .

3-Stemming and Lemmatization: In order to analyze words with related meanings as a single entity, stemming and lemmatization are procedures used to reduce words

to their root forms or base forms. This approach may further decrease the dimensionality of the text data and increase the accuracy of text mining analysis [77][.

4-Text Representation: Utilizing text representation methodologies such as the bag-of-words model, term frequency-inverse document frequency (TF-IDF), or word embeddings, textual data undergoes a conversion into numerical representations that are readily comprehensible by machine learning algorithms. This transformation allows for efficient analysis and processing of text data within machine learning frameworks.

5-Text Mining Techniques: Text classification, sentiment analysis, topic modelling, named entity identification, and connection extraction are just a few of the many text mining methods that may be used to evaluate and extract insights from textual data [78]..

3.3.1.2 Relevance to Higher Education Research

Text mining may be used for several purposes when analyzing research trends in higher education, including [79]:

1-Content Analysis: To uncover trends, patterns, and emerging topics in higher education research, text mining may be used to evaluate the content of research journals, conference proceedings, and other text-based resources.

2-Citation Analysis: By collecting and analyzing the citation data from research papers, text mining may provide light on the importance and impact of authors, organizations, and publications in the field of higher education.

3-Collaboration Analysis: Text mining can be used to study collaboration patterns among researchers, institutions, and countries by analyzing co-authorship networks, affiliations, and acknowledgements in research publications.

4-Sentiment Analysis: By analyzing the sentiment expressed in research publications, reviews, and other text-based resources, text mining can provide insights into the attitudes, opinions, and emotions surrounding various topics and trends in higher education research.

5-Knowledge Discovery: Text mining can facilitate the discovery of novel insights and knowledge by uncovering previously hidden relationships, patterns, and trends in higher education research data.

3.3.2. Distinguishing Text Mining and Data Mining

Even though the terms "text mining" and "data mining" are sometimes used interchangeably, they refer to different techniques for drawing out important insights from various kinds of data [72]. Understanding the differences between these two methods, as well as their potential for integration, is essential for researchers and practitioners interested in analyzing research trends in higher education. In this section, we will clarify the differences between text mining and data mining and discuss the potential for combining both approaches in higher education research analysis [80].

3.3.2.1 Differences between Text Mining and Data Mining

1-Types of Data: The kind of data they manage is the key distinction between text mining and data mining. Unstructured textual data, such as research papers, reports, and other text-based resources, are the subject of text mining. Data mining, on the other hand, deals with structured or semi-structured data, such as databases, spreadsheets, or transaction records[81] .

2-Techniques Employed: Text mining uses techniques from NLP, computational linguistics, and machine learning to process and analyze textual data. Data mining employs methods from fields like computer science, statistics, and machine learning to uncover patterns and connections in structured data[82] .

3-Goals of Each Method: Text mining aims to extract valuable information, patterns, and insights from unstructured text data, transforming it into structured, machine-readable formats for further analysis. Data mining focuses on discovering patterns, trends, and relationships in structured data to support decision-making and knowledge discovery.

3.3.2.2 Potential for Combining Text Mining and Data Mining in Higher Education Research

Combining text mining and data mining approaches might be highly useful for analyzing research trends in higher education [83]:

1. **Comprehensive Analysis:** By mixing text mining with data mining, researchers may undertake a thorough study of research trends, including both textual content and structured data, such as citation patterns, cooperation networks, and bibliometric indicators.
2. **Enhanced Insights:** The combination of text mining and data mining can lead to enhanced insights by enabling the discovery of previously hidden patterns, relationships, and trends that might not be apparent when analyzing textual or structured data separately.
3. **Multimodal Analysis:** The analysis of numerous data types and sources, including research articles, conference proceedings, bibliometric data, and social media data, is made possible by the integration of text mining and data mining approaches. This results in a more comprehensive picture of research trends in higher education.
4. **Methodological Innovation:** By fostering the development of cutting-edge methods and tools suited to the unique requirements and difficulties of higher education research analysis, text mining and data mining techniques together may spur methodological innovation.

3.3.3. Text Mining Process Steps

Text mining is a powerful technique for extracting meaningful information from unstructured text data in higher education research [72]. Understanding the process in general, as well as the challenges and considerations at each level, is crucial for academics and practitioners seeking to fully grasp the promise of text mining. The text mining process will be described in this section, along with the difficulties and factors to take into account at each stage, especially in the context of higher education research analysis[84] .

1-Data Collection: Collecting relevant textual material for analysis is the initial stage in the text mining process. This may include research articles, conference proceedings, reports, or other text-based resources related to higher education research. Challenges and considerations during this step include obtaining access to the data, ensuring data quality and representativeness, and dealing with large volumes of data [81].

2-Preprocessing: Preprocessing involves preparing the raw text data for analysis by removing noise, inconsistencies, and irrelevant information. Common preprocessing tasks include tokenization, stop word removal, stemming and lemmatization, and language detection. Challenges and considerations in this step include selecting the appropriate preprocessing techniques, handling multilingual data, and preserving meaningful information while removing noise [85].

3-Feature Extraction: Feature extraction entails converting preprocessed text data into a structured, machine-readable format that text mining algorithms can examine. For this, methods like the bag-of-words model, term TF-IDF, and word embeddings are often used. Choosing the best text representation approach, handling high-dimensional data, and ensuring the interpretability of the extracted features are challenges and factors to be taken into account throughout this process [86].

4-Analysis: To find patterns, trends, and insights, multiple text mining algorithms are used for the retrieved characteristics during the analysis process. This stage may make use of methods including relationship extraction, topic modelling, named entity recognition, sentiment analysis, and text categorization. Challenges and considerations during the analysis step include selecting the appropriate text mining techniques, ensuring the validity and reliability of the results, and addressing potential biases or ethical concerns in the analysis.

5- Interpretation: The results of the text mining method must be interpreted in order to draw meaningful conclusions from them. This may involve visualizing the results, validating the findings with domain experts, and integrating the insights with other research evaluation methods, such as bibliometrics or expert opinion. Assuring the

validity and reliability of the results, addressing any biases, and expressing the findings in a clear and intelligible way are challenges and factors to be taken into account during the interpretation process.

3.4. Exploring Topic Modeling for Analyzing Research Trends in Higher Education

With big document collections like academic publications, conference proceedings, and reports, topic modelling is an effective publishing text-mining technique that may be used to discover hidden subjects and trends. Researchers may get important insights into the underlying themes and trends influencing the discipline by using topic modelling for higher education research data[87] [92, [93]. This section will present the concept of topic modelling, along with its core concepts, techniques, and algorithms, as well as how they might be used in the examination of research trends in higher education.

3.4.1. Topic Modeling Definition and Concept

Topic modelling is a text-mining technique that aims to find hidden themes and patterns in a collection of texts [88]. Circumventing the necessity for manual categorization or labelling, this unsupervised machine learning approach empowers researchers to grasp the organization and substance of extensive document assemblages. It autonomously discerns latent motifs and groups of correlated terms or notions embedded within the textual data, thus facilitating a comprehensive understanding of the content.

3.4.1.1 Key Concepts and Methods in Topic Modeling

1-Latent Topics: The fundamental tenet of topic modelling is that each document in a collection is made up of a range of latent subjects, which are shown as word distributions. Topic modelling aims to identify these latent themes and their relationships [89].

2-Algorithms: LDA, Non-negative Matrix Factorization (NMF), and Correlated Topic Models are three of the most used topic modelling techniques (CTM). The

choice of the method relies on the particular needs of the study and each algorithm has strengths and disadvantages [90].

3-Parameter Estimation: For topic modelling methodologies, a prevalent requirement includes the estimation of model parameters, encompassing the number of topics, topic allocations for individual documents, and word distributions for each theme. Techniques such as Expectation-Maximization (EM), Gibbs sampling, and Variational Inference are frequently employed to achieve accurate parameter estimation [90].

4-Model Evaluation: Evaluating the quality of a topic model is an essential step in the topic modelling process[91] . Common evaluation metrics include perplexity, coherence, and human interpretability. Model evaluation can also involve visualizations, such as topic-word distributions and document-topic distributions, to help researchers understand the structure and content of the topics discovered.

3.4.1.2 Relevance to Higher Education Research

Topic modelling has numerous applications in the analysis of research trends in higher education, including:

1-Content Analysis: Topic modelling may be used to analyze the content of research papers, conference proceedings, and other text-based resources to find trends, patterns, and novel themes in higher education research[92] .

2-Collaboration Analysis: By examining the distribution of topics across authors, institutions, and countries, topic modelling can provide insights into the patterns of collaboration and knowledge exchange in higher education research [92].

3-Research Evaluation: To give a more thorough picture of research trends and effects in higher education, topic modelling may be used in combination with other research assessment techniques, such as bibliometrics [92][98].

4-Policy and Decision-Making: The knowledge acquired by topic modelling may assist higher education institutions, funding organizations, and other stakeholders in

making decisions and forming research agendas, allocating resources, and identifying strategic priority areas[93] .

3.4.2. The Crucial Role of Topic Modeling

Researchers may get important insights into the dynamics of research trends in higher education by using topic modelling to analyze the underlying structure and topics in large-scale text collections. The significance of topic modelling in spotting new trends, interdisciplinary linkages, and developing research fields in higher education will be covered in this section [94]:

1- **Identifying Emerging Trends:** Topic modelling can help researchers identify emerging trends and research directions in higher education by uncovering latent topics in large document collections. This allows researchers to track the evolution of research themes over time and anticipate future developments in the field.

2- **Interdisciplinary Connections:** By analyzing the distribution of topics across different research areas, topic modelling can reveal interdisciplinary connections and facilitate knowledge exchange between fields. This can lead to the identification of novel research questions and the development of innovative solutions to pressing challenges in higher education.

3- **Evolving Research Areas:** Topic modelling enables researchers to monitor the evolution of research areas and identify shifts in the research landscape. Researchers may learn about the creation of new study fields, the fall of others, and the general trajectory of higher education research by evaluating the variations in subject prevalence over time.

4- **Enhancing Research Evaluation:** By giving more detailed knowledge of research trends and their effect on higher education, topic modelling may supplement conventional research assessment techniques like bibliometrics. Topic modelling may identify the underlying themes and patterns that guide

research efforts and progress knowledge in the subject by examining the content of research papers and other text-based resources.

5- **Informing Policy and Decision-Making:** The insights gained from topic modelling can be valuable for higher education institutions, funding agencies, and other stakeholders in shaping research agendas, allocating resources, and identifying areas of strategic importance. By understanding the dynamics of research trends and the connections between different research areas, policymakers and decision-makers can make more informed choices and contribute to the ongoing development of higher education.

3.4.3. General Process of Topic Modeling

Among huge collections of documents, such as research papers and conference proceedings in higher education, topic modelling is a potent tool for revealing hidden topics and patterns[95] . For researchers hoping to fully use topic modelling's potential, it is essential to comprehend the process in general as well as the difficulties and factors to be taken into account at each stage. In this section, we will outline the topic modelling technique and go through the challenges and considerations that must be made at each level, particularly in the context of higher education research analysis.

1- **Data Preprocessing:** Since it gets the raw text data ready for analysis, preprocessing is a crucial stage in the topic modelling process [96]. Common preprocessing tasks include tokenization, stop word removal, stemming and lemmatization, and language detection. Challenges and considerations during this step include selecting appropriate preprocessing techniques, handling multilingual data, and preserving meaningful information while removing noise.

2- **Model Selection:** Choosing the right topic modelling algorithm is essential for producing accurate and interpretable results. Researchers must consider factors such as the size and complexity of the dataset, the desired level of granularity, and the interpretability of the model. Common algorithms include LDA, NMF,

and CTM. Challenges and considerations during this step include balancing computational efficiency, model complexity, and interpretability [95].

3- **Model Training:** Once a suitable algorithm has been selected, researchers must train the model on the preprocessed data. Involving the calculation of model parameters, this process encompasses determining the overall count of topics, ascertaining topic allocations for every document, and deducing word distributions for each respective topic. This systematic approach allows for a thorough understanding of the underlying structure within the data. Parameter estimation is often performed using techniques such as Expectation-Maximization (EM), Gibbs sampling, or Variational Inference. Challenges and considerations during this step include selecting appropriate parameter estimation techniques, tuning hyperparameters, and avoiding overfitting or underfitting [97].

4- **Interpretation of Results:** Interpreting the findings and drawing important conclusions from the study is the last phase in the topic modelling process. This may involve visualizing topic-word distributions and document-topic distributions, validating the findings with domain experts, and integrating the insights with other research evaluation methods, such as bibliometrics or expert opinion. Assuring the validity and dependability of the results, addressing any biases, and clearly and concisely explaining the findings are challenges and factors to be taken into account throughout this stage.

3.4.4. Comparing Basic Techniques of Topic Models

Various topic modelling techniques have been developed to uncover latent topics and patterns within large collections of documents, such as research publications and conference proceedings in higher education. In this section, we will go over three essential topic model techniques: Latent Semantic Analysis (LSA) [98], Probabilistic Latent Semantic Analysis (PLSA) [99], and LDA [100]. In the framework of higher education research analysis, we will go through their parallels, distinctions, advantages, and disadvantages.

1-Latent Semantic Analysis (LSA): LSA is a linear algebra-based technique that captures the underlying semantic structure of the text data by utilizing singular value decomposition (SVD) to lower the dimensionality of a term-document matrix. LSA attempts to identify latent themes by identifying the essential components of the term-document matrix and assuming that papers with related subjects have similar word use patterns [98].

Strengths: LSA is computationally efficient, easy to implement, and provides interpretable results.

Limitations: LSA does not explicitly model the probability distributions of topics and words, which can limit its performance compared to more advanced probabilistic techniques.

2-Probabilistic Latent Semantic Analysis (PLSA): In PLSA, a probabilistic variant of LSA, the relationships between documents, words, and themes are described using a mixture model. Each topic in PLSA is represented as a probability distribution over words, and each document is anticipated to include a variety of latent themes. PLSA uses maximum likelihood estimation and the Expectation-Maximization (EM) approach to determine the model parameters.

Strengths: PLSA provides a probabilistic framework that can capture more complex relationships between documents, words, and topics compared to LSA.

Limitations: PLSA's usefulness in certain situations is limited by overfitting caused by the high number of parameters and the lack of a generative model for fresh documents.

3- Latent Dirichlet Allocation (LDA): LDA, a generative probabilistic model, enhances PLSA by including priors on the topic distributions for documents and the word distributions for subjects.

LDA posits that the generation of each document entails an initial sampling of a topic distribution, followed by the sampling of words from the identified topics. To estimate the model parameters, techniques such as Gibbs sampling or Variational Inference are commonly employed, ensuring accurate representation and analysis of the document corpus.

Strengths: LDA provides a fully generative model, allowing it to better generalize to new documents and avoid overfitting issues associated with PLSA. LDA is widely used and has demonstrated strong performance in various applications.

Limitations: LDA can be more computationally intensive than LSA and PLSA, and the choice of hyperparameters can have a significant impact on the results.

Table 3.2: Topic Modeling Techniques: Comparing LSA, PLSA, and LDA - Strengths and Limitations.

Table 3-2:Topic Modeling Techniques: Comparing LSA, PLSA, and LDA - Strengths and Limitations

Method	Description	Strengths	Limitations
LSA	Latent Semantic Analysis	LSA is computationally efficient, easy to implement, and provides interpretable results.	LSA does not explicitly model the probability distributions of topics and words, which can limit its performance compared to more advanced probabilistic techniques.
PLSA	Probabilistic Latent Semantic Analysis	PLSA provides a probabilistic framework that can capture more complex relationships between documents, words, and topics compared to LSA.	PLSA can suffer from overfitting due to a large number of parameters, and it does not provide a generative model for new documents, limiting its applicability in some scenarios.
LDA	Latent Dirichlet Allocation	LDA provides a fully generative model, allowing it to better generalize to new documents and avoid overfitting issues associated with PLSA. LDA is widely used and has demonstrated strong performance in various applications.	LDA can be more computationally intensive than LSA and PLSA, and the choice of hyperparameters can have a significant impact on the results.

- **Similarities and Differences**

Similarities:

1- All three techniques aim to discover latent topics within document collections.
2- They assume that documents with similar topics exhibit similar word usage patterns.

Differences:

1- LSA is a linear algebra-based method, while PLSA and LDA are probabilistic techniques.
2- LSA does not model probability distributions of topics and words, whereas PLSA and LDA do.
3- PLSA is prone to overfitting and does not provide a generative model for new documents, while LDA offers a generative model and better generalization.

Table 3-3:Comparison of Basic Topic Modeling Techniques: LSA, PLSA, and LDA

Technique	**Similarities**	**Differences**	**Limitations**
Latent Semantic Analysis (LSA)	1. All three techniques aim to discover latent topics within document collections. 2. They assume that documents with similar topics exhibit similar word usage patterns.	1. LSA is a linear algebra-based method, while PLSA and LDA are probabilistic techniques. 2. LSA does not model probability distributions of topics and words, whereas PLSA and LDA do.	1. LSA does not explicitly model the probability distributions of topics and words, which can limit its performance compared to more advanced probabilistic techniques.
Probabilistic Latent Semantic Analysis (PLSA)		3. PLSA is prone to overfitting and does not provide a generative model for new documents, while LDA offers a generative model and better generalization.	1. PLSA can suffer from overfitting due to a large number of parameters. 2. It does not provide a generative model for new documents, limiting its applicability in some scenarios.

Table 3-3: (Continue)

Latent Dirichlet Allocation (LDA)			1. LDA can be more computationally intensive than LSA and PLSA. 2. The choice of hyperparameters can have a significant impact on the results.

3.4.5. Latent Dirichlet Allocation (LDA)

3.4.5.1. LDA Model and Underlying Assumptions

LDA, a generative probabilistic model for topic modelling, aims to uncover latent topics within a text corpus. This model is constructed based on multiple fundamental assumptions:

1- A limited number of topics are combined in each document in the collection.
2- A probability distribution across a predetermined lexicon of words is used to represent each subject.
3- The word probabilities for each subject and the topical distributions for each text are determined via Dirichlet distributions.

3.4.5.2. Key Concepts in LDA

- Dirichlet Distribution [101]: The Dirichlet distribution is a continuous probability distribution over probability vectors, which serves as the prior distribution for the topic proportions in a document and the word probabilities for each topic in LDA. It is parameterized by a set of positive real numbers called concentration parameters, which influence the shape and dispersion of the distribution.

- Generative Process: LDA assumes a generative process for creating each document in the collection:

- From a Dirichlet distribution, choose the subject proportions for the document.
- For each word in the document:

a. Select a topic according to the topic proportions.

b. Choose a word from the selected topic's word distribution.

3.4.5.3. Strengths and Limitations of LDA

- **Strengths**

1. LDA provides a fully generative model, which allows it to generalize well to new documents and avoid overfitting issues associated with some other topic modelling techniques.
2. LDA is widely used and has demonstrated strong performance in various applications, including the analysis of research trends in higher education.

- **Limitations**

1. LDA may need more computing power than certain other topic modelling techniques, like LSA and PLSA.
2. The choice of hyperparameters for the Dirichlet distributions can have a significant impact on the results, and their selection requires domain expertise or experimentation.

3.4.5.4. Applications of LDA in Higher Education Research Analysis

LDA has been widely applied in the analysis of research trends in higher education, including [102]:

1. Identifying research themes and trends: LDA can help uncover the main research topics and their evolution over time, providing insights into the development of the higher education field.
2. Analyzing interdisciplinary research: By discovering connections between topics, LDA can reveal interdisciplinary research areas and potential avenues for collaboration between different research communities.
3. Investigating research impact: By analyzing topic distributions in highly cited papers, LDA can help identify influential research areas and assess the impact of various research topics on the higher education field.
4. Informing research policy and funding decisions: LDA can provide valuable insights into the current state of higher education research, helping funding

agencies and policymakers make informed decisions regarding research priorities and resource allocation.

CHAPTER 4

METHODOLOGY

4.1. Introduction

This thesis examined the potential applications of text mining techniques in the field of education, particularly the use of topic-modelling techniques like LDA, to extract pertinent information from academic articles in higher education. The following chapter describes the methodology used in this study to address the research topics outlined in the introduction. The application of text mining techniques in the field of higher education was also investigated utilizing a range of analytical methods, with a focus on employing LDA topic modelling to extract important information from academic publications in this area. The study technique included a sequential analytical procedure that included both content analysis and bibliometric analysis. While the bibliometric analysis was performed to assess the state of the higher education domain, LDA topic modelling was used to identify current trends and subjects in research publications related to higher education. In the end, this research process enabled the researcher to fully understand the study challenges and assess the research objectives.

4.2. Research Methodology

After the literature review, a data collection process was carried out, which included the selection of relevant academic publications from various databases. The chosen articles were then pre-processed, and the LDA text-mining algorithm, is used to extract topics and trends. Bibliometric analytic techniques were also used to is used to view the descriptive data of the status of the higher education industry. The results were eventually analyzed and reported in the findings chapter, where conclusions were drawn in light of the information learned from the investigation.

4.2.1 Subject Comprehension

To identify the numerous terminology and expressions used to define research trends in the academic literature on higher education, a literature survey was conducted before the start of this study. Analytical techniques used to analyze research trends in higher education and their accompanying statistical patterns were also evaluated as part of the evaluation. In Chapter 2 of the research, this procedure is thoroughly detailed. The study question changed after the literature review to focus on examining research trends in higher education. To help with this endeavour, a domain expert was recruited to choose the best keywords for pulling out pertinent articles from the literature. The chosen keywords were picked for their applicability to the study issue and were utilized to locate relevant literature. This procedure assisted in concentrating the analysis on the studies that had the most relevance and influence in the area of higher education. Figure 4.1 proposes an insight into the steps that have been followed in this study.

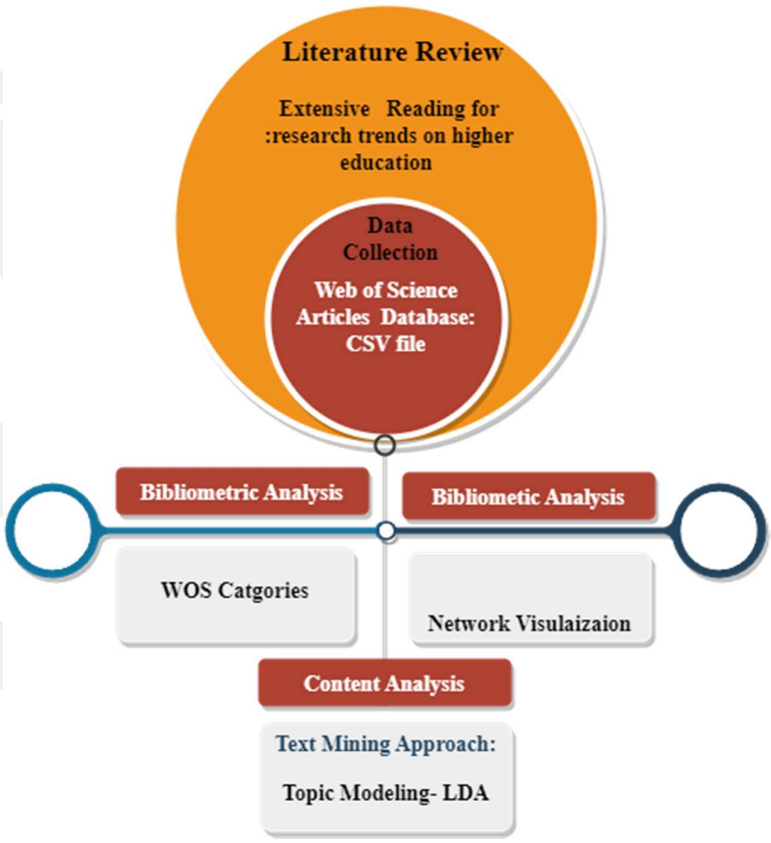

Figure 4-1:The Flow of The Research.

4.2.2 Data Collection/Data Extraction

The "Web of Science" database was used for this study because of its enormous collection of worldwide peer-reviewed papers, that counted over 68,000 papers. The database contains the number of indexes covering a wide range of subject areas, including the Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (AHCI), and Emerging Sources Citation Index (ESCI). Therefore, researchers studying a wide range of topics may gain a lot from the database. Hence, the following search query string was constructed using the identified keywords:

TITLE-ABS-KEY (“higher education" OR "Tertiary education" OR "university education")AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar"))AND (LIMIT-TO (LANGUAGE , "English")).

To locate pertinent peer-reviewed papers that meet the objectives and scope of this research, 68,459 related articles are downloaded from the Web of Science database. To obtain approximately 70,000 research papers from the Web of Science for the given query, an API is required to access the data. Unfortunately, Web of Science does not offer a public API; however, they do provide a "Web of Science Web Services Lite (WOSWSL)" API for their subscribers. Due to technical challenges that hindered acquiring this API, the data was downloaded manually, 10,000 documents at a time.
Quality journal articles and conference papers published in English were found in the Web of Science database with 68,459 results. The 45-year period covered by the publications was from May 2023 to 1978.
The careful cleaning procedures used during this part of the investigation included the removal of instances with "nan" (no available abstracts) labels and the eradication of duplicate items with the same abstracts. These measures were taken to ensure the dataset's quality and accuracy and to prevent any redundancy or duplication in the subsequent text-mining analysis.

4.2.3. Data Analysis

This study employed two analytical methodologies, bibliometric analysis and content analysis, to analyze the unique aspects of the dataset and generate relevant insights that would aid in solving the research challenges. The dataset was collected and preprocessed.

4.2.3.1 In Bibliometric Analysis

i. **Jupyter Notebook**[103] : The publishing analysis in the sector of higher education was completed using Jupyter Notebook [103]. Using an open-source web tool, users may create and share documents containing live code, math, graphics, and narrative content. Collaboration and repeatable interactive data processing, model creation and testing, and discovery discussion are all available in Jupyter Notebook. Among the supported programming languages are Python, R, and Julia. Jupyter Notebook is used extensively in scientific computing, data research, and education due to its adaptability and a broad variety of applications. The Python module was used to analyze higher education publications for this investigation.

ii. **Orange Datamining software** [104] The WOS dataset was examined in this thesis to draw conclusions and learn new things. To find important information, make wise judgments, and create prediction models to address certain challenges or concerns in research, analyzing a dataset corpus requires evaluating a collection of data in various formats, such as text, photographs, or numerical data. To find patterns, trends, and correlations in the data, this process may include several phases, such as data cleansing, exploratory data analysis, and the use of statistical and machine learning methods. Insights into the dataset and logical inferences are frequently the objectives of dataset corpus analysis. As will be mentioned in chapter 5, orange was utilized in this study to show unrelated papers and to identify whether the

downloaded documents from WOS journals are connected. A variety of statistical and machine learning methods are supported by Orange, a platform for data visualization and analysis. Users can interact with data, develop and evaluate models, and disseminate outcomes in a cooperative and reproducible fashion. This process fosters an environment conducive to effective collaboration and knowledge sharing.

4.2.3.2 Content Analysis Method

The broad text mining and topic modelling concepts discussed in Chapter 3 served as the foundation for the Content Analysis method used in this thesis. The text mining process was applied to the raw dataset using the LDA topic modelling approach. Using the Python Jupyter Notebook coding environment, the Gensim Library was utilized to construct the LDA technique. Unsupervised topic modelling was performed using Gensim, a popular open-source NLP tool. It handles a range of challenging problems by using modern statistical machine-learning algorithms. Gensim sets itself apart from rival topic modelling software by efficiently managing large text files without the need to load the entire document into memory, offering straightforward manual labelling and document tagging procedures, and offering reliable tools for developing and assessing excellent topic models and text preprocessing[105] .

4.2.4. Data Preprocessing

Reducing the dimensionality of text data and filtering out any extraneous noise are the main goals of text cleaning and preprocessing to enhance the content analysis process. This step is crucial since it speeds up computing and gets the data ready for qualitative analysis. Several processes are implemented to accomplish this objective, such as eliminating unnecessary attributes from the dataset and retaining only essential ones like Article Title, Source Title, Language, Document Type, Abstract, Publication Year, WoS Categories, Web of Science Index, and Research Areas that can assist in analyzing the articles as data samples. Additional tasks involve removing duplicate

rows with identical abstracts, numbers, regular expressions, punctuation, and whitespace, phrases like "no abstract available" and "introduction," substituting missing information with empty strings, and converting text to lowercase. Subsequently, an Excel file containing the preprocessed data is generated and preserved.

Once the preprocessed dataset has been handled, additional preparation is required before executing LDA analysis. The dataset should exclusively include words with a length between four and fifteen characters. The majority of the world, and, and, and even the future in and the epoch of the world in a novel age of the world. There are numerous data and data worldwide. a chart with data. Bigrams represent any two commonly appearing words in the corpus, whereas Trigrams signify any three. Phrase modelling is also employed to generate Bigram and Trigram models. These processes are essential for getting the dataset ready for LDA analysis, which will be covered in more detail in the next chapters.

To enhance the text data's quality, stop words like "a," "an," "the," "for," and "as" are removed. Following then, lemmatization is used to save just important phrases in the form of nouns, adjectives, verbs, or adverbs using predefined functions.

The next stage of text analysis, Data Transformation, entails vectorizing the documents by creating a Corpus and Dictionary. By applying the "id2word" function to the lemmatized data, the Dictionary gives a list of unique IDs for each term. Using the "id2word.doc2bow" function, the Term Document Frequency for each term is determined. Hence, the final corpus will be a mapping of each (word id, word frequency).

4.3. LDA Model Implementation and Fitting

The model requires training, evaluation, and presentation of the resulting topics using the multi-step LDA implementation procedure. Constructing an initial LDA model in the primary phase entails selecting suitable parameters, including the topic count, pass count, and *alpha* and *eta* hyperparameters. Topic Coherence Analysis serves as an assessment method for topic models, designed to determine the ideal quantity of topics for LDA-based content analysis. In this study, the optimal LDA parameter values—alpha and eta (both 0.0053)—were used to perform individual Coherence Scores and

Dominant Topic evaluations for each topic distribution, spanning from 5 to 30 topics (or beta). The investigation determined that 19 topics would be the ideal number. The study employed the Parallelized Latent Dirichlet Allocation (ldamulticore) from Python's Gensim library, leveraging all accessible CPU cores to expedite and parallelize the model training process.

The coherence score of the produced themes is used to assess the model's performance in the second stage. The model's accuracy in capturing the key themes and topics found in the text data is improved with the aid of this evaluation.

The last stage is to create a visually appealing representation of the generated topics. This may be done by producing visuals, such as word clouds and bar graphs, which are essential for conveying the findings from the LDA study to stakeholders and decision-makers.

In conclusion, the creation and assessment of models for the LDA implementation process must be done carefully and iteratively. To create an LDA model that properly captures the primary themes and subjects in the text data, perform the following technique

CHAPTER 5

RESULTS

This chapter presents the results of the analyses performed in response to the research questions outlined in Chapter 1. This study's major objective is to use dimensionality reduction and linear models for classification techniques to look at academic articles on the Web of Science (WOS) in the context of higher education. The two major portions of the results are bibliometric analysis and topic modelling analysis.

The bibliometric analysis aims to provide insights into the statistical distributions of WOS publications in higher education, including aspects such as the connection diagrams among publications by showcasing the co-occurrence of Author Keywords and collaborative efforts between Authors, Nations, and Institutions. Using Jupyter notebook, these networks use co-word analysis approaches while accounting for the volume of texts and their citations.

The topic modelling analysis, on the other hand, outlines the subjects and topics that have been the focus of studies as well as the evolution of research topics through time to identify research patterns and trends in the area of higher education.

This chapter presents the results in-depth, offering the findings with pertinent visualizations and tables to support them. To make linkages between the findings and their significance for the area of higher education research, these results will be further examined in the next chapter.

5.1 Bibliometric Analysis Results in Higher Education

By using python libraries and Jupyter notebook, the results of bibliometric analysis are obtained as will be shown in the following subsections.

5.1.1 The Distribution of Documents by Year

By using python Jupyter notebook, the Figure 5.1 is obtained. The line graph in Figure 5.1 provides crucial information on past patterns in publications in the field of higher education. It emphasizes the gradual increase of articles published over time, showing the growth and evolution of this area of research. The consistent volume of publications from 1975 to 1980-1981 points to a rather slow initial growth of research in this field. Nonetheless, the ensuing constant increase in publications published indicates the increased interest in and dedication to higher education research.

Since 2006, there has been a steady rise in publications, which is particularly noteworthy because it represents a large growth in higher education-related research activities. This trend persisted for more than ten years, with the number of papers steadily rising and reaching its peak in 2019–2020 with around 70,000 publications.

Despite this continued growth, the decline in publication counts shown in 2022 suggests a probable change in the research objectives in the area of higher education. The fluctuation in publication numbers highlights the necessity for continual inquiry and analysis to stay current with changes and fads in higher education, even if it is too soon to make definite conclusions about probable future trends in this field.

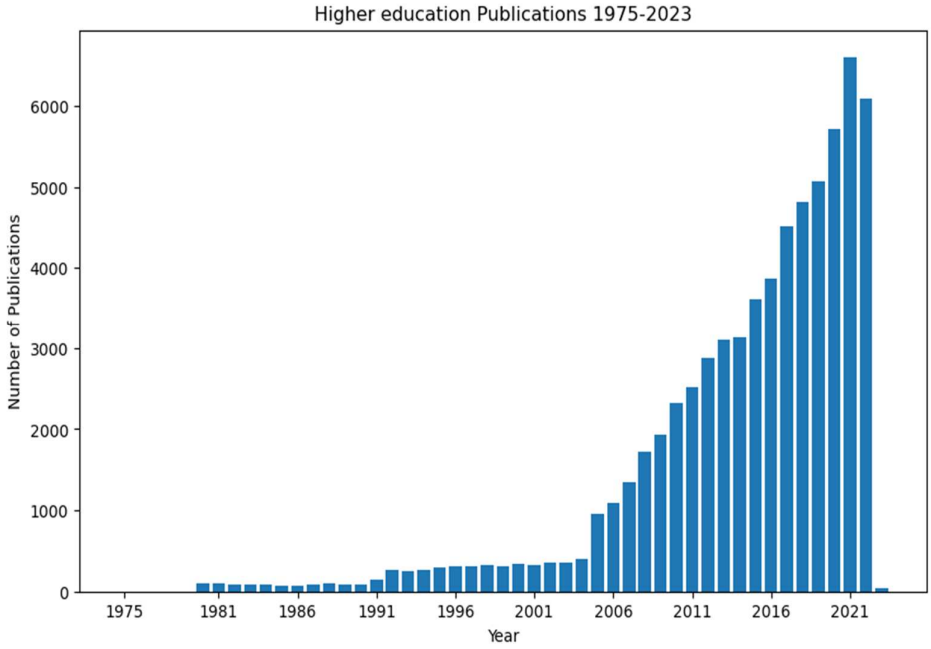

Figure 5-1: The Publications in Higher Education by Year

5.1.2. Distribution of Documents Across Subject Areas in Higher Education

Figure 5.2, and Table 5.1 are derived from Web of Science data (1975 - early 2023) with the help of jupyter notebook. The figure illustrates the distribution of documents by subject within the higher education field. Table 5.1 provides a detailed breakdown of these subject areas. Based on Figure 5.2 and Table 5.1, in general, Education & Educational Research is the most prominent subject in the WOS data under consideration accounting for 81.99% of publications, signifying the field's primary focus on educational research. The interdisciplinary character of higher education research is apparent, with various combinations of Education & Educational Research and other disciplines appearing in the table. The second largest subject area, Education & Educational Research; Psychology, constitutes a significantly smaller proportion (2.92%) than the first. The remaining subject areas each makeup less than 2% of the distribution, with interdisciplinary pairings such as Education & Educational Research; Linguistics (1.91%) and Business & Economics; Education & Educational Research (1.86%) being more prevalent.

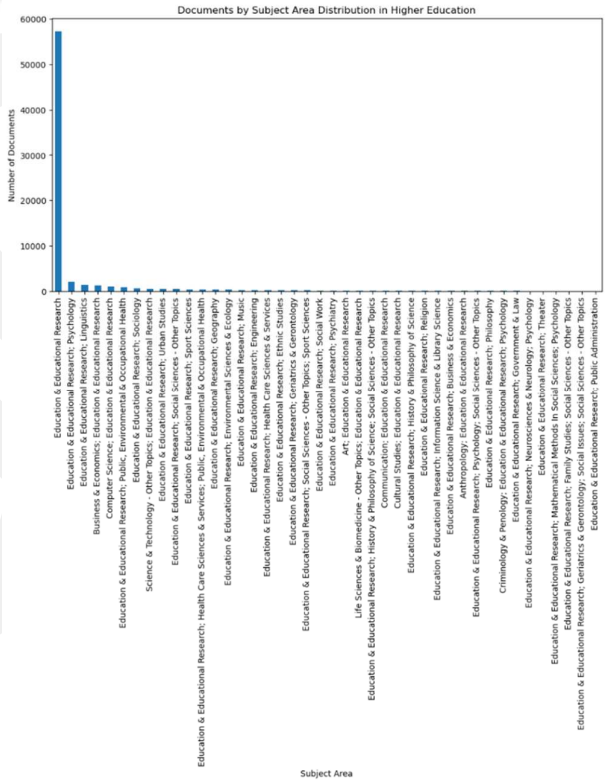

Figure 5-2:The Distribution of Documents By Subject in The Field of Higher Education.

In contrast, the least represented subject area is Anthropology; Education & Educational Research, at just 0.10% of publications.

This implies a limited incorporation of anthropological perspectives in higher education research. However, it is important to recognize that the value of a subject area should not be determined solely by its prevalence, as diverse perspectives, including anthropological insights, can contribute to a deeper understanding and progress in higher education research.

Table 5-1:The Detailed Distribution of Documents by Subject in the Field of Higher Education

Seq.	Subject Area	%
0	Education & Educational Research	81.98
1	Education & Educational Research; Psychology	2.91
2	Education & Educational Research; Linguistics	1.90
3	Business & Economics; Education & Educational Research	1.86
4	Computer Science; Education & Educational Research	1.47
5	Education & Educational Research; Public, Environmental & Occupational Health	1.25
6	Education & Educational Research; Sociology	0.86
7	Science & Technology - Other Topics; Education & Educational Research	0.728
8	Education & Educational Research; Urban Studies	0.65
9	Education & Educational Research; Social Sciences - Other Topics	0.62
10	Education & Educational Research; Sport Sciences	0.56
11	Education & Educational Research; Health Care Sciences & Services; Public, Environmental & Occupational Health	0.53
12	Education & Educational Research; Geography	0.47
13	Education & Educational Research; Environmental Sciences & Ecology	0.41
14	Education & Educational Research; Music	0.31
15	Education & Educational Research; Engineering	0.31
16	Education & Educational Research; Health Care Sciences & Services	0.27
17	Education & Educational Research; Ethnic Studies	0.27
18	Education & Educational Research; Geriatrics & Gerontology	0.23
19	Education & Educational Research; Social Sciences - Other Topics; Sport Sciences	0.23
20	Education & Educational Research; Social Work	0.19
21	Education & Educational Research; Psychiatry	0.17
22	Art; Education & Educational Research	0.17

Table 5 1:(Continued)

23	Life Sciences & Biomedicine - Other Topics; Education & Educational Research	0.16
24	Education & Educational Research; History & Philosophy of Science; Social Sciences - Other Topics	0.15
25	Communication; Education & Educational Research	0.14
26	Cultural Studies; Education & Educational Research	0.14
27	Education & Educational Research; History & Philosophy of Science	0.14
28	Education & Educational Research; Religion	0.14
29	Education & Educational Research; Information Science & Library Science	0.13
30	Education & Educational Research; Business & Economics	0.11
31	Anthropology; Education & Educational Research	0.10

5.1.3. The Distribution of Higher Education Documents Based on Type

By using Jupyter notebook the Figure 5.3 is obtained. Figure 5.3 illustrates the distribution of documents by type in the field of higher education. Table 5.2 provides a comprehensive breakdown of the document types and their corresponding percentages. The three most significant document types are Article (94.22%), Article; Early Access (4.80%), and Article; Proceedings Paper (0.96%).

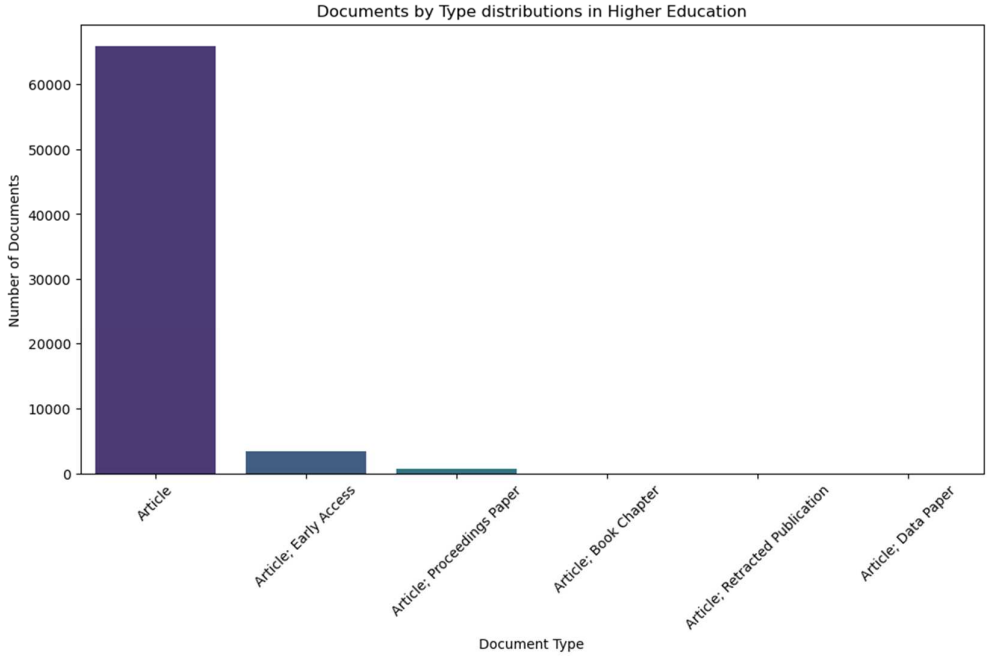

Figure 5-3:Distribution of Higher Education Documents Based on Their Type.

The substantial proportion of Articles underscores their predominant role as the primary medium for disseminating research findings in the higher education field. Early Access Articles, which constitute the second-largest category, allow for the rapid dissemination of research results, offering researchers the opportunity to access and build upon current findings. The third most important document type, Article; Proceedings Paper, combines conference proceedings with research articles, highlighting the value of conferences as a platform for sharing and discussing research within the field.

Conversely, the least represented document type is Article; Data Paper, with only two instances (0.003%). This low prevalence suggests that higher education research may not frequently prioritize the publication of data-focused papers.

Table 5-2:The Details Distribution of Higher Education Documents Based on Type.

Document Type	Number of Documents	%
Article	65872	94.22
Article; Early Access	3354	4.80
Article; Proceedings Paper	672	0.96
Article; Book Chapter	13	0.01
Article; Retracted Publication	3	0.00
Article; Data Paper	2	0.00

5.1.4 Documents by Country/Territory Distribution in Higher Education

Figure 5.4, using WOS data (1975 - early 2023), presents the distribution of higher education publications by the top 20 countries, with details in Table 5.3. The top three countries, England (20.33%), Australia (15.94%), and the People's Republic of China (7.42%), demonstrate a strong emphasis on higher education research. England's dominance is likely due to its renowned academic institutions and extensive research infrastructure.

Australia, the second-largest contributor, highlights the region's importance in higher education research. China's ranking showcases its growing influence in global higher education research. Conversely, Portugal, with the lowest representation (1.71%),

implies a lesser focus or limited resources for higher education research compared to more dominant countries.

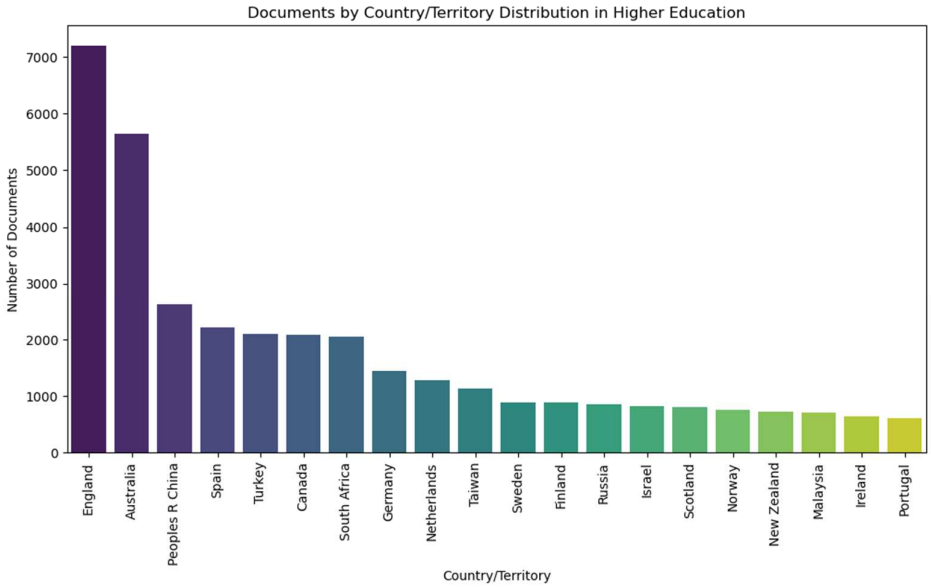

Figure 5-4: The Distribution of Higher Education Publications Based On Counties.

Table 5-3:The Detailed Distribution of Higher Education Publications Based on Counties.

Country	**No. of Documents**	**%**
England	7197	20.32
Australia	5642	15.93
Peoples R China	2626	7.41
Spain	2214	6.25
Turkey	2093	5.91
Canada	2078	5.86
South Africa	2039	5.75
Germany	1438	4.06
Netherlands	1284	3.62
Taiwan	1129	3.18
Sweden	891	2.51
Finland	882	2.49
Russia	851	2.40
Israel	813	2.29
Scotland	806	2.27
Norway	749	2.11

Table 5.3: (Continued)

New Zealand	722	2.03
Malaysia	710	2.00
Ireland	638	1.80
Portugal	604	1.70

5.1.5 Documents by Citations Distribution in Higher Education

Figure 5.5 presents a line graph illustrating the distribution of higher education citations by country, accompanied by numerical details in Table 5.4. This visualization emphasizes the varying contributions each country has made to the global discourse on higher education.

England leads with a remarkable 128,158 citations, showcasing its strong influence and commitment to higher education research. Australia follows closely with 93,635 citations, reflecting the country's significant contributions to the field. Peoples R China, ranking third with 30,372 citations, highlights its growing presence in higher education research.

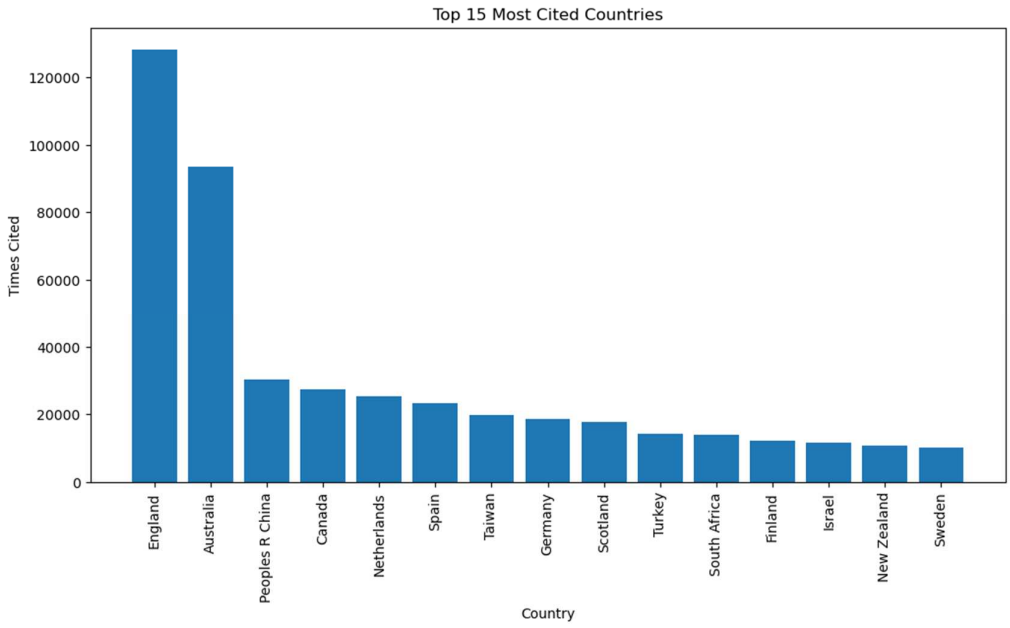

Figure 5-5:Distribution of Higher Education Citations by Countries

The graph also reveals a notable decrease in citations as we move from the top three countries to those ranked 4th to 15th. The remaining countries, such as Canada, the

Netherlands, Spain, and Taiwan, contribute a considerable amount to the total citations, demonstrating their dedication to higher education research.

Countries like Sweden, New Zealand, Israel, Finland, South Africa, and Turkey also make important contributions, emphasizing the global interest in higher education research and development.

Table 5-4: Documents by Citations Distribution in Higher Education

Country	No. of Citations
England	128158
Australia	93635
Peoples R China	30372
Canada	27432
Netherlands	25268
Spain	23274
Taiwan	19898
Germany	18640
Scotland	17866
Turkey	14346
South Africa	13996
Finland	12200
Israel	11668
New Zealand	10813
Sweden	10138

5.1.6. The Distribution of higher education documents based on Affiliation/Institution

Figure 5.6 displays the distribution of the top 20 higher education documents based on Affiliation/Institution, spanning 1975 to early 2023. Analyzing Table 5.5, the three most significant contributors to higher education publications are the University of California System (2.006%), the University of London (1.887%), and the University of North Carolina (1.727%).

The University of California System's leading position showcases its strong research capabilities and commitment to higher education. The University of London follows closely, reflecting its well-established academic reputation and extensive resources.

The University of North Carolina, ranking third, highlights its considerable research focus within higher education.

In contrast, the institution with the lowest percentage of higher education publications is the University of Toronto (0.779%). This may indicate that the University of Toronto has a relatively smaller focus on higher education research compared to other institutions.

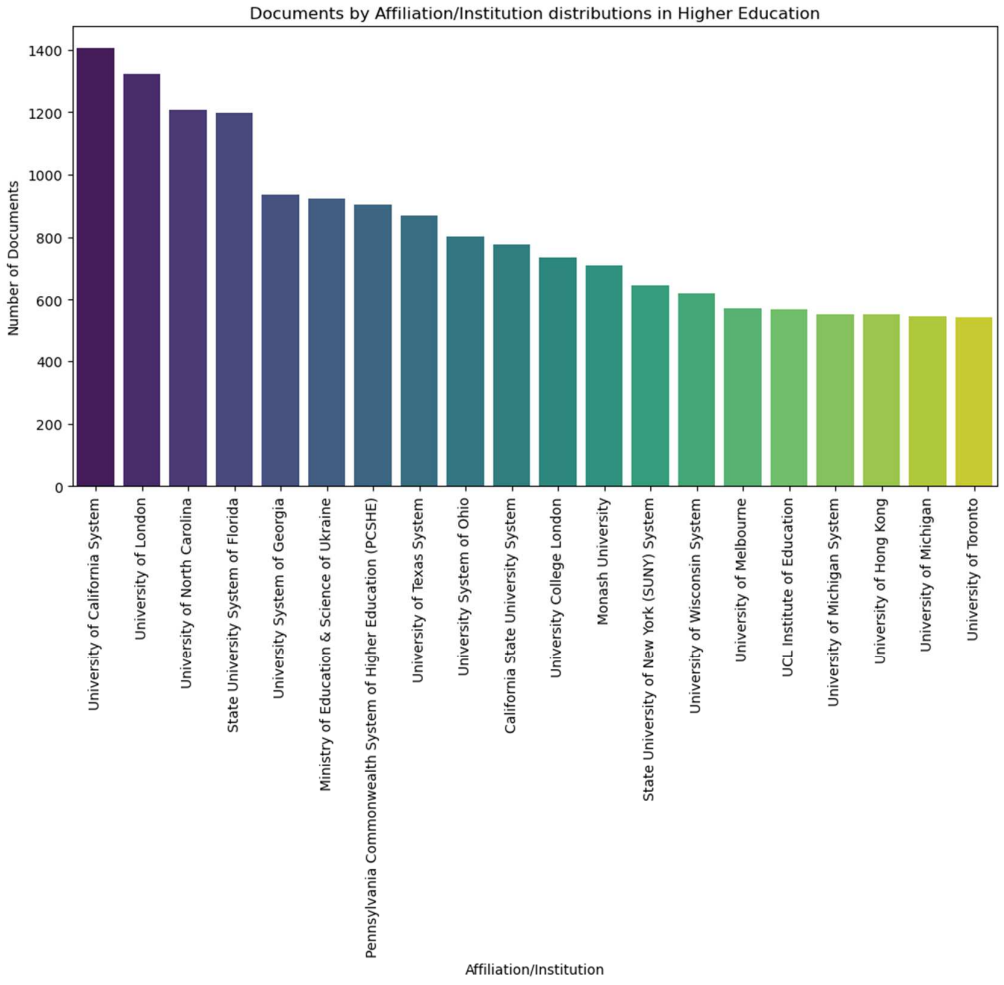

Figure 5-6:The Distribution of Higher Education Documents Based on The Top 20 Affiliation/Institution.

Table 5-5:The Numerical Details Distribution of Higher Education Documents Based on Affiliation/Institution

University/Institution	Count	%
The University of California System	1404	2.00
University of London	1321	1.88
University of North Carolina	1209	1.72
State University System of Florida	1197	1.71
University System of Georgia	937	1.33
Ministry of Education & Science of Ukraine	924	1.32
Pennsylvania Commonwealth System of Higher Education	903	1.29
The University of Texas System	870	1.24
University System of Ohio	803	1.14
California State University System	776	1.10
University College London	736	1.05
Monash University	711	1.01
State University of New York (SUNY) System	645	0.92
University of Wisconsin System	619	0.88
University of Melbourne	573	0.81
UCL Institute of Education	568	0.81
The University of Michigan System	554	0.79
University of Hong Kong	552	0.78
University of Michigan	547	0.78
University of Toronto	545	0.77

5.1.7. Co-occurrence of Author Keywords

Figure 5.7 presents a map of author keywords' co-occurrences for higher education publications from 1975 to early 2023. This Figure offers insights into the most frequently mentioned and interconnected subject areas in the field. The analysis of the top 20 author keywords, which are detailed in Table 5.6.

From the Figure, it is evident that the most dominant subject area in higher education publications is higher education itself, with 12111 occurrences and a link strength of 3458. Education and assessment follow as the second and third most important subject areas, respectively. These three subject areas represent a significant percentage of the overall publications in the field of higher education. On the other hand, the least important subject area among the top 20 author keywords is diversity, with 551 occurrences and a link strength of 281. This suggests that diversity is a less commonly studied topic in higher education publications, compared to the more dominant subject

areas. The figure provides valuable insights into the research trends in higher education. The dominance of higher education, education, and assessment highlights the importance of these areas in the field. However, it also suggests that there is room for growth and exploration in less commonly studied areas, such as diversity.

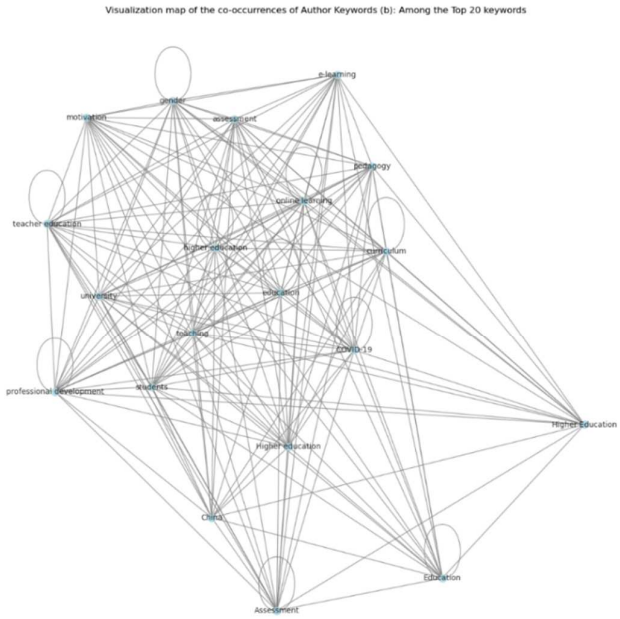

Figure 5-7: Map of Author Keywords' Co-Occurrences Among the Top 20 Author Keywords.

Table 5-6:Top 20 Author Keywords in Higher Education and Their Co-Occurrences.

Author Keywords	Occurrences	Link Strength
higher education	12111	3458
education	1532	540
assessment	1091	546
teacher education	1044	442
gender	921	421
online learning	893	724
e-learning	834	595
motivation	816	404
professional development	815	429
COVID-19	757	597
curriculum	741	396
students	704	501
pedagogy	676	442

Table 5-6: (Continued)

learning	653	570
university	630	390
teaching	607	502
distance education	579	394
blended learning	568	407
international students	563	225
diversity	551	281

5.1.8 Co-Authorship of Authors according to the Number of Papers

Table 5.7 showcases the top 20 co-authors in higher education based on the number of papers they have contributed to. This ranking serves as an indicator of their productivity and prominence within the field.

At the top of the list is Lee, J, with an impressive 132 documents to their name. Kim, S follows closely with 118 papers, while Henderson, C takes the third position with 113 documents. Liu, SY and Martin, L are not far behind, with 112 and 110 papers, respectively.

The remaining authors in the top 20 demonstrate a more gradual decline in the number of published documents. For instance, Hwang, GJ has contributed to 107 papers, while Zhang, Y has 106. Several authors, such as Jimenez, A and Burchinal, M, are tied at 104 documents. The list concludes with Wang, J and Gomez, J, each having 87 papers to their credit.

Overall, Table 5.6 effectively highlights the most prolific co-authors in the realm of higher education, offering insight into their scholarly contributions and impact on the field

Table 5-7:The Top 20 Co-Authors by The Number of Papers in Higher Education

Author Name	**No. of Documents**
Lee, J	132
Kim, S	118
Henderson, C	113

Table 5-7: (Continued)

Liu, SY	112
Martin, L	110
Hwang, GJ	107
Zhang, Y	106
Jimenez, A	104
Burchinal, M	104
Kim	101
Smith, S	96
Hernandez, C	96
Sharma, R	96
Lavonen, J	94
Tsai, CC	94
Campbell, M	90
Zhang, L	90
Gasevic, D	89
Wang, J	87
Gomez, J	87

5.1.9 Country Co-Authorship Based on Citation Count

Figure 5.8 presents a pie chart that illustrates the distribution of co-authorship by authors from the top 20 countries in higher education, as measured by the number of citations. The data in the chart can be further analyzed regarding Table 5.8, which provides specific numerical details and corresponding percentages for each country.

England leads the pack with a substantial 536,102 citations, accounting for 6.15% of the total. Australia closely follows, contributing 5.57% with 485,700 citations. Peoples R China occupies the third position with 257,858 citations, making up 2.96% of the total. The remaining countries in the top five include Spain (215,811 citations, 2.48%) and Canada (200,530 citations, 2.30%).

Countries ranking 6th to 20th exhibit a more gradual decrease in their respective shares. Notably, Turkey, Germany, the Netherlands, South Africa, and Taiwan each contribute over 1% to the total citations. The list concludes with Belgium, holding a 0.63% share with 55,167 citations. Overall, the pie chart and table effectively demonstrate the distribution of co-authorship in higher education among the top 20 countries by citation count.

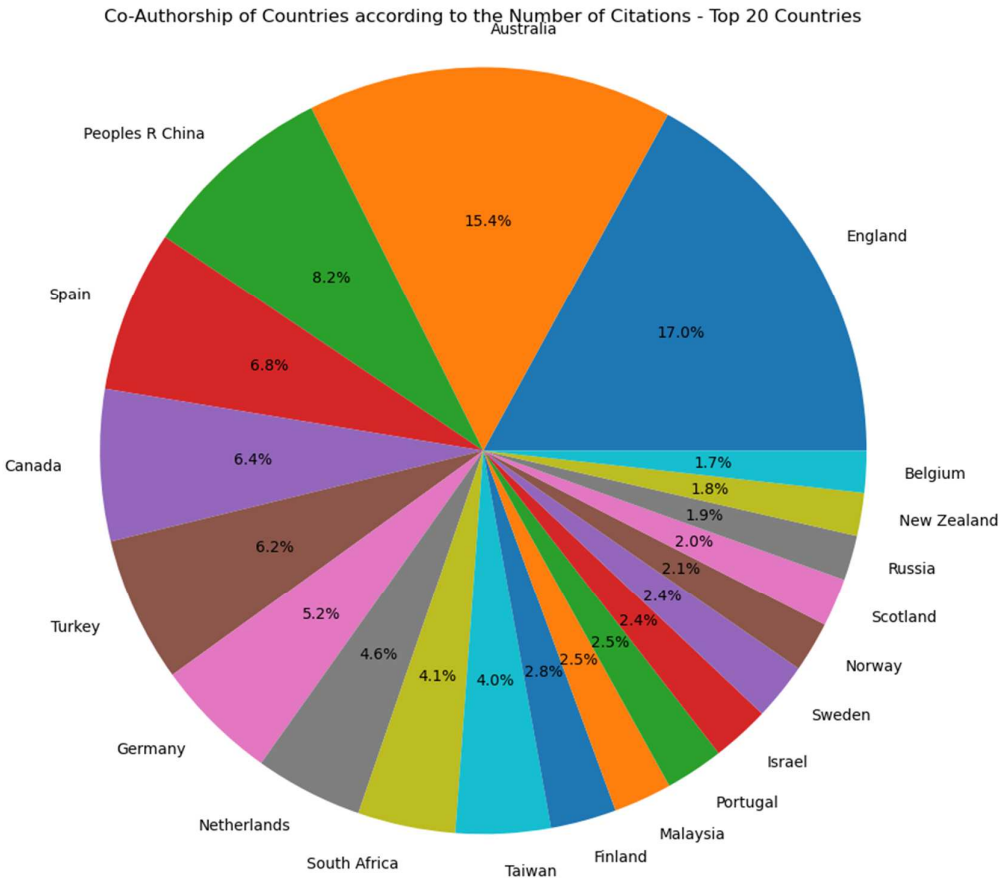

Figure 5-8: A Pie Chart of The Top 20 Countries' Co-Authorship Of Authors By The Number of Citations.

Table 5-8:The Top 20 Co-Authors Based on The Number of Citations in Higher Education.

Country	No. of Citations	Percentage
England	536102	6.15
Australia	485700	5.57
Peoples R China	257858	2.95
Spain	215811	2.47
Canada	200530	2.30
Turkey	194198	2.22
Germany	164448	1.88
Netherlands	143550	1.64
South Africa	129998	1.49
Taiwan	125959	1.44

Table 5 8:(Continued)

Malaysia	77511	0.88
Portugal	77384	0.88
Israel	76882	0.88
Sweden	76015	0.87
Norway	67080	0.76
Scotland	64526	0.74
Russia	59873	0.68
New Zealand	56928	0.65
Belgium	55167	0.63

5.1.10 Organizations Co-Authorship Based on Citation Count

Both Figure 5.9 and Figure 5.10 have been obtained by python jupyter notebook. These Figures illustrate the distribution of co-authorship among the top 20 organizations in higher education by the number of citations. The corresponding numerical details and percentages are provided in Table 5.9. The University of California System leads the chart with 29,213 citations, accounting for 9.52% of the total share.

The University of London follows closely with 27,080 citations, contributing to 8.83% of the total. The University of North Carolina ranks third, holding 20,711 citations or 6.75% of the total. Other prominent organizations in the top 20 include the Pennsylvania Commonwealth System of Higher Education (18,998 citations, 6.19%), the State University System of Florida (17,341 citations, 5.65%), and the University of Michigan System (16,851 citations, 5.49%).

The list continues with institutions such as the University of Texas System, the University System of Ohio, and the University System of Georgia, all contributing between 4.82% and 4.59% to the total citations. The top 20 concludes with the University of Sydney, which holds 10,518 citations, making up 3.43% of the total. Overall, the pie chart and table effectively demonstrate the distribution of co-authorship among leading organizations in higher education, highlighting their respective contributions and impact in the field.

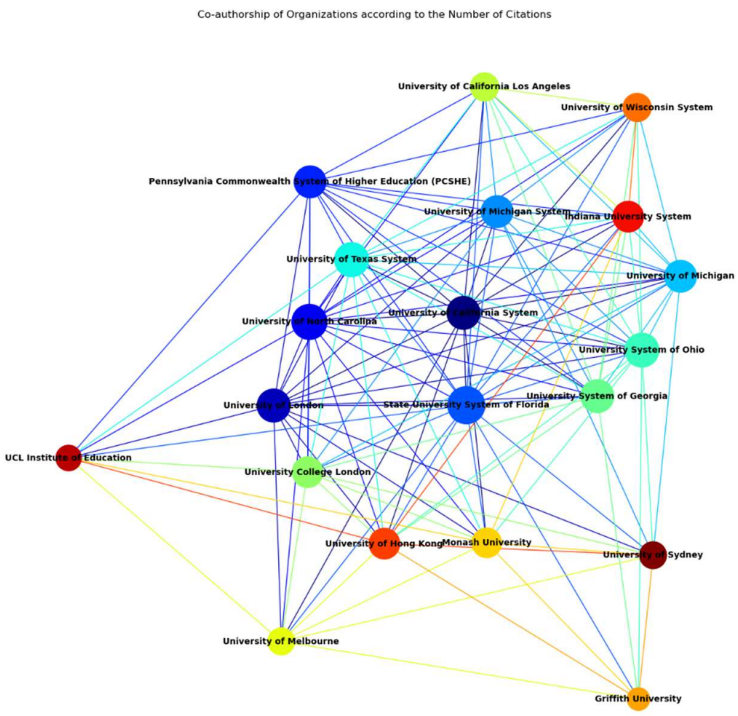

Figure 5-9: Map of the Co-authorship of Organizations by the Number of Citations in higher education.

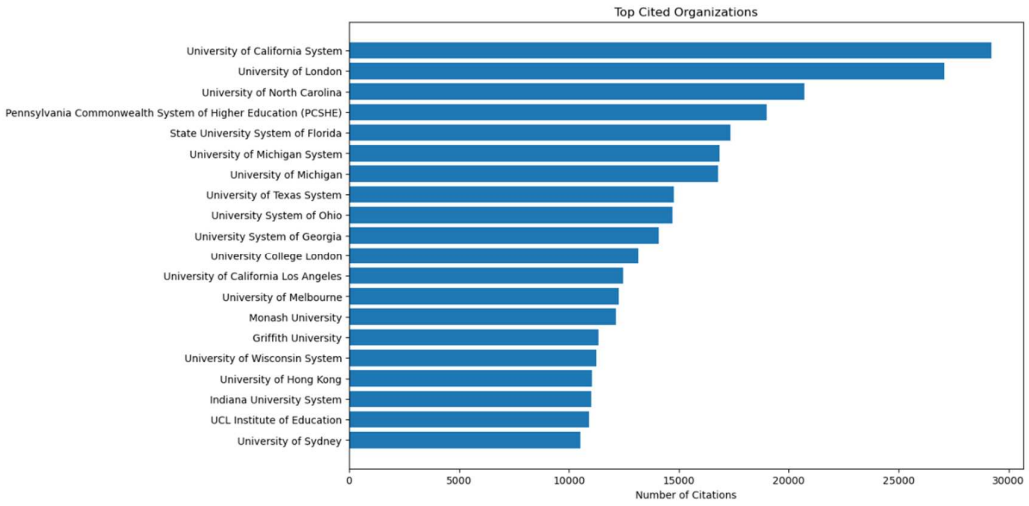

Figure 5-10: Co-authorship of the Top 20 Organizations by Citations in Higher Education.

Table 5-9: Table of Co-Authorship of The Top 20 Organizations by Citations in Higher Education

	Organization	Citations	%
0	The University of California System	29213	9.52
1	University of London	27080	8.83
2	University of North Carolina	20711	6.75
3	Pennsylvania Commonwealth System of Higher Education	18998	6.19
4	State University System of Florida	17341	5.65
5	The University of Michigan System	16851	5.49
6	University of Michigan	16796	5.48
7	The University of Texas System	14786	4.82
8	University System of Ohio	14717	4.80
9	University System of Georgia	14091	4.59
10	University College London	13147	4.29
11	The University of California Los Angeles	12475	4.07
12	University of Melbourne	12263	4.00
13	Monash University	12138	3.96
14	Griffith University	11356	3.70
15	University of Wisconsin System	11251	3.67
16	University of Hong Kong	11049	3.60
17	Indiana University System	11019	3.59
18	UCL Institute of Education	10923	3.56
19	University of Sydney	10518	3.43

5.2. Topic Modeling Results

In the initial hand study of the obtained CSV data set, it became clear that higher education research had a sluggish start, with only a few publications in the late 1970s. Only one piece was released annually in 1975, 1977, and 1978, with two articles being released in 1979. Nevertheless, this tendency was reversed in later years as the number of publications on the subject rapidly grew. The Web of Science data base's material was evaluated using LDA and Orange Data Mining software. However, The Orange tool is employed to ensure that each research paper falls within the scope of the study. However, in this thesis, all the downloaded papers were found to be pertinent. The Orange tool was utilized exclusively for this specific case., as shown in Figure 5.11. The Topic Modeling is conducting using Jupyter Notebook only. These results show that the raw data set used in this study includes all publications about trends in higher education research from 1975 to January 2023. This sizable data collection provides insightful historical patterns and topics in higher education research, enabling scholars

to recognize and assess changes and advances that have taken place in this area throughout time.

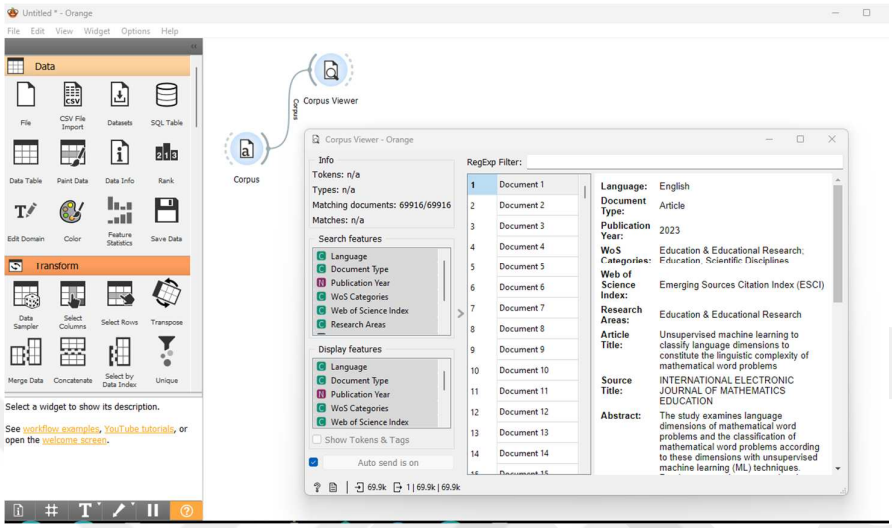

Figure 5-11: Analyzing Dataset Corpus in Orange Data Mining software [106].

By analyzing this data, it is possible to identify opportunities for further research and exploration as well as dig further into the important issues and themes that have inspired higher education research.

5.2.1 Descriptive Content Analysis Results

The original data collection included 69,916 items before text mining techniques like filtering and preprocessing were applied. This large data collection, which covered the period from 1975 to January 2023, contained a wealth of information concerning developments in higher education research. Nonetheless, certain measures were made to eliminate redundant and unnecessary data to assure data accuracy. The "NaN" values in the abstracts were first removed, and then duplicate abstracts that had been published in numerous places were taken down.

Using a Jupyter Notebook, several text preparation operations were performed to prepare the data set for the LDA method. These measures included doing descriptive content analysis and topic modelling on the corpus material. These techniques allow academics to identify major themes and patterns in contemporary higher education research, providing priceless insights into the development of the discipline. The

results of this analysis will be essential in influencing the future of higher education research and pointing researchers in new directions for discovery and study.

5.2.1.1 Documents Trend by Year Distribution (1975 - February 2023)

Table 5.10 displays the research trend in higher education by analyzing the distribution of published papers from 1975 to January 2023. Due to the presence of *NaN* values, missing data and mixed types of data, the original count of 69,916 documents were reduced to 66,562. The table highlights a consistent increase in the number of documents over the years, with percentages ranging from 0.001502% in the 1970s to a peak of 9.92% in 2021. The data reveals an upward trend in higher education research, with a noticeable acceleration in publication rates from 2005 onwards. The table provides a comprehensive overview of the growth and interest in higher education research over the past decades, culminating in a total of 100% of the analyzed documents.

Table 5-10:Papers in Higher Education Research Trend by Year Distribution (1975 – January 2023)

Year	Count	%
1975	1	0.00
1977	1	0.00
1978	1	0.00
1979	2	0.00
1980	98	0.14
1981	105	0.15
1982	93	0.13
1983	90	0.13
1984	83	0.12
1985	75	0.11
1986	75	0.11
1987	90	0.13
1988	99	0.14
1989	88	0.13
1990	93	0.13
1991	145	0.21
1992	260	0.39
1993	253	0.38

Table 5-10: (Continued)

1994	259	0.38
1995	302	0.45
1996	307	0.46
1997	309	0.46
1998	322	0.48
1999	309	0.46
2000	338	0.50
2001	332	0.49
2002	353	0.53
2003	361	0.54
2004	400	0.60
2005	963	1.44
2006	1092	1.64
2007	1346	2.02
2008	1721	2.58
2009	1939	2.91
2010	2317	3.48
2011	2522	3.78
2012	2874	4.31
2013	3096	4.65
2014	3139	4.71
2015	3607	5.41
2016	3869	5.81
2017	4513	6.78
2018	4816	7.23
2019	5070	7.61
2020	5713	8.58
2021	6603	9.92
2022	6084	9.14
2023 January	34	0.05
Sum	**66562**	**100**

5.2.1.2 Journals Predominantly Publishing Papers on Higher Education

The field of higher education research has experienced significant growth in recent years, with numerous academic journals emerging to address the diverse aspects of this discipline. This discussion focuses on the 10 journals listed in Table 5.11 and shown in Figure 5.12, which predominantly publish papers on higher education. These

journals vary in their publication frequency, research focus, and impact on the academic community.

1- **Higher Education:** With the highest publication frequency at 2998, Higher Education is a leading international journal in the field. It covers a broad range of topics, from policy and management to teaching and learning. The journal aims to promote a comprehensive understanding of higher education and contribute to its advancement worldwide.

2- **Studies in Higher Education:** With a publication frequency of 1471, Studies in Higher Education is another prominent journal in the field. This journal primarily focuses on theoretical and empirical research, aiming to enhance the understanding of higher education and facilitate its development.

3- **BMC Medical Education:** Though not exclusively focused on higher education, BMC Medical Education has a publication frequency of 1200 and is dedicated to publishing research on medical education at all levels, including undergraduate, postgraduate, and continuing education.

4- **Computers & Education:** This journal, with a publication frequency of 952, explores the role of technology and digital tools in education. It addresses the impact of technological advancements on teaching and learning practices in higher education institutions.

5- **Higher Education Research & Development:** With a publication frequency of 823, this journal focuses on research and development in higher education. It covers various topics, including curriculum design, teaching methodologies, and institutional management.

6- **Teaching in Higher Education:** With a publication frequency of 799, Teaching in Higher Education is a journal that addresses pedagogical issues and practices in higher education. It aims to promote effective teaching and improve the quality of learning in higher education institutions.

7- **Education Sciences:** With a similar publication frequency of 798, Education Sciences is an interdisciplinary journal that covers a wide range of topics related to education, including higher education research, policy, and practice.
8- **Education and Information Technologies:** This journal, with a publication frequency of 683, focuses on the intersection of education and information technology. It explores the use and impact of technology in higher education settings.
9- **Journal of Further and Higher Education:** With a publication frequency of 638, this journal covers research on further and higher education, addressing topics such as policy, curriculum, teaching, and learning.
10- **Economics of Education Review:** With a publication frequency of 636, the Economics of Education Review is a journal that explores the economic aspects of education, including higher education. It examines the relationship between education and economic growth, as well as the role of education in labour markets and public policy.

In conclusion, these ten journals play a significant role in advancing research and knowledge in the field of higher education. They vary in focus, from pedagogy and curriculum development to policy and economics, offering a comprehensive view of the multifaceted nature of higher education. Researchers and practitioners can benefit from these journals by staying informed of the latest developments and trends in higher education research.

Table 5-11: Top 10 Journals Predominantly Publishing Papers on Higher Education

Journal	Frequency
Higher Education	2998
Studies In Higher Education	1471
Bmc Medical Education	1200
Computers & Education	952
Higher Education Research & Development	823
Teaching In Higher Education	799
Education Sciences	798
Education And Information Technologies	683
Journal Of Further And Higher Education	638
Economics Of Education Review	636

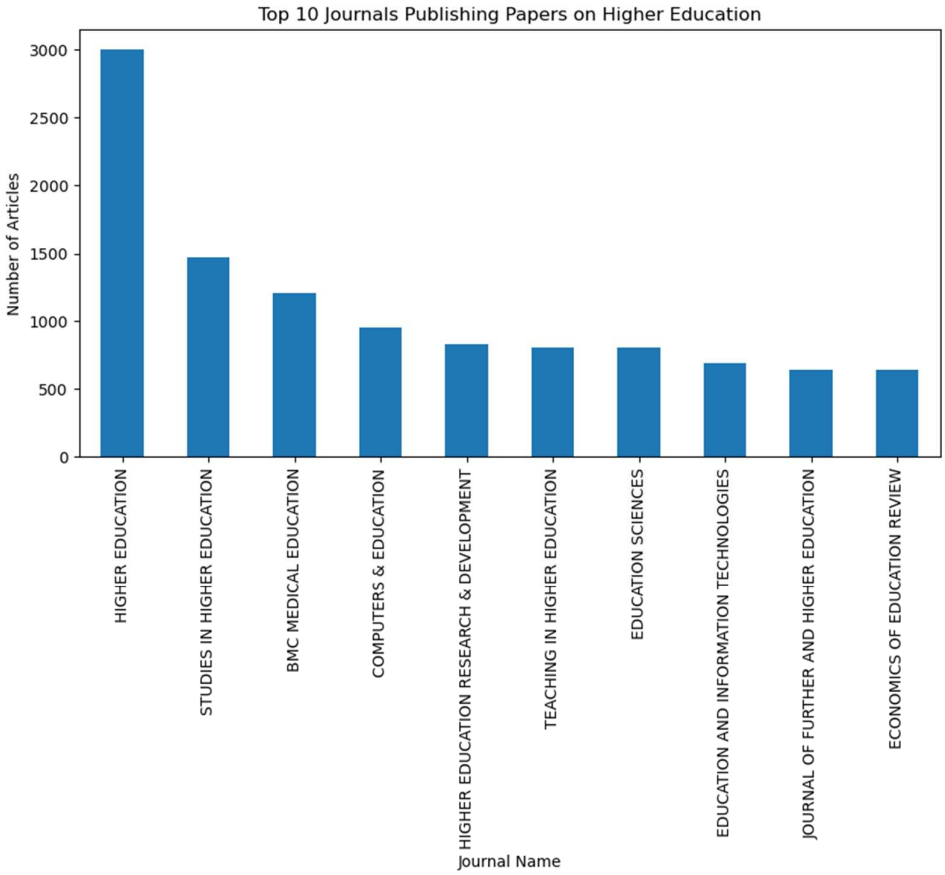

Figure 5-12:Top 10 Journals on Higher Education Publications.

5.2.2 Most Frequently Utilized Words in WOS Higher Education Articles

The "Most Frequently Utilized Words in WOS Higher Education Articles" section presents the results of the analysis of the most commonly used words in the corpus. The analysis shows the top terms and their relative weights, measured by TF-IDF. The visualization of the results is presented in Figure 5.13 as a Word Cloud, which shows the size of each term concerning its weight compared to other terms in the entire corpus. Table 5.12, give deeper insights into the results.

According to the analysis, the term 'higher education' appears to be the most frequent among the top terms, followed by 'study,' 'curriculum,' 'educational,' 'teacher,' and 'university.' These results are in line with the expected focus of research in the higher education domain, and they reflect the most frequently addressed and examined topics in this field.

Figure 5-13: Word Cloud Illustrating Key Terms in Higher Education from (1975-January 2023).

The Word Cloud visualization provides an effective way to quickly identify the most important terms in the corpus and their relative importance. It can be useful for researchers and practitioners to identify the key topics and concepts that are being discussed in the field and to gain a better understanding of the research trends and directions. the table of 10 top words in higher education from 1975 to January 2023 highlights some interesting insights.

The word "higher education" has the highest frequency, appearing 12964 times, which makes up 2.52% of the total data. This emphasizes the importance of higher education in the field.

The second most frequent word, "student," appears 5078 times, which shows the central role that students play in the higher education system. The words "teaching," "learning," and "university" also appear frequently, indicating the focus on education and instructional practices in higher education. The words "impact," "role," and "development" show the emphasis on the effects and outcomes of higher education on individuals and society.

Table 5-12: The 10 Top Words in Higher Education (1975-Jeuary 2023).

Word	Count	%
higher education	12964	2.52
student	5078	0.99
teaching	3418	0.66
learning	2915	0.57
university	2798	0.54
Effect	2496	0.49
High School	2296	0.45
impact	2177	0.42
role	1977	0.38
development	1938	0.38

5.2.3 Highly Cited Journals in the Web of Science Index

Table 5.13 presents the highly cited journals in the Web of Science Index for the field of higher education. All the journals listed have been indexed in the Social Science Citation Index, indicating their significance and impact in the field. The inclusion of these journals in the index implies that they have passed the rigorous standards of scientific publishing and have been cited by a considerable number of articles. It is noteworthy that the journals listed in the table cover a wide range of topics related to higher education, including research and development, sustainability, policy, science education, evaluation, and open and distributed learning. The inclusion of such diverse topics highlights the interdisciplinary nature of the field of higher education and the importance of research in addressing the multifaceted challenges facing higher education.

Overall, the table provides useful information for scholars and researchers interested in exploring the latest developments and trends in the field of higher education. The highly cited journals listed in the table can serve as a valuable resource for academics and researchers seeking to publish their work and stay up-to-date with the latest research findings in the field.

Table 5-13:Top Cited Journals in The Web of Science Index.

	Journal Name	Number of Articles	%
0	Higher Education	2998	4.29
1	Studies In Higher Education	1471	2.10
2	Bmc Medical Education	1200	1.72
3	Computers & Education	952	1.36
4	Higher Education Research & Development	823	1.18
5	Teaching In Higher Education	799	1.14
6	Education Sciences	798	1.14
7	Education And Information Technologies	683	0.98
8	Journal Of Further And Higher Education	638	0.91
9	Economics Of Education Review	636	0.91

5.2.4 The Most Common Words

Figure 5.14 offers valuable insights into the key themes and focus areas within the research field. Analyzing the top three most common words allows us to better understand the primary subjects and concerns in higher education research. To eliminate stop words, in the Python code, we employ the *nltk library's stopwords* list to filter out frequently used English words, leading to a more insightful visualization. Moreover, the function *WordNetLemmatizer,* is used to get more accurate root forms for words.

Figure 5.14 presents the ten most frequently occurring words in abstracts of higher education research papers. It is worth noting that the term "student" appears to be the most prevalent, with a frequency of 116,424, which underscores the centrality of students in the higher education discourse. The prominence of the word "education" (82,054 occurrences) and "higher" (64,847 occurrences) can be expected, as these words directly relate to the context of higher education.

Furthermore, the terms "study" and "learning" feature prominently, with 72,887 and 59,090 instances, respectively. These words reflect the primary focus of higher education research, which seeks to enhance the understanding of educational processes and learning outcomes. Additionally, the high frequency of "research" (44,333

occurrences) highlights the integral role of empirical investigations in driving advancements within this academic domain.

Institutions, such as "school" (38,116 occurrences) and "university" (37,818 occurrences), are also well-represented in the list of common words, as they form the backbone of the higher education system. Similarly, the presence of "teacher" (37,604 occurrences) emphasizes the vital contribution of educators in shaping the student learning experience. Lastly, the term "academic" (30,199 occurrences) serves as a reminder of the scholarly nature of higher education research and the importance of intellectual rigour in this field.

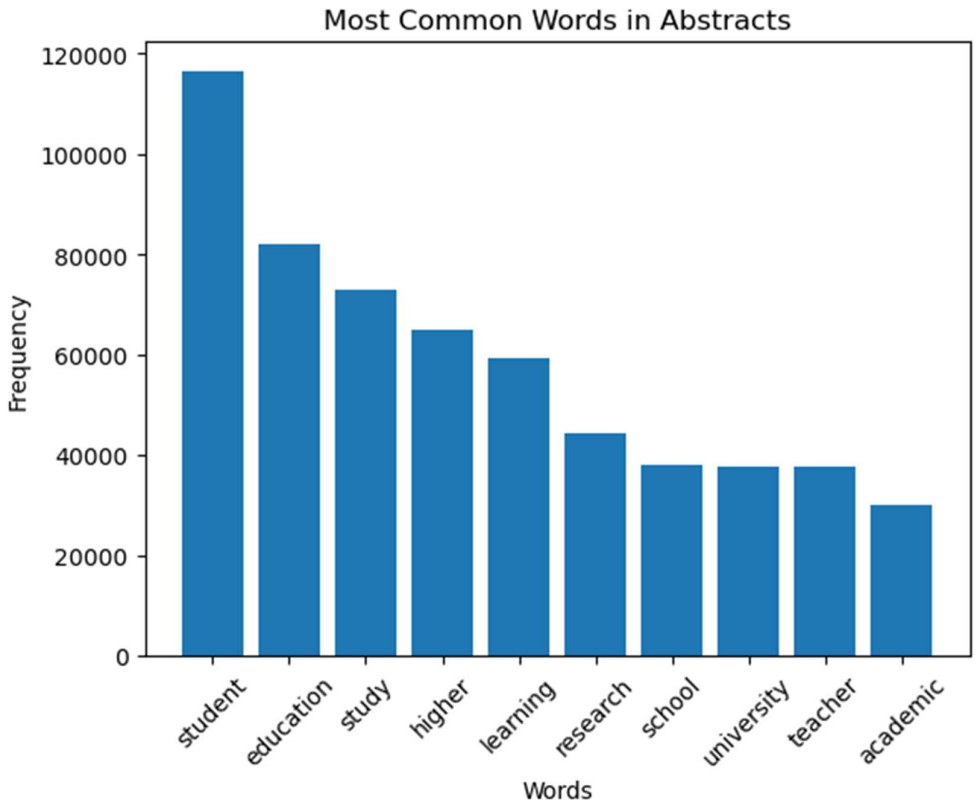

Figure 5-14:Most 10 Common Words in Higher Education Papers Abstracts.

5.4 Topic Modeling Analysis Results and Content Analysis using LDA

This section presents the results of the Topic Modeling analysis conducted on the WOS higher education dataset. The analysis was performed using Latent Dirichlet Allocation (LDA), a popular algorithm used for discovering topics within a set of documents. The main objective of this analysis was to uncover the main research trends, subjects, and frequently addressed topics within the field of higher education. The section presents the most prominent topics identified by the LDA model, along with the keywords that represent each topic. The section also explores the evolution of research topics over time, providing insights into how research in the field of higher education has evolved and changed. The results of this analysis provide valuable insights for researchers and policymakers looking to understand the current state of research in higher education and identify potential areas for future research.

5.4.1. Coherence Analysis and Dominant Topic Examination of Topics

Topic Coherence Analysis is an assessment technique used in topic models to establis h the ideal number of topics for LDA content analysis. The optimum values of the LD A parameters, *alpha* and *eta*, were used in this study's Coherence Score and Dominan t Topic analyses (shown in Figures 5.15 and 5.16) for each topic distribution spannin g from 5 to 30 topics (or *beta*). The investigation led to the conclusion that 19 topics would be the ideal number. To determine the optimal number of topics for LDA cont ent analysis, we performed a Coherence Score analysis for each topic distribution, ranging from 5 to 30 topics. The Coherence Score is a measure of how semantically similar the top words within a topic area are, and higher coherence scores generally indicate a more interpretable topic model. By examining and reviewing the Coherence *Sc-ore plot* (Figure 5.15) and *Dominant Topic plot* (Figure 5.16) for various numbers of topics, we observed that the coherence score was the highest or showed high improvement when the number of topics was set to 19. This suggests that having 19 topics provided the best balance between model interpretability and the granularity of the to pics, leading to the conclusion that 19 is the ideal number of topics for our LDA model.

It is important to highlight that the choice of the number of topics may be influenced by domain knowledge, interpretability of the results, and the specific goals of the research, which should be considered in conjunction with the coherence scores.

The Parallelized Latent Dirichlet Allocation (ldamulticore) from Python Gensim package was used in this work to parallelize and speed up model's training process[107]. Figure 5.17 shows the LDA Content Analysis Dominant Topic analysis.

Optimal LDA parameters calculation for Alpha and Eta for 19 topics:

eta = alpha = (1/Number of Topics); Therefore, alpha = eta = 1/19 = **0.0053**

```
Number of topics: 5  Coherence Score: 0.2980229621260011
Number of topics: 6  Coherence Score: 0.301789974135165
Number of topics: 7  Coherence Score: 0.30263095428445963
Number of topics: 8  Coherence Score: 0.31033564552862786
Number of topics: 9  Coherence Score: 0.3326318247456974
Number of topics: 10  Coherence Score: 0.29972943975513305
Number of topics: 11  Coherence Score: 0.31650390302494913
Number of topics: 12  Coherence Score: 0.3241535927556059
Number of topics: 13  Coherence Score: 0.38601818370968843
Number of topics: 14  Coherence Score: 0.38811779794383827
Number of topics: 15  Coherence Score: 0.3949002856957913
Number of topics: 16  Coherence Score: 0.3702529766226579
Number of topics: 17  Coherence Score: 0.38192737627530177
Number of topics: 18  Coherence Score: 0.39331100255275586
Number of topics: 19  Coherence Score: 0.39846282976264397
Number of topics: 20  Coherence Score: 0.38551251662501834
Number of topics: 21  Coherence Score: 0.4071714832281346
Number of topics: 22  Coherence Score: 0.3841890953920434
Number of topics: 23  Coherence Score: 0.3983820677096173
Number of topics: 24  Coherence Score: 0.3852165583552945
Number of topics: 25  Coherence Score: 0.3722571392273052
Number of topics: 26  Coherence Score: 0.3756513160001741
Number of topics: 27  Coherence Score: 0.39317327964402937
Number of topics: 28  Coherence Score: 0.39377057922196956
Number of topics: 29  Coherence Score: 0.409237162624061
Number of topics: 30  Coherence Score: 0.39679997354453683
```

Figure 5-15:Snippet of The Topics And Their Coherence Score.

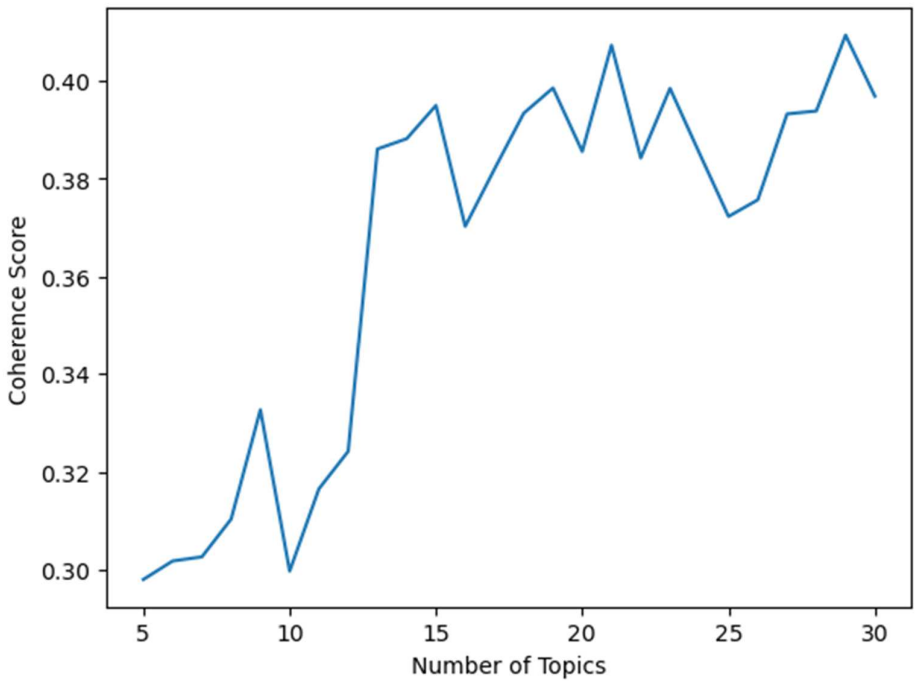

Figure 5-16:LDA Content Analysis Coherence Score Graph and Code Results.

10 topics

Dominant Topic	Count
0	1248
1	13819
2	4168
3	7140
4	13720
5	6313
6	3248
7	7493
8	3556
9	7510

11 topics

Dominant Topic	Count
0	1484
1	11779
2	3678
3	5970
4	12313
5	5571
6	3261
7	6960
8	3120
9	7376
10	6703

12 topics

Dominant Topic	Count
0	1443
1	10820
2	3378
3	6229
4	11163
5	5314
6	3339
7	6829
8	2997
9	7341
10	7012
11	2350

13 topics

Dominant Topic	Count
0	1098
1	10274
2	2908
3	5534
4	10092
5	5058
6	2929
7	6861
8	2426
9	7156
10	5501
11	2173
12	6205

14 topics

Dominant Topic	Count
0	1109
1	10109
2	3006
3	4564
4	9229
5	2824
6	2310
7	5727
8	2032
9	7028
10	6208
11	2279
12	5631
13	6159

15 topics

Dominant Topic	Count
0	950
1	10413
2	1078
3	4010
4	8489
5	3043
6	2135
7	5600
8	1962
9	7008
10	5217
11	2295
12	5816
13	6084
14	4115

(a)

16 topics

Dominant Topic	Count
0	962
1	10144
2	1050
3	3458
4	8192
5	2834
6	2739
7	4513
8	2032
9	6939
10	5647
11	2578
12	5554
13	5961
14	4112
15	1500

17 topics

Dominant Topic	Count
0	907
1	10014
2	1091
3	2791
4	5967
5	2443
6	2585
7	2890
8	2066
9	6425
10	5759
11	2204
12	4887
13	4653
14	3818
15	1447
16	8268

18 topics

Dominant Topic	Count
0	779
1	9797
2	1182
3	2807
4	6664
5	2516
6	2728
7	2982
8	1902
9	6434
10	5384
11	2347
12	4687
13	4127
14	3797
15	1418
16	7931
17	733

19 topics

Dominant Topic	Count
0	759
1	9344
2	1118
3	2595
4	5754
5	2633
6	2420
7	2882
8	1748
9	6376
10	5997
11	2414
12	4477
13	4071
14	3584
15	1384
16	7435
17	1123
18	2101

20 topics

Dominant Topic	Count
0	804
1	9214
2	1027
3	2935
4	6336
5	2354
6	1994
7	3271
8	1647
9	6231
10	5255
11	2067
12	4525
13	3795
14	3450
15	1462
16	7217
17	945
18	2129
19	1557

21 topics

Dominant Topic	Count
0	749
1	9007
2	1197
3	2700
4	5621
5	2504
6	1897
7	3111
8	1577
9	3963
10	5654
11	1993
12	4321
13	3587
14	3688
15	1384
16	7229
17	1034
18	1996
19	1440
20	3563

22 topics

Dominant Topic	Count
0	825
1	8812
2	1137
3	2759
4	6103
5	2040
6	2128
7	2949
8	825
9	4307
10	5413
11	2260
12	4401
13	4295
14	3383
15	1411
16	7063
17	426
18	2074
19	1423
20	2956
21	1225

(b)

(c)

Figure 5-17:LDA Content Analysis Dominant Topic Analysis shown in (a), (b) and (c).

5.4.2 LDA Topics Extraction by Dominant Topic Analysis

Figure Table 5.14 shows a table created by the appropriate code cell, displaying the key phrases of the 19 prominent topics identified by LDA. Table 5.14 shows Word Cloud representations of the LDA subjects with

Table 5-14:The Tabular Output of Dominant Topic Analysis Results.

T No.	Topic Top-Rated Keywords	Topic Label	No. of Papers	%
0	education, student, practice, learning, experience, article, higher, research, paper, within, way, approach, context, teaching, social	Topic 6	9051	13.26
1	education, higher, policy, university, system, article, institution, paper, country, public, change, new, development, state, reform	Topic 19	7524	11.02
2	learning, online, student, technology, course, use, education, teaching, study, digital, higher, environment, educational, learner, distance	Topic 3	6230	9.132
3	student, learning, group, study, approach, science, knowledge, design, skill, writing, course, research, teaching, result, strategy	Topic 7	5122	7.03
4	student, college, study, academic, course, performance, stem, effect, higher, result, achievement, year, data, grade, program	Topic 1	4591	6.73
5	university, academic, research, international, higher, institution, education, study, paper, staff, teaching, programme, sustainability, student, finding	Topic 9	4085	5.98
6	student, study, education, factor, university, higher, engagement, research, finding, relationship, satisfaction, model, social, result, perceived	Topic 4	3698	5.42
7	education, higher, educational, student, woman, social, study, policy, access, family, capital, participation, school, level, background	Topic 12	3495	5.12
8	program, faculty, professional, graduate, education, work, research, development, career, skill, study, experience, higher, member, university	Topic 16	3461	5.07

Table 5-14: (Continued)

9	assessment, quality, education, system, evaluation, process, higher, model, paper, development, research, educational, framework, approach, learning	Topic 18	3420	5.01
10	student, study, feedback, medical, result, score, group, scale, skill, assessment, course, learning, item, method, questionnaire	Topic 13	3111	4.56
11	teacher, language, teaching, english, study, classroom, school, education, practice, professional, instruction, data, high, finding, student	Topic 14	2659	3.89
12	student, skill, study, reading, activity, group, mathematics, level, game, result, child, effect, cognitive, education, test	Topic 2	2466	3.61
13	health, physical, student, education, study, female, gender, attitude, level, science, intervention, result, male, participant, difference	Topic 15	2244	3.28
14	student, social, study, identity, experience, diversity, medium, black, academic, campus, group, finding, peer, participant, university	Topic 11	2148	3.14
15	study, research, data, analysis, method, result, used, level, training, competency, education, information, using, ict, purpose	Topic 10	1938	2.84
16	child, leadership, early, care, parent, childhood, family, quality, education, leader, health, support, program, preschool, study	Topic 8	1200	1.75
17	school, high, public, state, secondary, district, student, urban, pupil, private, principal, education, primary, rural, data	Topic 5	1127	1.65
18	knowledge, science, education, music, sport, content, scientific, curriculum, coach, high, literacy, study, training, pedagogical, coaching	Topic 17	645	0.94

Figure 5-18:The Word Cloud representation of the 19 LDA Topics.

5.4.3 Visualizing and Interpreting LDA Topics

Visualization techniques are widely used to portray the Inter-Topic Distance Map in multidimensional space to more clearly understand the weights and relationships between the topics found by LDA, as shown in Figure 5.20. By extracting relevant phrases (keywords) from the created LDA topic model, pyLDAvis, a Python program for interactive topic modelling visualization, was used in this study to help with topic interpretation[108] . The size of the subject circles on the left shows the number of terms connected to each topic. The circles are arranged according to term frequency, which means that the more closely related the topics are, the more common words they share. The most prominent words are shown in the bar chart on the right. The bars show the total frequency of each phrase over the whole corpus[109] . Two main metrics are used in this visualization strategy: The Salience metric first find the phrases that are the most informative, representing subjects across all publications. A term's ability to identify a certain subject is increased by its prominence. Second, the Relevance measure ranks words inside subjects based on the probability distribution of their weights using a parameter known as λ (Lambda). The domain expert can determine the most

advantageous arrangement of words for labelling a given topic by modifying the topic terms shown in the bar chart using the Lambda slider. The topic labelling in this study was done by varying the Lambda values between 0.5 and 1, which gave the inferred topics the best order.

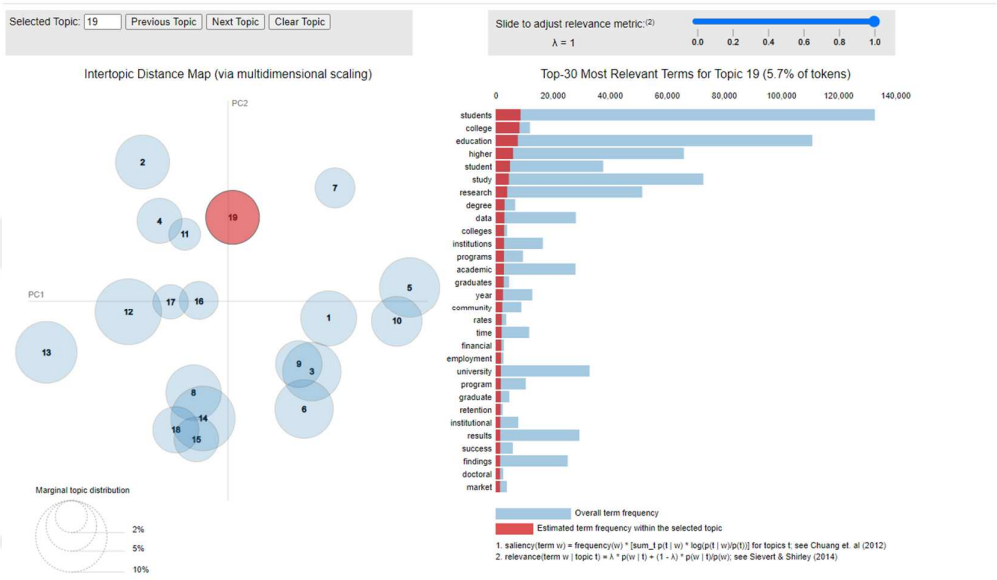

Figure 5-19:Map of Distance Between Topics and Top Most Relevant Terms for Each Topic.

5.4.4 Frequently Addressed and Examined Topics in WOS

Table 5 -14 that extracted by using jupyter notebook, highlights the top 19 dominant topics in higher education research between 1975 and January 2023. The topics, along with their top-rated keywords, are listed in descending order based on the number of papers and percentage of the total research corpus. The most dominant topic is "Higher education approach," accounting for 13.27% of the papers, followed by "Higher education reform" with 11.03%, and "Online learning technology" with 9.13%. Other notable topics in the list include "Learning approach strategy" (7.51%), "College performance study" (6.73%), "University research paper" (5.99%), and "University Study Engagement" (5.42%). The table illustrates the wide range of subject's researchers are exploring, from policy and reform to learning strategies and techniques, as well as

specific domains like "English teaching instruction" (3.90%), "Reading skill study" (3.62%), and "Diversity enriches campus" (3.15%).

This overview of the top 19 dominant topics in higher education research from 1975 to January 2023 showcases the diverse interests and concerns within the field, reflecting the evolving landscape of higher education and the multitude of factors that impact teaching, learning, and policy development.

Following is the detailed discussion for each topic result,

1. **Topic No. 6:** based on the given keywords, a possible topic label could be *Learning Experience in Higher Education*. This topic label encompasses several of the keywords provided, including education, student, learning experience, higher education, research paper, approach, teaching practice, social context, and experiential learning. The label suggests a focus on exploring how social context and experiential learning can impact teaching practices and student learning outcomes in higher education settings. This could include topics such as the role of social context in shaping teaching approaches and student experiences, the benefits and challenges of experiential learning, and the ways in which research can inform teaching practices. Additionally, the inclusion of terms like "article" and "paper" indicate a focus on exploring this topic through published research in the field of education.

2. **Topic No.19:** Based on the given keywords, a possible topic label could be *Higher Education Policy*. This topic label encompasses several of the keywords provided, including education, higher education, policy, university system, article, institution, paper, country, public, change, development, state, and reform. The label suggests a focus on exploring how higher education policy reforms and institutional changes are driven by state actors and public institutions in different countries. This could include topics such as the role of the state in shaping higher education policies and reforms, the impact of policy changes on universities and other institutions, and the development of new policies and programs to address emerging challenges in the higher education sector. Additionally, the inclusion of terms like "article" and

"paper" indicate a focus on exploring this topic through published research in the field of education.

3. **Topic No.3**: based on the given keywords, a possible topic label could be the *Use of Technology in Higher Education*. This topic label encompasses several of the keywords provided, including learning, online, student, technology, course, education, teaching, study, digital, higher education, educational, learner, and distance. The label suggests a focus on exploring the use of digital technologies in online and distance learning environments in higher education. This could include topics such as the effectiveness of digital learning environments for student learning and engagement, the use of technology to enhance teaching practices, and the impact of technology on the design and delivery of online courses. Additionally, the inclusion of terms like "study" and "learner" indicate a focus on understanding the experiences of students and learners in these digital environments.

4. **Topic No. 7:** Based on the given keywords, a possible topic label could be *Group Learning Strategies*. This topic label encompasses several of the keywords provided, including student, learning, group study, approach, science, knowledge, design, skill, writing, course, research, teaching, result, and strategy. The label suggests a focus on exploring group learning strategies for enhancing student learning in science courses, with a particular focus on developing writing and research skills. This could include topics such as the design and implementation of effective group learning strategies, the role of collaborative learning in science education, and the impact of these strategies on student learning outcomes. Additionally, the inclusion of terms like "knowledge" and "skill" indicate a focus on understanding how these group learning strategies can improve student mastery of key concepts and skills in the sciences.

5. **Topic No.1**: Based on the given keywords, a possible topic label could be *Effect of STEM Programs on College Student Academic Performance and Achievement.* This topic label encompasses several of the keywords provided, including student,

college, study, academic, course, performance, STEM, effect, higher education, result, achievement, year, data, grade, and program. The label suggests a focus on investigating the impact of STEM (Science, Technology, Engineering, and Mathematics) programs on student academic performance and achievement in college. This could include topics such as the effectiveness of STEM programs in improving student grades and retention rates, the impact of program design and implementation on student outcomes, and the long-term effects of STEM education on students' academic and professional trajectories. Additionally, the inclusion of terms like "data" and "result" indicate a focus on using empirical evidence to inform these investigations.

6. **Topic No. 9**: Based on the given keywords, a possible topic label could be *Sustainability in International Higher Education*. This topic label encompasses several of the keywords provided, including university, academic, research, international, higher education, institution, study, paper, staff, teaching, program, sustainability, student, and finding. The label suggests a focus on exploring how universities can promote sustainable practices in the context of international higher education. This could include topics such as the role of academic research in advancing sustainable development goals, the design and implementation of sustainable teaching programs, and the ways in which universities can engage with local communities to promote sustainable practices. Additionally, the inclusion of terms like "staff" and "student" indicate a focus on understanding the perspectives and experiences of different stakeholders in the university setting, and the ways in which they can contribute to sustainability initiatives.
7. **Topic No. 4:** based on the given keywords, a possible topic label could be *Relationship. between Student Engagement, Satisfaction, and Perceived Factors in Higher Education* .This topic label encompasses several of the keywords provided, including student, study, education, factor, university, higher education, engagement, research, finding, relationship, satisfaction, model, social, result, and perceived. The label suggests a focus on investigating the factors that contribute to student engagement and satisfaction in higher education, and how these factors are perceived by students. This could include topics such as the design and implementation of effective

engagement strategies in higher education, the role of social and environmental factors in shaping student experiences, and the impact of student engagement and satisfaction on academic achievement and retention. Additionally, the inclusion of terms like "model" and "result" indicate a focus on developing and testing theoretical frameworks to understand these relationships, and using empirical evidence to support these findings.

8. **Topic No. 12:** based on the provided keywords, a possible topic label could be *Gender and Socioeconomic Factors in Higher Education*. This topic label encompasses several of the keywords provided, including education, higher education, educational policy, student, woman, social capital, access, family background, and participation. The label suggests a focus on the ways in which gender and social background can affect students' access to and participation in higher education. The inclusion of terms such as "level" and "school" may indicate an interest in exploring differences in access and participation at different levels of education, such as primary, secondary, and tertiary education.
9. **Topic No. 16:** Based on the given keywords, a possible topic label could be *Graduate program and Career Skills in Higher Education*. This topic label encompasses several of the keywords provided, including program, faculty member, professional development, graduate education, research, work experience, career skills, and higher education. The label suggests a focus on the ways in which graduate education programs and faculty members can support the professional development and career skills of students. This could include exploring topics such as work experience opportunities, research projects, skill development programs, and career readiness initiatives. Additionally, the inclusion of terms like "university" and "higher education" indicate a focus on graduate-level education specifically, rather than undergraduate or vocational training.
10. **Topic No. 18:** Based on the given keywords, a possible topic label could be *Assessment and Evaluation of Educational Quality in Higher Education*. This topic label encompasses several of the keywords provided, including assessment, quality, education system, evaluation process, higher education, research paper, educational framework, and learning approach. The label suggests a focus on exploring the various models and frameworks that are used to assess and evaluate the quality of education in higher education systems. This could include topics such as evaluating the

effectiveness of teaching methods, assessing student learning outcomes, and evaluating the overall quality of educational programs. Additionally, the inclusion of terms like "development" and "research" indicate a focus on exploring new models and frameworks for assessing educational quality, and evaluating the effectiveness of existing models.

11. **Topic No. 13:** Based on the given keywords, a possible topic label could be *Feedback and Assessment in Medical Study*. This topic label encompasses several of the keywords provided, including student, study, feedback, medical education, assessment, learning outcomes, score, item, questionnaire, and assessment method. The label suggests a focus on exploring how questionnaires and learning outcome measures can be used to assess student learning in medical education. This could include topics such as developing effective assessment methods, using questionnaires to collect feedback from students, and analyzing assessment data to improve the quality of medical education programs. Additionally, the inclusion of terms like "skill" and "course" indicate a focus on evaluating the effectiveness of specific courses and programs, and assessing student learning outcomes in areas such as clinical skills and medical knowledge.
12. **Topic No. 14:** Based on the given keywords, a possible topic label could be *English Language Teaching Practices in High School Classroom*. This topic label encompasses several of the keywords provided, including teacher, language, teaching, English, study, classroom, school, education, practice, professional, instruction, data, high, finding, and student. The label suggests a focus on investigating effective English language teaching practices in high school classrooms, drawing on evidence from professional development programs and instructional data. This could include topics such as the design and implementation of effective professional development programs for English language teachers, the use of data to inform instructional decision-making and improve student learning outcomes, and the impact of effective teaching practices on student motivation, engagement, and achievement. Additionally, the inclusion of terms like "practice" and "instruction" indicate a focus on understanding how teachers can effectively apply research-based principles and pedagogies in their classroom teaching. Based on the given keywords, a possible topic label could be "Exploring Effective English Language Teaching Practices in High

School Classrooms: Evidence from Professional Development and Instructional Data". This topic label encompasses several of the keywords provided, including teacher, language, teaching, English, study, classroom, school, education, practice, professional, instruction, data, high, finding, and student.

13. **Topic No. 2:** Based on the given keywords, a possible topic label could be *Games on Cognitive Skills in Mathematics and Reading*. This topic label encompasses several of the keywords provided, including student, skill, study, reading, activity, group, mathematics, level, game, result, child, effect, cognitive, education, and test. The label suggests a focus on investigating the potential of educational games as a tool for developing cognitive skills, particularly in the areas of mathematics and reading, through group activities among children. This could include topics such as the design and implementation of effective educational games, the assessment of cognitive skills through tests and other measures, and the impact of group activities on student engagement and learning outcomes. Additionally, the inclusion of terms like "skill" and "level" indicate a focus on understanding how different levels of cognitive skills and competencies can be developed through educational games, and the potential implications of these findings for education policy and practice.
14. **Topic No.15:** Based on the given keywords, a possible topic label could be *Gender Differences in Attitudes towards Physical Education*. This topic label encompasses several of the keywords provided, including health, physical education, student, study, female, gender, attitude, level, science, intervention, result, male, participant, and difference. The label suggests a focus on investigating gender differences in attitudes towards physical education among female and male students, and how interventions can be used to address these differences. This could include topics such as the design and implementation of effective interventions to promote physical activity and healthy behaviors among students, the assessment of attitudes towards physical education through surveys or other measures, and the impact of gender on student engagement and learning outcomes in physical education. Additionally, the inclusion of terms like "level" and "science" indicate a focus on understanding how different levels of physical activity and scientific knowledge can influence attitudes and behaviors, and how these factors interact with gender to shape student experiences.

15. **Topic No.11:** Based on the given keywords, a possible topic label could be *Campus Diversity on Black Student Social Identity and Academic Experience*. This topic label encompasses several of the keywords provided, including student, social, study, identity, experience, diversity, medium, black, academic, campus, group, finding, peer, and participant. The label suggests a focus on investigating the impact of campus diversity on the social identity and academic experience of Black students at university, with a specific focus on how peer group mediums may influence these factors. This could include topics such as the experiences of Black students in predominantly White academic settings, the role of peer groups in shaping social identity and academic performance, and the impact of campus diversity initiatives on student outcomes. Additionally, the inclusion of terms like "finding" and "participant" indicate a focus on empirical research methods and the importance of centering the voices and experiences of Black students in this study.
16. **Topic No. 10:** Based on the given keywords, a possible topic label could be *Data Analysis Methods in Education Research*. This topic label encompasses several of the keywords provided, including study, data, analysis, research, result, used, method, education, level, training, assessment, test, scale, item, and using. The label suggests a focus on comparing and contrasting different data analysis methods that are commonly used in education research, particularly in the context of analyzing test results and training assessments. This could include topics such as the strengths and limitations of different statistical methods for analyzing educational data, the use of item response theory to analyze test results, and the use of scales and other assessment tools to measure student learning outcomes. Additionally, the inclusion of terms like "level" and "education" suggest a focus on exploring the impact of different data analysis methods on educational outcomes and student achievement.
17. **Topic No.8:** Based on the given keywords, a possible topic label could be *Early Childhood Education Programs*. This topic label encompasses several of the keywords provided, including child, leadership, early, care, parent, childhood, family, quality, education, health, support, program, preschool, and study.

    The label suggests a focus on investigating the impact of early childhood education programs, specifically quality preschool care, on child health and development. This could include topics such as the importance of family support and leadership in

promoting positive outcomes for children in early childhood education programs, the role of early childhood education in addressing health disparities and promoting wellness, and the impact of high-quality early childhood education on future academic and social success. Additionally, the inclusion of terms like "study" and "program" suggest an emphasis on empirical research and evaluation of early childhood education initiatives and interventions.

18. **Topic No.5 :** Based on the given keywords, a possible topic label could be *Public and Private School Education in Urban and Rural Districts.* This topic label encompasses several of the keywords provided, including school, high, public, state, secondary, district, student, urban, pupil, private, principal, education, primary, rural, and data. The label suggests a focus on comparing public and private school education in both urban and rural districts. This could include topics such as examining differences in academic performance and outcomes between public and private schools, analyzing the impact of district and school policies on student success, and investigating the role of school leadership in promoting positive outcomes for students. Additionally, the inclusion of terms like "data" and "analysis" suggests an emphasis on empirical research and the use of quantitative methods to evaluate and compare different educational settings.
19. **Topic No. 17:** Based on the given keywords, a possible topic label could *be Science, Literacy, and Pedagogy in Music and Sports Education.* This topic label encompasses several of the keywords provided, including knowledge, science, education, music, sport, content, scientific, curriculum, coach, high, literacy, study, training, pedagogical, and coaching. The label suggests a focus on exploring ways to incorporate scientific knowledge and literacy skills into music and sports education. This could involve developing new pedagogical approaches and training programs for coaches and educators, as well as revising existing curricula to integrate scientific and literacy content. The label also implies a focus on high-level education, suggesting a potential focus on advanced or specialized training programs for athletes and musicians. Additionally, the use of terms like "study" and "research" suggest an emphasis on empirical research and the evaluation of different pedagogical and training strategies.

Table 5-15:Top 19 Dominant Topics in Higher Education between (1975 – January 2023).

Topic No.	Topic Top-Rated Keywords	Topic Label	No.of Papers	%
6	education, student, practice, learning, experience, article, higher, research, paper, within, way, approach, context, teaching, social	Learning Experience in Higher Education	9051	13.27
19	education, higher, policy, university, system, article, institution, paper, country, public, change, new, development, state, reform	Higher Education Policy	7524	11.03
3	learning, online, student, technology, course, use, education, teaching, study, digital, higher, environment, educational, learner, distance	Use of Technology in Higher Education	6230	9.13
7	student, learning, group, study, approach, science, knowledge, design, skill, writing, course, research, teaching, result, strategy	Group Learning Strategies	5122	7.51
1	student, college, study, academic, course, performance, stem, effect, higher, result, achievement, year, data, grade, program	Effect of STEM Programs on College Student Academic Performance and Achievement	4591	6.73 %
9	university, academic, research, international, higher, institution, education, study, paper, staff, teaching, programme, sustainability, student, finding	Sustainability in International Higher Education:	4085	5.99
4	student, study, education, factor, university, higher, engagement, research, finding, relationship, satisfaction, model, social, result, perceived	Relationship between Student Engagement, Satisfaction, and Perceived Factors in Higher Education	3698	5.42

Table 5-15: (Continued)

12	education, higher, educational, student, woman, social, study, policy, access, family, capital, participation, school, level, background	Gender and Socioeconomic Factors in Higher Education	3495	5.12
16	program, faculty, professional, graduate, education, work, research, development, career, skill, study, experience, higher, member, university	Graduate program and Career Skills in Higher Education	3461	5.07
18	assessment, quality, education, system, evaluation, process, higher, model, paper, development, research, educational, framework, approach, learning	Assessment and Evaluation of Educational Quality in Higher Education	3420	5.01
13	student, study, feedback, medical, result, score, group, scale, skill, assessment, course, learning, item, method, questionnaire	Feedback and Assessment in Medical Study	3111	4.56
14	teacher, language, teaching, English, study, classroom, school, education, practice, professional, instruction, data, high, finding, student	English Language Teaching Practices in High School Classroom	2659	3.90
2	student, skill, study, reading, activity, group, mathematics, level, game, result, child, effect, cognitive, education, test	Games on Cognitive Skills in Mathematics and Reading	2466	3.62
15	health, physical, student, education, study, female, gender, attitude, level, science, intervention, result, male, participant, difference	Gender Differences in Attitudes towards Physical Education	2244	3.29

Table 5-15: (Continued)

11	student, social, study, identity, experience, diversity, medium, black, academic, campus, group, finding, peer, participant, university	Campus Diversity on Black Student Social Identity and Academic Experience	2148	3.15
10	study, data, analysis, research, result, used, method, education, level, training, assessment, test, scale, item, using	Data Analysis Methods in Education Research	1938	2.84
8	child, leadership, early, care, parent, childhood, family, quality, education, leader, health, support, program, preschool, study	Early Childhood Education Programs	1200	1.75
5	school, high, public, state, secondary, district, student, urban, pupil, private, principal, education, primary, rural, data	Public and Private School Education in Urban and Rural Districts	1127	1.65
17	knowledge, science, education, music, sport, content, scientific, curriculum, coach, high, literacy, study, training, pedagogical, coaching	Science, Literacy, and Pedagogy in Music and Sports Education	645	0.94

5.4.5 Research Trends in WOS Higher Education Domain

In Figure 5-20, we can observe several trends across the 19 topics over the years from 1975 to January 2023. Some topics show similar patterns, which can be grouped for discussion.

Topics 1, 5, and 7 demonstrate a steady increase in the percentage of papers published from 1975 to January 2023. These subjects may involve advances in technology, data analysis techniques, and interdisciplinary applications. The continuous growth in these topics, especially during the 2000s and 2010s, highlights their relevance and the potential for further discoveries and breakthroughs.

Some topics, like Topics 3, 6, and 16, experienced a significant rise in the number of papers published during specific periods, followed by a plateau or decline. For example, Topic 3 peaked around the late 1990s and early 2000s before gradually decreasing in interest. Similarly, Topic 6 saw a surge in the mid-2000s before levelling off, while Topic 16 gained attention in the 2010s before experiencing a decline. This pattern suggests that these topics may have been popular and groundbreaking at one point but have since reached a saturation point, with researchers shifting their focus to other areas.

On the other hand, Topics 2, 9, and 12 display a downward trend or a decline in interest over time. Topic 2 saw a steady decrease from the 1970s through the 2010s, while Topic 9 began declining in the early 2000s. Topic 12 showed a consistent downward trend from the late 1990s to 2023. These subjects may be related to traditional methods and approaches that were initially popular but have since been overshadowed by newer advancements or paradigm shifts in the respective fields.

Lastly, Topics 4, 10, 13, and 19 show fluctuations in interest over the years, with no clear upward or downward trend. Topic 4 had periods of increased interest in the early 1980s and the late 2000s, while Topic 10 experienced peaks in the mid-1990s and early 2010s. Topic 13 saw varying levels of interest across the entire period, with noticeable spikes in the early 1990s, mid-2000s, and early 2020s. Topic 19 exhibited a similar fluctuation pattern, with a peak in the mid-2000s.

This pattern could indicate that these topics are either niche areas with a smaller research community or areas where discoveries and advancements have been periodically revitalizing interest.

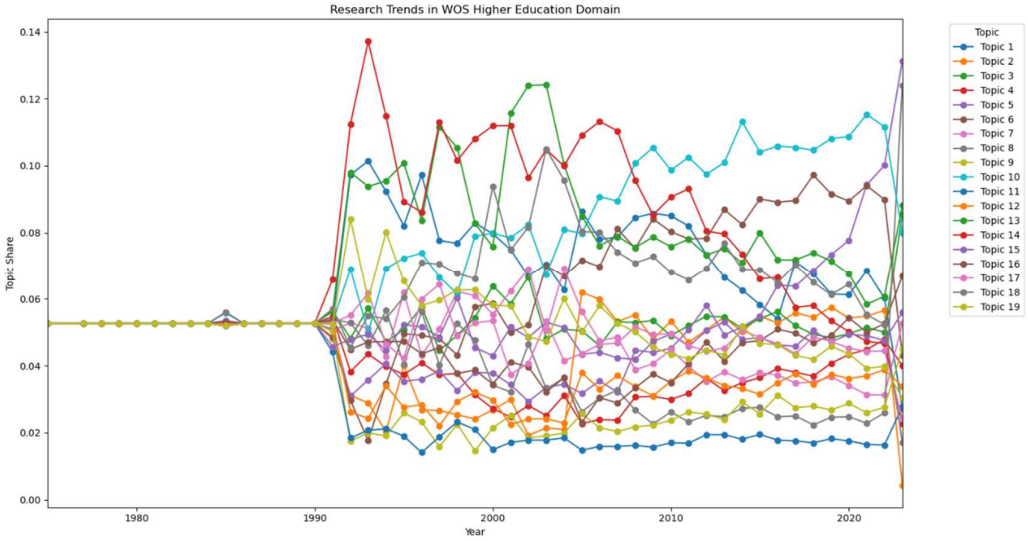

Figure 5-20:Research Trends in WOS Higher Education Domain.

5.4.6 Current State of the Dominant Topics and the Trend Topics Distribution

With help of LDA analysis tools, the trends in higher education have been obtained every five years.

The graph in Figure 5.21 displays the five-year distribution of publications on the most dominant topics based on the number of publications out of a total of 69,916 papers about higher education. The acceleration rates of the total number of publications relating to the 19 dominant topics may be found by looking at the graph. Spanning from 1985 to 2020, the data highlights the evolution of research interests and priorities in this field. Significant growth in the number of publications in the higher education domain can be seen over the years. The percentage of publications has risen sharply from a mere 0.01% in 1985 to a staggering 43.42% in 2020. This increase demonstrates the growing importance and relevance of higher education research, as well as the expansion of the academic community's focus on various aspects related to higher education. Furthermore, the data suggest that newer topics and research areas may have emerged in recent years, contributing to the observed increase in publications. Overall, the figure underscores the dynamic nature of higher education research and the continued need for exploration and analysis of dominant topics and emerging trends in the field.

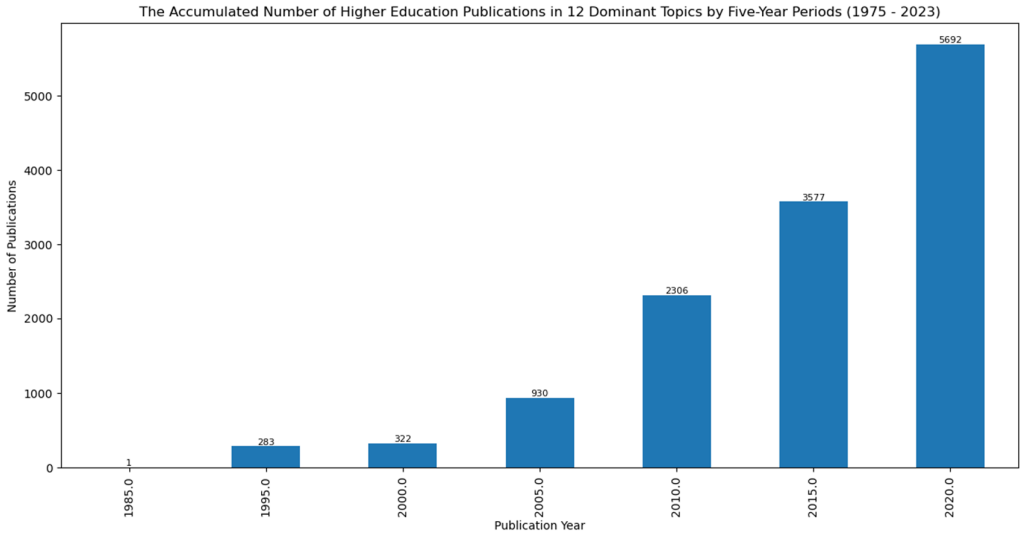

Figure 5-21:Distribution of Higher Education Dominant Topics Publications Every 5 Years.

By setting the LDA on finding the top three topics every 5 years, the resulting topics are showing in Table 5-16. The Table presents the top three higher education trend topics for five-year periods ranging from 2000 to January 2023 based on the number of publications. The number of publications on these 19 topics is the metric used to assign a rank to each topic, reflecting the focus of researchers on these topics over time.

The table highlights some interesting trends in higher education research. For instance, the first topic for the period 2000-2004 is "learning education students," indicating a focus on understanding how students learn and how to improve the educational experience. The topic "education higher students" is the top topic for the period 2020-2023, highlighting the impact of the COVID-19 pandemic on higher education and the need to address the challenges posed by remote learning.

Overall, Table 5-16 provides a useful overview of the trends in higher education research over the past two decades. The table highlights the importance of topics such as learning, higher education, and students, while also reflecting on the evolving nature

of higher education research in response to changing circumstances, such as the COVID-19 pandemic.

Table 5-16: Trend Topics Distribution by the Number of Higher Education.

Years	First Topic	Second Topic	Third Topic
2000-2004	Learning Experience in Higher Education	Higher education reform	Online learning, technology
2005-2009	Learning approach strategy	College performance study	University research paper
2010-2014	Student engagement and satisfaction	Gender and socioeconomic factors	Faculty graduate programs
2015-2019	Education quality assessment	Feedback in medical study	English language teaching practices
2020-2023	Games on cognitive skills	Gender differences in attitudes	Campus diversity

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1. Results Discussion

This study aimed to look at academic papers about higher education in around 70,000 documents from the Web of Science (WOS) using methods that make data simpler and easier to understand, as well as models to sort the information. The results, split into two parts - bibliometric analysis and topic modelling analysis, gave important information about the higher education research area. The study looked at papers published from 1975 to early 2023. So, the following is a summary of the most important findings related to the research questions, based on 'Results Chapter 5':

6.1.1. What is the Status of Publications in Higher Education from its existence to today? (Research Question 1)

6.1.1.1. The Distribution of Documents by Year (1975 - Jan 2023)

The line chart in Figure 5.1 provides insightful information on historical publishing trends in the field of higher education, highlighting the progressive rise in papers that marks the development and advancement of research through time. Similar to what was shown in Educational Technology[110] , the steady number of publications from 1975 to 1980-1981 points to a gradual beginning of research progress in this sector. The ensuing constant increase in articles published demonstrates the increased interest in and dedication to higher education research.

The steady rise in publications since 2006 is important since it reflects a major growth of research activities in higher education. The number of publications increased steadily over more than ten years, peaking in 2019–2020 with almost 70,000 publications. Despite the constant increases, the reduction in publications seen in 2022

raises the possibility of a change in the area of focus for research in higher education. This volatility highlights the necessity for continual study and analysis to keep current on the most recent advances and trends in higher education, as shown by relevant literature in other domains, even though it is still too early to draw firm conclusions about future patterns[111] .

6.1.1.2. Distribution of Documents Across Subject Areas in Higher Education (1975 - Jan 2023)

6.1.1.3. Documents by Type Distributions in Higher Education (1975 - Jan 2023)

Drawing from the data collection method employed in this study, three categories of higher education publications were selected (Journal articles, Reviews, and Conference papers). As depicted in Figure 5.2 and Table 5.1, the majority of publications fell under 'Journal Articles,' accounting for 70% (14,000 documents) of the total, followed by 'Conference Papers' at 25% (5000 documents). In contrast, the minority of publications were 'Reviews,' representing only 5% (1,000 documents). This result

6.1.2. Which are the Most Productive Countries in Higher Education? (Research Question 2)

6.1.2.1. Documents by Country/Territory Distribution in Higher Education

As illustrated in Figure 5.4 and detailed in Table 5.3, the top 20 countries with the highest higher education publication output, using WOS data (1975 - early 2023), saw England leading with 20.33% of publications, followed by Australia (15.94%) and People's Republic of China (7.42%) [112]. These countries' strong emphasis on higher education research can be attributed to renowned academic institutions, extensive research infrastructure, and growing global influence. Previous studies have also highlighted the significance of these countries in the research landscape [113]. In contrast, Portugal's lower representation (1.71%) suggests a lesser focus or limited resources for higher education research compared to the more dominant countries.

6.1.3. Which Countries/Regions and Institutions Were the Major Contributors? (Research Question 3)

6.1.3.1. Documents by Citations Distribution in Higher Education

Figure 5.8 displays a pie chart depicting the distribution of co-authorship among the top 20 countries in higher education based on citation count, with further details in Table 5.8. England ranks first with 6.15% of total citations (536,102)[114] , followed by Australia at 5.57% (485,700) and Peoples R. China at 2.96% (257,858). Spain and Canada complete the top five with 2.48% and 2.30%, respectively. Countries ranked 6th to 20th show a more gradual decline in shares, with Turkey, Germany, the Netherlands, South Africa, and Taiwan each contributing over 1% of total citations. Belgium ranks 20th with a 0.63% share (55,167 citations). The chart and table effectively highlight the co-authorship distribution among the top 20 countries in higher education by citation count.

6.1.3.2. Documents by Affiliation/Institution Distributions in Higher Education

Figure 5.6 and Table 5.5 display the distribution of higher education publications and citations among top countries, highlighting their varied contributions to the field. England [129], with 128,158 citations, and Australia, with 93,635 citations, lead the way, followed by Peoples R China with 30,372 citations. The graph shows a decline in citations from the top three countries to those ranked 4th to 15th. Despite this, countries like Canada, the Netherlands, Spain, Taiwan, Sweden, New Zealand, Israel, Finland, South Africa, and Turkey contribute significantly to the total citations, indicating global interest in higher education research and development. The leading of England in the field of Higher education proved again as shown in reference [115].

6.1.4. What were the Scientific Collaborations among Major Contributors like? (Research Question 4)

6.1.4.1. Country Co-Authorship Based on Citation Count

Figure 5.8 and Table 5.8 display the distribution of co-authorship by authors from the top 20 countries in higher education, measured by citation count. England takes the

lead with 536,102 citations, accounting for 6.15% of the total, followed closely by Australia with 485,700 citations (5.57%). Peoples R China ranks third, contributing 2.96% with 257,858 citations. Spain and Canada complete the top five, with 215,811 (2.48%) and 200,530 (2.30%) citations, respectively.

The countries ranked 6th to 20th show a gradual decline in their respective shares. Notably, Turkey, Germany, the Netherlands, South Africa, and Taiwan each contribute over 1% of total citations. Belgium, the final country on the list, holds a 0.63% share with 55,167 citations. These findings have been confirmed in research published in 2022 that shows England is a leading research collaboration[116].

6.1.4.2. Organizations Co-Authorship Based on Citation Count

Figure 5.9 and Figure 5.10, along with Table 5.9, showcase the co-authorship distribution among the top 20 organizations in higher education based on citation count. The University of California System takes the lead with 29,213 citations (9.52%), followed by the University of London with 27,080 citations (8.83%). The University of North Carolina ranks third with 20,711 citations (6.75%). Noteworthy organizations include the Pennsylvania Commonwealth System of Higher Education, the State University System of Florida, and the University of Michigan System. The list concludes with the University of Sydney, holding 10,518 citations (3.43%). A work published in 2021 in reference [115] validates this study resulting finding by showing the University of California as one of the leading organizations in higher education based on citation count.

6.1.5. In which Journals were Higher Education Studies Mainly Published? (Research Question 5)

6.1.5.1. Documents by Journal Distribution in Higher Education

Figures 5-12 and Table 5.13 highlight the top 10 journals on higher education publications. Leading the list are "Higher Education" with 2998 publications, and "Studies in Higher Education" with 1471 publications, both covering various aspects of higher education research and policy. However, to validate the obtained results, the

Scopus data confirm that the *Higher Education journal* is at the top of the list of journals that focus on higher education topics [117]. Other notable journals include "BMC Medical Education," "Computers & Education," "Higher Education Research & Development," "Teaching in Higher Education," "Education Sciences," "Education and Information Technologies," "Journal of Further and Higher Education," and "Economics of Education Review." These journals address diverse topics such as technology, pedagogy, curriculum design, and economics in higher education.

6.1.6. What Topics in Higher Education were commonly Discussed/Researched? (Research Question 6)

6.1.6.1. Dominant Topics in Higher Education Publications (Topic Modeling Analysis)

In order to discover the research topics in Higher education, a topic modeling analysis was conducted. The LDA-based topic-modeling analysis found 19 topics from the collected data containing around 70,0000 articles. Table 5.15 shows these topic labels with the keywords of each topic. The topic labels were given by considering the top-ranked keywords in each topic. In most cases, the first five keywords were combined in a meaningful manner to name each topic.

The table provided, titled "Table 5 15: Top 19 Dominant Topics in Higher Education between (1975 – January 2023)," presents an overview of the most prevalent research topics in higher education literature over the past five decades. The topics are ranked by their frequency in academic publications and are further characterized by their top-rated keywords, topic labels, number of papers, and percentage of the total papers.

The most dominant topic, with 13.27% of papers, is "Learning Experience in Higher Education," which focuses on the student experience, learning practices, and teaching contexts. Higher Education Policy follows as the second most dominant topic, with 11.03% of papers, examining policy issues, institutional changes, and development in higher education systems.

Other prevalent topics include the use of technology in higher education (9.13%), group learning strategies (7.51%), and the effect of STEM programs on college student academic performance and achievement (6.73%). The table also highlights research areas related to international education, gender and socioeconomic factors, graduate programs, assessment and evaluation, medical studies, language teaching, and cognitive skills.

Less prominent topics in higher education research include gender differences in attitudes towards physical education (3.29%), campus diversity and student identity (3.15%), data analysis methods in education research (2.84%), early childhood education programs (1.75%), and public and private school education in urban and rural districts (1.65%). The least dominant topic, with 0.94% of papers, centers around science, literacy, and pedagogy in music and sports education.

6.1.7. Is the Number of Articles related to these Topics increasing or decreasing? (Research Question 7)

6.1.7.1. Current State of the Dominant Topics and the Trend Topics Distribution (2002 - May 2021)

This section will examine Figure 5.21 to better understand the publishing status of popular trends in higher education. The graph shows the five-year distribution of dominating subject publications among the 69,916 total papers of higher education. We can calculate the growth rates of the total number of publications for the 19 major trends by looking at the graph. According to the data, the overall number of publications on dominating issues dropped from 496 to 428 by the end of 1980 but then increased to 1,410 in 1990. At the end of 2015, there were 21,875 publications in the field of higher education, up from 13,948 in 2010. In 2020, there was a little drop of 15.7% to 18,434 papers. These results offer insightful information on past patterns and modifications in the publications on the most popular subjects in higher education. This result is confirmed in reference[118] which was published in Sprinter Journal and sponsored by the German Federal Ministry of Education and Research.

6.1.8. How Did the Research Topics in Higher Education Evolve? (Research Question 8)

6.1.8.1. Trend Topics Distribution by the Number of Publications (1975 - Jan 2023)

The evolution of research topics in higher education has undergone significant changes over time, reflecting the shifting priorities and concerns in the field. By analyzing the data on dominant topics and their trends, we can observe the gradual transformation of research focus in higher education. Table 5.16 provides an overview of the trend topics in higher education research for various time periods, starting from 2000 to 2023. By analyzing the table, we can observe the evolving focus of research in higher education over the years. It is interesting to note that the research themes have shifted from broader aspects of higher education to more specialized and targeted areas.

In the first period (2000-2004), the focus was on the learning experience in higher education, higher education reform, and online learning and technology. These topics indicate that researchers were interested in understanding how students learn, the changing landscape of higher education, and the role of technology in education. This period also witnessed a decrease in the total number of dominant topics' publications from 496 to 428 documents by the end of 1980. However, the focus shifted, and the number of publications increased to 1,410 in 1990 as shown in Figure 5.21. From 2005 to 2009, the trend shifted towards learning approaches and strategies, college performance studies, and university research papers. This demonstrates a growing interest in understanding how different learning strategies affect student outcomes and the overall performance of higher education institutions.

During 2010-2014, the research topics became more focused on student engagement and satisfaction, gender and socioeconomic factors, and faculty graduate programs. This suggests a growing awareness of the importance of student well-being and the need to address disparities in higher education.

In the 2015-2019 period, the trend topics moved towards education quality assessment, feedback in medical studies, and English language teaching practices. These topics reflect an emphasis on assessing the quality of education, improving teaching practices, and understanding the effectiveness of feedback in specific fields such as medical studies.

Finally, in the most recent period (2020-2023), the focus has been on games and their impact on cognitive skills, gender differences in attitudes, and campus diversity. These topics highlight the interest in utilizing innovative approaches to learning, understanding gender dynamics, and promoting diversity and inclusion in higher education.

In summary, the table demonstrates a clear evolution of research interests in higher education, moving from broader concerns to more specialized areas. The shift towards targeted topics reflects a growing understanding of the complexity and interconnectedness of issues in higher education, as well as the need for research that can directly inform policy and practice as confirmed in the following references [119], [120] [121] [122].

6.2. Conclusion

The primary goal of this research was to identify the 'Research Themes and Trends in Higher Education' from its inception until the present. The main steps undertaken in this research included: First, conducting an extensive literature review on previous studies in higher education to assess the analysis methods employed and their findings. While various issues in the Higher Education context have been examined, only a few review studies have analyzed trend topics within Higher Education. Second, in addition to employing a bibliometric analysis approach, this research also utilized text mining techniques on Higher Education scientific literature content, extracted from a corpus database containing 69,916 documents, spanning publications from 1975 to Jan 2023. The results of this research offered valuable insights into the longitudinal research trends of higher education, investigated through the application of advanced

content analysis using the Latent Dirichlet Allocation (LDA) Topic Modeling technique, thereby contributing valuable insights to the higher education knowledge base, assisting in the conduct of targeted research studies, and the development of higher education frameworks in this promising domain for the future of adaptive educational techniques. In summary, this research showcased the benefits of supplementing traditional bibliometric analysis tools with text mining techniques, enabling researchers to uncover more research patterns, themes, and trends by conducting sophisticated content analyses based on different extensions of topic modelling algorithms, such as LDA, developed in deep learning techniques to construct more accurate and automated models for enhanced data pattern analysis and interpretation compared to conventional content analysis tools. Finally, this study found that the top 19 dominant topics in higher education research from 1975 to January 2023, revealing the most popular areas of study. The results demonstrates that "Learning Experience in Higher Education" was the most prominent theme, followed by "Higher Education Policy" and "Use of Technology in Higher Education." This distribution indicates the key interests and concerns in the field over this period, including student learning, policy development, and technological advancements

REFERENCES

[1] M. Hansen, A. Pomp, K. Erki, and T. Meisen, "Data-Driven Recognition and Extraction of PDF Document Elements," *Technologies*, vol.7,no.3, pp. 65-80, Sep. 2019.
[2] "What is Text Mining? | IBM", Internet: https://www.ibm.com/topics/text-mining [Jan. 01, 2023].

[3] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "The Application of Text Mining Methods in Innovation Research: Current State, Evolution Patterns, and Development Priorities," *R&D Management*, vol. 50, no. 3, pp. 329–351, Jun. 2020.

[4] D. M. Blei, "Probabilistic topic models," *Commun ACM, vol*. 55, no. 4, pp. 77–84, Apr. 2012.

[5] F. Gurcan, N. E. Cagiltay, and K. Cagiltay, "Mapping Human–Computer Interaction Research Themes and Trends from Its Existence to Today: A Topic Modeling-Based Review of past 60 Years," *International Journal of Human–Computer Interaction,* vol. 37, no. 3, pp. 267–280, 2020.

[6] R. Pfeiffer, *History of classical scholarship from the beginnings to the end of the Hellenistic age., 1st ed. Lond0n: Pfeiffer R. History of classical scholarship: from the beginnings to the end of the Hellenistic age*. Clarendon Press Oxford; 1968, pp111-73.

[7] W. J. Slater, "Grammarians and Handwashing," *Phoenix*, vol. 43, no. 2, p. 100, Summer 1989.

[8] D. B. Baker, J. W. Horiszny, and W. v. Metanomski, "History of Abstracting at Chemical Abstracts Service," *J. Chem. Inf. Comput. Sci*., vol. 20, no. 4, pp. 193–201, 1980.

[9] R. Klavans and K. W. Boyack, "Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?," *J Assoc Inf Sci Technol*, vol. 68, no. 4, pp. 984–998, Apr. 2017.

[10] T. Velden, K. W. Boyack, J. Gläser, R. Koopman, A. Scharnhorst, and S. Wang, "Comparison of Topic Extraction Approaches and Their Results," *Scientometrics*, vol. 111, no. 2, pp. 1169–1221, May 2017.

[11] M. Rivest, E. Vignola-Gagné, and É. Archambault, "Article-Level Classification of Scientific Publications: A Comparison Of Deep Learning, Direct Citation and Bibliographic Coupling," *PLoS One*, vol. 16, no. 5, pp.251-270, May 2021.

[12] T. Dien, H. Loc, and N. Thai-Nghe, "Article Classification using Natural Language Processing and Machine Learning," *Proceedings - 2019 International Conference on Advanced Computing and Applications, ACOMP*,2019, pp. 78–84.

[13] K. Thaoroijam, "A Study on Document Classification using Machine Learning Techniques*," IJCSI International Journal of Computer Science Issues*, vol. 11, no. 1, pp. 1694–0784, Mar. 2014.

[14] Y. Li, L. Zhang, Y. Xu, Y. Yao, R. Y. K. Lau, and Y. Wu, "Enhancing Binary Classification by Modeling Uncertain Boundary in Three-Way Decisions (extended abstract)," *Proceedings - IEEE 34th International Conference on Data Engineering, ICDE,* 2018, pp. 1827–1828.

[15] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[16] M. Koppel, N. Akiva, and I. Dagan, "Feature Instability As A Criterion For Selecting Potential Style Markers," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1519–1525, 2006.

[17] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text Mining in Education," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 9, no. 6, p. 1332, Nov. 2019.

[18] Kevin Chang, "The Powerful Benefits of Text Analysis (NLP) in Higher Education," Internet: https://www.kaianalytics.com/post/benefits-of-nlp, Nov. 10, 2020 [Jan. 11, 2023].

[19] E. Zaĭtseva, B. Tucker, and E. Santhanam," Analysing Student Feedback In Higher Education : Using Text-Mining To Interpret The Student Voice", *Routledge*, vol. 1. no.1, 2021.

[20] R. Cropanzano, "Writing Nonempirical Articles for Journal of Management: General Thoughts and Suggestions," *J Manage*, vol. 35, no. 6, pp. 1304–1311, 2009.

[21] S. Kunisch, M. Menz, J. M. Bartunek, L. B. Cardinal, and D. Denyer, "Feature Topic at Organizational Research Methods: How to Conduct Rigorous and Impactful Literature Reviews? Background and Motivation," *Organ Res Methods*, vol. 2, no. 3, pp. 519–523, 2018.

[22] M. K. Linnenluecke, M. Marrone, and A. K. Singh, "Conducting Systematic Literature Reviews and Bibliometric Analyses," *Australian Journal of Management*, vol. 45, no. 2, pp. 175–194, May 2020.

[23] S. Monteiro, V. de Oliveira, and M. de B. Pasquali, "Probiotics in Citrus Fruits Products: Health Benefits and Future Trends for the Production of Functional Foods— A Bibliometric Review," *Foods*, Vol. 11, no.4, pp. 1299-1314, 2022.

[24] Y. Okubo, "Bibliometric Indicators and Analysis of Research Systems Methods and Examples," *OECD Science, Technology and Industry Working Papers*, vol. 3, no. 1, pp. 230–176, 1997.

[25] J. Paul and M. Barari, "Meta-Analysis and Traditional Systematic Literature Reviews—What, why, when, where, and how?," *Psychol Mark*, vol. 39, no. 6, pp. 1099–1115, Jun. 2022.

[26] "Content Analysis Method and Examples , Columbia Public Health," Columbia School of Public Health, Internet:https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis,Apr. 04, 2020 [May 18, 2023].

[27] "Quantitative Data: What It Is, Types & Examples QuestionPro." Internet: https://www.questionpro.com/blog/quantitative-data/ Jul.12,2023[May 18, 2023].

[28] Hind Lahmami, "Social Science Research Methodology: The Sociology of Action in Relation to Values Interuniversity Electronic Journal of Teacher Training," *Revista Electrónica Interuniversitaria de Formación del Profesorado*, vol. 23, no. 1, pp. 59–73, 2020.

[29] F. Sebastiani, "Machine learning In Automated Text Categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[30] K. P. Ukey and A. S. Alvi, "Text Classification using Support Vector Machine," *International Journal of Engineering Research & Technology (IJERT),* vol. 1, no. 3, pp.234-456, May 2012.

[31] D. Granziol, B. Ru, S. Zohren, X. Dong, M. Osborne, and S. Roberts, "MEMe: An Accurate Maximum Entropy Method for Efficient Approximations in Large-Scale Machine Learning," *Entropy* , vol. 21, no.11, pp. 551-5530, May 2019.

[32] A. K. Uysal and S. Gunal, "The Impact of Pre-Processing on Text Classification," *Inf Process Manag*, pp. 104–112, Jan. 2014.

[33] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature Selection for Text Classification: A Review," *Multimed Tools Appl*, vol. 78, no. 3, pp. 3797–3816, Feb. 2019.

[34] N. Ringland, J. Nothman, T. Murphy, and J. R. Curran, "Classifying articles in English and German Wikipedia", *School of Information Technologies University of Sydney,2006.* Available: https://aclanthology.org/U09-1004.pdf [Mar.12,2023].

[35] H. Hasan, A. Abdulla, and W. Awad, "Text Classification of English News Articles using Graph Mining Techniques", *in Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART),*2022, vol.3, pp. 926-937.

[36] H. Li and Z. Li, "Text Classification Based on Machine Learning and Natural Language Processing Algorithms," *Wirel Commun Mob Comput*, vol.6,no.2, pp.324-295, July 2022.

[37] S. Bashir, I. U. Khattak, A. Khan, F. H. Khan, A. Gani, and M. Shiraz, "A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches," *Complexity*, vol. 20, no.1, pp.376-216, 2022.

[38] M.-Q. Nghiem, P. Baylis, A. Freitas, and S. Ananiadou, "Text Classification and Prediction in the Legal Domain*," Proceedings of the 13th Conference on Language Resources and Evaluation (LREC),*2022, pp. 4717–4722.

[39] J. Alejandro-Cruz, R. Rio-Belver, Y. Almanza-Arjona, and A. Rodriguez-Andara, "Towards a Science Map on Sustainability in Higher Education," *Sustainability*, vol. 11, no. 5, pp. 3521-221, Jun. 2019.

[40] Y. E. Park, "Uncovering trend-based research insights on teaching and learning in big data," *J Big Data*, vol. 7, no. 1, pp. 17–1, Dec. 2020.

[41] C. Shen and J. Ho, "Technology-Enhanced Learning in Higher Education: A Bibliometric Analysis with Latent Semantic Approach," *Comput Human Behav,* vol. 10, no.4, pp. 1061-77, Mar. 2020.

[42] S. Brika, K. Chergui, A. Algamdi, A. Musa, and R. Zouaghi, "E-Learning Research Trends in Higher Education in Light of COVID-19: A Bibliometric Analysis," *Front Psychol*, vol.12, no.6, pp. 67-17, Mar. 2022.

[43] J. Zilvinskis and G. Michalski, "Mining Text Data: Making Sense of What Students Tell Us," *Association for Institutional Research*, vol.10, no.2, pp.230-240, Sept. 2016.

[44] C. Yu, S. DiGangi, and A. Jannasch-Pennell, "Using Text Mining for Improving Student Experience Management in Higher Education", *IGI Global*, vol.1 no.1, pp. 196–213, 2011.

[45] R. Pazmiño, F. Badillo, M. C. González, and F. J. García-Peñalvo, "Ecuadorian Higher Education in COVID-19: A Sentiment Analysis," *ACM International Conference Proceeding Series*, 2020, pp. 758–764.

[46] S. Adhikari, "Absolute Impact Factor to Compare Journals and Research Publications of Different Subjects," *SRELS Journal of Information Management*, vol. 56, no. 3, pp. 129–134, 2019.

[47] T. Besley, "Research Quality, Bibliometric and The Republic of Science," *Assessing the Quality of Educational Research in Higher Education*, vol.33, no. 3, pp. 341–360, Jan. 2009.

[48] H. S. Zaghloul "Research Gaps and Future Trends in Educational Media and Educational Theater Research: Analytical Study in Scopus and Web of Science Databases," *Media Education (Mediaobrazovanie)*, vol. 18, no. 2, pp. 95-324, Jun. 2022.

[49] D. Yan, S. Chen, and B. Ma, "Analysis of China and ASEAN Scientific Research Cooperation Status based on Bibliometrics Methods," *China Academic Journal Electronic Publishing House*, vol.9, no.5, pp. 158–163, Aug. 2022.

[50] S. Mathieson, "Integrating research, teaching and practice in the context of new institutional policies: a social practice approach," *High Educ (Dordr),* vol. 78, no. 5, pp. 799–815, Nov. 2019.

[51] M. Moore and B. Griffin, "Identification of Factors That Influence Authorship Name Placement and Decisions to Collaborate in Peer-Reviewed, Education-Related Publications", *Studies in Educational Evaluation*, vol.32, no. 2, pp.125-135,2006.

[52] E. Nylander, L. Österlund, and A. Fejes, "The Use of Bibliometrics in Adult Education Research," *Doing Critical and Creative Research in Adult Education*, Vol.3, no.2, pp. 139–150, Apr. 2020.

[53] C. Belter, "Bibliometric indicators: opportunities and limits," *J Med Libr Assoc*, vol. 103, no. 4, pp. 219-176, Oct. 2015.

[54] G. Abramo, "Bibliometric evaluation of research performance: Where do we stand?," *Voprosy Obrazovaniya / Educational Studies Moscow*, vol. 2017, no. 1, pp. 112–127, 2017.

[55] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[56]. D. Feldman, "Knowledge discovery in Textual Databases (KDT) Proceedings of the First International Conference on Knowledge Discovery and Data Mining," *in KDD': Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 1995, pp. 112–117.

[57] V. Aguiar-Pulido, J. Seoane, M. Gestal, and J. Dorado, "Exploring Patterns of Epigenetic Information With Data Mining Techniques," *Curr Pharm Des*, vol. 19, no. 4, pp. 779–789, 2013.

[58] V. Kotu and B. Deshpande, "Knowledge Discovery in Database," *Data Science*, vol.32, no.5, pp. 1–18, 2019.

[59] P. Guleria and M. Sood, "Data Mining in Education : A Review on the Knowledge Discovery Perspective," *International Journal of Data Mining & Knowledge Management Process*, vol. 4, no. 5, pp. 47–60, Sep. 2014.

[60] A. Villanueva and L. G. Moreno, "Data Mining Techniques Applied in Educational Environments: Literature Review Andrés Villanueva Manjarres Data Mining Techniques Applied in Educational Environments: Literature Review," *Salinas Digital Education Review-Number*, vol. 33, no.8, pp.345-323, 2018.

[61] W. Zheng, "Cluster Analysis Algorithm in the Analysis of College Students' Mental Health Education," *Appl Bionics Biomech*, vol. 25, no.3, pp.193-2033, 2022.

[62] "Supervised vs Unsupervised Learning: Algorithms, Example, Difference." Internet: https://www.intellspot.com/unsupervised-vs-supervised-learning, Nov.5, 2021 [Mar. 19, 2023].

[63] J. Delua "What is Supervised Learning?" Internet: https://www.ibm.com/topics/supervised-learning, Mar. 12, 2021 [Mar. 19, 2023].

[64] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 1, pp. 241–266, 2013.

[65] A. Lora-Michiels, C. Salinesi, R. Mazo, R. A. Mazo, and R. Mazo, "A Method based on Association Rules to Construct Product Line Model A Method Based on Association Rules to Construct Product Line Models," *HAL open Sconce*, 2012.

[66] E. Hikmawati, N. U. Maulidevi, and K. Surendro, "Minimum Threshold Determination Method Based on Dataset Characteristics in Association Rule Mining*" ,J Big Data,* vol. 8, no. 1, pp. 1–17, Dec. 2021.

[67] P. Fournier "Sequential Mining: A Survey of Sequential Pattern Mining", *Data Science and Pattern Recognition,* vol.1, no.1, pp.54-71, 2017.

[68] M. R. Ullah, S. K. Shahzad, and M. R. Naqvi, "Challenges and Opportunities for Educational Data Mining in Pakistan," *International Conference on Engineering and Emerging Technologies, ICEET*, 2019, pp.218-227.

[69] M. Villanueva and L. Salenga, "Educational Data Mining: Successes and Challenges.", Internet: https://www.academia.edu/31889726/Educational_Data_Mining_Successes_and_Challenges [Mar. 19, 2023].

[70] N. Sudakova, T. Savina, A. Masalimova, M. Mikhaylovsky, L. Karandeeva, and S. Zhdanov, "Online Formative Assessment in Higher Education: Bibliometric Analysis," *Education Sciences*, vol. 12, no.2, pp. 209-253, Mar. 2022.

[71] B. Belcher and K. Hughes, "Understanding and evaluating the impact of integrated problem-oriented research programmes: Concepts and considerations," *Res Eval,* vol. 30, no. 2, pp. 154–168, Apr. 2021.

[72] Neelam Tyagi, "What is Text Mining? Text Mining Process, Methods and Applications Analytics Steps," Internet: https://www.analyticssteps.com/blogs/what-text-mining-process-methods-and-applications, May 03, 2021[Feb. 03, 2023].

[73] R. Feldman and J. Sanger, "*The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*," London, Cambridge University Press, 2007.

[74] "What is Text Mining, Text Analytics and Natural Language Processing? Linguamatics," Internet: https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing, Jan.13,2022 [Feb. 03, 2023].

[75]"Tokenization." Internet: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html, Aug.13,2019 [Mar. 20, 2023].

[76]"Stop Word Removal Documentation."Internet: https://www.ibm.com/docs/en/watson-explorer/11.0.0?topic=analytics-stop-word-removal Jan.22, 2020 [Mar. 20, 2023].

[77] "Stemming and Lemmatization." Internet: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html, Dec.6, 2021 [Mar. 20, 2023].

[78] "Basic Text Representation in NLP - Scaler Topics." Internet: https://www.scaler.com/topics/nlp/text-representation-in-nlp/, Dec.12, 2022 [Mar. 20, 2023].

[79] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 9, no. 6,pp.216-2033, Nov. 2019.

[80] L. Kumar and P. K. Bhatia, "Text Mining: Concepts, Process And Applications," *Journal of Global Research in Computer Sciences* , vol. 4, no. 3, pp. 36–39, Jan. 1970.

[81] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text Mining in Big Data Analytics," *Big Data and Cognitive Computing* , vol.1,no. 4,pp.110-230, Jan. 2020.

[82] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, pp.2001-233, 2016.

[83]"Text and Data Mining in Higher Education and Public Research." Internet: https://www.researchgate.net/publication/328768089_Text_and_data_mining_in_higher_education_and_public_research Jun.2,2020 [Mar. 20, 2023].

[84] H. Andreas, A. Nürnberger, and G. Paaß, "A brief survey of text mining." *Journal for Language Technology and Computational Linguistics*, vol.20,no.1,pp.19-62,2005.

[85] Lata Gohil, "Text Mining: Process and Techniques," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST),* vol. 3, no. 3, pp.70-95,2015.

[86]"BoW Model and TF-IDF For Creating Feature From Text." Internet: https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/ Feb.2,20120 [Apr. 26, 2023].

[87] S. Hong, T. Park, and J. Choi, "Analyzing Research Trends in University Student Experience Based on Topic Modeling," *Sustainability* , vol.12 ,no.9, pp. 3570-3900, Apr. 2020.

[88] A. Shadrova, "Topic Models Do Not Model Topics: Epistemological Remarks and Steps Towards Best Practices", *Journal of Data Mining & Digital Humanities*, vol.6, no.2, pp.35-51, 2021.

[89] P. Kherwa and P. Bansal, "Topic Modeling: A Comprehensive Review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, pp. 1–16, 2020.

[90] U. Chauhan and A. Shah, "Topic Modeling Using Latent Dirichlet allocation: A Survey," *ACM Comput Surv*, vol. 54, no. 7, pp.2013-230,Sep. 2022.

[91] A. Adewumi, S. Misra, and N. Omoregbe, "A Review of Models for Evaluating Quality in Open Source Software," *IERI Procedia*, vol. 4, no.1, pp. 88–92, Jan. 2013.

[92] J. Chuang , "Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations", *In Advances in Neural Information Processing Systems workshop on human-propelled machine learning*,vol.11,no.7, pp.1-9,2014.

[93] A. Piepenbrink and A. S. Gaur, “Topic Models as a Novel Approach to Identify Themes in Content Analysis: The Example of Organizational Research Methods,” *Annual Meeting of the Academy of Management, AOM* , vol. 2017-August, no.1. p.11335, 2017.

[94] O. Alkhnbashi, R. Nassr , “Topic Modelling and Sentimental Analysis of Students’ Reviews”, *Computers, Materials and Continua,*vol.74, no.3, pp. 6835-6848,2022.

[95] Y. Chen, B. X. Zhang, and Y. Yu, “Topic Modeling for Evaluating Students’ Reflective Writing: a Case Study of Pre-Service Teachers’ Journals,” *ACM International Conference Proceeding Series*,2016, pp. 1–5.

[96].Topic Modeling: A Complete Introductory Guide.” Internet: https://www.researchgate.net/publication/316660542_Topic_Modeling_A_Complete_Introductory_Guide, Jan.,19,2022 [20, 2023].

[97] M. Cutumisu and Q. Guo, “Using Topic Modeling to Extract Pre-Service Teachers’ Understandings of Computational Thinking from Their Coding Reflections,” *IEEE Transactions on Education*, vol. 62, no. 4, pp. 325–332, Nov. 2019.

[98] S. T. Dumais, “Latent Semantic Analysis,” *Annual Review of Information Science and Technology*, vol. 38, no.17, pp. 188–230, 2004.

[99] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” *Mach Learn*, vol. 42, no. 1–2, pp. 177–196, 2001.

[100] D. Blei, A. Ng, and J. Edu, “Latent Dirichlet Allocation Michael I. Jordan,” *Journal of Machine Learning Research*, vol. 3, no.1, pp. 993–1022, 2003.

[101] “Linear Discriminant Analysis (LDA) Concepts & Examples - Data Analytics.” Internet: https://vitalflux.com/linear-discriminant-analysis-lda-concepts-examples/ Sep.7, 2018[Mar. 20, 2023].

[102] B. Yin and C. H. Yuan, “Detecting Latent Topics and Trends in Blended Learning using LDA Topic Modeling,” *Educ Inf Technol (Dordr),* vol. 27, no. 9, pp. 12689–12712, Nov. 2022.

[103] “Project Jupyter, Home”, Internet: https://jupyter.org/ [Mar. 14, 2023].

[104] “Orange Data Mining - Data Mining.”, Internet: https://orangedatamining.com/ [Mar. 14, 2023].

[105] “What is Gensim?.”, Internet: https://radimrehurek.com/gensim/intro.html [Mar. 30, 2023].

[106] “Orange Data Mining - Data Mining.” Internet: https://orangedatamining.com/ [Mar. 12, 2023].

[107] “Gensim: Topic modelling for humans.”,Internet: https://radimrehurek.com/gensim/ [Mar. 12, 2023].

[108] C. Sievert and K. Shirley, “LDAvis: A Method for Visualizing and Interpreting Topics,” *In Proceedings of the workshop on interactive language learning, visualization, and interfaces*, Jun. 2014, pp. 63–70.

[109] H. Jelodar, “Latent Dirichlet allocation (LDA) and Topic Modeling: Models, Applications, a Survey,” *Multimed Tools Appl*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.

[110] “The Challenge to Higher Education Internationalisation.” Internet: https://www.universityworldnews.com/post.php?story=20180220091648602, Feb.22 [Mar. 30, 2023].

[111] R. Margare, "Educational Technology Research That Makes a Difference: Series introduction." *Contemporary Issues in Technology and Teacher Education*, vol. 5, no.2, pp.192-201,2005.

[112] J. Shin and G. Harman, “New Challenges for Higher Education: Global and Asia-Pacific Perspectives,” *Asia Pacific Education Review*, vol. 10, no. 1, pp. 1–13, Apr. 2009.

[113] S. Marginson and M. Mollis, “The door opens and he tiger leaps’: Theories and Reflexivity’s of Comparative Education for a Global Millennium,”, *Comparative education review*, vol.5. no.3. pp. 581-615.2001.

[114] J. Piercy, “Higher education research and innovation in facts and figures,” Universities Uk, Internet: https://www.universitiesuk.ac.uk/what-we-do/creating-voice-our-members/media-releases/higher-education-research-and-innovation, Feb. 17, 2023 [Apr. 20, 2023].

[115] B. S. Lancho-Barrantes and F. J. Cantu-Ortiz, “Quantifying the publication preferences of leading research universities,” *Scientometrics*, vol. 126, no. 3, pp. 2269–2310, Mar. 2021.

[116] Y. Fu, M. Marques, Y. Tseng, J. Powell, and D. Baker, “An evolving international research collaboration network: spatial and thematic developments in co-authored higher education research, 1998–2018,” *Scientometrics,* vol. 127, no. 3, pp. 1403–1429, Mar. 2022.

[117]“Journal of Higher Education - Impact Factor, Overall Ranking, Rating, h-index, Call For Paper, Publisher, ISSN, Scientific Journal Ranking (SJR), Abbreviation, other

Important Details, Resurchify." Internet: https://www.resurchify.com/impact/details/19490, Jul.18, 2022 [Apr. 20, 2023].

[118] D. Orr, "Springer briefs in Education Higher Education Landscape 2030 A Trend Analysis Based on the AHEAD International Horizon Scanning", Internet: http://www.springer.com/series/8914, May.12, 2022 [Apr.20,2023].

[119] "The Evolution of Higher Education." Internet: https://www.watermarkinsights.com/resources/blog/evolution-of-higher-ed Mar.18, 2022 [Apr. 20, 2023].

[120] H. Kanuka, "Trends in Higher Education," *Higher Education Res*., vol. 1, no.1 pp. 56-57, Dec. 2022.

[121] Austin, Ian, and Glen A. Jones, "Emerging trends in higher education governance: Reflecting on performance, accountability and transparency." *Research handbook on quality, performance and accountability in higher education*, vol.7, no.1, pp. 536-547, Jul. 2018.

[122] The Lancet, "Research and higher education in the time of COVID-19," *The Lancet*, vol. 396, no. 10251, pp. 583-643, Aug. 2020.