S. ALGABSI

ATILIM UNIVERSITY 2021

DETECTING RESEARCH THEMES AND TRENDS IN MOBILE LEARNING FROM ITS EXISTENCE TO TODAY, USING TOPIC MODELING

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ATILIM UNIVERSITY

SALAH EDDIN ALGABSI

A MASTER OF SCIENCE THESIS
IN
THE DEPARTMENT OF INFORMATION SYSTEMS ENGINEERING

OCTOBER 2021

DETECTING RESEARCH THEMES AND TRENDS IN MOBILE LEARNING FROM ITS EXISTENCE TO TODAY, USING TOPIC MODELING

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

BY

SALAH EDDIN ALGABSI

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

OCTOBER 2021

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

_____

Prof. Dr. Ender KESKİNKILIÇ
Director

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

_____

Prof. Dr. Murat KOYUNCU
Head of Department

This is to certify that we have read the thesis DETECTING RESEARCH THEMES AND TRENDS IN MOBILE LEARNING FROM ITS EXISTENCE TO TODAY, USING TOPIC MODELING submitted by SALAH EDDIN ALGABSI 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, Atilim University _____

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

Asst. Prof. Dr. Zafer Kadirhan
Department of Management Information Systems,
Kirsehir Ahi Evran University _____

Date: 14.10.2021

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.

Name, Last Name: Salah Eddin Algabsi

Signature:

ABSTRACT

DETECTING RESEARCH THEMES AND TRENDS IN MOBILE LEARNING FROM ITS EXISTENCE TO TODAY, USING TOPIC MODELING

Algabsi, Salah Eddin
MSc., Department of Information Systems Engineering
Supervisor: Asst. Prof. Dr. Cansu Çiğdem EKİN

October 2021, 92 pages

Nowadays, a mobile phone is more than just a phone; it is a smart device that can be used to educate individuals, enabling enhanced features for learning, and leading to an increased interest by scholars in leveraging mobile devices to develop a new paradigm of modern education. Recently, with the spread out of the Covid-19 pandemic, countries worldwide have become more orientated towards developing distance learning technologies, specifically online/e-learning, resulting in the emergence of *Mobile Learning* domain (m-learning) that becomes attractive and pervasive globally. This study aims to detect the longitudinal research trends of mobile learning from its existence to today. First, the study conducted a wide-spectrum literature review on the former studies in m-learning to evaluate the former state of the domain knowledge. Second, the study applied text mining techniques using *Latent Dirichlet Allocation* (*LDA*) Topic Modeling to the mobile learning publications content. The main findings of this study concluded 12 dominant topics in m-learning. The Top Three Topics were: "Learn with Mobile Technology", "Student Language Learn", and "Learning Design".

Keywords: Mobile Learning, M-Learning, Text Mining, Topic Modeling, Latent Dirichlet Allocation (LDA), Machine Learning.

ÖZ

KONU MODELLEME İLE VARLIĞINDAN BÜGÜNE KADAR MOBİL ÖĞRENME ALANINDAKİ ARAŞTIRMALARIN TEMA VE TRENDLERİNİN TESPİTİ

Algabsi, Salah Eddin
Yüksek Lisans, Bilgi Sistemleri Mühendisliği Bölümü
Tez Yöneticisi : Dr. Öğr. Üy. Cansu Çiğdem EKİN

Ekim 2021, 92 sayfa

Günümüzde mobil telefonlar sadece bir telefon olarak değil, aynı zamanda bireyleri eğitmek için kullanılabilecek, öğrenme için gelişmiş özellikler sağlayan ve bilim adamlarının yeni bir modern eğitim paradigması geliştirmek için mobil cihazlardan yararlanma konusundaki ilgisinin artmasına yol açan akıllı cihazlardır. Son zamanlarda, Covid-19 pandemisinin etkisiyle, dünya çapındaki ülkeler, özellikle çevrimiçi/e-öğrenme olmak üzere uzaktan öğrenme teknolojileri geliştirmeye daha fazla yönelerek Mobil Öğrenme (m-öğrenme) ortamlarının daha çekici ve yaygın olmasını sağladı. Bu çalışma, mobil öğrenmenin varlığından günümüze kadar olan tüm araştırma eğilimlerini tespit etmeyi amaçlamaktadır. İlk olarak, bu çalışmada, alan bilgisinin önceki durumunu değerlendirmek için m-öğrenmede önceki çalışmalar hakkında geniş spektrumlu bir literatür taraması gerçekleştirilmiştir. İkinci olarak, çalışma, mobil öğrenme yayınlarının içeriğine Latent Dirichlet Allocation (LDA) Konu Modellemesi ile metin madenciliği teknikleri uygulamıştır. Uygulanan analiz sonucunda, m-öğrenmede 12 baskın konu tespit edilmiştir. İlk üç konu: “Mobil Teknoloji ile Öğrenim”, “Öğrenci Dil Öğrenim” ve “Öğrenme Tasarımı” dır.

Anahtar Kelimeler: Mobil Öğrenme, M-Öğrenme, Metin Madenciliği, Konu Modelleme, Latent Dirichlet Algoritması (LDA),Makine Öğrenmesi.

To

My Family,

My "Mother & Father", may God bless your souls in Heaven - AMEN.

ACKNOWLEDGMENTS

I would like to start by expressing my sincere gratefulness to my supervisor, Asst. Prof. Dr. *Cansu Çiğdem EKİN*. It is just straightforward to say this work would not have seen the light without her patience, support, guidance, and extreme help.

Special thanks to: The administration and professors of ATILIM UNIVERSITY, including THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES & THE DEPARTMENT OF INFORMATION SYSTEMS ENGINEERING; for the good treatment and collaboration.

I shall not forget to be grateful to all my family members for their continued support and for being there whenever I needed them, *God bless you all*.

Thanks to my colleagues and friends I had in this journey and throughout my life, specifically the ones in Ankara - Turkey, for their support and hospitality.

Finally, I find it my duty also to thank:
- The Libyan Embassy - Cultural Affairs office in Ankara - Turkey.
- The Ministry of Higher Education - Libya.
- The National Agency for Scientific Research - Libya.
- The Electronic Systems and Programming Research Center - Libya.

I would express my sincere gratitude to my country LIBYA, for enabling me to follow my graduate study. I hope that God will help me in serving and promoting it faithfully.

"Praise be to God, thank you very much, good and blessed."

TABLE OF CONTENTS

ABSTRACT iii

ÖZ iv

ACKNOWLEDGMENTS vi

TABLE OF CONTENTS vii

LIST OF TABLES xii

LIST OF FIGURES xiii

LIST OF ABBREVIATIONS xv

CHAPTER 1 1

INTRODUCTION 1

1.1. Statement of the Problem 2

1.2. Purpose of the Study 2

1.3. Significance of The Study 3

1.4. Research Questions 3

1.5. The Structure of this Thesis' chapters 3

CHAPTER 2 4

LITERATURE REVIEW 4

2.1. Background 4

2.2. Literature review on Mobile Learning Studies that used Traditional Methods. 5

2.3. Literature review on M-Learning Studies that used Text Mining Techniques . 8

2.4. Summary 10

CHAPTER 3 11

BACKGROUND OF THE STUDY 11

3.1. From Distance Learning to Mobile Learning 11
3.1.1 The Concept of Digital Learning 11
3.1.2. Distance Education/Learning (d-learning) 11
3.1.3. Electronic/Online Learning (e-Learning) 12
3.1.4. Mobile Learning (m-learning) 12
3.2. Data Mining (DM) and Knowledge Discovery from Databases (KDD) 13
3.2.1. Overview of Knowledge Discovery from Databases (KDD) 13
3.2.2. Overview of Big Data and Machine Learning 16
3.2.3. Data Mining Tasks, Techniques, and Applications 18
3.3. Text Mining 21
3.3.1. Text Mining definition and concept 21
3.3.2. Text Mining vs Data Mining 22
3.3.3. Text Mining Process 23
3.4. Topic Modeling 28
3.4.1. Topic Modeling definition and concept 28
3.4.2. Importance of Topic Modeling 29
3.4.3. General Process of Topic Modeling 29
3.4.4. Basic Techniques of Topic Models 34
3.4.5. Latent Dirichlet Allocation (LDA) 35
CHAPTER 4 40
METHODOLOGY 40
4.1. Introduction 40
4.2. Design of the Study 40
4.2.1 Understanding the Domain 41

4.2.2. Data Collection/Data Extraction 41
4.2.3. Data Analysis 42
4.2.3. Data Preprocessing 43
4.3. LDA Model Implementation and Fitting 44
CHAPTER 5 45
RESULTS 45
5.1. Bibliometric Analysis – Mobile Learning (1968 - May 2021) – Scopus 45
5.1.1. Documents by Year Distribution in Mobile Learning 45
5.1.2. Documents by Subject Area Distribution in Mobile Learning 46
5.1.3. Documents by Type distributions in Mobile Learning 47
5.1.4. Documents by Country/Territory Distribution in Mobile Learning 47
5.1.5. Documents by Affiliation/Institution distributions in Mobile Learning ... 48
5.2. Bibliometric Analysis – Mobile Learning (1968 - May 2021) – VOSviewer. 49
5.2.1. Co-occurrence of Author Keywords 49
5.2.2. Co-Authorship of Authors according to the Number of Papers 50
5.2.3. Co-Authorship of Authors according to the Number of Citations 52
5.2.4. Co-Authorship of Countries according to the Number of Citations 53
5.2.5. Co-authorship of Organizations according to the Number of Citations ... 55
5.3. Topic Modeling Results 57
5.3.1. Descriptive Content Analysis Results 58
5.3.2. LDA Content Analysis Results 63
CHAPTER 6 75
DISCUSSION AND CONCLUSION 75
6.1. Results Discussion 75

6.1.1. What is the Status of the Publications in Mobile Learning from its existence to today? (Research Question 1) 75

6.1.1.1. Documents by Year Distribution in Mobile Learning (1984 - May 2021) 75

6.1.1.2. Documents by Subject Area Distribution in M-Learning (1984 - May 2021) 76

6.1.1.3. Documents by Type distributions in Mobile Learning (1984 - May 2021) 76

6.1.2. Which are the Most Productive Countries in Mobile Learning? (Research Question 2) 76

6.1.2.1. Documents by Country Distribution in Mobile Learning 76

6.1.3. Which Countries/Regions and Institutions were the Major Contributors? (Research Question 3) 77

6.1.3.1. Documents by Citations Distribution in Mobile Learning 77

6.1.3.2. Documents by Affiliation/Institution distributions in Mobile Learning 77

6.1.4. What were the Scientific Collaborations among Major Contributors like? (Research Question 4) 77

6.1.4.1. Co-Authorship of Countries according to the Number of Citations 77

6.1.4.2. Co-authorship of Organizations according to the Number of Citations 78

6.1.5. In which Journals were Mobile Learning Studies Mainly Published? (Research Question 5) 78

6.1.5.1. Documents by Journal Distribution in Mobile Learning 78

6.1.6. What Topics in Mobile Learning were commonly Discussed/Researched? (Research Question 6) 78

6.1.6.1. Dominant Topics in M-Learning Publications (Topic Modeling Analysis) 78

6.1.7. Is the Number of Articles related to these Topics increasing or decreasing? (Research Question 7) 79

6.1.7.1. The Status of Dominant Topics Articles in M-Learning (2002 - May 2021) 79

6.1.8. How did the Research Topics in Mobile Learning evolve over time? (Research Question 8) 79

6.1.8.1. Trend Topics Distribution by the number of Publications (2002 - May 2021) 79

6.2. Conclusion 81

REFERENCES 82

LIST OF TABLES

TABLES

Table 5.1: Co-occurrences of the Top 20 author keywords in Mobile Learning 49
Table 5.2: The Top 20 co-authors by the number of Papers in Mobile Learning. 51
Table 5.3: The Top 20 co-authors by the number of Citations in Mobile Learning .. 52
Table 5.4: Top 20 Co-authorship Countries by the number of Citations in Mobile Learning. 54
Table 5.5: Co-authorship of the Top 20 Organizations by Citations in M-Learning. 56
Table 5.6: Papers in Mobile Learning by Year distribution (1984 - May 2021) 59
Table 5.7: Types of Publications in Mobile Learning (1984 - May 2021). 60
Table 5.8: Top 15 Journals with the most articles in Mobile Learning (1984-May 2021) 61
Table 5.9: Top 12 Dominant Topics in Mobile Learning between (1984 - May 2021). 68
Table 5.10: The Top Three Mobile Learning Trend Topics, for Five-Year periods (2002 - May 2021). 74

LIST OF FIGURES

FIGURES

Figure 3.1: Knowledge Discovery from Databases (KDD Process) [57]. 14
Figure 3.2: The position of Topic Modeling in Machine Learning [87]..... 28
Figure 3.3: Topic Modeling Process [84]. 30
Figure 3.4: An example of representing text documents using BoW and VSM models [93]. 31
Figure 3.5: Outputs of topic model training [92]. 34
Figure 3.6: Graphical model of LDA with plate notation [98]. 37
Figure 3.7: General Process of LDA Topic Modelling [105]. 39
Figure 4.1: The Main Steps of the Research Process..... 40
Figure 5.2: Visualization of documents by subject area distribution in M-Learning. 46
Figure 5.3: Visualization of documents by type distribution in Mobile Learning..... 47
Figure 5.4: Visualization of documents by country distribution in Mobile Learning. 48
Figure 5.5: Visualization of documents by affiliation distribution in Mobile Learning 48
Figure 5.6: Visualization map of the co-occurrences of Author Keywords. (a): Among all author keywords, (b): Among the Top 20 keywords. 50
Figure 5.7: Visualization map of the Co-authorship of Authors by the Number of Papers in Mobile Learning. 51
Figure 5.8: Visualization map of the Co-authorship of Authors by the Number of Papers in Mobile Learning 53
Figure 5.9: Visualization map of the Co-authorship of Countries by the Number of Citations in Mobile Learning. 55
Figure 5.10: Visualization map of the Co-authorship of Organizations by the Number of Citations in Mobile Learning. 55

Figure 5.11: Analyzing Dataset Corpus in Orange Data Mining Software [109, 110] 57
Figure 5.12: Distribution of M-Learning Publications by Countries (1984-May-2021). 62
Figure 5.13: Distribution of M-Learning Citations by Countries (1984-May 2021). 62
Figure 5.14: Word Cloud representation of Mobile Learning Top Terms (1984 - May 2021). 63
Figure 5.15: LDA Content Analysis Coherence Score Graph and Code Results. 64
Figure 5.16: LDA Content Analysis Dominant Topic Analysis. 64
Figure 5.17: The Tabular Output of Dominant Topic Analysis Results. 65
Figure 5.18: The Word Cloud representation of the 12 LDA Topics. 65
Figure 5.19: Inter-Topic Distance Map and Top of Most Relevant Terms per Topic. 67
Figure 5.20: The Distribution of M-Learning Publications in the Dominant Topics, by Five-Year periods (1984 - May 2021). 69
Figure 5.21: The Distribution of M-Learning Publications (%) by Dominant Topics (1984 - May 2021). 70
Figure 5.22: Acceleration Analysis of Trend Topics in M-Learning (2002 - May 2021). 72
Figure 5.23: The Distribution of M-Learning Publications in the 12 Dominant 73

LIST OF ABBREVIATIONS

LDA	Latent Dirichlet Allocation
MOOCs	Massive Open Online Courses
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
WSD	Word Sense Disambiguous
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

We live in the information age characterized by faster and easier access to online content sources than ever, leading to rapid growth in the voluminous data being collected, stored, and made available in digital formats every day.

While structured data is essential, unstructured data is even more valuable, especially to scholars, businesses, and decision-makers; and when appropriately analyzed, it provides a wealth of critical insights for many organizations. Besides rich-data formats, such as images and videos, most unstructured data is stored in text documents, such as articles, Web pages, social media contents, and books [1, 2].

The developments in digital information publishing and online-accessing have contributed to broader access to knowledge content resources, making it challenging, if not impossible, for scholars to review and analyze the literature in order to recognize the thematic patterns and evolution in academic research disciplines [3, 4]. A research literature analysis can be conducted using automated Text Mining Techniques in the field of Data Mining to discover *latent* (i.e., hidden) semantic topics in a collection of documents; thus, providing valuable insights into how the research themes and trends have evolved over time, in addition to efficiently summarizing vast volumes of text documents. Text Mining techniques are explained in detail in chapter 2 of this study.

The late 20th century witnessed the great invention of mobile communication technology that emerged in many applications; recently, it has changed almost every aspect of people's lifestyle, resulted in a significant impact on the traditional education systems [5]. The Internet nowadays is more accessible and probably less expensive than ever; in addition, the successive advancements in smartphones features, such as wireless connectivity coupled with their ubiquity, have led to an increased interest in leveraging mobile devices to develop a new paradigm of modern education [6].

In the past two decades, the fields of learning/education gained more competencies by utilizing the digital mobile technologies that enable ubiquitous learning features; anywhere and anytime [7]. As a result, the ways we use to learn and teach have been dramatically influenced by these technological advancements [8]. Earlier in the 2000s, fewer businesses and individuals began to use the electronic/online learning (e-learning) [9]. Since then, e-learning has become more common and essential than ever, with an incrementing number of individuals realizing the benefits of online learning.

Recently, with the rapid development of mobile technologies towards the need for more adaptive learning, a new way of e-learning, termed *Mobile Learning* (m-learning), is becoming attractive and pervasive globally, and nowadays is considered an essential learning style in the future [6, 10].

Mobile learning (m-learning) is emerged as a natural evolution and upgrade of distance learning (d-Learning) and electronic learning (e-Learning), as it affords portability and mobility features for the mobile devices and learning content; thus, replacing books and notes with small devices, filled with tailor-made learning contents; therefore, it is believed that "*Mobile learning encourages a pedagogic shift from a teacher-centric model to a learner-centric approach*." [11]

The intersection of mobile computing and e-learning can lead to providing ubiquitous virtual learning experience m-learning. Hence, sharing is almost instantaneous among everyone using the same content, which improves learning performance through the reception of instant feedback and tips [6]. Because of the growing penetration rate of mobile devices globally, many scholars argue that m-learning should be differentiated from online learning and distance learning due to its own features and functions." [12]

Nowadays, a cell phone is more than just a phone; it is a smart device that can be used to educate individuals based on oneself's willingness and convenience. Furthermore, mobile technologies enable new and enhanced opportunities for learning, such as personalization and adaptability, context awareness and ubiquity, interactivity, communication, and collaboration among learners, and seamless bridging between contexts in both formal and informal education [13, 14]. According to Annual Reports by [15], as of Jan 2021, the rate of mobile phone users in the world has overstepped 66% of the global population (i.e., over 5 billion), and the active use of social media

in mobile usage has overstepped 53%. These statistics show that mobile phones have become a part of people's lives; around 4.2 billion mobile users globally spend, on average, about seven hours a day on the Internet.

1.1. Statement of the Problem

Based on what was concluded (in Chapter 2) from the literature review of most of the previous review studies in the field of mobile learning and their adopted analysis techniques, the following limitations were identified:

- The majority of the studies adopted traditional quantitative techniques often supplemented by manual content analysis, which can be quite time-consuming and labor-intensive; additionally, the number of publications analyzed was relatively limited;
- The number of publications in the field of m-learning is increasing, and it is becoming more difficult to conduct a traditional analysis on such big data over a long period; and
- Most of the studies discussed the technical aspects of m-learning in a quite limited period, much more than its topic trends and their overall temporal evolution.

Therefore, researchers need more efficient research techniques, such as text mining to conduct faster and more automated content analysis and categorization. However, although few review studies have already adopted the text mining approach, the research study on m-learning requires adopting sophisticated text mining techniques in order for the scholars to gather much more applicable insights for developing and enhancing the different aspects related to this style of education in the future.

1.2. Purpose of the Study

The main research objective of this study is to examine the status of academic publications in the field of mobile learning from its existence to the present time, using two different data analysis methods to extract valuable knowledge and insights regarding this domain. In brief, the main research objectives of this study can be summarized as follows:

- Using *Bibliometric Analysis Tools* to analyze the statistical distributions of m-learning publications with a broad-spectrum literature review.
- Using *Text Mining Techniques* to reveal the research themes and trends, providing a broad perspective on better understanding the m-learning domain.

1.3. Significance of The Study

- This study can contribute to the literature review on mobile learning by unveiling the current situation of the domain and identifying the research gaps related to the use of text mining techniques in this regard.
- The findings of this study contribute to the mobile learning body of knowledge to develop a *Systematic Taxonomy* reflecting the mobile learning research landscape.

1.4. Research Questions

Using a hybrid analysis approach composed of *Bibliometric* and *Text Mining* analysis techniques, this study aims to answers to the following research questions:

1. What is the status of the publications in m-learning 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 mobile learning studies mainly published?
6. What topics were commonly discussed/researched in mobile learning?
7. Is the number of articles concerning these topics increasing or decreasing?
8. How did research topics evolve over time?

1.5. The Structure of this Thesis' chapters

- Chapter 1: Introduction.
- Chapter 2: Literature Review on Mobile Learning.
- Chapter 3: Background of the study.
- Chapter 4: Methodology.
- Chapter 5: Results.
- Chapter 6: Discussion and Conclusion.

CHAPTER 2

LITERATURE REVIEW

2.1. Background

Literature reviews play a vital role in academic research in order to understand existing knowledge and clarify the status of the field being studied in the scope of the research questions being answered by the researcher [16, 17]. There are traditional methods used to review and analyze the literature, which can be summarized as follows:

Systematic reviews utilize a repeatable, research-based, and intelligible process to achieve a precise and explicit goal of reducing systematic errors and biases through extensive reviews of relevant literature; hence, inspecting the reviewer's procedure and conclusion to answer a particular research question [18].

Bibliometric analysis method is used to assess the research-based literature to determine specific patterns by summarizing publications and quantifying them using quantitative data statistics, such as the yearly growth of studies and their citations [19]. Scientists use bibliometric analyses to determine the impact of published academic papers and to contrast the contribution of individual authors, organizations, and nations [20].

Meta-Analysis statistically analyzes the results of several studies handling a similar research question with independent reporting measures anticipated to have a degree of inaccuracy. The main advantage of this approach is the aggregation of information resulting in statistically powered and more accurate point estimation than is possible in any individual study [21].

Content analysis is a research method for making systematically evaluating topics, texts, documents or ideas in qualitative results. Researchers can use content analysis by interpreting and coding the meanings and relationships of such specific words, themes, or concepts [22].

Trend analysis is the task of gathering information to discover patterns in a dominant field of study [23]. The purpose of applying trend analysis is to present the results of literary articles and studies in a qualitative, organized and distinctly interpreted manner, such as applying the percentage form and repetition [24].

In this study, the literature review on Mobile Learning will be conducted as follows: At first, general findings of some of the significant studies will be summarized, most of which followed the aforementioned traditional research methods. Afterward, the review will summarize the findings of two former review studies that pursued research methods based on implementing Text Mining techniques, therefore comparing their results with this thesis study in the 'Discussion Chapter 6'.

2.2. Literature review on Mobile Learning Studies that used Traditional Methods

In the past two decades, scholars have relied on several different aspects to review the literature related to mobile learning and ubiquitous learning. In these studies, diverse outcomes were concluded using different research methods, such as systematic reviews, bibliometric analysis, content analysis, and meta-analysis. Traditionally, these approaches are conducted using techniques such as surveys, experiments, and statistical tools.

Mobile learning emerged at the end of the twentieth century as a new paradigm in education and gained the interest of scholars over time since the year 2000 with the spread of mobile devices [25]. This paper, followed by [26], set out a fundamental theory and framework for a technology-mediated 'lifelong learning' project to design a new educational technique, utilizing PC systems (hand-held or wearable) that enable "*ubiquitous learning all through a lifetime*".

In the literature, one of the most notable early studies in the field of mobile learning covering an earlier period was conducted by [27]. In this study, samples and manual data coding and analysis were used to examine patents related to m-learning between 1976 - 2013 and created a detailed view of m-learning trends from the aspect of patents using top authoritative social sciences databases: CNIPR, USPTO, and Espacenet.

[28] conducted a systematic review of journals in the *Social Science Citation Index* database (SSCI) between 2001 and 2010, identifying 154 articles on mobile and

ubiquitous learning as well as stating the number of publications, study sample groups chosen, learning areas studied, and publishing countries. Among the key results were:

- The number of publications on the field of *Mobile and Ubiquitous Learning* significantly increased between 2006 – 2010; and
- The majority of studies focused on examining the students' encouragements, intuitions, and preferences toward mobile and ubiquitous learning, as well as oriented-courses for engineering, languages, arts, and science.

In their study of the literature reviews on mobile learning, [29] adopted a meta-analysis process to systematically review and synthesize 164 research articles between 2003 - 2010, providing an extensive analysis of former studies in terms of the significant research purposes and methodologies. The key results of this study included:

- Among many of literature studies analyzed, most of them focused on performance, then mobile learning system design; and
- The most utilized research methods were primarily surveys and experiments.

Similar findings of this study were subsequently arrived at by [30] utilizing a systematic process to conduct a literature review study, providing detailed research and synthesis for 100 of the highly referred publications on the design and implementation of mobile learning in different contexts, between 2003 and 2014. Among the key results of this study were:

- Most of former studies adopted empirical approach as primal research methods;
- The results focused on appraisals of m-learning systems and framework designs.

Another study was conducted by [31] to summarize the research findings on mobile learning in the relevant literature, published between 2013 and 2017 using extensive content analysis to examine trend topics. The study found that:

- Among many of the studies analyzed, quantitative methods were preferred;
- Some popular topics in the mobile learning research field were: Learners' or educators' perceptions of m-learning and technologies; Quality issues of m-learning; Students' acceptance of m-learning; Review of m-learning; Connectivity of m-learning; and Sustainability of mobile learning.

[32] used a bibliometric mapping of mobile learning in 5167 bibliometric studies by following the Web of Science WoS database. In this study, co-authorship, bibliographic association, co-existence, and citation analysis were used to reveal mobile learning trends from 2015 to 2019, considering authors, publications, keywords, journals, countries, universities, and citations variables. The main findings of the study were:

- Respectively, the most productive countries in mobile learning were: *Taiwan*, the *United States*, *China*, and *England*.
- According to the keyword frequency analysis, '*mobile devices, higher education, mobile technologies, tablets, and smartphones*' had emerged in the field of m-learning.
- Popular topics were *Educational Technologies*, specifically, 'tablets and mobile phones, MOOCs, and learning strategies.'

According to [33], a total of 1023 selected articles related to mobile learning, published between 2016 and 2019 were analyzed and classified based on 'research model, sample size, sample level, learning fields, subject-area classification, data collection tool, data analysis technique, mobile device, dependent variable, independent variable, number of authors, and publication year.'

According to the study results, 40% of articles adopted *quantitative approaches*, 13% were *literature reviews*, and 18% followed *mixed methods*. This study adopted a *descriptive analysis technique* to analyze the findings of the published articles, revealing the '*effectiveness*, *efficiency*, and *superiority*' of mobile learning based on different databases, year ranges, and research problems.

However, other issues and aspects of mobile learning were addressed in the literature during the past three decades. For instance, [34] indicated that most teachers supported the use of 'cell-phones' in classrooms and claimed that student participation and encouragement are the major gains; whereas [35] declared that teachers' ages impacted their views of using mobile phones in classrooms. The findings emphasized that still have more to research in mobile learning as a new field of education.

[36] studied the application and effects of mobile technology-enhanced learning and concluded that m-learning was significant to improve the learning performance and motivation of students' interests.

In their systematic review, [37] examined 90 articles in mobile technology-assisted cooperative learning between 2007 and 2016. Their results found that the number of studies in mobile cooperative learning expanded, and the connection between new mobile technologies and cooperative learning activities consolidated; these studies, in general, focused on university students more than teachers and adults.

While various issues in the mobile learning context have been discussed and published in many research papers and articles, few of which discussed m-learning topic trends, which is vital to provide actionable insights to conduct and develop better future research studies and applications in the mobile learning field [38].

Trend analysis can be accomplished using traditional bibliometric methods enhanced by content analysis, though this process is often time-consuming and labor-intensive. Therefore, researchers need more practical research tools for faster content analysis and classification, such as text-mining [39].

2.3. Literature review on M-Learning Studies that used Text Mining Techniques

The following is a summary of the two research studies that investigated trend topics in the literature review of m-learning using different tools of text mining techniques:

The first study was conducted by [39]; based on their review of the previous studies in m-learning literature, the authors claimed that due to the relatively short history of m-learning, the authors could not find any studies focused on its longitudinal research trends, since it can be regarded as a part of e-learning. SAS Enterprise Miner, a text-mining tool, was used in this study to investigate the longitudinal trends of academic articles in mobile learning from 2003 to 2008, through retrieving, then analyzing a total of 119 refereed journal articles and proceedings papers from the SCI/SSCI database. The taxonomies of m-learning publications were grouped into 12 clusters (representing the topics) and four domains based on analyzing the abstract parts using the text mining tool. The results included:

- Basic bibliometric statistics,

- Trends in the frequency of each topic over time,
- The predominance in each topic by country, and
- Preferences for each topic by the journal.

The authors concluded the following key findings of this study:

- Mobile learning publications had increased from 8 in 2003 to 36 in 2008,
- Trend topics in m-learning: effectiveness, evaluation, and personalized systems,
- Studies on strategies and frameworks of m-learning were expected to increased,
- Taiwan was as the most contributing country and university regarding journal publications on m-learning.

The study provided a quick, comprehensive overview for scholars interested in mobile learning publications, and identified the topics and areas that need further research since that m-learning articles have appeared in journals of both computer science and education schemes; consequently, the interdisciplinary approaches need more research and development of m-learning to synthesize knowledge from both disciplines.

Overall, this study illustrates the power of combining bibliometric and text mining techniques that enable scholars to discover research patterns, themes, and trends; therefore, these techniques provide better processing tools for data interpretation and pattern analysis than traditional information processing tools such as content analysis.

The second review study that used text-mining was conducted by [38]. A total of 146 papers from the publication in mobile learning between 2007 and 2018 were retrieved and analyzed using the LDA algorithm topic modeling approach applied to the abstracts of the journal articles. A total of 50 prominent topics were discovered and categorized into three groups representing mobile learning dimensions: technology, learning, and learners.

According to the topic modeling results: more than 50% of publications discussed the technology side of mobile learning, as follows:

- The first topic group is made up of the majority of terms (keywords) which are relevant to the technological aspects of mobile learning applications, such as adaptation and recommendations, which may be interpreted as the possibility to use these two technologies to develop mobile applications that are adaptable to

the needs of each student. In addition, keywords such as gesture and size are also recurrent in this topic group, indicating the need for mobile learning applications to consider taking advantage of specific capabilities of mobile technology that are distinct from desktop computers.

- The second topic group contains fewer keywords related to the features or content of the mobile learning application; in this group, games, stories, discussions, and sharing are the main features that should characterize m-learning applications, making it different from traditional learning method.
- The third topic group has the least number of keywords, paid attention to learners and their membership of application, meaning that all the parties involved in m-learning, i.e., both students and teachers, can be incorporated into a membership mechanism, thus improving their participation engagement with the application.

Based on the trend analysis of topic modeling in m-learning publications from 2007 to 2018, the study concluded that research paradigm had started to shift towards optimizing the use of mobile devices in the learning process. Therefore, the future of research can be oriented to developing theories or learning frameworks using mobile devices, ranging from the early design of the related mobile application up to evaluating its implementation.

2.4. Summary

In general, when examining the literature of research studies and articles in the field of mobile learning, it is noted that many studies adopted the quantitative analysis methods and pursued the systematic review or meta-analysis. However, fewer studies adopted other review analysis approaches, such as extensive-bibliometric mapping, content analysis, and topic trend analysis. Almost all these studies show that research on mobile learning has increased recently, revealing that the mobile learning field is becoming more significant in developing and improving learning performance.

Traditional bibliometrics methodologies are often supplemented by content analysis, though it can be quite time-consuming and labor-intensive. Therefore, scholars need more efficient research techniques, such as text-mining, in order to perform faster and even more automated content analysis and categorization.

CHAPTER 3

BACKGROUND OF THE STUDY

3.1. From Distance Learning to Mobile Learning

3.1.1 The Concept of Digital Learning

Digital Learning can be defined as **"**any learning type that is facilitated by technology, or by an instructional practice that makes effective use of technology;**"** [40] and **"**it comprises the application of a wide spectrum of practices across all curriculum learning areas, including blended and virtual learning.**"** [41]

The domain of digital learning has witnessed essential technological developments since the 1970s, specifically in *Alan Kay's Dynabook project*: The first portable and hands-on personal computer [42]; however, the modern style of distance education practically took place in the late 1980s, at which the advancements in the desktop computer provided means to uses to integrate texts, images, videos, audios, and virtual environments; to interact others, in order to teach or learn efficiently.

Later on, the evolution and development of the Internet, followed by mobile technologies during the late 1990s unto early 2000s, enabled online learning and mobile learning to emerge into distance learning consecutively [43].

3.1.2. Distance Education/Learning (d-learning)

Distance Education, also referred to as Distance Learning, is "the education of students who may not always be physically present at schools." [44] Traditionally, before the digital era, this required the students to join *Correspondence Courses* and communicate with the institutions via mail. Nowadays, d-learning also includes Online Education, and can be redefined as "a form of teaching/learning in which the participants are separated by physical distance, time and resources" [45].

A distance learning program can also combine distance learning with traditional classroom teaching, known as *hybrid* or *blended* education; however, other names like *Distributed Learning*, *e-Learning*, *Online Learning*, and *Virtual Classrooms*, are often used as synonyms for distance learning [44].

A recent application of d-learning is the Massive Open Online Courses (MOOCs): "are educational modes in distance education, offering large-scale interactive participation and open access through the World Wide Web or other network technologies" [46].

3.1.3. Electronic/Online Learning (e-Learning)

E-learning has roots in distance education due to the growth in new media technologies via the Internet, providing the capabilities to advance learning [47].

The *e-learning* education-setting utilizes *Information and Communication Technologies* (ICT) as a platform for educational activities; it is defined as "*pedagogy empowered by digital technology to access educational curriculum outside of a traditional classroom*." [6] E-learning also indicates courses, programs, or degrees offered entirely online, not in traditional classrooms. It is interactive such that learners may interact in real-time during online lectures with the instructors or other classmates in their virtual class; further, it can sometimes be a prerecorded lesson.

3.1.4. Mobile Learning (m-learning)

Almost every research in this developing domain attempts to form a consistent definition of 'Mobile Learning'. In literature, since the early 2000s, several definitions have been proposed for this relatively new domain of education in its early existence, seeing some technological breakthroughs since ever; for instance, [48] summarized a composite set of Mobile Learning definitions formed by earlier studies as follows:

- "a natural consequence of e-learning evolution, a method that intersects mobile computing and e-learning;
- a method that adopts the use of mobile technology to achieve anytime, anywhere, ubiquitous learning; and
- a method that emphasizes learner's mobility and personalized learning."

More common definitions and terms related to this field can also be concluded as follows:

- Mobile Learning is "the use of mobile devices as mediator in the process of learning and teaching" [49]. Mobile devices (e.g., mobile Smartphones, personal digital assistants PDAs, Notebooks, Tablet and Laptop PCs) are used in an educational setting, for mobile learning adoption to deliver learning content and facilitate academic activities [50, 51].
- Mobile Learning is "a subset of e-Learning that uses wireless, portable and handheld technologies including laptops, table computers, smartphones and other wireless computing devices to provide learning experience in more dynamic environments" [42].

Overall, d-learning, e-learning, and m-learning are considered nearly associated, with few distinctions between them as pre-described above; m-learning is an extension or subdomain of e-learning, whereas d-learning combines several learning techniques, including the two [40].

3.2. Data Mining (DM) and Knowledge Discovery from Databases (KDD)

3.2.1. Overview of Knowledge Discovery from Databases (KDD)

The daily amount of data being gathered is becoming larger due to numerous web services, business activities, social media, images, videos, scientific research, publications data, among other sources. Therefore, it is necessary to use intelligent systems to analyze the raw data on hand, then automatically produce reports, evaluations, or data summaries to improve decision-making [52].

Data Mining an interdisciplinary field and data analyzing approach that may be described in different ways. Conceptually, "Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data" [52]; in practical terms, "Data Mining is a set of methods often applies to voluminous yet complex databases to reduce randomness and discover (extract) hidden patterns using computationally intensive methods (i.e., algorithms)" [53]. Algorithms are 'computer programs that combine Statistics, Artificial Intelligence with Database Management'; data sources may include Databases, Data Warehouses, or the Web [54].

In the literature, the following are the most basic definitions for DM and KDD:

Knowledge Discovery in Databases KDD (aka Knowledge Discovery Process KDP) is “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns structure in data”, while Data Mining DM is “a step in the KDD process consisting of applying computational techniques that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data.” [55]

Knowledge Discovery in Databases is the entire procedure of turning the ‘raw data’ into meaningful and actionable facts. Data Mining is regarded as the ‘core step’ within KDD process, intended for applying computational techniques (i.e., data mining algorithms as programming codes or software tools) on the data to identify existing patterns of interest. The steps before DM are intended to organize, clean, and preprocess the raw data for better mining results, while the latter steps are used to evaluate and present the discovered patterns [56].

Figure 3.1: Knowledge Discovery from Databases (KDD Process) [57].

The output of KDD process is valuable information that typically includes knowledge, leading to decisions and actions based on the results. The main concept of the KDD process utilized by many organizations is simplified in figure 3.1 as an interactive and repetitive sequence (i.e., feedback loops are applicable if needed), constituted by the following steps [52, 58, 59]:

1. *Setting and Developing the Business Objectives*: In Data Mining projects, data scientists and business stakeholders often need to work together to:
    - Understand the application domain and the problem to be solved.
    - learn about the relevant prior knowledge and data types needed.

2. *Data Selection and Integration (Creating a target dataset)*: Selecting the suitable dataset, data attributes, and data samples, on which the knowledge discovery is to be performed; and thus, will help to answer the relevant questions to the business.

   In data integration phase, heterogeneous data from varied sources are combined into a common source using particular techniques such as data migration tools, data synchronization tools, and load-transformation process.

3. *Data Cleaning and Preprocessing*: This step deals with irrelevant (inconsistent) data in the collection in order to prepare a clean data set for data mining analysis. The activities in this step mainly include:

   - Removing any noisy data and outliers;
   - handling duplicates and missing values.

   Data outliers: are unusual/far values from the majority of data points in a dataset.

   Noisy Data: are often large amounts of meaningless information that machines cannot interpret well, such as unstructured text. Noisy data can also be generated due to improper data collection, data entry errors, et cetera.

4. *Data Transformation* (*reduction and projection*): Converting the preprocessed data into a suitable structure for data mining. This process comprises the following steps:

   - Data Mapping: Finding valuable features describing the data to a set goal.
   - Dimension Reductions: Performing summary/aggregation operations to combine the variables' space into a smaller number of dimensions.
   - Code Generation: Creating the actual program that fits for purpose.

5. *Data Mining*: Sophisticated AI techniques are implemented to identify the highly related patterns of interest. The main steps involved in DM are:

   - Determining the task of data mining:
     - Based on the overall purpose of KDD process, choosing whether the goal is data classification, clustering, summarizing, or trend analysis, etc.
   - Choosing data mining algorithm(s):

  - Choosing the most suitable search method(s) for pattern recognition;
  - Deciding upon the appropriate models and their parameterization; and
  - Harmonizing a specific DM technique to the main goal of KDD process.
- Data mining:
  - Identifying the patterns of interest within a specific mode of representation, such as classification rules, clustering, decision trees, association rules, trends, and regression models; and
  - Transforming data features related to a particular DM task into patterns.

6. *Patterns Evaluation and Interpretation*:
   - Identifying interesting scores for mined (i.e., extracted) patterns that represent knowledge based on given measures; and
   - interpreting and visualization of patterns based on the extracted models to make data valid, novel, useful, and understandable by users.
7. *Consolidating of Discovered Knowledge*:
   - Using visualization techniques to represent DM results to end-users.
   - The discovered patterns can be used in a system analyzed by the KDD process to create documentation, including tables and reports.

3.2.2. Overview of Big Data and Machine Learning

Big Data is a collection of data with complex contents and huge volume that is growing exponentially over time, making traditional data management tools incapable of storing or processing it efficiently [60]. The essential characteristics of big data include relatively-high *Volume*, *Velocity*, *Variety*, *Veracity*, and *Value* of the data [61].

Big Data Analytics involves using advanced 'Artificial Intelligence techniques' for voluminous, diverse data sets that include structured, unstructured, and semi-structured data, from numerous sources and at different sizes. Both unstructured and semi-structured data require preprocessing to extract meanings and support metadata. The following is a brief explanation of these data types [1, 62]:

- ***Structured Data*** is a quantitative data that is stored and processed in tabular format with many rows and columns, containing organized inputs, such as names,

addresses, dates, variable values, and ID numbers; making it easier to analyze by simple tools and machine learning algorithms, such as data pattern analyses in DM. Examples of structured data: *Relational databases*, *Excel* and *Access files*.

- ***Unstructured Data*** is a qualitative data that is not organized in a relational database; thus, it is much harder to analyze since it may include heavy-text or rich-media formats from different sources like text documents, journals articles, social media, and e-mails. Unstructured data must be preprocessed before applying machine learning algorithms. Examples of unstructured data: Html, images, audio, PDF, and Word document files.
- ***Semi-structured Data*** is a data type that has mixed characteristics of both structured and unstructured data types; it has some organizational format (non-relational database). Examples of semi-structured data: IoT data, XML, JSON, LOG, and CSV files.

Machine Learning (ML) is a subfield of *Artificial Intelligence* and *Computer Science* specializes in applying algorithms on data to mimic the human learning. Machine learning algorithms are used to build statistical models based on training the data to make classifications or predictions for data patterns; thus, drawing conclusions to identify key insights in data mining plans that can support decision-making [63].

In general, ML systems are designed to gradually update and improve their accuracy over time through the learning experience, with no/minimal human intervention. ML Models can be classified into two principal methods: *Supervised* and *Unsupervised Learning*, in addition to a hybrid method from both, *Reinforcement Learning* [64]:

- ***Supervised Machine Learning***: Supervised learning models use labeled input datasets to train ML algorithms for classifying the data or predicting predefined results accurately. Examples of supervised ML algorithms: Regression and Classification algorithms. Supervised ML models are used to solve several real-world problems at scale, like:
  - Classifying spam messages into a separate folder from the email inbox.
  - Recommendation systems and facial recognition techniques.

- ***Unsupervised Machine Learning***: Unsupervised learning models use ML algorithms designed to analyze and cluster unlabeled input datasets (i.e., no predefined outcomes). Examples of unsupervised ML algorithms: Probabilistic Clustering, Neural Networks. Unsupervised ML models are used to solve specific real-world problems, such as:
  - Dimensionality reduction process to reduce the number of features in models.
  - Exploratory data analysis (e.g., data visualization, academic research review).

3.2.3. Data Mining Tasks, Techniques, and Applications

Data mining transforms voluminous data into valuable information by utilizing various algorithms and techniques. In general, data mining has the following primary goals:

- Prediction: Using data variable-values to forecast current or expected values.
- Description: Identifying human-interpretable data patterns for representation.

Data mining algorithms and techniques have already been developed and applied in a variety of research fields such as *Statistics*, *Mathematics*, *Pattern Recognition*, *Artificial Intelligence* AI, *Machine Learning* ML, and *Database Management*; each one makes use of specialized methods from the domain of applicability in question [57, 59, 65-68]. Among various data mining techniques, the most used ones are:

1. *Classification*: The process of learning and developing a functional algorithm in order to classify a dataset item into one of many pre-defined categories (classes). *Classification techniques* are used in various application domains, including bioinformatics, financial services, social networks analysis, text processing, documents classification, etcetera.
2. *Regression*: The general process of learning a functional algorithm for mapping a dataset item into a real-valued forecast variable (numerical value). *Regression techniques* are commonly applied in economics, market trends analysis, ecological studies, climatology, and epidemiology.
3. *Clustering*: A typical descriptive activity that involves identifying limited sets of groups or clusters among data points in order to describe the data. *Clustering techniques* are used in many applications such as networks intrusion analysis, gene analysis, medicinal imaging, and text mining.

4. *Summarization*: The task that involves the use of specific techniques for finding a concise yet informative description for a sub-set of data. For unstructured data, summarization methods commonly comprise classifying text and grouping documents that share the same features. *Summarization methods* are widely utilized in the interactional analysis of big data and automated reporting (e.g., network traffic monitoring, financial reporting).
5. *Dependency Modeling*: The task that aims at building a model for describing meaningful dependencies between data variables. *Dependency modeling techniques* are utilized in software development processes, retail and business operation management.
6. *Association Rule Discovery*: A rule-based data mining technique that aims to find interesting association relationships between two or more data objects that co-occur in records of a given dataset. *Association discovery* can improve decision-making actions in many applications, including web page access analysis, market basket analysis, catalog design, communication network usage, credit card services, and medical diagnosis.
7. *Outlier* (or *Anomaly*) *Detection*: A data analysis technique used to recognize any data item incompatible with or widely deflect from other items or clusters in the dataset. While an outlier is deemed an error or noise, outliers can be interesting as they may hold vital information to some applications, such as fraud detection. *Outlier detection techniques* are utilized in many application areas, such as credit card fraud detection, clinical tests, network intrusion detection, and systems performance analysis.
8. *Evolution and Deviation Analysis*: The task pertains to studying the most notable variations in the data over time from initially measured standard values. Evolution analysis models comprise 'Change and Deviation Detection' techniques to capture the evolutionary trends in a data set; thus, they help in characterizing, classifying, or clustering of time-related data. *Evolution and deviation techniques* are used by many applications, such as time series analysis, trend estimation, network intrusion detection, website tracking, email spam filtering, and medical diagnostics.

9. *Artificial Neural Network* (ANN): A data mining technique often associated with AI and deep learning. A neural network, in general, is a computational system of hardware and (or) software that simulates neurons operation in the human brain by using a set of linked input/output nodes, where each link has its own weight present. ANN solves issues that would be impossible or difficult by statistical standards. ANN has self-learning abilities to produce better results while more data is available.
10. *Text Mining*: A data mining technique, specifically designed for text data to extract valuable information from unstructured or semi-structured text documents of various formats, such as document files, textual databases, emails, HTML web files, and user online text (e.g., comments, posts, product reviews, feedbacks). Text mining is a multidisciplinary field that incorporates various techniques such as data mining, information retrieval, computational linguistics, machine learning, statistics, and natural language processing (NLP). Text data mining techniques may include association/link analysis, visualization, and predictive analytics. Typical text mining tasks include document processing, summarization, indexing, clustering, classifications, generating granular taxonomies, parts-of-speech tagging, sentiment analysis, and concept extraction (or topic modeling). Text Mining has been utilized in numerous fields such as fraud detection, spam filtering, security issues (e.g., cybercrime prevention), bioinformatics applications (e.g., biomedicine and drug discovery), marketing (e.g., contextual advertising), business intelligence (e.g., Risk Management, Customer Service), social media analysis (e.g., sentiment analysis), and Knowledge Management (e.g., improving academic research performance and quality).

3.3. Text Mining

3.3.1. Text Mining definition and concept

In the literature, Text Mining (aka Text Data Mining TDM or Knowledge Discovery from Text databases KDT) refers to “the process of extracting meaningful, nontrivial patterns or knowledge from a set of unstructured texts.” [69] A broader definition is “a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools.” [70]

In practical terms, *Text Mining* (or *Text Analytics*) is “an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms.” [71].

Similar to text analytics, text mining is “*the process of deriving high-quality information from written natural language*.” *High-quality information* refers to “information that is new, relevant, and of interest for the project at hand.” [72]

The unifying theme behind text mining and text analytics techniques is the need to ‘turning text into numbers’ by using sophisticated analytical algorithms on huge document datasets after transforming the text to a structured, numerical representation using text handling techniques, ranging from single words to whole document databases [73].

According to IBM [1], "almost 80% of data in the world resides in an unstructured format"; within databases, text is considered one of the most common contents of data available in different formats, from different sources, such as document files, published news/scientific articles, books, digital libraries, customer reviews, and social media contents. As a result, text mining is becoming a valuable practice for data analysis within organizations with the goal of deriving ‘High-Quality’ information from text by identifying patterns and trends, using techniques such as statistical pattern learning, language modeling, and topic modeling [52].

The Six fields and Seven Practice Areas of Text Mining/Analytics:

Text Mining is an interdisciplinary field that emerged from a group of other related but distinct six fields: Statistics, Computational Linguistics (CL), Artificial

Intelligence (AI) and Machine Learning (ML), Data Mining, Library and information science, and Databases [73]. Conducting a typical text-mining task often requires several techniques from various domains; in general, there are seven distinct but highly interrelated text mining practice areas. These areas represent the main intersection points between text mining with the other six fields that contribute to it. In brief, "the seven practice areas of text mining are:

1. *Search and Information Retrieval (IR)*: Storage and retrieval of text documents, including search engines and keyword search;
2. *Document Clustering*: Grouping and categorizing of words (aka terms), snippets, paragraphs, or documents, using data mining clustering methods;
3. *Document classification*: Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples;
4. *Web Mining*: Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web;
5. *Information extraction (IE)*: Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text;
6. *Natural language processing (NLP)*: Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics; and
7. *Concept Extraction/Topic Modeling*: Grouping of related words and phrases into semantically similar groups." [73].

3.3.2. Text Mining vs Data Mining

In principle, text mining is regarded as a sub-field of data mining, specialized more on structuring the unstructured data required for text analysis; to extract novel and valuable insights. Text mining uses techniques that are data mining methods but instead, text mining techniques are used specifically for unstructured textual data analysis [1].

The concept of text mining step in *KDT* process is similar to data mining step in *KDD* process [74] that seeks to uncover meaningful patterns and knowledge in huge volumes of structured data, such as rational databases and spreadsheets (as pre-explained in section 3.2.1). In contrast, text mining tools are specifically designed to extract hidden patterns and knowledge from natural language text stored in massive amounts of semi-structured and unstructured data formats, such as email messages, text documents, social media, and HTML files [75-77].

3.3.3. Text Mining Process

To find useful knowledge in a group of text documents, text mining typically comprises five main consecutive steps: Data selection, Data cleaning, Data transformation, Data mining, and Results' evaluation and interpretation [57].
Based on this concept, the following is an overview of KDT process basic steps that constitute a common text mining research in any domain [78-81]:

1. Data Gathering and Selection: In any text mining research project, the initial step is to gather the textual data that 'fits for the analysis purposes' in unstructured or semi-structured formats (e.g., document files, CSV files); and from different sources like databases/archives, or by scraping data from websites or social media contents. After data collection, some parts of documents are examined to select the most significant attributes that represent the entire document concept based on the purpose of the research study, such as selecting only titles, abstracts, and keywords fields in the collection dataset. Using the complete information available on every document usually produces ambiguities in analysis results due to the vast volume of data contained in text files. The resultant data collections are considered semi-structured documents since they follow specific layout rules of relational databases.

2. Text Preprocessing (Data Cleaning): Text preprocessing is an essential step in many text mining algorithms that aims to prepare the unstructured or semi-structured text for analysis using natural language processing tools. The main tasks of text preprocessing can be abridged as follows:

- *Text Cleanup*: This task is often used before the main text preprocessing phase to eliminate parts from the text not related to the linguistic processing, such as advertisements in web pages, tables, figures, etcetera.
- *Tokenization*: Text preprocessing usually starts with tokenization task that aims to break out a sequence of characters of the entire text into separate words or sentences, called *Tokens*. The resultant list of tokens is used for next processing.
- *Case Normalization*: The text must be normalized to the same case level before any further processing using several related tasks, such as:
  - Converting all text to the same lower-case or upper-case characters.
  - Converting number words to the numeric form.
  - Removing HTML tags, numbers, and extra whitespaces.
  - Removing special characters/punctuation marks like (!”#$&()*+. /:;<=).
- *Filtering*: Document filtering is the task of removing specific words in the text. A commonly used filtering option is the removal of “Stop-Words”, which are the words that carry little (or no) relevant content-information, such as Determiners (a, an, the), Conjunctions, Prepositions, Pronouns, etcetera. Scholars may use one of the published stop words lists, and also, they can create domain-specific stop words (or phrases) to make better quality content analysis. Similarly, *"frequent-words"* that are quite often occurring, or *"infrequent-words"* that very rarely occurring in the text, are both considered to have little significant information and, thus, can be filtered out from the document.
- *Linguistic Processing*: This task in NLP involves three sub-tasks as follows:
  1. *Part-of-Speech (POS) tagging*: Determines the linguistic category of the words by assigning a word class to each token (noun, pronoun, adjective, verb, adverb, preposition, conjunction, and interjection).
  2. *Word Sense Disambiguous (WSD)*: Determine if a particular word is unclear (ambiguous) in a text. Essentially, it automatically assigns the most suitable meaning to a polysemous (i.e., multiple meanings) word in a given context.

3. *Semantic Structure*: Involves full parsing and partial parsing methods to represent the knowledge on the meaning of lexical items (words or phrases) that construct the context of sentences.

- *Lemmatization:* Lemmatization task takes into consideration the morphological analysis (different structures) of words by combining the many versions of a word to be treated as a unique term; i.e., lemmatization approaches attempt to assign the various verb modes to a base-infinitive tense, and nouns to a singular mode. Lemmatizing a document requires that the process must first understand the context by determining the Part-of-Speech (POS) of each word in the document, and then finds the word lemma (e.g., "go" is the lemma of goes, gone, going, went).
- *Stemming:* Stemming is the process of obtaining the *stem* (or root) of its derived words, i.e., transforming a group of different forms of related words into their normalized root-form (e.g., the word “go” is the stem of goes, going, and gone). In the case of topic modeling, a lemmatization technique is preferred over stemming because it generates more readable and understandable terms.

Note: Text preprocessing activities might be executed in iterations to refine the text preprocessing outcomes and improve the final results of research analysis.

3. Text Transformation: To typically analyze the text data, all word tokens (terms) produced by the text preprocessing step need to be transmuted into a vector representation, compatible with various text mining algorithms. Text transformation process performs two main tasks: *Feature generation*, followed by *Feature selection*.

- *Feature (Attribute) Generation*: The two basic approaches used to represent text documents are: *Bag-of-Words model* (BoW), and *Vector Space Model* (VSM). These methods will be discussed a bit more in *Topic Modeling Process* section.
- *Feature (Attribute) Selection*: The feature selection task can now be used to select a subset of the most significant features of text in order to reduce the dimensionality of data features by eliminating redundant or irrelevant

features for analysis objectives. In general, to build optimal text mining models, two main techniques are used in feature selection: *Supervised* and *Unsupervised* methods.

4. Content Analysis: The main computer-assisted content analysis approaches are: Dictionary-based methods originated from linguistics, and algorithmic methods originated from statistics and computer science.
    1. *Computational Content Analysis with Dictionary-Based Methods*: Dictionary-based text mining analysis is also known as *Sentiment Analysis* or *Automated Content Analysis*, based on the research goals. In computer coding, *Dictionaries* are lists of terms, words, or phrases used by text analysis programs or tools to conduct a content analysis. Scholars may use the pre-existing dictionaries developed in former researches; or may need to develop their private lexicons to match their research needs.
    2. *Computational Content Analysis with Algorithmic Methods*: From a machine-learning perspective, algorithmic content analytical techniques are typically classified into classification or supervised techniques, and clustering or unsupervised techniques.
        - *Classification or Supervised Techniques*: Scholars mostly attempt to map text-data entities (words or documents) to pre-defined classes by generating 'Tags,' a meta-data type, summarizing the category in which a text belongs. Classifiers are supervised ML algorithms that require human pre-understanding when used to analyze textual data. Classifiers are trained on pre-categorized (labeled) data; the common approach of these algorithms is that they are designed to distinguish multiple outcomes or categories. For instance, multiple-binary classifiers typically compare input data with each pre-defined class. A simple binary classifier can distinguish categories producing a binary outcome, similar to a yes/no decision; an example of binary classifiers is the *Spam Filter*, utilized by the majority of email service providers. In general, based on the classifier algorithm

setups and parameters, the result will classify the new document's class membership.

- *Clustering or Unsupervised Techniques*: In contrast to classification, clustering methods use unsupervised ML algorithms that do not require human pre-understanding when used to analyze textual data, but rather categorize text according to contents similarities. Clustering methods belong to the wider category of data dimensionality-reduction techniques.
  - Most clustering approaches rely on similarity as a distance measure between the documents being grouped in a meaningful way into a text corpus.
  - Alternatively, non-distance-based algorithms use a probabilistic modeling approach to determine the similarity, which is, here, is calculated as a probability of the membership in a cluster.

  To typically work with clustering results, the discovered clusters should be labeled. This process normally includes a manual task based on either approach:
  - Inspecting the documents at the center of each final clusters; or
  - Inspecting the most recurring terms and phrases of a cluster.

  However, these techniques may be merged throughout the labeling process. Furthermore, another finer-grained approach to clustering documents or other textual data can be achieved by two essential steps:
  - Adding another layer of analysis that takes words into account;
  - Clustering words rather than documents.

  This approach is based on the assumption that "*words that co-occur across documents are used to express a certain latent topic*." [78]

  Finally, based on the topics, document similarity can be determined.

3.4. Topic Modeling

3.4.1. Topic Modeling definition and concept

In natural language understanding, *meanings* can be learned from a text document via its hierarchical levels, from individual words to whole sentences and paragraphs; "*At a document level, a useful way to understand a text is by analyzing its topics*." [82] The method of identifying and deriving topics in a corpus is termed '*Topic Modeling*'.

In ML and NLP, a Topic Model refers to as *"a type of statistical model for discovering the abstract 'topics' that occur in a collection of documents."* [83]; accordingly, a Topic can be defined as 'a group of clustered words (or terms) that often occurring together, and probably appearing within the same context' [38, 84].

Topic Modeling is a smart text mining technique that seeks to find latent (hidden) semantic structures (topics) in a set of unstructured textual data for interpretation; thus, it can be used to understand, organize (categorize), and summarize large collections of documents, at a scale beyond human annotation capabilities [3, 85].

Topic Models, in general, are *statistical, unsupervised machine learning algorithms* (Figure 3.2), designed to find patterns of words by analyzing the original texts in a large unstructured document collection. The discovered patterns often reflect the underlying topics, their relationships, and how they evolved over time [86].

Typically, topic modeling algorithms do not require priorly annotating or labeling the documents; the topics arise as a result of the textual analysis [3].

Figure 3.2: The position of Topic Modeling in Machine Learning [87].

Topic models can associate synonymous terms with alike meanings, and discriminate among the different utilizations of polysemous terms having multiple meanings; thus, finding patterns and relationships of the terms used in documents to link the topics between documents [88].

3.4.2. Importance of Topic Modeling

Topic modeling is considered a powerful tool that, besides serving as a classification and clustering task, it is capable of modeling objects in a corpus of documents as *Latent Topics*, often reflecting the notion of the corpus being analyzed [84].

Topic modeling techniques can perform faster and automated content analysis and categorization in conducting academic literature reviews, and have become a vital tool for scholars in various domains to study the longitudinal trends of academic articles and research studies in specific fields, to clarify its existing thematic knowledge and possible future topic trends [8, 89].

Once the themes are identified, different types of analyses can be conducted in order to uncover more insights. For instance, some of the research questions that scholars could proceed for more analyses are: [90, 91]

- Which authors are connected with the topics in question?
- Which organizations/countries are the most prominent in particular topics?
- How have the topics changed/evolved over time?

Therefore, Topic Modeling may act as a *platform* for further research analyses.

3.4.3. General Process of Topic Modeling

Figure 3.3 depicts the main steps in any *Topic Modeling Process* that often involves the same sequence concept in Text Mining Process, with few differences based on the model used in the analysis and output steps. However, the generic process of topic modeling approaches is briefly described as follows [84, 90]:

Figure 3.3: Topic Modeling Process [84].

1. *Text Input:* This step represents data gathering and selection of textual data as a collection of documents (corpus).
2. *Text Preprocessing:* In this step, several text mining techniques are applied, including Cleanup, Tokenization, Case Normalization, Filtering , Linguistic Processing, Lemmatization, and Stemming; thus, producing a list of words (aka Dictionary) that contains all 'abstracted' vocabulary (terms/tokens) that represent the documents within a corpus.
3. *Text Transformation*: All dictionary tokens produced by text preprocessing will be transformed into a vector representation, input to the topic modeling algorithm. The main tasks in text transformation are *feature generation* followed by *feature selection*. In the feature generation task, the two basic approaches used to represent text documents are the *Bag-of-Words* (*BoW*) and *Vector Space* (*VSM*) models. BoW model is a traditional approach used in Information Retrieval and NLP to represent the preprocessed text documents in a simple 'word–frequency' table that contains the terms (aka tokens) and their frequency values in the corpus. There are two main limitations to using this approach:

    - Since the order and sequence of terms are not considered in the BoW model, this representation disregards the linguistic (semantic) structure within a document.

- This method is not suitable to represent large documents' collections (corpora) due to the voluminous size (number of dimensions) of the resultant matrix.

The most common way to represent documents for analysis by machine learning algorithms is to convert them from *Bag-of-Words* representation into numeric vectors; this representation is called the Vector Space Model (VSM), "*a simple algebraic model directly based on the term-document matrix.*" [92] In VSM representation, a variable replaces each term/token with a numerical value, indicating the weight (importance) of this term/token in the document. VSM model is the generalization form of the BoW model that is widely used in various text mining algorithms for an efficient analysis of corpora. The terms in BoW are represented in a matrix called Term Document Matrix (TDM), where rows represent specific terms/tokens and columns represent documents; the matrix' cells will hold the frequency values of terms in their respective document, as in the example figure 3.4:

Figure 3.4: An example of representing text documents using BoW and VSM models [93].

If ***n*** documents and *w* words are represented in the dictionary as an input, then the BoW of the corpus is a *matrix* of size ***n*×*w***. The value $w_{i,j}$ in the matrix denotes the number of times that the ***i****-th* term appears in the ***j****-th* document; therefore, the numbers in TDM matrix cells represent the frequency values of the terms in their respective document. Computing the weights (importance) of

the terms in TDM is performed as follows: The DTM model is regarded as a simplified representation of documents as an input to the topic model algorithm step, when computing the average weights of terms. Although various methods are used to compute the weights of terms in the corpus, the most used one is Term Frequency-Inverse Document Frequency (*TF-IDF*), a statistical model used to estimate the relative importance of a specific term in a document within a corpus. TF-IDF finds the important (relevant) data terms as follows:

- Term Frequency (TF) value computed at the document level to allocate higher weights to usual terms by scoring 'how frequent' the word is in the document.
- Inverse Document Frequency (IDF) value computed at the corpus level to allocate higher weights to unusual terms by scoring 'how rare' the word is in the corpus.

The underlying concept of *TF-IDF* measure is to value those terms that, in comparison to other terms, are more frequent in a specific document while being less frequent in the entire corpus (i.e., their TF and IDF values are relatively higher) [93]. The feature selection task can now be used to select a subset of the most significant features of the text in order to reduce the TDM dimensionality by eliminating any redundant or irrelevant features according to the topic modeling analysis goals.

4. *Model Training*: Based on the exchangeability of words in the BoW representation of documents in a corpus, *the semantic structures* of the words are ruled by unseen '*latent variables*'; hence, topic modeling was primarily used to discover those variables, which indicate '*themes*' running through documents and shape their context meanings [3]. The underlying concept of topic modeling is '*documents exhibit various topics'*, where a topic is regarded as '*a distribution of terms across a dictionary of fixed vocabulary'* that describes the entire corpus [90].

   The common probabilistic assumptions underlying the most topic models are:

   - *Every document is made up of a blend of topics.*
   - *Every topic is made up of a cluster of terms.*

Practically, the size of the TDM resulted from the BoW of the corpus can be extremely high. Therefore, in order to discover the latent themes by analyzing the words of the texts, this task must involve applying a topic model algorithm that 'fits for the analysis purpose' on the TDM to reduce matrix dimensionality. Subsequently, the topics can be discovered during the training process of models described by the model type such as LSA, PLSA, and LDA, the fundamental types of topic models (explained in section 3.4.4).

In general, topic models need to mimic the '*Generative Process*' of the documents, where every word in a document is supposed to be selected by choosing a topic assignment first, then choosing the word from the analogous topic. A generative model can be evaluated using conventional statistical techniques like "*complexity control, model testing, and cross-validation*" [92].

5. Model Output: Assuming that in a corpus, there is a number of $\boldsymbol{k}$ topics, constructing $\boldsymbol{n}$ documents with a dictionary of $\boldsymbol{w}$ words as an input to a topic model used to identify the latent topics; figure 3.5 illustrates the concept of factorizing the matrix TDM with a higher dimension ($\boldsymbol{n}\times\boldsymbol{w}$) during the model training process, and convert it into two matrices with lower dimensions in the model output, as follows:
   - ($\boldsymbol{n}\times\boldsymbol{k}$) matrix represents 'Probability distributions' of *topics over documents*.
   - ($\boldsymbol{k}\times\boldsymbol{w}$) matrix represents 'Probability distributions' of *words over topics*.

   Regardless of the specific technique used by the model type, this process aims to reduce the TDM dimensionality by converting the vector space representation of the corpus from a *word-space model* ($\boldsymbol{n}\times\boldsymbol{w}$) into *topic-space* ($\boldsymbol{n}\times\boldsymbol{k}$ and $\boldsymbol{k}\times\boldsymbol{w}$) because, in practice, the words in documents typically outnumber the topics ($\boldsymbol{k}<\boldsymbol{w}$) [94].

Figure 3.5: Outputs of topic model training [92].

The topic model output reveals some important characteristics that may become useful to several applications, and can be summarized as follows:

- Analyzing documents at a lower-dimension topic-level rather than at term-level can assist in conducting faster text analysis to find latent semantic relationships, providing meaningful structures from the documents;
- Besides the primary goal of 'uncovering latent topics', topic models can also be used to cluster the corpus into groups based on the similarity of the *topic probability distributions over documents*; moreover, the ability for interpreting the clustering results using the *word probability distributions over topics*; and
- In topic models, data can be gathered from different topics — not just a single one.

3.4.4. Basic Techniques of Topic Models

Although it is not a topic model by itself, "*Vector Space Model VSM is the basis for many advanced Information Retrieval techniques and topic models.*" [92]
Another information retrieval model, *Latent Semantic Indexing* (LSI), introduced by [95]; also referred to as *Latent Semantic Analysis* (LSA). It essentially extended the VSM model by reducing the TDM dimensionality (as previously explained) utilizing a matrix algebra technique termed as *Singular Value Decomposition* SVD. The goal of LSA was to find relevant documents from search words, thus finding the hidden meaning of terms based on their occurrences in documents [84].

Although it is not an authentic 'probabilistic model', LSA provided a foundation for the later-developed topic modeling algorithms [92].

Based on the LSI model, a *Probabilistic Latent Semantic Indexing* (PLSI), also referred to as *Probabilistic Latent Semantic Analysis* (PLSA), is a 'fundamental topic model' proposed by [96] as a probabilistic generative model to address the 'statistical unsoundness' issue of LSA model. PLSA model does not use SVD method as in LSA; instead, it introduced the basic framework of the '*Generative Process of documents*' based on the statistical '*Aspect Model*' that utilizes latent variables for data co-occurrence to represent the 'topics' in documents [92]. PLSA models the *likelihood* of each co-occurrence of words and documents (***w,d***) as a combination of *multinomial-distributions* with conditional independence [97].

PLSA model aims to identify and distinguish between different context usages of a word without resorting to a dictionary or thesaurus; PLSA model involves two important implications:

- *It allows disambiguating of polysemy and synonymy*; and
- *discloses thematic similarities by grouping the words with a common context*.

PLSA has been used in several real-world applications such as *Recommender Systems* and *Computer Vision*. However, the model suffers from the *overfitting* dilemmas since its number of estimated parameters grows linearly against corpus size; in addition, the parameters cannot be applied easily to new, unseen documents outside the training set.

3.4.5. Latent Dirichlet Allocation (LDA)

3.4.5.1. Theoretical Overview

LDA model, was developed by [98] as an extension of the PLSA model toward generalizing a complete *probabilistic generative topic model* that overcame the overfitting problem of PLSA, based on the statistical models of the *Bayesian conditional probability theory* [3, 88]. The Dirichlet processes are "*a family of stochastic processes whose realizations are probability distributions*." [99]

LDA is one of the primary '*state-of-the-art*' approaches that made topic modeling tasks easier to use, and preferred by many scholars; the model is also extendable for

numerous applications in different research areas. Various LDA-based probabilistic topic models have been developed to support specific applications; thus, understanding LDA is 'vital' for the topic modeling extended application [90, 100].

According to [3], the underlying intuition of LDA model is that 'text documents exhibit various topics, and every document exhibits these topics at different proportions', indicating that there is a set of topics that can describe the entire corpus. Each document may involve one or several of these topics; besides, each word in the entire repository can be included in more than one of the topics [92].

LDA represents the topics by their related word probabilities. The highest probability words in any topic often provide a good idea about the topic [101].

LDA is well-supported in many coding languages and software tools. Many open-source tools can also be found in different online resources, such as 'Python, R-project, MALLET, and Orange data mining', in addition to programming developer communities, such as GitHub, which includes various open-source projects for topic modeling [102].

LDA is considered a fully generative model of documents in a corpus, assuming that **'**each document is modeled as a blend of distributed probabilities for a set of hidden topics**'**, while **'**each hidden topic can be modeled as a blend of distributed probabilities for a set of fixed vocabulary**'**; the words in documents are observed, while the topics are considered yet to be discovered [92, 98]. In brief, the data produced by LDA model can be considered of as coming from a generative process, set by a combined probability distribution upon what is **'**observed**'** and **'**hidden**'** [101].

3.4.5.2. LDA Parameters and Generative Process

Technically, Latent Dirichlet Allocation model hypothesizes that topics are specified first before the documents are generated [3]. As shown in figure 3.6, "LDA model is represented as a *Probabilistic Graphical Model* using a *Plate Notation*." [98]

The *Graphical Representation* of probabilistic models is a graph-based model used to express the conditional dependency structure between random variables of the model.

The *Plate Notation* is the method used in conditional probability inference to represent the variables iterating in probabilistic graphical models; so, instead of individually

drawing each repeated variable, box shapes called 'Plates' are used to group these variables into subgraphs, which will iterate their variables for several times given by the number drawn on the plate corner. With plate notation, the dependencies between the model variables can be gleaned concisely [103, 104].

The LDA graphical model with plate notation describes the basic *Probabilistic Generative Process* of the documents in a corpus as follows [98]:

Figure 3.6: Graphical model of LDA with plate notation [98].

- Model variables are shown as nodes (shaded node is observed; others are latent).
- The plates represent iterations for the repeated entities (variables).
- The outer plate represents the corpus (documents).
- The internal plate represents the positions of the repeated word in a given document; each position is linked with the choice of a topic and word.

The variables representing the input parameters to LDA model are defined as follows:

K is the number of topics to be inferred (optimal K is obtained during model training);

M is the number of all documents in the corpus D to be analyzed;

N is the number of words in a given document;

α is per-document topic distribution, to control the number of topics in the documents;

β is per-topic word distribution, controlling the number of words in the topics;

φ_k is the word distribution for the topic k, where ($k = 1,2,\ldots, K$).

θ_d is the topic distribution for the document d having N_d words, where ($d = 1,2,\ldots, M$).

$w_{d,n}$ is the n-th position word in the d-th document, where (n = 1,2,…, N).

$z_{d,n}$ is the topic assignment for the specific word w_{dn}.

φ and θ are Dirichlet distributions.

w and z are multinomial (aka categorical) formulas.

α and β are Dirichlet-prior concentrating parameters at a corpus-level.

In practice:

- A high α value will lead to more similar documents in terms of the topics they include.
- A high β value will result in more similar topics regarding the words they include.

LDA algorithm will model the corpus D based on the following generative process; in this process, words in documents are only observed variables, others are latent:

1. Selecting a multinomial distribution φ_k for topic k from Dirichlet $Dir(\beta)$.
2. Selecting a multinomial distribution θ_d for Document d from a $Dir(\alpha)$.
3. For a specific Word $w_{n,d}$ in document d:
   i. First Selecting a topic $z_{n,d}$ from θ_d; then
   ii. Selecting a word w_{nd} from φ_{znd}
4. The LDA algorithm will iteratively process the input parameters and train the model, realizing three primary outcomes, as shown in figure 3.7:
   i. Frequency of words by topic, i.e., topic-words probability distribution;
   ii. Cluster of words by topic, i.e., distribution of words per each topic k;
   iii. Cluster of documents by topic, i.e., distribution of topics per document d.

Figure 3.7: General Process of LDA Topic Modelling [105].

General Notes:

- The probabilistic details of the LDA generative process [98], including the optimal values of the model hyper-parameters, are *out of the scope* of this thesis.
- Deciding the optimal values of LDA parameters to adopt when applying the model to a given dataset is a crucial factor for improving the model outcomes. Many researchers from various communities agree that no general setting can work well for any dataset [94]. However, different heuristic approaches have been proposed by scholars to best tune LDA parameters to optimal levels according to the given task, evaluation metrics, and implemented programming library or software tool (e.g., *MALLET* software, *R, and Gensim* packages). This study mainly used the 'Gensim' package in Python.

CHAPTER 4

METHODOLOGY

4.1. Introduction

This study aims to investigate the use of text-mining techniques in an educational context, specifically, topic-modeling with LDA to extract knowledge from m- learning publications. This chapter illustrates the steps taken throughout the research process in order to answer the main research questions in the 'Introduction Chapter 1'. Therefore, to accomplish this task, the research questions 1 through 8 were investigated and reported later in the 'Results Chapter 5' using the following analysis approaches:

- Bibliometric analysis tools to examine the patterns of m-leaning domain status;
- LDA topic modeling to discover the trend topics of m-learning.

4.2. Design of the Study

Figure 4.1 summarizes the main steps of the research process of this study, as follows:

Figure 4.1: The Main Steps of the Research Process.

4.2.1 Understanding the Domain

At first, a broad-spectrum Literature Review was conducted to determine the synonyms and expressions used to describe mobile learning in the literature and evaluating the methods used for analysis with their results regarding the patterns and statistics of mobile learning (Chapter 2). Afterward, a domain expert was consulted to determine the specific keywords suitable for extracting the related articles within the literature, as follows: mobile learning, m-learning, and mlearning.

4.2.2. Data Collection/Data Extraction

For Data Gathering, this study considered using 'Scopus Database' to retrieve peer-reviewed publications that fit the scope of the study, since this database includes over 5000 publishers globally, including Elsevier, Emerald, IEEE, Sage, Springer, Taylor & Francis, and Wiley Blackwell [10]. Therefore, the keywords identified were used in the follows searching query string:

> TITLE-ABS-KEY ("mobile learning" OR "m-learning" OR "mlearning") AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))

In search results, a total of 8191 documents were gathered, covering quality publications of Journal Articles and Conference Papers published in English, in more than 50 years, between 1968-May 2021; the obtained documents were extracted and stored in a semi-structured, comma-separated value CSV dataset, ready to use for Text Mining Analysis. For quality analysis results, a manual preprocessing was performed on the dataset by the domain expert of this study, comprising a set of quick reviewing and cleaning steps, such as:

- Removing the 'nan' (no abstract available) parts.
- Removing duplicated items (having the same abstract).
- Removing unrelated articles that might be gathered due to keyword matching with other domains than mobile learning, specifically in earlier period of 1980s.

The results of these steps and further, are explained in the 'Results Chapter'.

4.2.3. Data Analysis

After the dataset was gathered and manually preprocessed, two methods were adopted: Bibliometric and Content analysis, as follows:

1. *Bibliometric Analysis*: In this method, two analysis tools were used as follows:
   - *Scopus Analysis*: Scopus database provides a quich yet easy to use analysis utility that demonstrates the preliminary search results, divided into separate tabs by: *Year*, *Source*, *Author*, *Affiliation*, *Country*, *Document type*, and *Subject area* [106].
   - *VOS viewer Analysis*: VOSviewer is a computer application used to construct and visualize Bibliometric Networks that may include scholars, or individual publications, based on *citation*, *bibliographic coupling*, *co-citation*, or *co-authorship* relations. VOSviewer also includes a text-mining feature that enables the creation and visualization of Co-occurrence Networks for key terms retrieved from a corpus of scientific literature [107].
2. *Content Analysis*: In this study, the adopted content analysis method was conceptually based on the general process of text mining and topic modeling (Chapter 3). Latent Dirichlet allocation LDA method in topic modeling was used to apply the text-mining approach on the raw dataset, implemented by Gensim Library in 'Python Jupyter Notebook' coding environment. Gensim was adopted due to its popularity as an open-source NLP package employed for unsupervised topic modeling. It performs a variety of complicated tasks by using leading academic and up-to-date statistical machine learning models; thus, Gensim library is favored among other topic modeling packages in several issues, including:
   - It allows handling huge text files without loading a whole file into memory;
   - Does not involve complex labeling or manual document tagging; and
   - Provides adequate and reliable facilities for building and evaluating high-quality topic models and text preprocessing [108].
   - For Software methodology, the Waterfall software development approach was suitable in this thesis since the software code for the proposed LDA topic

model was relatively small, simple, and there were no significant changes expected in the pre-defined requirements. Other reasons for adopting the Waterfall approach are:

- It is straightforward planning and designing process;
- A step-by-step phased (sequential) process;
- Each stage of the proposed model life cycle occurs in a sequence, one single cycle and single release of deliverables;
- The project has well-defined features within limited budget and timeline.

4.2.3. Data Preprocessing

1. *Text Cleaning and Preprocessing*: In this critical stage, the dimensionality of the text data will be reduced, and the noise in data will be filtered out as much as possible to enhance the content analysis process since the activities in this stage can help in reducing the computational time and preparing the data for further qualitative analysis. The main activities implemented in this stage were as follows:
   - Removing redundant features in the dataset (i.e., columns) that may not serve the purpose of the study, keeping on the necessary features: Title, Abstract, Authors Keywords, Year, Source Title, and Document Type, that will help in analyzing the articles, presented as data samples (i.e., rows).
   - Removing duplicated rows by abstracts (having similar meaning).
   - Removing numbers, regular expressions, punctuation marks, and whitespaces.
   - Removing unnecessary terms, such as no Abstract available, Introduction.
   - Replacing all NaN (not a number or missing data) values with empty strings.
   - Combining the contents of features: Title, Abstract, and Authors Keywords.
   - Converting text to lower case to avoid the distinction of those terms written in mixed cases.
   - The preprocessed data was saved as an Excel file.
2. Now the dataset is assumed cleaned and preprocessed enough to apply a set of *Descriptive Statistics* and *Quantitative Analysis* functions ready in the code, like:
   - Total number of documents published by year or country,
   - Number of citations by countries, and

- Top journals with the most articles in the scope of mobile learning.

However, the analysis results can be more clarified by iterating the steps in stage 1 to further cleaning the text in the preprocessed dataset.

3. To prepare the dataset (already cleaned from previous step) for LDA analysis, further specific preprocessing steps was required as follows:
    - Words length: Keeping words with more than 3 characters and less than 15.
    - Tokenization: Converting each sentence into individual words list (*Tokens*).
    - Phrase modeling: Building the Bigram and Trigram models. Bigrams are any two words that frequently co-occurring in the corpus, while Trigrams are any three words that frequently co-occurring in the corpus. Examples from this thesis' dataset, for instance (code output samples):
        - Bigram: 'augmented_reality'.
        - Trigram: 'artificial_neural_network'.
    - Removing unnecessary Stopwords (e.g., a, an, the, of, as) from previous text, then performing Lemmatization on the text using pre-defined functions; hence, only meaningful terms will be kept in forms of: nouns, adjectives, verbs, or adverbs.
    - Data Transformation: This step vectorizes the documents by creating a Dictionary and Corpus. The Dictionary represents a set of unique ids for each term, obtained by applying the function 'id2word' on the lemmatized data. Next step is to find the Term Document Frequency for each term using 'id2word.doc2bow' function; therefore, the final corpus will be the mapping of each (word_id, word_frequency).

4.3. LDA Model Implementation and Fitting

After preprocessing, the row count of data samples was reduced from 8191 to 7829.

The main steps involved in LDA implementation throughout the code were as follows:

1. Creating a base LDA model for training.
2. Evaluating LDA model.
3. Visualizing the topics.

These steps are explained in (Results Chapter 5); LDA part of Topic Modeling Results.

CHAPTER 5

RESULTS

The results of this study present two types of data analysis to the extracted dataset:
In the first part, Preliminary Bibliometric Analysis Results are used to descriptively present two types of analysis in order to answer the research questions 1 through 4, using the following methods:

1. The Statistical Distributions of documents, based on years, subject areas, publication types, countries, and affiliations, using Scopus analysis tool.
2. The Bibliometric Networks that visualize the relationship maps between documents, such as the co-occurrences of Author Keywords and co-authorships of Authors, Countries, and Organizations; based on the co-word analysis, with the number of documents and their citations, using VOSviewer analysis tool.

In the second part, Topic Modeling Analysis Results are used to evaluate the current status of the research topics in mobile learning literature and their evolution in order to answer the research questions 5 through 8; thus, providing a closer look at the content trends of the studies related to this domain, using LDA analysis approach.

5.1. Bibliometric Analysis – Mobile Learning (1968 - May 2021) – Scopus

5.1.1. Documents by Year Distribution in Mobile Learning

As shown by the line chart in figure 5.1, the search results in Scopus database related to mobile learning publications exhibited that 1968 was the beginning of the articles published in this field with one document. However, this number had almost remained steady without a significant change until 2002, when it started to increase slightly with just over ten articles in total. Later on, in 2004, the number of publications increased (almost linearly) with more than 50 documents, reaching its peak in 2019-2020 with nearly 750 documents. Thereafter, until this study was conducted in May 2021, there had been a significant drop in the number of publications to less than 400.

Figure 5.1: Visualization of documents by year distribution in Mobile Learning.

5.1.2. Documents by Subject Area Distribution in Mobile Learning

As demonstrated in figure 5.2, within the analyzed documents in the corpus, the top three subject areas in which most of their publications addressed topics related to m-learning are Computer Science (39.6%), Social Science (27.1%), and Engineering (12.5%), with a total proportion of 79.2% of all documents. Whereas the bottommost three subject areas whose publications discussed m-learning topics are Medicine (1.6%), Psychology (1.5%), and Physics and Astronomy (1.4).

Figure 5.2: Visualization of documents by subject area distribution in M-Learning.

5.1.3. Documents by Type distributions in Mobile Learning

Figure 5.3 shows that among the three selected types of publications related to mobile learning (journal articles, reviews, conference reviews, and conference papers) in the data collection step (Chapter 4), the majority of extracted publications were conference papers, led by 56.2%; then, article papers by 42.1%; whereas the minority of publications were conference reviews by 1.7% only.

Figure 5.3: Visualization of documents by type distribution in Mobile Learning

5.1.4. Documents by Country/Territory Distribution in Mobile Learning

The bar chart in figure 5.4 presents the top 10 of most productive countries or territories in terms of m-learning publications between 1968 and May 2021; in the first and second positions, China and the United States, respectively, led by more than 750 documents for each, while Taiwan came in third with just under 700.

Figure 5.4: Visualization of documents by country distribution in Mobile Learning.

5.1.5. Documents by Affiliation/Institution distributions in Mobile Learning

Figure 5.5 shows the top 10 affiliations to which the most m-learning publications belonged. The *National Taiwan University of Science and Technology* (Taiwan) was first with nearly 100 documents (about 1.28% of total articles), followed by *Athabasca University* (Canada) with nearly 85 documents (about 1.1%) and *National Central University* (Taiwan) with nearly 80 documents (about 1.0%), respectively.

Figure 5.5: Visualization of documents by affiliation distribution in Mobile Learning

5.2. Bibliometric Analysis – Mobile Learning (1968 - May 2021) – VOSviewer

5.2.1. Co-occurrence of Author Keywords

When analyzed, in total, there were 12187 different Author Keywords in the dataset. After limiting the number of occurrences threshold to 20, only 170 keywords are left. The top 20 author keywords having the highest number of co-occurrences with other keywords within the corpus on the dataset are shown in table 5.1.

Figure 5.6 (a) illustrates the co-occurrence networks of all the 170 author keywords. Figure 5.6 (b) illustrates the co-occurrence networks of only the top 20 keywords, showing the coupling relations among these most co-occurred author keywords.

Table 5.1: Co-occurrences of the Top 20 author keywords in Mobile Learning

Author Keywords	Occurrences	Link Strength
mobile learning	3975	4880
m-learning	1165	1808
e-learning	544	1115
mobile devices	291	575
higher education	285	573
augmented reality	257	452
education	206	455
mlearning	188	315
ubiquitous learning	186	385
mobile technology	183	336
collaborative learning	156	273
blended learning	137	268
mobile application	124	214
learning	122	235
game-based learning	120	226
informal learning	96	182
usability	95	188
cloud computing	94	167
gamification	94	197
mobile technologies	94	184

(a)

(b)

Figure 5.6: Visualization map of the co-occurrences of Author Keywords. (a): Among all author keywords, (b): Among the Top 20 keywords.

5.2.2. Co-Authorship of Authors according to the Number of Papers

When analyzed, in total, there were 14427 different *Authors* in the dataset.

By limiting the papers'/documents' threshold number to 10, only 140 authors are left.

Among the 140, only co-authors were taken into consideration for this analysis.

The Top 20 co-authors having at least 10 papers per author are listed in table 5.2.

Among the 140, figure 5.7 illustrates the co-authorship networks, showing the co-authorships cooperation among the different groups of co-authors within the corpus.

Figure 5.7: Visualization map of the Co-authorship of Authors by the Number of Papers in Mobile Learning.

Table 5.2: The Top 20 co-authors by the number of Papers in Mobile Learning.

Authors	# Of Documents
hwang, g.-j.	85
huang, y.-m.	45
ogata, h.	41
wang, j.	38
looi, c.-k.	35
specht, m.	31
milrad, m.	29
parsons, d.	25
yano, y.	25
seow, p.	22
so, h.-j.	22
wang, m.	22
yin, c.	22
hwang, w.-y.	21
li, y.	21
botticki, i.	20
castro, m.	19
liu, y.	19
wong, l.-h.	19
li, m.	18

5.2.3. Co-Authorship of Authors according to the Number of Citations

When analyzed, in total, there were 14427 different *Authors* in the dataset. After limiting the papers and citations' threshold to 10, only 183 authors are left. Among the 138, only co-authors were taken into consideration for this analysis. The Top 20 co-authors according to the number of citations are listed in table 5.3. Among the 138, figure 5.8 illustrates the co-authorship networks, showing the co-citations cooperation among the different groups of co-authors within the corpus.

Table 5.3: The Top 20 co-authors by the number of Citations in Mobile Learning

Authors	# Of Documents	# Of Citations
hwang, g.-j.	85	3652
kukulska-hulme, a.	16	1330
looi, c.-k.	35	1276
sharples, m.	15	1239
wong, l.-h.	19	1096
huang, y.-m.	45	1068
seow, p.	22	967
liu, t.-c.	16	870
chen, w.	12	831
chu, h.-c.	16	765
so, h.-j.	22	744
yang, s.j.h.	13	717
chang, k.-e.	10	714
sung, y.-t.	10	705
tsai, c.-c.	15	693
zhang, b.	15	652
wu, p.-h.	15	646
ogata, h.	41	554
wang, m.	22	533
milrad, m.	29	528

Figure 5.8: Visualization map of the Co-authorship of Authors by the Number of Papers in Mobile Learning

5.2.4. Co-Authorship of Countries according to the Number of Citations

When analyzed, in total, there were 148 different *Countries* in the dataset.
After limiting the papers and citations' threshold to 10, only 83 countries are left.
Among the 83, only countries with co-authorships were taken into consideration.
The Top 20 co-authorship countries by Citations' number are listed in table 5.4.
Among the 83, figure 5.9 illustrates the co-authorship networks, showing the co-citations cooperation among the different groups of countries within the corpus.

Table 5.4: Top 20 Co-authorship Countries by the number of Citations in Mobile Learning.

Country	# Of Documents	# Of Citations
taiwan	650	14725
united states	718	13563
united kingdom	573	10794
australia	370	4285
china	696	3737
spain	391	3727
turkey	161	2728
malaysia	470	2697
canada	228	2622
greece	169	2168
japan	207	2120
south korea	134	2104
singapore	112	2086
new zealand	128	1660
finland	146	1433
germany	280	1350
sweden	128	1277
netherlands	75	1093
saudi arabia	117	1030
switzerland	59	1023

Figure 5.9: Visualization map of the Co-authorship of Countries by the Number of Citations in Mobile Learning.

5.2.5. Co-authorship of Organizations according to the Number of Citations

When analyzed, in total, there were 3934 different *Organizations* in the dataset. After limiting the papers and citations' threshold to 20, only 79 organizations are left. Among the 79, only organizations with co-authorships were taken into consideration. The Top 15 co-authorship countries by Citations' number are listed in table 5.5.

Among the 79, figure 5.10 illustrates the co-authorship networks, showing the co-citations cooperation among the different groups of Organizations within the corpus.

Figure 5.10: Visualization map of the Co-authorship of Organizations by the Number of Citations in Mobile Learning.

Table 5.5: Co-authorship of the Top 20 Organizations by Citations in M-Learning

Organization	# Of Documents	# Of Citations
national taiwan university of science and technology taipei, taiwan	92	4613
national central university taiwan chung-li, taiwan	79	2889
national university of tainan taiwan tainan, taiwan	65	2698
national taiwan normal university taipei, taiwan	73	2379
the open university milton keynes, united kingdom	58	2271
national cheng kung university tainan, taiwan	62	1628
national institute of education singapore city, singapore	56	1606
athabasca university athabasca, canada	79	1094
university of technology sydney sydney, australia	42	870
university of wolverhampton wolverhampton, united kingdom	22	787
national sun yat-sen university taiwan kaohsiung, taiwan	21	727
national kaohsiung university of science and technology kaohsiung, taiwan	24	632
university of southampton southampton, united kingdom	21	586
university of hull hull, united kingdom	20	574
auckland university of technology auckland, new zealand	44	555
the university of hong kong pokfulam, hong kong	40	532
university of north texas denton, united states	25	528
nanyang technological university singapore city, singapore	23	500
university of wollongong wollongong, australia	24	486
tokushima university tokushima, japan	34	463

5.3. Topic Modeling Results

During the primary manual analysis of the collected CSV dataset contents, only one article was published in 1968, followed by three articles in 1984. *Orange Data Mining* software (Figure 5.11) was used to analyze the abstract of the article in 1968, which was found not related to the scope of this study; and was included in *Scopus* database due to a similarity in a keyword search term (m. learning), though this article discussed another issue in learning before the era of digital learning. As a result, this article was removed from the dataset before applying text mining to the rest of the articles, published since 1984, which were found 'related to the mobile learning concept in its foundation stage'. Accordingly, the raw dataset used in this study will include all publications related to *Mobile Learning*, covering the period from 1984 to May 2021.

Figure 5.11: Analyzing Dataset Corpus in Orange Data Mining Software [109, 110]

5.3.1. Descriptive Content Analysis Results

The total row count (Number of documents) in the raw dataset before applying text mining techniques (e.g., filtering and preprocessing) was **8191** (as data samples).

After filtering out 'NaN' values in abstracts (Not a Number or missing abstracts) and the duplicated abstracts (e.g., published in multiple sources), the new row count will be reduced to **7829** documents; the number of all unique **'**Tokens**'** becomes **12363**, and the total number of citations of all publications (Total Cited By) is **94764**.

Before applying the LDA algorithm to the corpus contents, the following are main descriptive analysis results after text preprocessing steps, implemented by *Python v3*:

5.3.1.1. Documents in Mobile Learning by Year Distribution (1984 - May 2021)

As demonstrated in table 5.6, the number of publications in mobile learning noticeably started to increase (albeit slightly) in 2002 by 12 documents, compared to 9 documents (in total) in the period from 1984 to 2001. Afterward, the number continued to increase at almost the same rate (below 0.5% annually), then it relatively began to increase more to reach 116 documents in 2006, followed by 228 documents in 2007; whereas, the peak number of publications in m-learning was consecutively recorded in 2019 by 738 documents, followed by 740 in 2020.

During the first third of 2021, until this study's dataset was collected in May 2021, there had been a noticeable decline in the number of publications, reaching only 222 documents. In general, the distribution of the total number of publications in mobile learning in the past two decades was as follows:

- During the period from 2001 to 2010: The total number of publications reached 1521 documents, which is equivalent to 19.42% of total publications.
- During the period from 2011 to 2020: The total number of publications reached 6079 documents, which is equivalent to 77.64% of total publications.

Table 5.6: Papers in Mobile Learning by Year distribution (1984 - May 2021)

Years	# Of Papers	%
2021	222	2.84%
2020	740	9.45%
2019	738	9.43%
2018	708	9.04%
2017	611	7.80%
2016	602	7.69%
2015	620	7.92%
2014	611	7.80%
2013	493	6.30%
2012	526	6.72%
2011	430	5.49%
2010	430	5.49%
2009	325	4.15%
2008	254	3.24%
2007	228	2.91%
2006	116	1.48%
2005	86	1.10%
2004	53	0.68%
2003	15	0.19%
2002	12	0.15%
2001	2	0.03%
2000	2	0.03%
1998	1	0.01%
1992	1	0.01%
1991	1	0.01%
1984	2	0.03%
	7829	**100%**

5.3.1.2. Documents by Type Distribution in M-Learning (1984 - May 2021)

Table 5.7 shows that among the 7829 documents published in m-learning, the majority were 'Conference Papers' with 55.78% of the total publications (4367 documents), followed by 'Journal Articles' with 42.50% (3327 documents); whereas the minority were 'Review Articles' with 1.72% (135 documents).

Table 5.7: Types of Publications in Mobile Learning (1984 - May 2021).

Source Type	# Of Papers	%
Conference Paper	4367	55.78%
Journal Article	3327	42.50%
Review Article	135	1.72%
Total	**7829**	**100.00%**

5.3.1.3. Documents by Journal Distribution in M-Learning (1984 - May 2021)

Regarding the source titles, table 5.8 lists the Top 15 Journals with the most articles in m-learning with 1931 documents in total (24.7% of all). Among these sources, the *Lecture Notes in Computer Science (LNCS)* came first with 321 documents, followed by the *International Conference Proceedings Series (ICPS)* with 216, and the *International Journal of Mobile Learning and Organisation (IJMLO)* with 184 articles.

Table 5.8: Top 15 Journals with the most articles in Mobile Learning (1984-May 2021)

Source Name	# Of Papers
Lecture Notes in Computer Science	321
ACM International Conference Proceeding Series	216
International Journal of Mobile Learning and Organisation	184
International Journal of Interactive Mobile Technologies	161
Communications in Computer and Information Science	155
Advances in Intelligent Systems and Computing	146
Computers and Education	120
International Journal of Mobile and Blended Learning	118
CEUR Workshop Proceedings	104
Journal of Physics: Conference Series	89
Educational Technology and Society	76
Education and Information Technologies	67
Interactive Learning Environments	60
British Journal of Educational Technology	58
International Review of Research in Open and Distance Learning	56

5.3.1.4. Documents by Country Distribution in M-Learning (1984-May 2021)

Figure 5.12 shows the distribution of the number of publications in mobile learning among the Top 15 most productive countries in this regard, with a total of 5561 documents, representing 71% of all publications. The *United States* came first with 718, followed by *China* and *Taiwan* with 695 and 650 documents, respectively.

Figure 5.12: Distribution of M-Learning Publications by Countries (1984-May-2021).

5.3.1.5. Documents by Citations Distribution in M-Learning (1984-May 2021)

Figure 5.13 shows the distribution of the total number of citations in m-learning among the Top 15 most cited countries with a total of 70431 citations, representing 74.32% of all publications citations. *Taiwan* was first with 14725 documents, followed by the *United States* and *United Kingdom* with 13563 and 10794 documents, respectively.

Figure 5.13: Distribution of M-Learning Citations by Countries (1984-May 2021).

5.3.1.6. Most Frequent Terms in Mobile Learning Documents (1984 - May 2021)

Figure 5.14 depicts the Word Cloud representation of the most frequent terms in the term-document matrix of the corpus after cleaning and preprocessing steps in the topic modeling process (chapter 3), such as the removal *Stop Words*. In this technique, the size of the term reflects its weight (TF-IDF) compared to other terms in the entire corpus. It can be seen that among the top frequent terms are '*mobile_learn*', '*study*', '*design*', '*student*', '*education*', '*mobile_device*', etcetera.

Figure 5.14: Word Cloud representation of Mobile Learning Top Terms (1984 - May 2021).

5.3.2. LDA Content Analysis Results

5.3.2.1. Topics' Coherence and Dominant Topic Analyses

Topic Coherence Analysis is an evaluation technique in topic models, used to estimate the optimal number of topics to be used for LDA content analysis. In this study, for each number of topic distributions from 5 to 30 topics, *Coherence Score and Dominant Topic* analyses (shown in figures 5.15 and 5.16) were performed separately using the optimal values of LDA parameters: ***alpha*** and ***eta*** (or ***beta***). As a result of the analysis, the most appropriate number of topics was **12**. In this study, the *Parallelized Latent Dirichlet Allocation* (ldamulticore) in *Python's Gensim* package was used, utilizing all available CPU cores to parallelize and speed up the model training process [111].

The optimal LDA parameters calculation for Alpha and Eta for 12 topics was performed as follows:

eta = alpha = (1/Number of Topics); Therefore, alpha = eta = 1/12 = **0.083**

```
Num Topics = 5  is having Coherence Value of 0.3458
Num Topics = 6  is having Coherence Value of 0.3737
Num Topics = 7  is having Coherence Value of 0.3656
Num Topics = 8  is having Coherence Value of 0.3723
Num Topics = 9  is having Coherence Value of 0.3631
Num Topics = 10  is having Coherence Value of 0.3519
Num Topics = 11  is having Coherence Value of 0.372
Num Topics = 12  is having Coherence Value of 0.3521
Num Topics = 13  is having Coherence Value of 0.3573
Num Topics = 14  is having Coherence Value of 0.3537
Num Topics = 15  is having Coherence Value of 0.3573
Num Topics = 16  is having Coherence Value of 0.3565
Num Topics = 17  is having Coherence Value of 0.3441
Num Topics = 18  is having Coherence Value of 0.3496
Num Topics = 19  is having Coherence Value of 0.3467
Num Topics = 20  is having Coherence Value of 0.3518
Num Topics = 21  is having Coherence Value of 0.3465
Num Topics = 22  is having Coherence Value of 0.342
Num Topics = 23  is having Coherence Value of 0.354
Num Topics = 24  is having Coherence Value of 0.3588
Num Topics = 25  is having Coherence Value of 0.3425
```

Figure 5.15: LDA Content Analysis Coherence Score Graph and Code Results.

```
Dominant_Topic 10
0     2148
1        5
2      333
3      113
4     2159
5       78
6     1104
7      448
8        2
9     1439
```

```
Dominant_Topic 11
0      243
1     1209
2       35
3        1
4       10
5      173
6     1731
7      678
8     2892
9      803
10      54
```

```
Dominant_Topic
12
0      158
1       33
2      380
3      149
4     1961
5      682
6     1025
7     1957
8      431
9      692
10      28
11     333
```

```
Dominant_Topic
13
0      134
1       22
2      207
3      361
4     1498
5       70
6       90
8        6
9       58
10     403
11      70
12    4910
```

```
ominant_Topic14
0       94
1       29
2        2
3      442
4       74
5     3410
6        6
7       82
8     1570
9        1
10      14
11     990
12    1098
13      17
```

```
Dominant_Topic
15
0     1193
1       30
2        5
3      208
5      168
6     3893
7       11
8        7
9      436
10     430
11       1
12     385
13    1042
14      20
```

```
Dominant_Topic
16
0     2353
1      453
2      744
4      369
5      884
6       31
7       27
8      198
9      511
10      23
11     919
12      98
13    1190
14      10
15      19
```

```
Dominant_Topic17
0      787
1      155
2       20
3     2729
4      493
5     1459
6      167
7      878
8       10
9        3
10       2
11     121
13      99
14     365
15       1
16     540
```

```
Dominant_Topic18
0       32
1      836
2      243
3     2217
4       23
5        2
6       45
8     1794
9        9
10     871
11     656
12      51
13     404
14      17
15       3
16     422
17     204
```

```
Dominant_Topic19
0       22
1      106
2       12
3     1221
4       13
5       13
6       11
7       54
8      192
9        6
11    1121
12     261
13     400
14     810
15    1375
16      49
18    2163
```

```
Dominant_Topic 20
0      382
1     1311
2-3    138
5      391
6       11
7       48
8     2486
9       39
10      19
11    2179
12       4
13      15
14       2
15       1
16     288
17       1
18     511
19       3
```

```
Dominant_Topic
21
0        2
1      388
2        3
3      165
4       90
5     1074
6      224
7      581
8      136
10     145
11     832
12       3
13       4
14     605
15     674
16     348
17       2
18     512
19      20
20    2021
```

```
Dominant_Topic
22
0        6
1      577
2       69
3        4
4       34
5      218
6     1929
7      434
8       59
10      47
11      29
12       1
13     245
14       4
15    1112
16      53
17     131
18     454
19    1362
20    1053
21       8
```

```
Dominant_Topic23
0     1086
1        6
2      702
3      715
4       20
5      303
6      516
7       75
8      121
9      186
10     841
12     243
13     242
14     229
15      14
16       1
17    1702
18      21
20       3
21       3
22     800
```

Figure 5.16: LDA Content Analysis Dominant Topic Analysis.

5.3.2.2. LDA Topics in Mobile Learning inferred by Dominant Topic Analysis

Figure 5.17 shows a screenshot of the table resulting from the relevant code cell, illustrating the most representative sentences of the 12 dominant topics inferred by LDA. The Word Cloud representations of the LDA topics are shown in Figure 5.18. The preliminary analysis results obtained by the code were used with a help of the domain expert to answer research question 6, related to 'the topics were commonly researched in m-learning'; a list of the topics is organized in detail in section (5.3.2.4).

Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0.0	0.7916	study, mobile_learne, review, library, literature, education, analysis, finding, field, identify	['study', 'identify', 'publication', 'education', 'describing', 'feature', 'study', 'web', 'science', 'database', 'search', 'publication', 'education', 'bibliograp...
1.0	0.4672	training, knowledge, medical, agent, player, elearning, train, management, intervention, skill	['skill', 'management', 'postpartum', 'hemorrhage', 'neonatal', 'resuscitation', 'preØpost', 'intervention', 'study', 'globally', 'mobile_learne', 'mlearning', 'too...
2.0	0.7109	game, application, design, user, app, mobile, child, usability, develop, base	['usability', 'kindergarten', 'malay', 'language', 'reading', 'emphasize', 'potential', 'mobile', 'educational', 'application', 'enhance', 'read', 'kindergartener',...
3.0	0.6992	base, network, architecture, platform, mobile_learne, resource, service, datum, intelligent, design	['design', 'fragment', 'mobile_learne', 'resource', 'base', 'interactive', 'approach', 'quality', 'mobile_learne', 'resource', 'directly', 'efficiency', 'mobile_lea...
4.0	0.9752	learn, student, study, learning, language, base, mobile_learne, method, skill, improve	['acquisition', 'language', 'compare', 'independent', 'study', 'student', 'experimental', 'mobile', 'student', 'control', 'student', 'form', 'class', 'design', 's...
5.0	0.8166	education, technology, teacher, school, project, digital, educational, study, teaching, practice	['mobile_learne', 'socio', 'materiality', 'classroom', 'practice']
6.0	0.9817	learn, learning, design, support, mobile_learne, activity, mobile, context, environment, paper	['assist', 'space', 'mat', 'potential', 'learn', 'paper', 'outline', 'space', 'theory', 'support', 'teach', 'learn', 'context', 'educational', 'technology', 'lea...
7.0	0.9427	learn, mobile, technology, mobile_learne, learning, application, paper, education, service, development	['stick', 'era', 'stride', 'mobile_learne', 'paper_discusse', 'sluggish', 'growth', 'rate', 'mobile_learne', 'phone', 'propose', 'application', 'mobile', 'technolo...
8.0	0.9129	study, mobile_learne, factor, learn, student, acceptance, influence, adoption, technology, perceive	['factor', 'predict', 'online', 'university', 'study', 'analyze', 'relationship', 'factor', 'predict', 'online', 'university', 'studentso', 'mobile_learne', 'manag...
9.0	0.8252	student, mobile, phone, device, mobile_device, learn, access, support, classroom, wireless	['mobile', 'phone', 'teacher', 'day', 'class', 'mobile', 'phone', 'mobile_device', 'ban', 'disruptive', 'technology', 'classroom', 'investigation', 'mobile', 'pho...
10.0	0.5487	model, medium, evaluation, assessment, base, technique, expert, profile, vocational, quality	['investigation', 'primary', 'component', 'factor', 'analysis', 'assess', 'standard', 'primary', 'component', 'factor', 'advantage', 'mobile_learne', 'terminal', '...
11.0	0.7665	learner, learn, learning, content, environment, ubiquitous, user, adaptive, object, base	['personalized', 'recommendation', 'mobile', 'base', 'learning', 'experience', 'learner', 'learn', 'anytime', 'portable', 'mobile', 'device', 'vast', 'educational'...

Figure 5.17: The Tabular Output of Dominant Topic Analysis Results.

Figure 5.18: The Word Cloud representation of the 12 LDA Topics.

5.3.2.3. LDA Topics Visualization and Interpretation

To better understand the weights and interrelations of the topics inferred by LDA, visualization tools are commonly used to depict the *Inter-Topic Distance Map* in a multidimensional space, as shown in figure 5.19. In this study, *pyLDAvis*, a Python package for an interactive topic modeling visualization, was used to help interpret the topics by extracting the relevant terms (keywords) from the fitted LDA topic model to form an interactive web-based visualization [112].

On the left side, the topics circles' area reflects the term count associated with each topic. The circles are arranged by term frequency, so the closer the topics are to each other, the more words in common they have. On the right side, the bar chart depicts the most salient terms. The bars represent the overall frequency of the term throughout the corpus [113]. In this visualization technique, two main metrics are used: First, the *Salient* measure identifies the most informative terms, representing topics across all texts. The more saliency the term has, the more beneficial it is for identifying a particular topic. Second, the *Relevance* measure utilizes a parameter named (Lambda) to rank the terms within topics based on the probability distribution of their weights. The Lambda slider is used to alter the topic terms displayed in the bar chart so the domain expert can distinguish the most helpful order of terms in order to label a particular topic. This study performed topic labeling by changing the Lambda values between (0.5 - 1), which gave the most appropriate order of the inferred topics.

Figure 5.19: Inter-Topic Distance Map and Top of Most Relevant Terms per Topic.

5.3.2.4. Most Researched Topics in M-Learning Publications (1984 - May 2021)

The LDA-based topic modeling analysis identified 12 dominant topics extracted from a corpus database of 7829 published documents in mobile learning. These topics and their topmost frequent keyword terms representing each topic are illustrated in table 5.8. In this study, the topic labels were determined based on the highest order of most salient relevant terms associated with each topic, as mentioned in the previous section. The order of the keywords of each topic was considered when the domain expert assigned names (labels) to the topics, often by combining the top five keywords to describe a meaningful name for each topic. The 12 dominant topics in mobile learning publications from 1984 to May 2021 were classified in table 5.9 using *LDA Topic Modeling Analysis* based on the number of documents they represented in the entire corpus database, containing 7829 documents in total. As demonstrated in table 5.9, the topmost three discussed topics in the documents were: "*Student Language Learn*" was first with a total of 1961 documents, representing over 25.0% of the total documents; followed by "Learn with Mobile Technology" as second with 1957 documents (25.0%); and "*Learning Design*" as third with 1024 documents (13.1%). Whereas the least three discussed topics, together representing 2.7% of the total documents, were as follows: "*Network Architecture*" with 149 documents; followed by "*Training Medical Knowledge*" with 33; and "*Model/Medium Evaluation*" with 29 documents.

Table 5.9: Top 12 Dominant Topics in Mobile Learning between (1984 - May 2021).

T #	Topic Top-Rated Keywords	Topic Label	# Of Papers	%
T5	student, language, learn, study, motivation, performance, english, skill, effectiveness, method, experimental, class, experiment, improve,	Student Language Learn	1961	**25.0 %**
T8	learn, mobile, technology, mobile_learne, learning, application, paper, education, service, development, user, mobile_device, system,	Learn with Mobile Technology	1957	**25.0 %**
T 7	learn, learning, design, support, mobile_learne, activity, mobile, context, environment, paper, base, technology, experience, collaborative, approach	Learning Design	1024	**13.1 %**
T10	student, mobile, phone, device, mobile_device, learn, access, support, classroom, wireless, lecture, experience, smartphone, question, online	Student Mobile Phone	692	**8.8 %**
T6	education, teacher, school, digital, project, mlearne, technology, professional, literacy, practice, country, adult, educational, teaching,	Teacher Education	682	**8.7 %**
T9	factor, acceptance, study, influence, mobile_learne, perceive, adoption, attitude, intention, perception, investigate, survey, modeling, variable, proposal	Acceptance Model /Acceptance of Mobile	431	**5.5 %**
T3	game, application, design, user, app, mobile, child, usability, develop, base, prototype, educational, evaluation, test, paper	Game Application Design	380	**4.9 %**
T12	learner, learn, learning, content, environment, ubiquitous, user, adaptive, object, base, material, multimedia, context, style, provides	Learning Content	333	**4.3 %**
T1	study, mobile_learne, review, library, literature, education, analysis, finding, field, identify, technology, paper, article, trend, issue	Mobile Learning Review	158	**2.0 %**
T4	base, network, architecture, platform, mobile_learne, resource, service, datum, intelligent, design, mode, propose, function,	Network Architecture	149	**1.9** %
T2	training, knowledge, medical, agent, player, elearning, train, management, intervention, skill, health, project, tutor, regulate, risk	Training Medical Knowledge	33	**0.4** %
T11	model, medium, evaluation, assessment, base, technique, expert, profile, vocational, quality, method, criterion, analysis, apply, level	Model/Medium Evaluation	29	**0.4** %
			7829	**100**

5.3.2.5. The Status of Dominant Topics Publications in M-Learning (1984-2021)

Due to the relatively young age of the mobile learning domain characterized by fast paces of development in technology, the analyses in this section and beyond were taken every five years to trace the changes in the dominant topics commonly researched. This section provides a clear picture regarding the status of m-learning publications related to these topics by analyzing figures 5.20 and 5.21 as follows:

First, the change rates in the total number of m-learning publications related to the 12 dominant topics can be determined from the graph in figure 5.20, which demonstrates five-year distributions of the dominant topics' publications among all 7829 documents. The analysis shows that, by the end of 2011, the total number of dominant topics' publications had increased significantly by a rate of 22.0%, from 83 to 1804 documents. Then, by the end of 2016, the number had increased by 15.0%, from 1804 to 2973 documents, reaching its most peak at 38% of the total publications; then, it had slightly decreased by 0.16%, from 2973 to 2960 documents in May 2021.

Figure 5.20: The Distribution of M-Learning Publications in the Dominant Topics, by Five-Year periods (1984 - May 2021).

Second, figure 5.21 depicts the distribution of total publications in each of the 12 dominant topics between 1984 and May 2021, showing that among these topics, “Student Language Learn” and “Learn with Mobile Technology” were at the topmost of research interests, each with 25.0% of total publications. The second level was held by “Learning Design” with 13.1%, followed by both “Student Mobile Phone” and “Teacher Education”, at almost comparable levels with 8.9% and 8.7%, respectively. In contrast, the lowermost topics of research interests were “Model/Medium Evaluation” and “Training Medical Knowledge”, each with 0.4% of total publications, followed by both “Network Architecture” and “Mobile Learning Review”, at almost comparable levels, with 1.9% and 2.0% respectively. Therefore, as demonstrated in figure 5.21 and table 5.9, the *Top Three Dominant Topics* among the most researched topics in M-Learning between 1984 - May 2021, based on the total publications, were: “Learn with Mobile Technology, Student Language Learn, and Learning Design”.

Figure 5.21: The Distribution of M-Learning Publications (%) by Dominant Topics (1984 - May 2021).

5.3.2.5. Trend Topics Distribution by the Number of Publications for each Topic

In general concept, the *Topic Evolution* indicates ‘how a topic is changing over time’. In this study, although the publications in mobile learning already started in 1984, the trend analysis conducted for the 12 dominant topics and their evolutions considered

only the publications period since 2002 that witnessed the first remarkable rise in the number of publications ever, from 2 to 12 documents, as previously explained in both the Bibliometric and Topic Modeling results (tables 5.1 and 5.6). This analysis aims to provide a holistic perspective of the identified trend topics in mobile learning between 2002 and May 2021, in terms of the following two criteria:

First, an Acceleration Analysis of each dominant topic was manually performed in 'Excel' by the domain expert to investigate the acceleration rate for each topic, measured for the five-year periods between 2002 and May 2021, as demonstrated in figure 5.22; by subtracting the topic's publications percentage of the previous period's value from the percentage of the current period's value and dividing the result by 5. Therefore, the difference values (Delta) for all dominant topics, representing their acceleration rates, are plotted as graph charts, shown in figure 5.22. In general, the purpose of the acceleration rates graph is to identify the topics of 'Research Interests' within a given period (regardless of the number of publications discussing this topic) by analyzing the value of 'delta' for each topic. For instance, in the period (2002-2006), although topic T5 (*Student Language Learn*) reached the highest number of articles in this period with 38.6% of total publications, the rate of acceleration was low (-2.29), indicating that within this period, the research interest in this topic decreased compared to the previous period (1997-2001), when it reached 50% of total publications. Contrarily, topic T8 (*Learn with Mobile Technology*) reached 24% of total publications between (2002-2006), while its acceleration rate (5.78), indicating that the research interest in this topic increased compared to the previous period (0%).

Figure 5.22: Acceleration Analysis of Trend Topics in M-Learning (2002 - May 2021).

Second, an *Evolution Analysis* of the top three trend topics among the 12 dominant topics in mobile learning was then conducted based on the distribution of the total publications in the dominant topics by five-year periods between 2002 and May 2021, as demonstrated in figure 5.23. For instance, in the first period (2002 - 2006), the total number of m-learning publications in all dominant topics was relatively small (below 100 documents); at the top rank, most publications discussed the "*Student Language Learn*" topic, followed by "*Learn with Mobile Technology*", whereas, at the third rank, the "*Learning Design*" and "*Teacher Education*" topics were discussed evenly. Later on, in the following periods (2007 - 2011) and (2012 - 2016), the "*Learn with Mobile Technology*" topic came first with about 600 then 800 documents, respectively; whereas the second and third ranks were alternately held by "*Learning Design*" and "*Student Language Learn*" topics. Recently, in the period (2017 - May 2021), the "*Student Language Learn*" topic came back to the first rank of most discussed topics, in more than 1000 documents, followed by "*Learn with Mobile Technology*" with about 550, then "*Teacher Education*" with less than 300 documents.

Figure 5.23: The Distribution of M-Learning Publications in the 12 Dominant Topics, by Five-Year periods (2002 - May 2021).

As a result of this analysis, table 5.10 lists the order of *Trend Topics* in mobile learning for every five-year period, compared to the other periods between 2002 and May 2021. The rank given to each trending topic was determined by considering the number of publications discussing this topic to learn '*how these trend topics evolved*' in terms of the research interests of scholars in such topics over time.

In general, among the 12 dominant topics in m-learning, only four topics interchanged the ranks of the top three trend topics among each other within the five-year periods between 2002 and May 2021. The first period (2002 - 2006) witnessed the initial rise in m-learning publications, specifically in "*Student Language Learn*" as the most researched topic, followed by "*Learn with Mobile Technology*", and the third rank of the research trend was interestingly shared between "*Teacher Education*" and "*Learning Design*" topics. However, until this study was conducted, the trending order had alternately changed among the 4 topics in the later periods, as shown in table 5.10.

Table 5.10: The Top Three Mobile Learning Trend Topics, for Five-Year periods (2002 - May 2021).

	First Topic	***Second Topic***	***Third Topic***
2002 – 2006	Student Language Learn	Learn with Mobile Technology	Teacher Education**;** Learning Design
2007 – 2011	Learn with Mobile Technology	Learning Design	Student Language Learn
2012 – 2016	Learn with Mobile Technology	Student Language Learn	Learning Design
2017 – 2021	Student Language Learn	Learn with Mobile Technology	Teacher Education

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1. Results Discussion

This study aimed to investigate the status of the Mobile Learning field in the scientific literature with two analysis approaches, *Bibliometric* and *Topic Modeling* techniques. Throughout the analysis process, this study sought to answer a set of research questions 1 through 8 composed in the 'Introduction Chapter 1' by analyzing a corpus of 7829 documents extracted from *Elsevier's Scopus Database*.

As pre-mentioned in Topic Modeling Results (Section 5.3), although one document was found in the Scopus databases results related to the 'keyword search query' used in 'Methodology Chapter 4' between 1968 and 1983, this document was excluded from the analysis due to its content that when manually analyzed by the domain expert, found not related to the scope of this study. Therefore, for better quality analysis, the range of publications related to Mobile Learning included by this study will only cover the period between 1984 - May 2021. Accordingly, the following is a summary of the most relevant findings to the research questions, based on the 'Results Chapter 5':

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

6.1.1.1. Documents by Year Distribution in Mobile Learning (1984 - May 2021)

Regarding the number of publications discussing topics related to mobile learning, it can be concluded that the volume of articles started to increase slightly in 2002, exceeding ten articles (12 documents in total), as demonstrated by figure 5.1 and table 5.6. These results support the related literature for this period, where m-learning was still considered a "very young and newly emerging topic" in the field of distance education and e-learning [10].

Since 2004 it was noted that the number of publications started to increase noticeably, exceeding 100 documents in 2006 and 400 documen in 2010, to reach its peak with 750 documents in 2020. These results are in agreement with the statistics reported by two former studies in the literature review on mobile learning, one of which used a systematic review approach [28] and the other that used text mining [39] though both studies were within a short period and therefore a limited number of relevant publications. However, the number of m-learning publications decreased during the first third of 2021 to reach 222 documents.

6.1.1.2. Documents by Subject Area Distribution in M-Learning (1984 - May 2021)

As demonstrated in figure 5.2, within the analyzed documents in the corpus, the top three subject areas in which most of their publications addressed topics related to m-learning are Computer Science (39.6%), Social Science (27.1%), and Engineering (12.5%), with a total proportion of 79.2% of all documents. Whereas the bottom-most three subject areas whose publications discussed m-learning topics are Medicine (1.6%), Psychology (1.5%), and Physics and Astronomy (1.4).

6.1.1.3. Documents by Type distributions in Mobile Learning (1984 - May 2021)

Based on the data collection method used in this study, three types of publications related to m-learning were selected (Journal articles, Reviews, Conference reviews, and Conference papers). As demonstrated in figure 5.3 and table 5.7, among these types, the majority of publications belonged to 'Conference Papers' by more than 55% (4367 documents), followed by 'Article Papers' by 42.5% (3327 documents), whereas the minority of publications were 'Conference Reviews' by 1.72% (135 documents).

6.1.2. Which are the Most Productive Countries in Mobile Learning? (Research Question 2)

6.1.2.1. Documents by Country Distribution in Mobile Learning

As demonstrated in figures 5.4 and 5.12, among the Top 15 most productive countries/regions in m-learning publications, the *United States* came first with 718 documents, followed by *China* and *Taiwan* with 695 and 650 documents, respectively.

These countries were mentioned in this regard by two former studies in the literature review, one of which used a traditional bibliometric approach [32] and the other that used text mining [39] though both were within a short period and limited publications.

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

6.1.3.1. Documents by Citations Distribution in Mobile Learning

Figure 5.13 shows the distribution of the total number of citations in m-learning among the Top 15 most cited countries with a total of 70431 citations, representing 74.32% of all publications citations. *Taiwan* was first with 14725 documents, followed by the *United States* and *United Kingdom* with 13563 and 10794 documents, respectively.

6.1.3.2. Documents by Affiliation/Institution distributions in Mobile Learning

Figure 5.5 shows the top 10 affiliations to which the most m-learning publications belonged. Among these institutions, the *National Taiwan University of Science and Technology* (Taiwan) was first with nearly 100 documents (1.28% of total articles), followed by *Athabasca University* (Canada) with nearly 85 documents (1.1%) and *National Central University* (Taiwan) with nearly 80 documents (1.0%), respectively.

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

6.1.4.1. Co-Authorship of Countries according to the Number of Citations

Among 148 contries in the raw dataset, only 83 countries having a minimum of 10 papers and citations were analyzed with VOSviewer Co-Authorship Networks, as depicted by the visualization map in figure 5.9, showing the co-citation cooperation among the diferent groups of countries within the corpus. Table 5.4 lists the Top 20 *Co-Authorship Countries* according to the number of citation in mobile learning. These results show that *Taiwan* was the major contributor in m-learning publications' co-citation with other co-authorship countries, with 14725 citations, followed by the *United States* with 13563, and the *United Kingdom* with 10794 citations.

6.1.4.2. Co-authorship of Organizations according to the Number of Citations

Among 3934 organizations in the raw dataset, only 79 having a minimum of 20 papers and citations were analyzed with VOSviewer Co-Authorship Networks, as depicted by the visualization map in figure 5.10, showing the co-citations cooperation among the different groups of organizations within the corpus. Table 5.5 lists the Top 15 *Co-Authorship Organizations* according to the number of citations in mobile learning. The results show that the top three co-authorship organizations contributed in m-learning publications' co-citation with other co-authorship organizations, were the *National Taiwan University of Science and Technology* (Taiwan) with 4613 citations, followed by the *National Central University* (Taiwan) with 2889, and the *National University of Tainan* (Taiwan) with 2698 citations.

6.1.5. In which Journals were Mobile Learning Studies Mainly Published? (Research Question 5)

6.1.5.1. Documents by Journal Distribution in Mobile Learning

Regarding the source titles, table 5.8 lists the Top 15 Journals with the most articles in m-learning with 1931 documents in total (24.7% of all). Among these sources, the *Lecture Notes in Computer Science (LNCS)* came first with 321 documents, followed by the *International Conference Proceedings Series (ICPS)* with 216, and the *International Journal of Mobile Learning and Organisation (IJMLO)* with 184 articles.

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

6.1.6.1. Dominant Topics in M-Learning Publications (Topic Modeling Analysis)

The 12 dominant topics in mobile learning publications from 1984 to May 2021 were ranked in table 5.8 based on the number of documents they represented in the corpus; the topics were labeled by the top frequent keywords that best describing each topic. The most studied m-learning topics in the documents were: '*Student Language Learn*', '*Learn with Mobile Technology*', and '*Learning Design*'.

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

6.1.7.1. The Status of Dominant Topics Articles in M-Learning (2002 - May 2021)

First, the change rates in the total number of m-learning publications related to the 12 dominant topics were determined from the graph in figure 5.20, demonstrating five-year distributions of the total publications in the dominant topics. The analysis shows that, by the end of 2011, the total number of dominant topics' publications had increased significantly by a rate of 22.0%, from 83 to 1804 documents. Then, by the end of 2016, the number had increased by 15.0%, from 1804 to 2973 documents, reaching its most peak at 38% of the total publications; then, it had slightly decreased by 0.16%, from 2973 to 2960 documents in May 2021.

Second, the distribution of total publications in each of the 12 dominant topics is depicted in figure 5.21, showing that the '*Student Language Learn*' and '*Learn with Mobile Technology*' were at the topmost of research interests, each with 25.0% of total publications, while '*Learning Design*' came second with 13.1%, followed by '*Student Mobile Phone*' and '*Teacher Education*', at almost comparable levels with 8.9% and 8.7%, respectively. Therefore, as demonstrated in figure 5.21 and table 5.9, the *Top Three Dominant Topics* among the most researched topics in Mobile Learning between 1984 and May 2021, based on the total number of related publications, were: "Learn with Mobile Technology", "Student Language Learn", and "Learning Design".

6.1.8. How did the Research Topics in Mobile Learning evolve over time? (Research Question 8)

6.1.8.1. Trend Topics Distribution by the number of Publications (2002 - May 2021)

In general concept, the *Topic Evolution* indicates 'how a topic is changing over time'. This study conducted a trend analysis for the dominant topics in mobile learning and their evolutions for five-year periods, between 2002 and May 2021, as follows:

First, an Acceleration Analysis of each dominant topic was manually performed to determine the acceleration rate for each topic publications, as illustrated in figure 5.22. The purpose of the acceleration analysis is to identify the topics of 'Research Interests' within a given period (regardless of the number of publications discussing this topic).

Second, an *Evolution Analysis* of the top three trend topics among the 12 dominant topics in mobile learning was then conducted based on the distribution of the total publications in the dominant topics by five-year periods, as illustrated in figure 5.23.

By comparing the results of the two analyses (Topic Acceleration and Evaluation), depicted in figure 5.22 and 5.23, respectively; it is noted that the effect of a topic acceleration in a given period, not always reflected on the number of publications related to this topic within the next period. As a result of these analyses, Table 5.10 lists the order of the *Three Trend Topics* in mobile learning for every five-year period, compared to the other periods between 2002 and May 2021. In general, among the 12 dominant topics in mobile learning, only four topics interchanged the ranks of the top three trend topics among each other within the five-year periods between 2002 and May 2021, as follows:

- In the first period (2002 - 2006): The initial rise in m-learning publications, specifically in '*Student Language Learn*', was the most researched topic, followed by '*Learn with Mobile Technology*', while a third rank of the trend was interestingly shared among '*Teacher Education*' and '*Learning Design*' topics.
- In the second period (2007 - 2011): '*Learn with Mobile Technology*' topped the research trend ranking, followed by '*Learning Design*', as it moved up to the second place, whereas '*Student Language Learn*' fell behind to the third place, and '*Teacher Education*' left the ranking below the third.
- In the third period (2012 - 2016): The '*Learn with Mobile Technology*' topic remained a trending topic though it had started to decrease in the publications as a topic of 'research interests', whereas both '*Learning Design*' and '*Student Language Learn*' topics interchanged their places in the trend ranking as second and third with each other, respectively.
- In the recent period (2017 - May 2021): '*Student Language Learn*' climbed back to the top of the 'research interest' trend, followed by '*Learn with Mobile Technology*', as it fell behind to the second place, whereas, in the third-place '*Teacher Education*' was back to the trend ranking, and '*Learning Design*' left the ranking below the third.

6.2. Conclusion

The main objective of this study was to detect the '*Research Themes and Trends in Mobile Learning*' from its existence to today. The main steps taken by this study were:

First, the study conducted a wide-spectrum *Literature Review* on the former studies in m-learning to evaluate the analysis methods used in these studies and their findings. Although various issues in the mobile learning context have been discussed, few review studies analyzed the trend topics of m-learning. Most of these studies were conducted for short periods and a limited number of articles; therefore, no review study recently focused on the longitudinal research trends of m-learning since its existence.

Second, besides conducting a bibliometric analysis approach, this study also applied text mining techniques to m-learning scientific literature contents, extracted in a corpus database of 7829 documents, covering the publications from 1984 to May 2021.

The findings of this study provided valuable patterns on the longitudinal research trends of mobile learning, explored by applying sophisticated content analysis with *Latent Dirichlet Allocation* (LDA) Topic Modeling technique; thus, providing actionable insights to the mobile learning body of knowledge, concerned with conducting oriented research studies and developing mobile application frameworks in this promising domain for the future of adaptative educational techniques.

Overall, this study demonstrated the power of supporting traditional bibliometric analysis tools with text mining techniques, enabling scholars to discover more research patterns, themes, and trends, by conducting sophisticated content analyses based on the different extensions of topic modeling algorithms, such as LDA, developed in deep learning techniques to build more accurate and automated models; for better use in data patterns analysis and interpretation than traditional content analysis tools.

REFERENCES

[1] I. C. Education. "What is Text Mining?" IBM Cloud Learn Hub. https://www.ibm.com/cloud/learn/text-mining (accessed June, 2021).

[2] O. A. Ise, "Integration and analysis of unstructured data for decision making: Text analytics approach," *International Journal of Open Information Technologies,* vol. 4, no. 10, 2016.

[3] D. M. Blei, "Probabilistic topic models," *Communications of the ACM,* vol. 55, no. 4, pp. 77-84, 2012.

[4] 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, 2021.

[5] M. Sarwar and T. R. Soomro, "Impact of smartphone's on society," *European journal of scientific research,* vol. 98, no. 2, pp. 216-226, 2013.

[6] R. Cobcroft, S. Towers, J. Smith, and A. Bruns, "Mobile learning in review: Opportunities and challenges for learners, teachers and institutions," in *Learning on the Move: Proceedings of the Online Learning and Teaching Conference 2006*, 2006: Queensland University of Technology, pp. 21-30.

[7] G. Conole, "Research through the generations: Reflecting on the past, present and future," *Irish Journal of Technology Enhanced Learning,* vol. 2, no. 1, 2017.

[8] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education," *Computers & Education,* vol. 151, p. 103855, 2020.

[9] Z. Bezovski and S. Poorani, "The evolution of e-learning and new trends," in *Information and Knowledge Management*, 2016, vol. 6, no. 3: IISTE, pp. 50-57.

[10] F. Gurcan, O. Ozyurt, and N. E. Cagitay, "Investigation of Emerging Trends in the E-Learning Field Using Latent Dirichlet Allocation," *The International Review of Research in Open and Distributed Learning,* vol. 22, no. 2, pp. 1-18, 2021.

[11] X. Zhou, L.-H. Chen, and C.-L. Chen, "Collaborative learning by teaching: A pedagogy between learner-centered and learner-driven," *Sustainability,* vol. 11, no. 4, p. 1174, 2019.

[12] A. Zhang and D. Cristol, *Handbook of mobile teaching and learning*. Singapore: Springer, 2015, pp. 13-14.

[13] M. I. Qureshi, N. Khan, S. M. A. Hassan Gillani, and H. Raza, "A Systematic Review of Past Decade of Mobile Learning: What we Learned and Where to Go," *International Journal of Interactive Mobile Technologies,* vol. 14, no. 6, 2020.

[14] S. A. Nikou and A. A. Economides, "Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018," *Computers & Education,* vol. 125, pp. 101-119, 2018.

[15] S. Kemp, "Digital 2021: Global Digital Overview," in "DataReportal: Global," 2021. Accessed: May 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-global-overview-report

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

[17] 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?," *Organizational Research Methods,* vol. 21, no. 3, pp. 519-523, 2018.

[18] 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, 2020.

[19] G. Keshaval and M. Gowda, "ACM transaction on information systems (1989-2006): A bibliometric study," *Information Studies,* vol. 14, no. 4, pp. 223-234, 2008.

[20] Durieux, V., & Gevenois, “Bibliometric Indicators: Quality Measurements of Scientific Publication”. Radiology, vol. 255(2), pp. 342–351, May 2010.

[21] W. Contributors. "Meta-Analysis." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Meta-analysis&oldid=1054260397 (accessed May, 2021).

[22] M. Palmquist. "Content Analysis." Columbia University Mailman School of Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis (accessed June, 2021).

[23] B. T. Erford, E. M. Miller, K. Duncan, and B. M. Erford, "Submission patterns: Measurement and Evaluation in Counseling and Development author and article characteristics from 1990 to 2009," *Measurement and Evaluation in Counseling and Development,* vol. 42, no. 4, pp. 296-307, 2010.

[24] A. Yıldırım and H. Şimşek, "Qualitative research methods in social sciences," *Ankara: Seçkin Publishing,* pp. 24-32, 2005.

[25] M. Sharples, "The design of personal mobile technologies for lifelong learning," *Computers & education,* vol. 34, no. 3-4, pp. 177-193, 2000.

[26] M. Sharples, D. Corlett, and O. Westmancott, "The design and implementation of a mobile learning resource," *Personal and Ubiquitous computing,* vol. 6, no. 3, pp. 220-234, 2002.

[27] F. K. Chiang, G. Zhu, Q. Wang, Z. Cui, S. Cai, and S. Yu, "Research and trends in mobile learning from 1976 to 2013: A content analysis of patents in selected databases," *British journal of educational technology,* vol. 47, no. 6, pp. 1006-1019, 2016.

[28] G. J. Hwang and C. C. Tsai, "Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010," *British Journal of Educational Technology,* vol. 42, no. 4, pp. E65-E70, 2011.

[29] W.-H. Wu, Y.-C. J. Wu, C.-Y. Chen, H.-Y. Kao, C.-H. Lin, and S.-H. Huang, "Review of trends from mobile learning studies: A meta-analysis," *Computers & education,* vol. 59, no. 2, pp. 817-827, 2012.

[30] S. Al Saleh and S. A. Bhat, "Mobile learning: A systematic review," *International Journal of Computer Applications,* vol. 114, no. 11, 2015.

[31] A. Sönmez, L. Göçmez, D. Uygun, and M. Ataizi, "A review of current studies of mobile learning," *Journal of Educational Technology and Online Learning,* vol. 1, no. 1, pp. 12-27, 2018.

[32] I. Goksu, "Bibliometric mapping of mobile learning," *Telematics and Informatics,* vol. 56, p. 101491, 2021.

[33] G. Yıldız, A. Yıldırım, B. A. Akça, A. Kök, A. Özer, and S. Karataş, "Research trends in mobile learning," *International Review of Research in Open and Distributed Learning,* vol. 21, no. 3, pp. 175-196, 2020.

[34] K. M. Thomas, B. W. O'Bannon, and N. Bolton, "Cell phones in the classroom: Teachers' perspectives of inclusion, benefits, and barriers," *Computers in the Schools,* vol. 30, no. 4, pp. 295-308, 2013.

[35] B. W. O'bannon and K. Thomas, "Teacher perceptions of using mobile phones in the classroom: Age matters!," *Computers & Education,* vol. 74, pp. 15-25, 2014.

[36] W.-Y. Hwang, Y.-M. Huang, R. Shadiev, S.-Y. Wu, and S.-L. Chen, "Effects of using mobile devices on English listening diversity and speaking for EFL elementary students," *Australasian journal of educational technology,* vol. 30, no. 5, 2014.

[37] Q.-K. Fu and G.-J. Hwang, "Trends in mobile technology-supported collaborative learning: A systematic review of journal publications from 2007 to 2016," *Computers & Education,* vol. 119, pp. 129-143, 2018.

[38] Hamzah, A., Hidayatullah, A. F., & Persada, "Discovering Trends of Mobile Learning Research Using Topic Modelling Approach". *International Journal of Interactive Mobile Technologies (IJIM),* vol. 14(09), pp. 4-14, June 2020.

[39] J.-L. Hung and K. Zhang, "Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques," *Journal of Computing in Higher education,* vol. 24, no. 1, pp. 1-17, 2012.

[40] S. Kumar Basak, M. Wotto, and P. Belanger, "E-learning, M-learning and D-learning: Conceptual definition and comparative analysis," *E-learning and Digital Media,* vol. 15, no. 4, pp. 191-216, 2018.

[41] W. Contributors. "Digital Learning." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Digital_learning&oldid=1031424916 (accessed May, 2021).

[42] W. Contributors. "M-learning." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=M-learning&oldid=1039938111 (accessed May, 2021).

[43] U. Demiray and A. İşman, "History of distance education". *Sakarya Üniversitesi Eğitim Fakültesi Dergisi,* vol. 1(1), pp. 88-108, 2001.

[44] W. Contributors. "Distance education." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Distance_education&oldid=1053893122 (accessed May, 2021).

[45] S. I. Wains and W. Mahmood, "Integrating m-learning with e-learning," in *Proceedings of the 9th ACM SIGITE conference on Information technology education*, 2008, pp. 31-38.

[46] A. M. Kaplan and M. Haenlein, "Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster," *Business Horizons*, vol. 59, no. 4, pp. 441-450, 2016.

[47] M. Nichols, "E-learning in context," *E-Primer series*, vol. 1, pp. 1-28, 2008.

[48] L. F. M. G. Pedro, C. M. M. de Oliveira Barbosa, and C. M. das Neves Santos, "A critical review of mobile learning integration in formal educational contexts," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, pp. 1-15, 2018.

[49] D. S. Al Hamdani, "Mobile learning: A good practice," *Procedia-Social and Behavioral Sciences*, vol. 103, pp. 665-674, 2013.

[50] S. Y. Park, M. W. Nam, and S. B. Cha, "University students' behavioral intention to use mobile learning: Evaluating the technology acceptance model," *British journal of educational technology*, vol. 43, no. 4, pp. 592-605, 2012.

[51] A. Hamzah and N. F. Muchlis, "The exploration through the factors affecting students' adoption on m-learning technologies," in *AIP Conference Proceedings*, 2018, vol. 1977, no. 1: AIP Publishing LLC, p. 020023.

[52] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Waltham: Elsevier, 2012, pp. 1,8,597.

[53] D. Team. "Introduction to Data Mining (Complete Guide)." Data-Flair.Training. https://data-flair.training/blogs/data-mining-tutorial/ (accessed July, 2021).

[54] V. Aguiar-Pulido, J. A Seoane, M. Gestal, and J. Dorado, "Exploring patterns of epigenetic information with data mining techniques," *Current pharmaceutical design*, vol. 19, no. 4, pp. 779-789, 2013.

[55] U. Fayyad, "Knowledge discovery in databases" in *Relational Data Mining*, 1st ed. S. Džeroski and N. Lavrač, Ed. New York: Springer, 2001, pp. 28-45.

[56] S. Džeroski, "Data Mining in a Nutshell" in *Relational Data Mining*, 1st ed. S. Džeroski and N. Lavrač, Ed. New York: Springer, 2001, pp. 3-26.

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

[58] P. Skoda and F. Adam, *Knowledge Discovery in Big Data from Astronomy and Earth Observation: Astrogeoinformatics*. St. Louis, Missouri: Elsevier, 2020, pp. 5.

[59] I. C. Education. "What is Data Mining?" IBM Cloud Learn Hub. https://www.ibm.com/cloud/learn/data-mining (accessed July, 2021).

[60] D. Taylor. "What is BIG DATA?" Guru99 for Free Education. https://www.guru99.com/what-is-big-data.html (accessed July, 2021).

[61] S. Gutta. "Data Science: The 5 V's of Big Data." Medium. https://medium.com/analytics-vidhya/the-5-vs-of-big-data-2758bfcc51d (accessed July, 2021).

[62] M. Smallcombe. "Structured vs Unstructured Data: 5 Key Differences." Xplenty Platform. https://www.xplenty.com/blog/structured-vs-unstructured-data-key-differences/ (accessed July, 2021).

[63] I. C. Education. "What is Machine Learning?" IBM Cloud Learn Hub. https://www.ibm.com/cloud/learn/machine-learning#toc-what-is-ma-qhM6PX35 (accessed August, 2021).

[64] M. T. Jones. "Models for machine learning." IBM Developer. https://developer.ibm.com/articles/cc-models-machine-learning/ (accessed August, 2021).

[65] H. A. Madni, Z. Anwar, and M. A. Shah, "Data mining techniques and applications—a decade review," in *2017 23rd International Conference on Automation and Computing (ICAC)*, 2017: IEEE, pp. 1-7.

[66] D. Talia, P. Trunfio, and F. Marozzo, "Introduction to Data Mining" in *Data analysis in the cloud: models, techniques and applications*, 1st ed. Ed. Oxford: Elsevier, 2015, pp. 1-25.

[67] R. Tamilselvi and S. Kalaiselvi, "An overview of data mining techniques and applications," *International Journal of Science and Research (IJSR), India Online ISSN*, pp. 2319-7064, 2013.

[68] E. Burns and J. Burke. "What is a neural network? Explanation and examples." TechTarget. https://searchenterpriseai.techtarget.com/definition/neural-network (accessed August, 2021).

[69] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," in *KDD*, 1995, vol. 95, pp. 112-117.

[70] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press, 2006, pp 1.

[71] Linguamatics. "What is Text Mining, Text Analytics and Natural Language Processing? ." Linguamatics IQVIA. https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing (accessed June, 2021).

[72] R. Morikawa. "What is text mining? Applications and preprocessing techniques." TELUS International. https://www.telusinternational.com/articles/what-is-text-mining?INTCMP=ti_lbai (accessed Ocotober, 2021).

[73] G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, and D. Delen, “The Seven Practice Areas of Text Analytics” in *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, 1st ed, Gary Miner, Ed. Oxford: Academic Press, 2012, pp. 29-41.

[74] J. Paralic and P. Bednar, "Text mining for document annotation and ontology support," *Intelligent Systems at the Service of Mankind*, pp. 237-248, 2003.

[75] L. Kumar and P. K. Bhatia, "Text mining: concepts, process and applications," *Journal of Global Research in Computer Science*, vol. 4, no. 3, pp. 36-39, 2013.

[76] 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. 414-418, 2016.

[77] 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. 4, no. 1, p. 1, 2020.

[78] 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, 2020.

[79] M. Allahyari *et al.*, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[80] L. Gohil, "Text Mining: Process and Techniques," *International Journal of Innovative Research in Computer Science & Technology*, 2015.

[81] N. Tyagi. "What is Text Mining? Process, Methods and Applications." AnalyticsSteps. https://www.analyticssteps.com/blogs/what-text-mining-process-methods-and-applications (accessed August, 2021).

[82] J. Xu. "Topic Modeling with LSA, PLSA, LDA & lda2Vec." Medium. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 (accessed July, 2021).

[83] W. Contributors. "Topic Model." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Topic_model&oldid=1053102044 (accessed March, 2021).

[84] B. V. Barde and A. M. Bainwad, "An overview of topic modeling methods and tools," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2017: IEEE, pp. 745-750.

[85] N. Pathik and P. Shukla, “Simulated Annealing Based Algorithm for Tuning LDA Hyper Parameters,” in *Proc. SoCTA 2019*, 2020, pp. 515-521.

[86] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113-120.

[87] A. Naskar. "Latent Dirichlet Allocation for Beginners: A high level overview." ThinkInfi https://thinkinfi.com/latent-dirichlet-allocation-for-beginners-a-high-level-overview/ (accessed September, 2021).

[88] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 6, no. 1, 2015.

[89] F. Gurcan and N. E. Cagiltay, "Research trends on distance learning: a text mining-based literature review from 2008 to 2018," *Interactive Learning Environments*, pp. 1-22, 2020.

[90] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, pp. 1-22, 2016.

[91] A. Alyahya. "What is topic modelling." ANIVATION. https://anivation.wordpress.com/2018/05/02/tuberculosis-topic-modeling/ (accessed September, 2021).

[92] T.-H. Chen, S. W. Thomas, and A. E. Hassan, "A survey on the use of topic models when mining software repositories," *Empirical Software Engineering*, vol. 21, no. 5, pp. 1843-1919, 2016.

[93] H. Girdher. "TDM (Term Document Matrix) and DTM (Document Term Matrix)." Medium. https://medium.com/analytics-vidhya/tdm-term-document-matrix-and-dtm-document-term-matrix-8b07c58957e2 (accessed September, 2021).

[94] A. Panichella, "A systematic comparison of search algorithms for topic modelling—a study on duplicate bug report identification," in *International Symposium on Search Based Software Engineering*, 2019: Springer, pp. 11-26.

[95] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science,* vol. 41, no. 6, pp. 391-407, 1990.

[96] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning,* vol. 42, no. 1, pp. 177-196, 2001.

[97] W. Contributors. "Probabilistic Latent Semantic Analysis." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Probabilistic_latent_semantic_analysis&oldid=1019081895 (accessed April, 2021).

[98] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[99] W. Contributors. "Dirichlet Process." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Dirichlet_process&oldid=1050840964 (accessed September, 2021).

[100] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *Journal of Big Data,* vol. 6, no. 1, pp. 1-18, 2019.

[101] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications,* vol. 78, no. 11, pp. 15169-15211, 2019.

[102] A. O. Source. "Top of Topic Modeling Open Source Projects on Github " AwesomeOpenSource.com. https://awesomeopensource.com/projects/topic-modeling (accessed September, 2021).

[103] W. Contributors. "Graphical Model." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Graphical_model&oldid=1048104624 (accessed September, 2021).

[104] W. Contributors. "Plate Notation." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Plate_notation&oldid=957279049 (accessed September, 2021).

[105] H. Nabli, R. B. Djemaa, and I. A. B. Amor, "Efficient cloud service discovery approach based on LDA topic modeling," *Journal of Systems and Software,* vol. 146, pp. 233-248, 2018.

[106] E. B.V. "How do I use the Analyze search results function?" Elsevier. https://service.elsevier.com/app/answers/detail/a_id/14181/supporthub/scopus / (accessed October, 2021).

[107] N. J. Van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *scientometrics,* vol. 84, no. 2, pp. 523-538, 2010.

[108] T. Point. "Gensim - Introduction." Tutorials Point. https://www.tutorialspoint.com/gensim/gensim_introduction.htm (accessed October, 2021).

[109] J. Demšar *et al.*, "Orange: data mining toolbox in Python," *the Journal of machine Learning research,* vol. 14, no. 1, pp. 2349-2353, 2013.

[110] *Orange Data Mining, Fruitful and Fun*. (2021). University of Ljubljana. [Online]. Available: https://orangedatamining.com/

[111] R. Řehůřek. "models.ldamulticore – parallelized Latent Dirichlet Allocation." RadimRehurek. http://man.hubwiz.com/docset/gensim.docset/Contents/Resources/Documents /radimrehurek.com/gensim/models/ldamulticore.html (accessed November, 2021).

[112] 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*, 2014, pp. 63-70.

[113] S. Firmin. "Topic Modeling | Part 3 - Interpreting the Visualization." Alteryx Community. https://community.alteryx.com/t5/Data-Science/Getting-to-the-Point-with-Topic-Modeling-Part-3-Interpreting-the/ba-p/614992 (accessed November, 2021).