Date of publication DEC-31, 2021, date of current version AUG-28, 2021.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.20.4.2437

Using Latent Dirichlet Allocation and Text Mining Techniques for Understanding Medical Literature

SAADAT M. ALHASHMI[1], MOHAMMED MAREE[2], ZAINA SAADEDDIN[3]

[1]Department of Information Systems, College of Computing and Informatics, University of Sharjah, P.O.Box: 27272 Sharjah, UAE,
(e-mail: salhashmi@sharjah.ac.ae)
[2]Department of Multimedia Technology, Faculty of Engineering and Information Technology, Arab American University,
P.O. Box 240 Jenin, 13 Zababdeh, Palestine, (e-mail: mohammed.maree@aaup.edu)
[3]Department of Computer Science, Faculty of Engineering and Information Technology, Arab American University, Palestine,
(e-mail: z.saadeddin@student.aaup.edu)

Corresponding authors: Saadat M. Alhashmi, Mohammed Maree (e-mail: salhashmi@sharjah.ac.ae, mohammed.maree@aaup.edu).

ABSTRACT Over the past few years, numerous studies and research articles have been published in the medical literature review domain. The topics covered by these researches included medical information retrieval, disease statistics, drug analysis, and many other fields and application domains. In this paper, we employ various text mining and data analysis techniques in an attempt to discover trending topics and topic concordance in the healthcare literature and data mining field. This analysis focuses on healthcare literature and bibliometric data and word association rules applied to 1945 research articles that had been published between the years 2006 and 2019. Our aim in this context is to assist saving time and effort required for manually summarizing large-scale amounts of information in such a broad and multi-disciplinary domain. To carry out this task, we employ topic modeling techniques through the utilization of Latent Dirichlet Allocation (LDA), in addition to various document and word embedding and clustering approaches. Findings reveal that since 2010 the interest in the healthcare big data analysis has increased significantly, as demonstrated by the five most commonly used topics in this domain.

KEYWORDS text mining; data analysis; medical domain; trending topics; word association rules.

I. INTRODUCTION

OVER the past few years, research data in the medical domain has increased significantly [1-3]. Handling such an enormous amount of data is one of the crucial challenges for various stakeholders in this domain. For instance, extracting appropriate knowledge for supporting decisions of medical organizations is one of the most important factors that can assist in analyzing past, current and future risks [4]. Given the exponential growth of the published literature in the medical domain, the challenge of identifying and extracting important and meaningful medical data and associated literature-based trends becomes even more pronounced. This challenge gets amplified, knowing that 80% of online data is in unstructured format, making it difficult to process and analyze by computers [5, 6]. Therefore, the need to apply new techniques that can seamlessly make medical literature information more accessible and useful becomes more critical than ever before. Recently, more studies have started to focus on medical bibliometric and literature analysis for a better understanding of the research trends in this domain. Also, with the emergence of text mining tools and techniques, the number of studies that use text mining applications has grown tremendously [7]. As reported in [8], the utilization of such tools and methods has proved to advance information and knowledge extraction across multiple application domains. As such, we exploit text mining as a potential solution to address the problem of deriving important knowledge from historical medical data. Since this domain is interdisciplinary and attracts researchers from the fields of information retrieval, machine learning, statistics, computational linguistics and especially data mining as reported in [9] and [10], our purpose of this study is to explore several text mining methods to better understand and analyze earlier attempts and research works in this domain. Accordingly, our study presents a literature and bibliometric review based analysis, associated with medical words embedding and several other techniques that we applied on a dataset that comprises 1945 research articles that were published between 2006 and 2019. The first part of the analysis refers to quantitative analysis wherein we present an overview on the distribution of medical big data analysis papers over the time, keyword frequency by year, merging keyword frequencies, and topic detection by keyword frequency statistics.

In the second part, we apply clustering and Latent Dirichlet Allocation (LDA) techniques on the abstracts of the articles to detect the topics and to know the most important research areas which discuss medical data analysis. It is important to point out that by conducting this research work, our aim is mainly to assist improving the understanding of important topics and trends in the medical domain, as well as to summarize the topics that have been detected with the aim of identifying their main characteristics across similar references.

The rest of this paper is organized as follows. First, we discuss state-of-the-art medical literature-based analysis approaches that have been proposed in this domain. In addition, we present our methodology for medical-related bibliometric and literature data analysis. Second, we introduce our analysis of medical research articles and highlight the main findings. Then, we discuss the results and draw conclusions. Finally, we present the extensions of current work.

II. RELATED WORK

Text mining is a knowledge-intensive process where a user analyzes a collection of documents through applying a suite of analysis tools [11]. Conventionally, this process is applied on data sources that are in unstructured formats from where it is possible to extract useful information through the identification and exploration of interesting data patterns [11, 12]. As reported by Dang et al., there are five basic text mining steps that are normally followed [5], these are:

a) Collecting information from unstructured data.
b) Converting this information into structured formats.
c) Identifying patterns from structured data.
d) Analyzing the patterns.
e) Extracting valuable information and storing it in the database.

In this context, text mining techniques transform unstructured content of a collection of documents into a more manageable and understandable format that can be further processed and analyzed [11, 13]. It can discover hidden knowledge from large volumes of data across various areas such as bioinformatics, marketing and business, to name a few [14-16]. Text mining approaches rely on algorithmic and heuristic techniques to analyze distributions, frequencies and associations between the terms of the document collection, in an attempt to answer questions, such as: What is the general trend of the topic in a period of time? What are the most frequent words? Are these maintained over time? How are the words related? What could these relationships mean?

The different analysis techniques are used to compare the frequencies of the concepts in the documents and these are the discovery of ephemeral associations and the detection of deviations [11].

Over the past years, literature and bibliometric analysis has been conducted by reviewing the literature manually. The articles were collected, classified, and reviewed to identify important information that can be further used by various stakeholders of the domain of interest. In this context, a brief reading of the literature needs to be carried out, giving more attention to abstracts of articles as they provide summaries from which readers can decide its relevance to the domain under investigation [17]. For instance, DiMatteo developed a quantitative review of 50 years of medical research on patient adherence to treatment. In this article, he explained that he used an electronic database to search the literature by keywords while the citations and abstracts were examined by the author and research assistants [18, 19].

Similarly, the authors of works [20, 21] carried out a systematic literature review of factors affecting outcomes in older medical patients admitted to hospital. In these researches, it is reported that the analysis was conducted using statistical methods where an independent statistician was consulted. Currently carrying out a literature analysis or a bibliometric study with manual methods has become almost impossible due to the ever increasing amount of published literature in the domain. For example, reviewing and analyzing a dataset with 1945 research articles in the medical domain would take a very long time of manual exploration and revision, and will call for many specialists to accomplish this task. As such, given the exponential growth of literature published in the medical domain, exploiting existing data mining tools to carry out the review task has become more indispensable. Nowadays, many researchers are turning to use various text mining tools for bibliometric analysis. In 2015, Hsu et al. carried out a research on the digital archives, where they applied techniques of text mining for co-word analysis and clustering to detect the most talked about topics [22]. Bach et al. carried out a qualitative analysis of literature using a systematic literature review, citation and co-citation analysis in the financial sector with text mining [23]. In a similar line of research, Amado et al. presented a research literature analysis based on a text mining semi-automated approach with the goal of identifying the main trends in marketing using the LDA techniques [4]. Similarly, Moro et al. analyzed recent literature in the search for trends in business intelligence applications for the banking industry LDA [24]. In 2018, Liao et al. developed a bibliometric analysis on medical big data research where they analyzed the types of documents and their frequencies, as well as frequencies of keywords, the most mentioned magazines, and the relationships between the extracted keywords [1].

III. MATHEMATICAL FORMULATION

Latent Dirichlet allocation (LDA) is a widely-used generative probabilistic model that has been implemented to identify topics for collections of discrete data, among which textual data are the most common. It is formed based on a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities as described in the below equation [25]:

$$p(\mathcal{D}|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d .$$

As such, and in this context, each topic probabilities are employed to form an explicit topic representation of a document. Accordingly, each topic will be represented using a set of words to which all the documents are mapped under the hypothesis that words that frequently appear together are likely to be close in meaning.

IV. DATA ANALYSIS AND INFORMATION EXTRACTION

In this study, we analyze and visualize medical literature related data using one of the most popular data mining tools; that is RapidMiner[1] which is a comprehensive data analysis tool that provides various operators that can be applied for the analysis of large amounts of data. The primary Natural Language Processing (NLP) steps of the text mining procedure are: Tokenize, Transform Cases, Filter Stopwords, Generate n-Grams and Filter Tokens. These steps are depicted in Fig. 1.

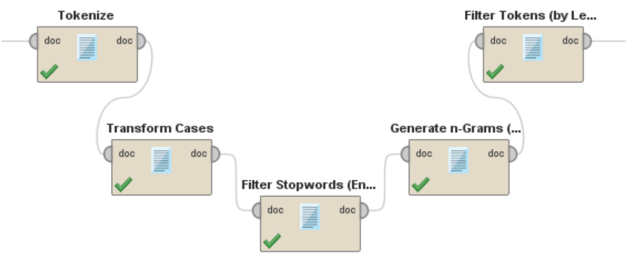


Figure 1. Five basic NLP operators

We have used VOSviewer[2] tool to create the network and density maps in order to visualize the main topics and the co-occurrence of the words.

First, we analyzed the frequency of published papers by year to know the growth of the literature of text mining in the medical domain as shown in Figure 2. We can see increasing growth in research publications, mainly since the year 2012.

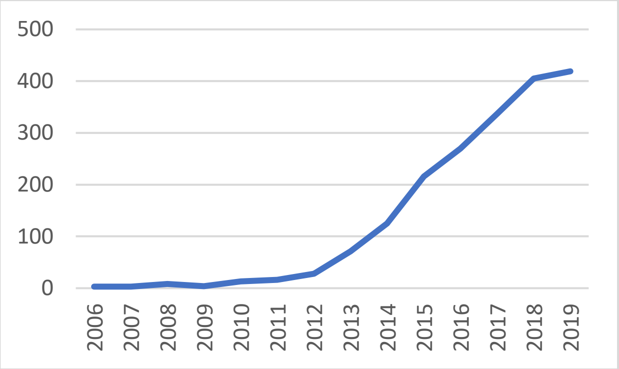


Figure 2. Published Papers in the Medical Domain by Year

For the frequency analysis of keywords, we extracted the keywords per year to observe the trend that exists in the topics discussed. In this analysis, we used the Process Document operator of RapidMiner with which morphological analysis steps, such as tokenize, convert to lowercase and remove stop words were applied. Figure 3 shows the most frequent terms over the last years.

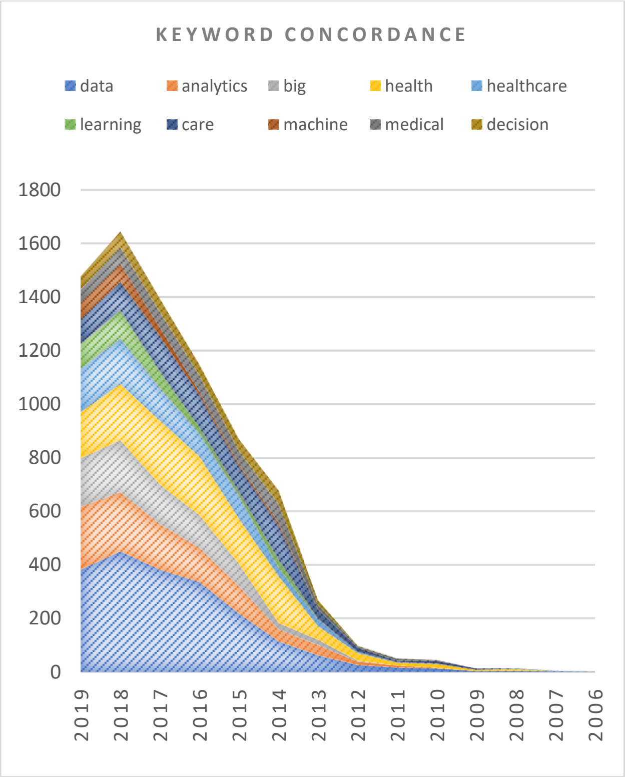


Figure 3. Medical Keyword Frequencies

Since 2010 the researchers have been addressing challenges related to data analysis in the Healthcare domain. In 2013, the term Big Data was increasingly used, and in 2017 researchers started to pay greater attention to the utilization of machine learning techniques. So, we can clearly notice that the

[1] www.rapidminer.com

[2] www.vosviewer.com

trend has shifted to focus on the exploitation of Big Data and Machine Learning approaches.

A. DETECTING TOPICS BY MERGING TOKENS AMONG KEYWORD SECTIONS

The RapidMiner “Process Document” operator was used for the analysis during this phase. The objective of this phase is to detect the merging words to identify the topics of references by year. The number of related terms (n-grams) was configured not to exceed three terms, and this process was also configured to allow the words (token) to have at least three letters.

After configuring the settings, we executed the process using several datasets corresponding to each year. Our goal in this context is to obtain the occurrence of the merging keywords. That is to define the topic of each dataset (each year)—the most frequent words taken into account because these are the ones that represent the dataset. For instance, Table 1 shows the merging keywords that were detected as the most frequent during the year 2019. Accordingly, from the set of extracted merging keywords, it is possible to conclude that the topics were mainly about Healthcare, Big Data, Data Analytics and Machine Learning.

Table 1. Merging Keywords detected in 2019

Words	Frequency
Health care	224
Big data	177
Data analytics	104
Machine learning	117
Big data analytics	93
Data mining	80
Internet things	78
Decision making	74
Predictive analytics	68
Electronic health	57

As shown in Table 1, we can figure out that the topics for this year (2019) are medical analysis, medical data and machine learning. The same procedure was applied from 2006 to 2019, concluding the following topics in Table 2:

Table 2. Merging Keywords detected from 2006 to 2019

Topics	Year
Big Data, Data Analytics, Machine Learning and Healthcare	2019
Big Data, Machine Learning and Healthcare	2018
Big Data, Machine Learning and Healthcare	2017
Big Data, Decision Support and Healthcare	2016
Big Data, Decision Support and Healthcare	2015
Big Data, Decision Support and Healthcare	2014
Big Data, Decision Support and Healthcare	2013
Big Data, Decision Support and Healthcare	2012
Data Analytics and Healthcare	2011
Data Analytics and Healthcare	2010
Data Management and Healthcare	2009
Data Mining and Healthcare	2008
Databases and Data Representation	2007
System Management	2006

B. DETECTING TOPICS BY MERGING TOKENS AMONG ABSTRACT SECTIONS

Clustering by abstract similarities

For analyzing the 1929 articles, complete abstract references were selected. K-Medoids algorithm was applied to the references by similarities to detect the most discussed topics. This algorithm selected k data items randomly as initial medoids to represent the k cluster and included in a cluster which has its medoid closest to them. The K-Medoids apply the following mathematical equations [26]:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \text{ , } j = 1\,,2\,,3\,,\ldots,n, \qquad (1)$$

where d_{ij} is the distance between object *i* and object *j*. When the pre-determined number of clusters is *k*, the first *k* smallest set is selected from the ordered v_j as the initial medoids. To visualize the clusters, the VOSviewer was used. Fig. 4 shows the topics that were most frequently discussed in the cluster with the largest number of papers grouped. As depicted by Fig. 4, the most relevant topics were: cloud, prediction, internet, cost, etc.

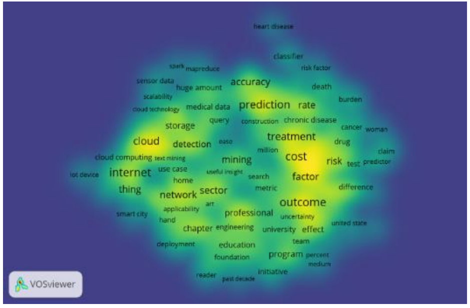


Figure 4. Most frequent words

Fig. 5 shows the Network Graph that exposes the co-occurrence of the words. References were clustered into four groups where each colour represents a cluster. Also, it is possible to understand that the topics are most relevant by the size of the bubble. Some of the most pertinent terms are *prediction, cost, accuracy, cloud, internet, treatment.*

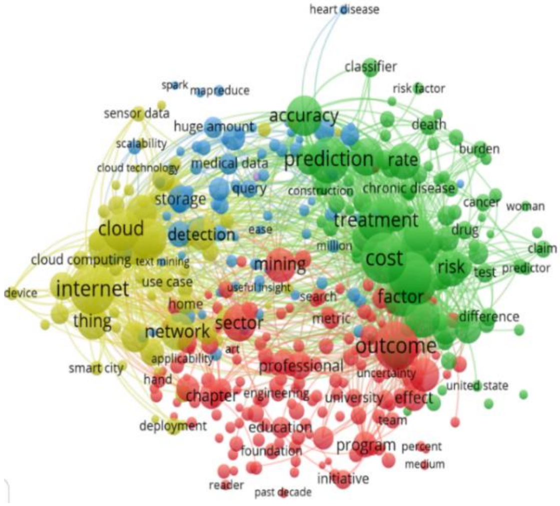


Figure 5. Co-occurring words

Topics Detected by Clustering

The analysis applied during this phase is mainly concerned with the utilization of clustering techniques to group the references by term similarities. To do this, we have applied the operators of RapidMiner that are depicted in Figure 3. Based on the results obtained during this phase, it is possible to identify the topic of each cluster, for a collection of 142 references grouped by similarities among their abstracts as shown in Table 3.

Table 3. Words Frequency

Words	Occurrence	Documents
data	373	122
patients	281	77
healthcare	231	115
health	174	81
care	146	66
analytics	132	112
study	131	80
patient	126	56
analysis	104	65

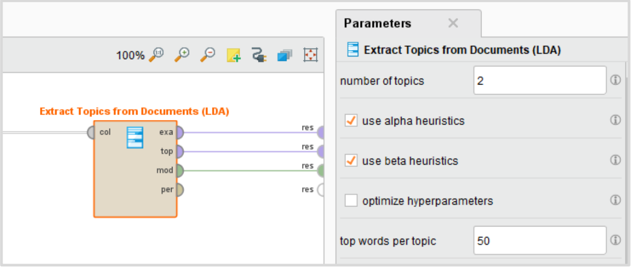
After analyzing the dataset and identifying topic clusters, it is possible to conclude that the focus has been on Data, Patients and Healthcare topics. This is indeed what the results confirm in Table 4.

Table 4. Merging Detected Words

Merging Words	Frequency
Health care	36
Health analytics	34
Healthcare costs	26
Claims data	24
Analytics marketscan	23
Data analytics	23
Big data	20
Data analysis	17

Using Latent Dirichlet Allocation for Topic Modeling

Latent Dirichlet Allocation algorithm allows for topic modeling through the inference of the latent structure behind a collection of documents. The main goal in this context is to deliver a “thematic summary” of a set of documents; allowing to answer the question: What themes are these documents discussing? [27].



Parameters

Extract Topics from Documents (LDA)

number of topics: 2

☑ use alpha heuristics

☑ use beta heuristics

□ optimize hyperparameters

top words per topic: 50

Figure 6. Latent Dirichlet Allocation Process

As shown in Figure 6, we have implemented LDA technique as offered by the RapidMiner; carrying out several tests on the used dataset. Figure 6 shows the configuration made that we have set for the LDA algorithm, where we initialized the algorithm with two topics and 50 words per topic. Table 5 shows the results obtained from this process.

After analyzing the words of each topic in Table 5, it is possible to answer the question: What themes are these documents discussing? So, we can say that Topic 1, in general, is about Systems Management and Data Analytics, while Topic 2 is about Big Data, Data Analysis and Healthcare.

Table 5. Topics detected with LDA

TOPIC #1		TOPIC #2	
Words	**Weight**	**Words**	**Weight**
based	418	data	1844
using	289	healthcare	807
system	287	analytics	463
systems	268	big	433
management	264	big_data	410
analytics	208	research	318
information	205	data_analytics	253
model	180	technology	214
learning	175	applications	172
time	175	information	152
business	171	science	150
approach	162	school	145
process	161	analysis	142
support	139	engineering	138
social	135	technologies	123
design	135	challenges	112
framework	133	predictive	108
include	122	improve	106
mining	117	monitoring	106
computing	113	application	105
performance	112	computer	100
paper	110	services	94
service	108	health_care	84

V. DISCUSSION AND ANALYSIS

With the constant growth in the number of publications in the medical domain, it has become more challenging to manually carry out a comprehensive literature review covering a wide range of articles over a long period. That is why we must resort to new technologies that allow the automatic treatment of the large volume of articles published in any domain, including the medical domain that we are addressing in this work. Recent text mining approaches incorporate various techniques that can assist in addressing large-scale and heterogeneous data analysis.

This study showed that text mining techniques can be successfully applied to better understand the trending topics that researchers had covered over the period from 2006 to 2019. It is demonstrated that the most frequent keywords, the relationships between keywords and the detection of topics of thousands of publications in the medical literature can be automatically processed and analyzed with high precision. Although there are already several text mining investigations, there are a few that are associated with the medical literature and healthcare domain. Accordingly, this research work contributes to the knowledge of various text mining techniques that can be involved as part of automated literature data analysis, specifically in the area of healthcare.

Regarding the analysis, findings demonstrated that the text mining process applied in RapidMiner allows knowing the frequency of the keywords and topic modeling and analysis over time. For example, it was possible to find out that since 2013 the term Big Data has been introduced in the Healthcare domain. Additional operators and configuration of the employed techniques have led to better explore and identify term co-occurrence statistics and automatic detection of the keywords that have been most frequently used across a huge number of publications.

The analysis made by clustering techniques allows creating several groups according to the data similarities. As such, in this investigation, it was possible to observe that this technique adapts perfectly to the study area. Visualizations made in VOSviewer verified that particular characteristics define each group. It helped in the explanation of the topic in each group. LDA technique on the other hand has helped to detect topics, where it was possible to verify and effectively detect the topics related to the studied articles. It was also possible to confirm by making a comparison between the other techniques applied. Accordingly, this study showed the basic steps that must be followed in Rapidminer to start a text analysis.

From the present study, it is possible to carry out further future research tasks such as:

- Semantically-enhanced text mining roles and challenges in the analysis of medical literature.
- Comparative analysis between existing text mining techniques that have been explored for medical literature related data.

VI. CONCLUSIONS AND FUTURE WORK

This study analyzed several Text Mining techniques that can be utilized for the automatic analysis of medical literature and other related bibliometric data analysis. One of the most critical issues is the huge amount of data available in the domain. The study exploited several text mining techniques applied to the medical literature analysis. Our initial findings demonstrate that Text Mining is a useful tool to understanding medical literature and bibliometric information. We conducted this approach to the healthcare domain to better understand research progress and evolution of topics over years. We found that healthcare topics are changing from Hospital and Health Management to Medical Analysis, Medical Big Data and Machine Learning. Also, it was possible to understand that the main issues are Patient Healthcare, Systems and Big Data. Text Mining also allows an exhaustive automatic analysis of the literature. It helps in obtaining clear and concise results for the development of bibliometric research in the studied domain. This work is a proof of concept on a dataset that comprised articles published between 2006-2019, and results are very encouraging.

Future work will involve expanding the number of papers and identify scaling up issues. Besides, the plan is to explore the role of semantic resources in the context of medical literature analysis.

References

[1] H. Liao, M. Tang, L. Luo, C. Li, F. Chiclana, and X.-J. Zeng, "A bibliometric analysis and visualization of medical big data research," *Sustainability*, vol. 10, no. 1, p. 166, 2018. https://doi.org/10.3390/su10010166.

[2] G. K. Savova *et al.*, "Use of natural language processing to extract clinical cancer phenotypes from electronic medical records," *Cancer Research*, vol. 79, no. 21, pp. 5463-5470, 2019. https://doi.org/10.1158/0008-5472.CAN-19-0579

[3] T. Hulsen *et al.*, "From big data to precision medicine," *Frontiers in Medicine*, vol. 6, p. 34, 2019. https://doi.org/10.3389/fmed.2019.00034.

[4] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on big data in marketing: A text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1-7, 2018. https://doi.org/10.1016/j.iedeen.2017.06.002.

[5] S. Dang and P. H. Ahmad, "Text mining: Techniques and its application," *International Journal of Engineering & Technology Innovations*, vol. 1, no. 4, pp. 866-2348, 2014.

[6] X. Liu, P. V. Singh, and K. Srinivasan, "A structured analysis of unstructured big data by leveraging cloud computing," *Marketing Science*, vol. 35, no. 3, pp. 363-388, 2016. https://doi.org/10.1287/mksc.2015.0972.

[7] M. Maree, I. Noor, K. Rabayah, M. Belkhatir, and S. M. Alhashmi, "Head concepts selection for verbose medical queries expansion," *IEEE Access*, vol. 8, pp. 93987-93999, 2020. https://doi.org/10.1109/ACCESS.2020.2987568.

[8] E. W. Ngai and P. T. Y. Lee, "A review of the literature on applications of text mining in policy making," *PACIS 2016 Proceedings*, 2016, 343. https://aisel.aisnet.org/pacis2016/343.

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

[10] S. S. Tandel, A. Jamadar, and S. Dudugu, "A survey on text mining techniques," *Proceedings of the 2019 5th IEEE International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 1022-1026. https://doi.org/10.1109/ICACCS.2019.8728547.

[11] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007. https://doi.org/10.1017/CBO9780511546914

[12] R. L. Patibandla and N. Veeranjaneyulu, "Survey on clustering algorithms for unstructured data," in *Intelligent Engineering Informatics*: Springer, 2018, pp. 421-429. https://doi.org/10.1007/978-981-10-7566-7_41.

[13] M. Maree, "Semantics-based key concepts identification for documents indexing and retrieval on the web," *International Journal of Innovative Computing and Applications*, vol. 12, no. 1, pp. 1-12, 2021. https://doi.org/10.1504/IJICA.2021.113608.

[14] Y. Xue, Y. Zhou, and S. Dasgupta, "Mining competitive intelligence from social media: A case study of IBM," in *PACIS 2018 Proceedings*, 2018, p. 313.

[15] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.

[16] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464-472, 2013. https://doi.org/10.1016/j.ijinfomgt.2013.01.001.

[17] A. Ramdhani, M. A. Ramdhani, and A. S. Amin, "Writing a literature review research paper: A step-by-step approach," *International Journal of Basic and Applied Science*, vol. 3, no. 1, pp. 47-56, 2014.

[18] M. R. DiMatteo, "Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research," *Medical care*, pp. 200-209, 2004. https://doi.org/10.1097/01.mlr.0000114908.90348.f9.

[19] P. Kokol, H. Blažun Vošner, and J. Završnik, "Application of bibliometrics in medicine: a historical bibliometrics analysis," *Health Information & Libraries Journal*, vol. 38, no. 2, pp. 125-138, 2021. https://doi.org/10.1111/hir.12295.

[20] S. E. Campbell, D. G. Seymour, and W. R. Primrose, "A systematic literature review of factors affecting outcome in older medical patients admitted to hospital," *Age and Ageing*, vol. 33, no. 2, pp. 110-115, 2004. https://doi.org/10.1093/ageing/afh036.

[21] C. Fogg, P. Griffiths, P. Meredith, and J. Bridges, "Hospital outcomes of older people with cognitive impairment: an integrative review," *International Journal of Geriatric Psychiatry*, vol. 33, no. 9, pp. 1177-1197, 2018. https://doi.org/10.1002/gps.4919.

[22] F.-M. Hsu, C.-M. Lin, and C.-T. Fang, "The trend and intellectual structure of digital archives research," in *PACIS 2015 Proceedings*, 2015, p. 128.

[23] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, p. 1277, 2019. https://doi.org/10.3390/su11051277.

[24] S. Moro, P. Cortez, and P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1314-1324, 2015. https://doi.org/10.1016/j.eswa.2014.09.024.

[25] 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.

[26] A. Bhat, "K-medoids clustering using partitioning around medoids for performing face recognition," *International Journal of Soft Computing, Mathematics and Control*, vol. 3, no. 3, pp. 1-12, 2014. https://doi.org/10.14810/ijscmc.2014.3301.

[27] 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. https://doi.org/10.1007/s11042-018-6894-4.

SAADAT M. ALHASHMI received the PhD degree from Sheffield Hallam University, Sheffield, U.K. Over the years, he has supervised several PhD students and published extensively in various high impact journals and conferences. Currently, he is Associate Professor of Information Systems at the University of Sharjah, Sharjah, UAE.

MOHAMMED MAREE received the PhD degree in Information Technology from Monash University. He began his career as a Research and Development Manager with gSoft Technology Solution, Inc. He was the Director of research and Q.A. with Dimensions Consulting Company. Subsequently, he joined the Faculty of Engineering and Information Technology (EIT), Arab American University, Palestine (AAUP), as a full-time Lecturer. From September 2014 to August 2016, he was the Head of the Multimedia Technology Department, and from September 2016 to August 2018, he was the Head of the Information Technology Department. In addition to his work at AAUP, he worked as a Consultant for SocialDice and Dimensions Consulting companies. He is currently the Head of the Multimedia Technology Department, Faculty of Engineering and Information Technology, AAUP. He has published articles in various high-impact journals and conferences, such as ICTAI, Knowledge-Based Systems, IEEE Access and the Journal of Information Science. He is currently a Committee Member and a Reviewer of several conferences and journals. He has s*upervised a number of master's* students in the fields of knowledge engineering, data analysis, information retrieval, natural language processing, and hybrid intelligent systems.

ZAINA SAADEDDIN works with the A.I. (Artificial Intelligence) team in Open Development & Education. Besides her work with the ODE team, she teaches for Code For Palestine (C4P), a uniquely designed program for the top-scoring students in the West Bank and Gaza, where students are given access to world-class groundbreaking skills in technology, design thinking, coding, and leadership. In 2014, she graduated with a bachelor's degree in computer engineering from An-Najah National University, Nablus, *Palestine (ANNU). Currently pursuing a master's degree in* computer science from Arab American University, Palestine (AAUP). Researcher in the Computational Neuroscience Unit in the Palestinian Neuroscience Initiative (PNI). Her research interest in Natural Language Processing, Machine Learning, and Artificial Intelligence.