Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1350–1360, DOI 10.12694/scpe.v25i3.2360

USE OF TOPIC ANALYSIS FOR ENHANCING HEALTHCARE TECHNOLOGIES

USHA PATEL,[*] PREETI KATHIRIA,[†] CHAND SAHIL MANSURI,[‡] SHRIYA MADHVANI,[§] AND VIRANCHI PARIKH[¶]

Abstract. Nowadays, technology has played a vital role in the advancement of the healthcare sector. Various healthcare datasets are available on the web in the form of patents, scientific papers, articles, textual feedback, chatlogs, abstracts of papers, medical reports, and social media posts. It is a tedious task for the stakeholders to find hidden crucial knowledge on the discussed topic from this content, which if utilized optimally can lead to the rapid development of the healthcare sector. Topic analysis concepts are very effective in extracting meaningful topics from the data. Here, frequently applied Topic modeling methods -Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Correlated Topics Model, and Non-negative matrix factorization are surveyed along with their benefits and drawbacks. Insights on new innovative topic modeling techniques used in healthcare with their objective, opportunities, and challenges are provided, which can help the researchers for the enhancement of healthcare facilities.

Key words: Topic Analysis, Topic Modeling, Health Care 5.0 , Stress Analysis, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Correlated Topics Model, Non-negative matrix factorization

1. Introduction.

In today's connected world, technology has become vital support in every business industry. Healthcare technology is one of the most critical industries where technology plays a key role. As new and improved technology is finding its applications in the healthcare industry, stakeholders in the health sector are getting more reliant on it for saving countless lives worldwide. It helps doctors to improve their practice by giving an advanced diagnosis to enhance patient care. Healthcare technology includes IT tools or software which is designed for better hospital administration, provide new insights in the field of medicines and treatment, or enhance the complete quality of care. Recent technology development includes Artificial intelligence in healthcare like Natural Language(NLP) technology and Machine learning [28].

The topic analysis is one of the natural language processing techniques which enable us to extract meaningful words from the input text by detecting topics based on their occurrence in the textual data. The topic analysis is the process of analysis that establishes a particular topic structure. This can be of assistance in pointing out which topics are present in the input text and what significance it brings in understanding the predicament at hand. The topic analysis is found useful in diverse applications related to text processing. One such example includes the assistance of topic analysis in the formation of an automatic indexing technique to infer information. It benefits in grasping the main topics and subtopics from the complete input textual data and also provides the information as to where exactly these subtopics are used in the data. The two renowned approaches for using machine learning to achieve topic analysis are topic modeling and topic classification. Topic classification is a supervised machine learning approach, it requires labels to classify the data into different classes. In the medical field, it is difficult to always get the labelled data. Therefore, with the unlabelled data, unsupervised learning - clustering is needed [25], and further topic modeling takes place.

1.1. Importance of topic analysis in Healthcare.

Most of the knowledge and information are collectively digitized and stored in the form of scientific articles, books, news, blogs, web pages, and social networks. It becomes somewhat difficult to collect the information in need efficiently and effectively. All this information is in an unstructured way, needed to convert into an organized and structured form. Computational methods

[*] Institute of Technology, Nirma University, India (ushapatel@nirmauni.ac.in)
[†] Institute of Technology, Nirma University, India (Corresponding author, preeti.kathiria@nirmauni.ac.in)
[‡] Institute of Technology, Nirma University, India (19mca016@nirmauni.ac.in)
[§] Institute of Technology, Nirma University, India (shriyu1304@gmail.com)
[¶] Institute of Technology, Nirma University, India (19mced08@nirmauni.ac.in)

and tools are needed to convert it into structured form and also to understand and search these vast amounts of information.[7]

Apps and portals are organizing online health communities which help in patient-doctor interactions and feedback of the patient regarding their overall experience with the organization and their caregivers. The electronic health record system is gaining popularity with physicians and patients. During the COVID-19 pandemic, more telephonic services were used by doctors, medical specialists, patients, and health systems. These practices provide a huge amount of structured as well as unstructured data which leaves an immense possibility for analyzing and understanding the data. The biggest challenge for health and medical data science research is to develop effective methods for finding the concealed meaning in considerable complex medical and healthcare datasets and using them to respond to the questions about that data[35].

Two popular methods utilized in analyzing medical text are bag-of-words and topic modeling. Bag-of-words technique acts on data as documents on the basis of frequency of the words like a matrix [24]. Topic modeling on the other hand is used for obtaining topics from the collection of documents.

The data for analysis not only includes the patient records but can also include various other means of information that can help in understanding the overall medical industry and help in enhancing the technological and administrative practices. The data for analysis can also include the various previous researches and patents information or current news articles or social media discussions on the current medical situations. The previous research and patients help in improving the innovation and also helps in understanding the impact of it in the current commercial market[13]. Scientific knowledge is transformed into new technology and that knowledge becomes the basis of further technological innovations[1].

1.2. Contribution of the paper. This paper discusses the importance of Topic modeling in Healthcare. Well-known Topic modeling techniques with their basic mathematical model, advantages, and disadvantages are discussed. Various forms of data used for topic modeling along with possible types of outcomes are analyzed from the different state-of-the-art publications.

Also, innovative topic models with their objectives used in healthcare are also included. At last major challenges and applications of topic modeling in healthcare are also given.

1.3. Organization of paper. Figure 1.1 reflects the structure of the survey. section - 1 gives the introduction about Topic analysis and its importance in healthcare along with the motivation and contribution of the paper. Section 2 explains Healthcare 5.0 and well-known Topic Modeling methods. Section 3 compares the various forms of data used in the healthcare domain on which topic analysis can be applied. Section 4 represents the advanced methods opted by the topic analysis in the healthcare domain. Section 5 describes the application areas, challenges, and research opportunities in this field.

2. Topic Modelling and Healthcare 5.0. The combination of Topic Modeling and Healthcare 5.0 can prove to be miraculous in the healthcare sector. For instance, we have multiple datasets of various patients and their medical records, which with the help of topic modeling can be used to identify the category of the disease of the future patients based on patterns, findings, and provide diagnosis for the current patients. This can lead the healthcare sector to advancement and help the stakeholders in the healthcare sector to improve healthcare services.

2.1. Healthcare 5.0. The use of technology is driving the health industry into a significant innovative transformation. Using technological devices regularly to monitor one's health helps in enhancing the standard of life. This stage of the utilization of cutting-edge technologies to benefit people in enhancing the functioning in the healthcare industry is referred to as Healthcare 5.0. The fundamental features of healthcare 5.0 include the use of Internet of Things (IoT), 5G communications, and Artificial Intelligence (AI). The research by Mohanta et al. [34] mainly focuses on Healthcare 5.0. It emphasizes the importance of 5G as the fundamental network infrastructure for enabling smart healthcare. Furthermore, Ambient Assisted Living (AAL) technology based on IoT provides variety of resolutions for improving people's quality of life, assisting stakeholders by providing impairments. It provides the products which helps in monitoring day to day health conditions.[43]

Nowadays, a wide variety of healthcare digital data is available in the form of sensor readings, patient detail records, social media, and news articles. Collecting all the data and finding out insights is also a challenging

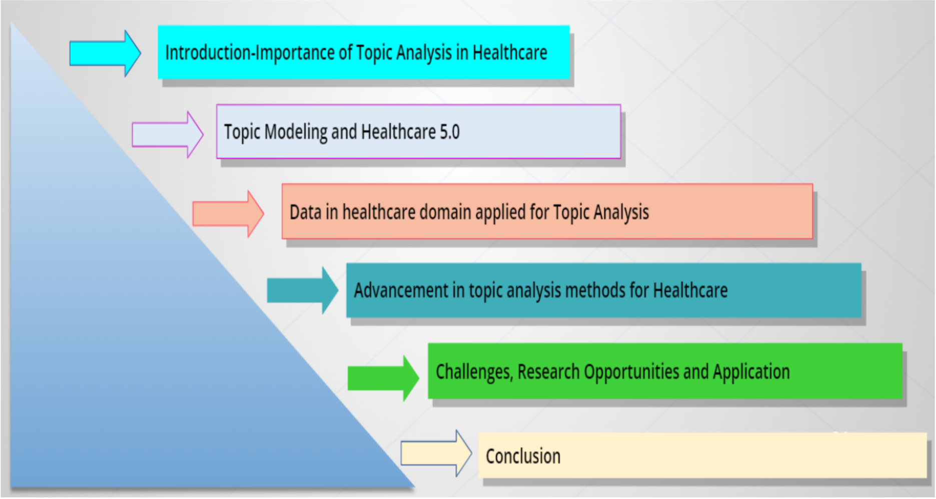
Introduction-Importance of Topic Analysis in Healthcare
Topic Modeling and Healthcare 5.0
Data in healthcare domain applied for Topic Analysis
Advancement in topic analysis methods for Healthcare
Challenges, Research Opportunities and Application
Conclusion

Fig. 1.1: Organization of the Paper

task for the big data analytics researcher. To fulfill the challenge one of the ways - Topic modeling, from health care data, may be useful.

2.2. Topic Modeling. Topic modeling is a very popular topic analysis technique which follows an unsupervised machine learning approach that scans documents to detect words and phrases patterns within them and produces the clusters of words and expressions based on their similarity. Topic modeling defines a document like a probability distribution over topics and each topic as the probability distribution over the words [23]. Topic modeling is applied in any field including software engineering, crime, medical, geographical, political, and linguistic sciences [21]. Recent examples of topic modeling in healthcare include extracting knowledge from electronic health records [3] [39] and analyzing user comments and online reviews [16]. It is applied by the authors Kathiria et. el in their research work for finding out the recent trends of topics from the abstract of the research papers[26].

The two primary approaches for topic analysis are supervised learning and unsupervised learning. Supervised learning approach contains labeled datasets and the method helps in disclosing the hidden structure from the dataset. Various Word embedding models Doc2Vec, Word2Vec, Glove, BERT can help in it [27]. Unsupervised learning approach has unlabeled data and it works in discovering the pattern from it to find some insights. The widely used supervised learning technique - Classification works to train the corpus with already available labeled dataset and based on that classify a new document accordingly [33]. The other widely used unsupervised technique - clustering assigns each document of a corpus to a respective cluster as per the similarity between the documents.

Figure 2.1 relfects the steps used for topic modeling process. First, the data needs to be collected for which the topic needs to be extracted. Since the data can be in the form of unstructured textual form it needs pre-processing. The pre-processing includes tokenization, removing stop words, words being lemmatized and stemmed. Once all the unwanted words and characters are filtered out of the data it is ready for topic modeling. Then the appropriate topic model algorithm is used. The results are then visualized by suitable means.

2.2.1. Well known Topic Models used in Literature.

LSA: Latent Semantic Analysis (also known as Latent Semantic Index, LSI) is one of the popular techniques for topic modeling. LSA makes use of the bag-of-words model which helps in generating a term-document matrix that demonstrates the occurrence of terms in the document [49]. LSA finds out latent topics by carrying out matrix decomposition over the term-document matrix using Singular Value Decomposition. In short, LSA acts as the dimension reduction approach.

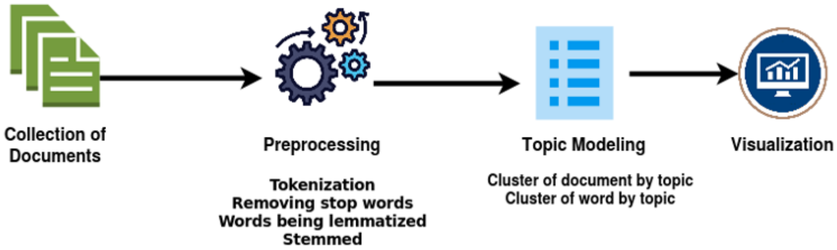

Fig. 2.1: Topic Modeling

PLSA: Probabilistic Latent Semantic Analysis is the probabilistic technique used to model the data to find the topics by making sense of the context of the text data. It evolved from Latent Semantic Analysis where the topics are hidden variables. It is used in applications that involve natural language processing, information retrieval, and filtering, applying machine learning on textual data, and in other related areas. The traditional Latent Semantic Technique makes use of linear algebra and executes Singular Value Decomposition of co-occurrence table while PLSA method makes use of the latent class model to derive mixture decomposition which includes strong statistical concepts [17].

LDA: Blei, Ng, and Jordan in 2003 [38] had proposed an unsupervised generative probabilistic model - Latent Dirichlet Allocation (LDA) which is opted to calculate the similarity between the given text files, moreover achieving their respective distributions of each document over topics. LDA is based on a three-level hierarchical Bayesian model. LDA follows a basic notion in which the documents correspond to random mixtures over latent topics, where a topic is set apart by distribution over words.

LDA can be said as a distinguished tool for latent topic distribution for a sizable corpus. Due to this, it inhibits the ability to recognize sub-topics for a technology area compiled of many patents, representing each of those patents in an array of topic distribution. Using LDA a vocabulary is generated which is then applied to discover hidden topics. There are several methods given to estimate LDA parameters, such as Gibbs sampling, variational method, and expectation propagation.

In health and medical science, LDA also serves in a variety of applications like the use of the knowledge obtained from the literature to predict protein-protein relationships [5], dig up relevant medical concepts and structures from health records of the patients [3], detecting patterns of medical events in a group of brain cancer patients [2], etc. LDA denotes more precise meaningful words as compared to LSA that's why it provides better accuracy and results [50].

CTM: Correlated Topic Model is the extension of LDA which on further evolution can be useful for creating more advanced topic models. Although LDA is the most popular topic modeling method, it has some limitations. LDA is not able to correlate the topics because it uses the Dirichlet distribution to model the unevenness amid the topic proportions. CTM makes use of logistic normal distribution to demonstrate correlation in the topic proportions 1[8].

NMF: Non-negative matrix factorization (also known as non-negative matrix approximation) represents a set of algorithms in linear algebra and multivariate analysis in which a matrix X is decomposed into two matrices W and H making sure that these are non-negative values. The problem cannot give exact values in which case approximation of numerical values is done. The application of non-negative matrix factorization can be found in fields such as document clustering, computer vision, audio signal processing, astronomy, recommender systems, missing data imputations, and other similar areas. In Table 2.1 the majorly used topic models with their advantages and disadvantages are discussed. As far as healthcare and medical topic analysis are concerned, from Table 2.1 , LDA is performing better than most of the topic modeling methods.

3. Usage of Various Forms of Data in Healthcare Domain Applied for Topic Analysis. Various forms of data present in healthcare domain can provide insights for the enhancement of the medical field in terms of diagnosis, treatment, administration, and providing other healthcare facilities. The dataset can include patents, scientific papers or articles, textual feedback, chat logs from the chat groups, abstracts of the scientific

Table 2.1: Various topic modeling techniques with pros and cons

Models	Advantages	Limitations
Latent Semantic Analysis (LSA)	- Using a single value decomposition for reducing the dimensionality of tf-idf - Statistical background is not robust - Helps in extracting synonyms of words	-Estimating the number of topics is difficult - In some cases, labeling a topic seems difficult using the words in the topic
Probabilistic Latent Semantic Analysis (PLSA)	- To some extent PLSA handles polysemy - Each word is generated from the single topic; - Different words can be generated from different topics in a document	- At the level of documents there is no probabilistic model
Latent Dirichlet Allocation (LDA)	- Can manifest nouns and adjectives in topics - Long-length documents can be handled - Provides complete generative model including distribution (i.e., multinomial) for words in dirichlet distribution over topics and other topics	-Not able to model relations amid topics
Correlated Topics Model (CTM)	-Logistic normal distribution is used for relations among opics -Helps in forming topic graphs -In other topics the appearance of the word is allowed	- Occurrences of general words in topics is allowed - Requires complex computations
Non-negative matrix Factorization (NMF)	- The use of positive values turns out easier inspection of resulting matrices	- Since it has a constraint of positive values, it can lose more information when truncating.

papers, medical reports of the patients, and social media posts.

Some of the recent papers were referred to understand what kinds of datasets are being used which is reflected in Table 3.1. From table 3.1, it can be observed that the use of social media data is more frequent in recent years. Social media data can include posts, comments, and messages. The main purpose is to explore the data and find the topics using topic modeling to understand the data correctly and find some trends, patterns, or directions for better research in the future [26]. The aim of each research can be unique and their social media data may also be unique based on its aim. But the method of solving the problem involves the exploratory analysis using the concepts of topic modeling here. One such research paper's dataset Dreaddit [20] in the next subsection is considered for the Categorical Stress Analysis on Dreaddit Dataset.

3.1. Categorical Stress Analysis on Dreaddit Dataset. Stress is omnipresent in one form or another. There can be various reasons for a person to feel stressed; it can be due to home lessness, relationship problem, domestic violence, and many others. Many surveys are conducted every year to know the actual cause of stress and provide assistance to the ones who are the victim of it. People commit illegal actions and can also lead to harming themselves if are not helped at the right time. Stress does not only affect the life of the person suffering from it, but it also has a significant impact on the people around the victim. This matter is also observable in social media as most people write multiple posts each day in which when closely observed; one can notice that the particular person is suffering from stress and even know the specific category of stress. This observation can be done with the help of topic analysis. Topic Analysis is a technique in which topics to text data automatically. Analyzation of unstructured text is done using Topic Analysis; such as social media interactions and emails.

A great study regarding stress is shown in paper [20] where the authors have presented Dreaddit; a large social media data from multiple domains for identifying stress in people. This dataset contains 1,90,000 posts which are collected from five different communities on Reddit. They additionally labelled 3500 segments (total) which were taken from 3000 posts with the help of Amazon Mechanical Turk.

Table 3.1: Types of data used in recent studies (X indictes the presence of the data type)

Author	Type of Data used						Purpose
	Scientific paper, abstracts, patents	Textual feedback	Chat logs	Patients clinical records	Social media posts	News articles	
[13]	x						Forecasting commercial viability or sustainability of healthcare innovations
[23]	x	x			x	x	Proposing a new method in topic modeling
[12]	x						Build an understanding for future reference
[47]	x						Analysis of use of "personality" and "mental health"
[19]		x					Insights for improvement of quality healthcare by analysis the patients reviews for their physicians
[48]			x				Insights for improving healthcare for pregnant women based on their social opinions
[45],[40]			x		x		Analyzing the feedbacks and patient opinions for improvement
[9]			x				Mental Health insights
[18],[44]				x			Predict clinical risk of a patient
[32]				x			Predict depression prior to clinical diagnosis
[3]				x			Case-based information retrieval of similar patients using patient's clinical records
[11]					x	x	Insights for improvement of the public health communication strategies
[4], [11], [46], [31], [14] [22], [37]					x		Discover and understand trends
[42],[42]					x		Clustering using topic models
[30]						x	Insights for improvement of health inequality issues in Korea
[10]					x		Analysis of use of social media in healthcare research

On the same dataset using the four specific features namely: confidence, anxious, angry, sad; which show the feelings of the person who is the victim of stress, stress analysis is done. Figure 3.1 shows a bar graph which depicts the four features and their significance on different categories of stress such as ptsd, assistance, relationships, survivorsofabuse, domestic violence, anxiety, homeless, stress, almosthomeless, and food pantry. For instance, people are equally angry, anxious, and sad at the same time due to relationship problems. Also, due to anxiety a person can feel stressed as shown in the graph wherein anxiety, the anxiousness of a person increases on a large scale which can lead to stress.

There is another type of data that is most commonly used in recent studies is the scientific papers, their abstracts, patents, books, or grey literature. There is already an ample amount of research papers published. It

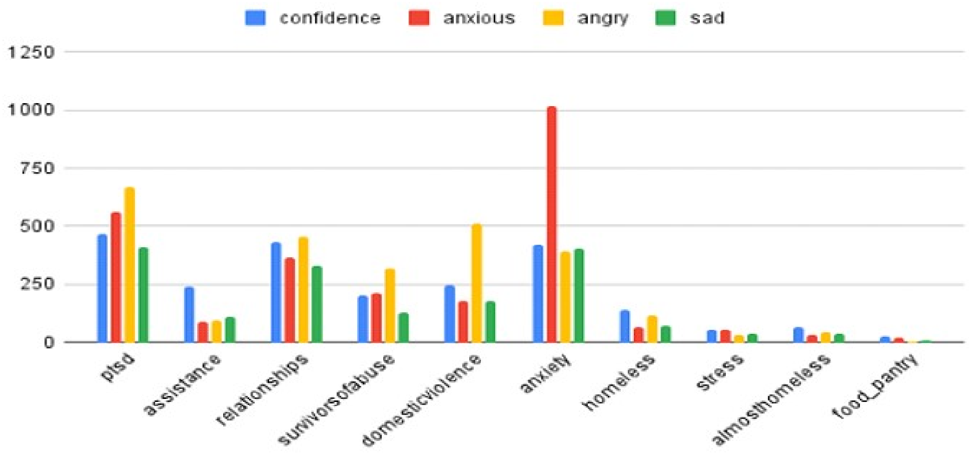

Fig. 3.1: Categorical Stress Analysis on Dreaddit Dataset

is difficult to read them all and infer knowledge from them which requires both time and effort. Topic modeling can help in analyzing the data and finding the relevant information in less time.

4. Advancement in Topic Analysis Methods for Healthcare. LDA is one of the well known topic modeling approaches among all. Apart from LDA, some research trends in topic modeling to follow and a new proposed improved framework for topic modeling are given in Table 4.1.

A study on understanding the different health topics related to social media is done by the authors' Paul et. al. [30]. A new statistical topic model was proposed by them referred to as Ailment Topic Aspect Model (abbr. ATAM). Since LDA is able to discover not only topics related to health around ailments but also other frequent topics which are not relevant as ailments and can be symptoms or something else. This noise needs to be filtered out to have just the health-related topics. ATAM was developed to explicitly label each tweet according to its ailment category; the model can also incorporate treatment and symptoms information. It is based on LDA where the ailment model will have three separate word distributions for symptoms, treatment, and general words. Thus it has a structurally different distribution.

Another study describes that the common statistical topic analysis approaches are not practical for rapidly processing the ever increasing online data. They proposed an alternative approach of automatic topic detection on the basis of document clustering for the extraction of topics related to health from online communities [30, 24]. The use of Expectation Maximization (EM) Clustering is applied in the same work. EM Clustering is a category of probability clustering method that allocates each occurrence with a probability which denotes that they belong to each cluster.

Apart from LDA, there are other techniques like correlated, dynamic, and hierarchical topic models. In recent years, another technique that is gaining popularity is Structural Topic Modeling (STM) [14]. STM encompasses metadata related to the text such as where, when, by whom the text was written, etc. For the estimation of Correlations, the Dirichlet distribution in the standard LDA can be replaced with the logistic normal distribution as it is in the Correlation Topic Models [8].

Likewise, there are different requirements to solve different problems at hand, and to do so we might need to boost the power of LDA accordingly. Thus, there are different variants possible and researchers tend to manage to improve it by combining it with other methods to form a hybrid topic model [42], to make it more effective and as per the requirements.

Table 4.1: New proposed improved framework for topic modeling

Paper	Year	Innovative Model	Objective
[6]	2023	The study examines deep learning (DL) and machine learning (ML) methods for healthcare prediction.	The paper assesses predictive analytics in healthcare, focusing on ML and DL techniques, emphasizing accurate disease prediction.
[15]	2022	Comparative Analysis and Classification of Topic Modeling Algorithms.	This study aims to provide a comprehensive overview of topic modeling in healthcare, including a comparative analysis of algorithms and their applications.
[32]	2020	Hierarchical Clinical Embeddings with Topic Modeling	Predicting depression
[46]	2020	Structural Topic Model	Study the contribution of social bots in the COVID-19 discussions on twitter
[31]	2020	Clustering Newman Algorithm	Social media based health disparity for COVID-19
[42]	2019	Visual Non-negative Matrix Factorization (VNMF), Visual Latent Dirichlet Allocation (VLDA), Visual Probabilistic Latent Schematic Indexing (VPLSI), Visual Latent Schematic Indexing (VLSI)	Using Hybrid Topics models by integrating topic models with VAT, for visualizing the health tendency and the topic clouds in the document collection.
[14]	2019	Structural Topic Model	To analyze tweets of stroke Survivors; their reactions based on their gender
[23]	2018	Fuzzy Latent Semantic Analysis (FLSA)	A better topic modeling approach compared to LDA is proposed in medical domain
[29]	2016	Application and Development of Topic Models in Bioinformatics	This paper reviews bioinformatics applications of topic models, categorizes studies, and underscores the need for tailored models to optimize biological data interpretation.
[48]	2016	Topic Interest Model	Use online healthcare chat logs to extract topics and infer user interests.
[18]	2015	Probabilistic Risk Stratification Model (PRSM)	Predict patients clinical risk to strategize the treatment accordingly
[41]	2014	Ailment Topic Aspect Model (ATAM)	Obtaining topics related to health from the tweets including its symptoms and treatment
[44]	2013	Co-occurrence Based Clustering, Dirichlet Process Mixture Model	To estimate patient disease risk

5. Challenges, Research opportunities, and Application. Various challenges, research opportunities and application of topic modeling techniques are discussed in this section.

5.1. Challenges. Figure 5.1 describes that there are many challenges in healthcare related to topic modelling. These can be categorized as technical, social and authenticity related challenges. Technical challenges involve security-related and topic modeling related challenges. Security-related challenges include problems related to advanced encryption and authentication. While, topic modeling-related challenges include problems related to used topic modeling approaches. Furthermore, Social challenges involve challenges related to mental insecurity and privacy of patient records. Moreover, authenticity related challenges include medical data related difficulties which consists of gathering labeled data and social media challenges which can be further questioned for authenticity. The limitation of the popular topic modeling approach LDA involves the difficulty in recognizing the numerical value of topics. Therefore, the general judgment of researchers can be used to find out the numerical value of topics.

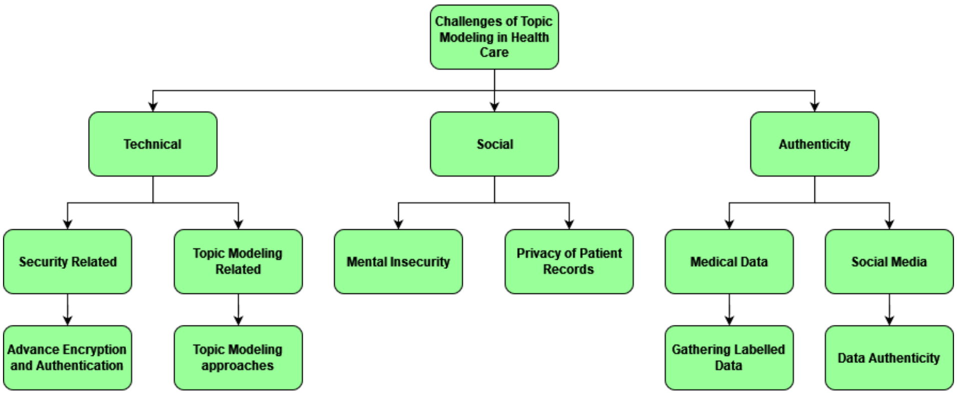

Fig. 5.1: Challenges of Topic Modelling in Healthcare

The topic analysis in healthcare often requires patient records. The protection of a patient's privacy and confidentiality is critical as well as challenging. The violations of privacy can not only hurt the patients but also the reputation of the healthcare firms. These violations can result in legal battles along with eroding the patient's trust and affect the long-term viability of healthcare firms.

5.2. Research Opportunities and Application. Research opportunities and Applications in this area of research are as follows.

The visualization of the health tendency can be found based on past data which can help for better planning of nutritious food for better health. The topic modeling applied to scientific research articles can give insight into the advancement of medical diagnosis, tools, and technologies.

The vast amount of data available on social networks is a significant resource for analyzing the pandemic's diverse impact on society. Using approaches such as topic modeling, this data can be evaluated to identify common themes, feelings, and worries voiced by people from various demographics and geographic areas. Armed with these insights, big organizations and government agencies can create educated policies and focused awareness guidelines to meet unique social requirements and issues caused by the epidemic.

Furthermore, combining patient health records and social network data allows healthcare practitioners to provide tailored medical interventions, modifying treatment programs, and prescribing medications based on specific patient needs and circumstances. This comprehensive strategy not only improves patient outcomes but also optimizes resource use within the healthcare system ([36]), emphasizing the importance of social network data analysis and patient health records in addressing the pandemic's impact at both the societal and individual levels.

The enormous data on social media allows for the prediction of the relationship between physical health and socio-cultural factors, revealing how cultural norms and online interactions influence behaviors and outcomes. By examining this data, researchers can develop personalized interventions and policies to alleviate regional health disparities. Initiatives such as "Healthy People 2030" in the United States demonstrate how data-driven tactics may generate substantial change and promote health equity across varied geographic locations.

6. Conclusion. Digitalization has offered us with various unstructured health-related data, which contains information and knowledge of great potential. Topic Modeling from this unstructured health-related data can easily identify the hidden pattern which can prove to be helpful in improving the treatment of the patients and the healthcare facility. In this paper, we surveyed various research papers and addressed the work of those researchers. We also did an analysis on Dreaddit dataset, categorizing the type of cause which can lead to stress and presented the way, in which topic modeling can be used to benefit the health sector. Many topic modeling techniques such as LDA, ATAM, FLSA, CTM, NMF, and LSA have been discussed throughout the paper and it was observed that LDA is superior to all the techniques discussed in the paper as it is easily able to manifest nouns and adjectives in a topic and its capability of handling very large documents. Also, we then discussed the challenges researchers can face in the healthcare sector while using topic modeling techniques. Moreover, future research opportunities and applicability of topic modeling techniques in sectors other than healthcare, such as bioinformatics, IT (Predication and Recommendation system), and Social Network Analysis have been discussed, which can help the researchers in understanding the requirement of topic modeling and the areas where it can be used. The main aim of conducting this research was to understand the work done in the aspect of topic analysis in the healthcare sector and use it to form a conclusion, to improvise the strategy to deliver a better treatment and healthcare facility.

REFERENCES

[1] A. Abbas, L. Zhang, and S. U. Khan, *A literature review on the state-of-the-art in patent analysis*, World Patent Information, 37 (2014), pp. 3–13.

[2] C. Arnold and W. Speier, *A topic model of clinical reports*, in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1031–1032.

[3] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira, *Clinical case-based retrieval using latent topic analysis*, in AMIA annual symposium proceedings, vol. 2010, American Medical Informatics Association, 2010, p. 26.

[4] M. Asghari, D. Sierra-Sosa, and A. Elmaghraby, *Trends on health in social media: Analysis using twitter topic modeling*, in 2018 IEEE international symposium on signal processing and information technology (ISSPIT), IEEE, 2018, pp. 558–563.

[5] T. Asou and K. Eguchi, *Predicting protein-protein relationships from literature using collapsed variational latent dirichlet allocation*, in Proceedings of the 2nd international workshop on Data and text mining in bioinformatics, 2008, pp. 77–80.

[6] M. Badawy, N. Ramadan, and H. A. Hefny, *Healthcare predictive analytics using machine learning and deep learning techniques: a survey*, Journal of Electrical Systems and Information Technology, 10 (2023), p. 40.

[7] D. Blei, L. Carin, and D. Dunson, *Probabilistic topic models*, IEEE signal processing magazine, 27 (2010), pp. 55–65.

[8] D. M. Blei and J. D. Lafferty, *A correlated topic model of science*, The annals of applied statistics, 1 (2007), pp. 17–35.

[9] B. Carron-Arthur, J. Reynolds, K. Bennett, A. Bennett, and K. M. Griffiths, *What's all the talk about? topic modelling in a mental health internet support group*, BMC psychiatry, 16 (2016), pp. 1–12.

[10] X. Chen, Y. Lun, J. Yan, T. Hao, and H. Weng, *Discovering thematic change and evolution of utilizing social media for healthcare research*, BMC Medical Informatics and Decision Making, 19 (2019), pp. 39–53.

[11] W. Chipidza, E. Akbaripourdibazar, T. Gwanzura, and N. M. Gatto, *Topic analysis of traditional and social media news coverage of the early covid-19 pandemic and implications for public health communication*, Disaster medicine and public health preparedness, (2021), pp. 1–8.

[12] R. Dantu, I. Dissanayake, and S. Nerur, *Exploratory analysis of internet of things (iot) in healthcare: A topic modeling approach*, (2019).

[13] S. S. Erzurumlu and D. Pachamanova, *Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations*, Technological Forecasting and Social Change, 156 (2020), p. 120041.

[14] A. Garcia-Rudolph, S. Laxe, J. Saurí, M. B. Guitart, et al., *Stroke survivors on twitter: sentiment and topic analysis from a gender perspective*, Journal of medical Internet research, 21 (2019), p. e14077.

[15] A. Gupta and H. Fatima, *Topic modeling in healthcare: A survey study*, NEUROQUANTOLOGY, 20 (2022), pp. 6214–6221.

[16] H. Hao, K. Zhang, et al., *The voice of chinese health consumers: a text mining approach to web-based physician reviews*, Journal of medical Internet research, 18 (2016), p. e4430.

[17] T. Hofmann, *Probabilistic latent semantic analysis*, arXiv preprint arXiv:1301.6705, (2013).

[18] Z. Huang, W. Dong, and H. Duan, *A probabilistic topic model for clinical risk stratification from electronic health records*, Journal of Biomedical Informatics, 58 (2015), pp. 28–36.

[19] T. L. James, E. D. V. Calderon, and D. F. Cook, *Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback*, Expert Systems with Applications, 71 (2017), pp. 479–492.

[20] L. JD, *A correlated topic model of science. the annals of applied statistics 2007*, 17 (2007), pp. 17–35.

[21] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, *Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. arxiv*, arXiv preprint arXiv:1711.04305, (2017).

[22] I. Kagashe, Z. Yan, I. Suheryani, et al., *Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using twitter data*, Journal of medical Internet research, 19 (2017), p. e7393.

[23] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, *Fuzzy approach topic discovery in health and medical corpora*, International Journal of Fuzzy Systems, 20 (2018), pp. 1334–1345.

[24] P. Kathiria and H. Arolkar, *Study of different document representation models for finding phrase-based similarity*, in Information and Communication Technology for Intelligent Systems, Springer, 2019, pp. 455–464.

[25] P. Kathiria and H. Arolkar, *Document clustering based on phrase and single term similarity using neo4j*, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9 (2020), pp. 3188–3192.

[26] P. Kathiria and H. Arolkar, *Trend analysis and forecasting of publication activities by indian computer science researchers during the period of 2010–23*, Expert Systems, 39 (2022), p. e13070.

[27] P. Kathiria, U. Patel, and N. Kansara, *Document classification using deep neural network with different word embedding techniques*, International Journal of Web Engineering and Technology, 17 (2022), pp. 203–222.

[28] P. Kathiria, U. Patel, S. Madhwani, and C. S. Mansuri, *Smart crop recommendation system: A machine learning*

approach for precision agriculture, in Machine Intelligence Techniques for Data Analysis and Signal Processing, D. S. Sisodia, L. Garg, R. B. Pachori, and M. Tanveer, eds., Singapore, 2023, Springer Nature Singapore, pp. 841–850.

[29] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, *An overview of topic modeling and its current applications in bioinformatics*, SpringerPlus, 5 (2016), pp. 1–22.

[30] Y. Lu, P. Zhang, J. Liu, J. Li, and S. Deng, *Health-related hot topic detection in online communities using text clustering*, Plos one, 8 (2013), p. e56221.

[31] J. Mantas et al., *Application of topic modeling to tweets as the foundation for health disparity research for covid-19*, The Importance of health Informatics in Public Health during a Pandemic, 272 (2020), p. 24.

[32] Y. Meng, W. Speier, M. Ong, and C. W. Arnold, *Hcet: Hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression*, IEEE Journal of Biomedical and Health Informatics, 25 (2020), pp. 1265–1272.

[33] T. M. Mitchell and T. M. Mitchell, *Machine learning*, vol. 1, McGraw-hill New York, 1997.

[34] B. Mohanta, P. Das, and S. Patnaik, *Healthcare 5.0: A paradigm shift in digital healthcare system using artificial intelligence, iot and 5g communication*, in 2019 International Conference on Applied Machine Learning (ICAML), Los Alamitos, CA, USA, may 2019, IEEE Computer Society, pp. 191–196.

[35] E. National Academies of Sciences, Medicine, et al., *Future directions for NSF advanced computing infrastructure to support US science and engineering in 2017-2020*, National Academies Press, 2016.

[36] S. Neely, C. Eldredge, and R. Sanders, *Health information seeking behaviors on social media during the covid-19 pandemic among american social networking site users: Survey study*, J Med Internet Res, 23 (2021), p. e29802.

[37] M. D. T. Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz, *What patients can tell us: topic analysis for social media on breast cancer*, JMIR medical informatics, 5 (2017), p. e7779.

[38] K. Odongo, *Uncovering consumer preferences for a novel apple variety using latent dirichlet allocation*, (2022).

[39] P. C.-I. Pang and S. Chang, *The twitter adventure of# myhealthrecord: an analysis of different user groups during the opt-out period*, Studies in Health Technology and Informatics, 266 (2019), pp. 142–148.

[40] P. C.-I. Pang and L. Liu, *Why do consumers review doctors online? topic modeling analysis of positive and negative reviews on an online health community in china*, (2020).

[41] M. J. Paul and M. Dredze, *Discovering health topics in social media using topic models*, PloS one, 9 (2014), p. e103408.

[42] K. R. Prasad, M. Mohammed, and R. Noorullah, *Hybrid topic cluster models for social healthcare data*, International Journal of Advanced Computer Science and Applications, 10 (2019).

[43] P. Purohit, P. Khanpara, U. Patel, and P. Kathiria, *Iot based ambient assisted living technologies for healthcare: Concepts and design challenges*, in 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2022, pp. 111–116.

[44] A. K. Rider and N. V. Chawla, *An ensemble topic model for sharing healthcare data and predicting disease risk*, in Proceedings of the international conference on bioinformatics, computational biology and biomedical informatics, 2013, pp. 333–340.

[45] P. Sampath, G. Packiriswamy, N. Pradeep Kumar, V. Shanmuganathan, O.-Y. Song, U. Tariq, and R. Nawaz, *Iot based health—related topic recognition from emerging online health community (med help) using machine learning technique*, Electronics, 9 (2020), p. 1469.

[46] W. Shi, D. Liu, J. Yang, J. Zhang, S. Wen, and J. Su, *Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter*, International Journal of Environmental Research and Public Health, 17 (2020), p. 8701.

[47] R. Sperandeo, G. Messina, D. Iennaco, F. Sessa, V. Russo, R. Polito, V. Monda, M. Monda, A. Messina, L. L. Mosca, et al., *What does personality mean in the context of mental health? a topic modeling approach based on abstracts published in pubmed over the last 5 years*, Frontiers in psychiatry, 10 (2020), p. 938.

[48] T. Wang, Z. Huang, and C. Gan, *On mining latent topics from healthcare chat logs*, Journal of biomedical informatics, 61 (2016), pp. 247–259.

[49] C. Wenli, *Application research on latent semantic analysis for information retrieval*, in 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2016, pp. 118–121.

[50] S. Yang, G. Huang, and B. Cai, *Discovering topic representative terms for short text clustering*, IEEE Access, 7 (2019), pp. 92037–92047.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Jun 22, 2023
Accepted: Mar 14, 2024