INTENSIF, Vol.8 No.1 February 2024
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.22058

Unveiling Insights: A Knowledge Discovery Approach to Comparing Topic Modeling Techniques in Digital Health Research

Received:
21 January 2024
Accepted:
6 February 2024
Published:
10 February 2024

[1*]Siti Rohajawati, [2] Puji Rahayu, [3]Afny Tazkiyatul Misky,
[4]Khansha Nafi' Rasyidatus Sholehah, [5]Normala Rahim,
[6]R.R. Hutanti Setyodewi
[1]Sistem Informasi, Universitas Bakrie
[2-4]Teknik Informatika Universitas Mercubuana
[5]Fakulti Informatik dan Komputeran, Universiti Sultan Zainal Abidin, Malaysia
[6]DR. Gerard sp. z o.o., Industries, Poland
E-mail: [1]siti.rohajawati@bakrie.ac.id,
[2]puji.rahayu@mercubuana.ac.id,
[3]518110172@student.mercubuana.ac.id,
[4]41518110152@student.mercubuana.ac.id ,
[5]normalarahim@unisza.edu.my, [6]hutanti.setyodewi@gmail.com
*Corresponding Author

Abstract— This paper introduces a knowledge discovery approach focused on comparing topic modeling techniques within the realm of digital health research. Knowledge discovery has been applied in massive data repositories (databases) and also in various field studies, which use these techniques for finding patterns in the data, determining which models and parameters might be suitable, and looking for patterns of interest in a specific representational. Unfortunately, the investigation delves into the utilization of Latent Dirichlet Allocation (LDA) and Pachinko Allocation Models (PAM) as generative probabilistic models in knowledge discovery, which is still limited. The study's findings position PAM as the superior technique, showcasing the greatest number of distinctive tokens per topic and the fastest processing time. Notably, PAM identifies 87 unique tokens across 10 topics, surpassing LDA Gensim's identification of only 27 unique tokens. Furthermore, PAM demonstrates remarkable efficiency by swiftly processing 404 documents within an incredibly short span of 0.000118970870 seconds, in contrast to LDA Gensim's considerably longer processing time of 0.368770837783 seconds. Ultimately, PAM emerges as the optimum method for digital health research's topic modeling, boasting unmatched efficiency in analyzing extensive digital health text data.

Keywords—Knowledge Discovery; Topic Modeling; Digital Health

Corresponding Author:

Siti Rohajawati,
Sistem Informasi,
Universitas Bakrie,
Email: siti.rohajawati@bakrie.ac.id,
Orchid ID: http://orcid.org/0000-0002-6775-8997

I. INTRODUCTION

The automatic, exploratory examination and modeling of massive data repositories is known as Knowledge Discovery in Databases (KDD). KDD is the systematic process of locating reliable, unique, practical, and intelligible patterns in big data and complicated data sets. Technology development has made it more affordable to gather vast amounts of data through data collection channels. People are beginning to understand that a substantial amount of data can provide valuable knowledge that can inform decision-making. For many sophisticated data studies, an online analytical processing (OLAP) or basic SQL query may not be adequate. Currently, KDD has thus become critically important to modern civilization [1].

Finding reliable, unique, maybe helpful, and eventually intelligible patterns in data is the process of knowledge discovery. It entails the difficult extraction of information from data that is implicit, unknown, and possibly valuable. Choosing techniques for finding patterns in the data, determining which models and parameters might be suitable, looking for patterns of interest in a specific representational form or a set of such topics modeling, regression, clustering, and other techniques, interpreting mined patterns, and compiling found knowledge are all part of the iterative KDD process. Using a database and any necessary preprocessing, subsampling, and other techniques, the KDD method aims to extract knowledge from data in the context of big datasets. By analyzing and modeling information in a visual format, modeling techniques are frequently utilized to uncover new information [1].

This study used the knowledge discovery framework (KDF) [2] [3], where we delved into the intricacies of comparing two topic modeling [4] techniques (LDA and PAM) within the context of digital health [5] [6]. Moreover, employing topic modeling as an analytical method for clustering topics within a given corpus [7], LDA and PAM are considered generative probabilistic models suitable for such analyses in corpus-like data. LDA utilizes hierarchical Bayesian analysis to identify topics within text corpora and determine topic proportions [8] [9]. Conversely, the PAM utilizes a directed acyclic graph to represent and learn topic correlations, providing a more flexible and comprehensive approach to capturing topics and word correlations [9]. In addition, the KDD also used to substantiate prior claims that the PAM surpasses LDA. The goal is to offer valuable insights to aid researchers in selecting the most efficient method for topic modeling in digital health research.

In the rapidly evolving landscape of digital health research, the quest for valuable insights has become increasingly complex and vital. Digital health, encompassing areas such as wearable device data, electronic health records [10], and telemedicine [11], has generated an unprecedented volume of information. Navigating through this sea of data requires sophisticated analytical methods to uncover meaningful patterns, trends, and insights. This research was employing a KDF [3][12][13] to compare and contrast different approaches to topic modeling[7]. According to the previous research, the KDF has been applied to encounter topic modeling such detecting terrorism [14], multi-agent systems [15], predicting wastewater [3], corporate bankruptcies [16], genderless fashion [17], bioinformatics [18], even music radio [2].

Digital health, known as telehealth, is a technological advancement focused on addressing health issues by offering long-distance health services using information technology and communication [11]. Its utilization is on the rise annually, encompassing a variety of health services such as diagnosis [6], prevention, treatment [19], and lifestyle enhancements [20] [17]. However, despite the increasing adoption of digital health, especially in Asia, research in this field encounters several challenges. The complexity arises from the vast diversity of topics and the inherent complexities associated with implementing these technologies, which are relatively unfamiliar to the general public [7]. To aid researchers in selecting relevant and popular topics, a digital health function mapping has been developed to assist in decision-making regarding research focus [21].

By adopting a KDF, we aim to not only assess the effectiveness of various topic modeling in digital health, but also to reveal nuanced insights that may have otherwise remained hidden. As we embark on this intellectual journey, our goal is to contribute to the refinement of methodologies in digital health research, offering a nuanced understanding of the strengths and limitations of different topic modeling approaches [22]. Through this KDF and comparative analysis, we aspire to empower researchers, practitioners, and stakeholders in the digital health domain with the knowledge needed to make informed decisions and drive advancements in the field.

II. RESEARCH METHOD

This study used the KDF [3][14][23][24], that provides a systematic and structured approach for uncovering valuable insights and patterns from large datasets following the phases (Fig. 1):

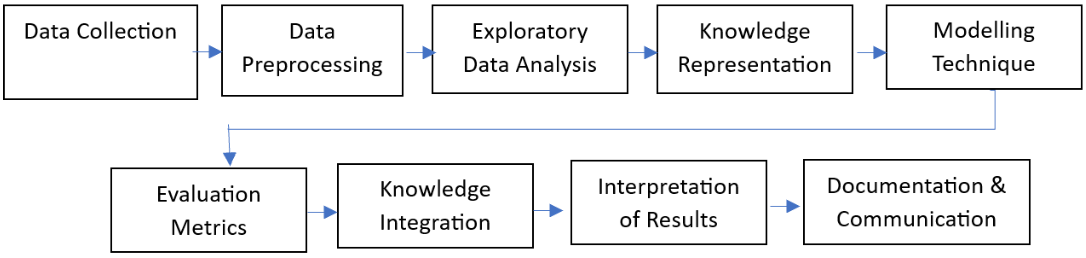


Fig 1. The Knowledge Discovery Framework Phases

A. Data Collection [24]: Identify and gather relevant datasets from digital health sources. This step involves acquiring structured or unstructured data that aligns with the research objectives.
B. Data Preprocessing [25]: Cleanse and preprocess the data to ensure its quality and suitability for analysis. This may include handling missing values, normalizing data, and transforming it into a format suitable for analysis.
C. Exploratory Data Analysis (EDA) [26] [27] [28]: Conduct exploratory data analysis to gain initial insights into the dataset. Visualizations and statistical summaries can aid in understanding the distribution of data and identifying potential patterns.
D. Knowledge Representation [29][30]: Choose an appropriate representation for the data. In the context of topic modeling, this step might involve transforming textual data into a format suitable for the selected modeling techniques.
E. Modeling Techniques [4]: Implement and compare different topic modeling techniques include algorithms LDA and PAM. Evaluate the performance of each technique in uncovering meaningful topics.
F. Evaluation Metrics [31][32]: Define and use appropriate evaluation metrics to assess the effectiveness of each topic modeling technique. Common metrics include coherence, perplexity, or other domain-specific measures.
G. Interpretation of Results [12]: Interpret the results of the topic modeling techniques and compare their outcomes. This step involves extracting meaningful insights and patterns that contribute to answering the research questions of topic modeling in digital health.
H. Knowledge Integration [33][34]: Integrate the discovered knowledge into the broader context of digital health research. Relate findings to existing literature and theories, highlighting the implications for the field.

Documentation and Communication: Document the entire process, including methodologies, results, and insights. Communicate findings effectively through reports, visualizations [35], or presentations to share knowledge with relevant stakeholders. Adapting this framework will involve tailoring each step to reveal the most topic research intensive in digital health research.

III. RESULT AND DISCUSSION

A. Data Collecting

Text documents that were scraped from open source online journals like journals.sagepub.com, ejournal.unisayogya.ac.id, and sinta.kemdikbud.go.id/ serve as the study's data source. The terms "digital health," "e-health," "modern health," etc. were utilized in the document selection process. During the Python scraping procedure [27], 443 documents were gathered into Mendeley Desktop

Tools. Due to their comprehensiveness, 404 documents were found to be appropriate study material; the remaining documents were excluded due to inaccurate information, such as missing the publication year or description. Following data collection, the documents will have their titles, years, and descriptions converted to CSV format. Table I presents the final document summary.

Table 1. Final Documents Summary

Publish Year	Article's Summary
2018	42
2019	36
2020	166
2021	125
2022	35
TOTAL	**404**

B. Data Preprocessing

First, Punctuation Removal, this procedure to eliminate punctuation marks within the data, including symbols like commas (,), backslashes (), periods (.), and others. The resulting outcome can be observed into: 1) Before, Drug-Drug Interaction (DDI) alert overrides, or the practice of users clicking past alerts without acting on recommendations, have been the subject of numerous studies. These studies have focused on a number of issues, including the appropriateness of alert overrides, variations in override rates among physicians, patient cohorts, and contexts (such as departments), and the harm that can result from patients overriding alerts. Then, 2) After, numerous studies have examined DDI alert overrides, which occur when users click past alerts without following recommendations. Of particular interest are differences in override rates among patient cohorts of physicians and contexts (such as departments), the appropriateness of alert overrides, and the harm that results from patients overriding alerts.

Second, Lowerization, it is converting uppercase letters to lowercase. The outcome is displayed into: 1) Before, numerous research endeavors have focused on the overrides of DDI alerts, specifically examining differences in override rates among physicians' patient groups and settings, the validity of these overrides, and the resulting patient harm caused by dismissed alerts. Then, 2) After, numerous research studies have explored the phenomenon of DDI alert overrides, wherein users dismiss alerts without implementing the provided recommendations. These studies have specifically examined the differences in override rates across physicians' patient groups and various contexts, such as different departments, assessing the validity of these overrides and the resulting harm to patients when alerts are disregarded.

Third, Stopword Removal, the elimination of stopwords involves removing common words like "and," "of," "then," etc., from the text. The outcome of this procedure is depicted into: 1) Before, Numerous research studies have delved into the phenomenon of DDI alert overrides, wherein users bypass alerts without following the recommended actions. These studies focus on understanding the variances in override rates across physicians' patient groups and diverse contexts (e.g., departments), assessing the suitability of these overrides, and examining the resulting harm to patients when alerts are ignored. Then, 2) After, Numerous studies have explored the phenomenon of DDI alert overrides, which involve users bypassing alerts without taking recommended actions. These studies specifically examine the differences in override rates among physicians' patient groups within various contexts (e.g., departments), evaluating the appropriateness of alert overrides and the subsequent harm to patients when alerts are disregarded. Below is an array containing 30 sample data entries that represent the overall outcome of preprocessing: ['accelerating', 'academiaedu', 'analysing', 'apa', 'citation', 'cite', 'chicago', 'downloaded', 'formation', 'get', 'health', 'international', 'mla', 'patterns', 'public', 'pulmonary', 'rainfall', 'research', 'related', 'spatial', 'visual', 'hendra', 'rohman', 'science', 'paper', 'styles', 'tuberculosis', 'papers', 'world']

C. Exploratory Data Analysis (EDA) [36]**:** Calculating Coherence Score

In this phase, the selection of the number of topics is based on the coherence score. The processing weighting of word analysis using Term Frequency-Inverse Document Frequency technique to reduce unnecessary words, vocabulary and eliminating noise [22]. which indicates the model's capacity to present data in a comprehensible manner for humans. A higher coherence score signifies better performance. The coherence score is derived through LDA Gensim training across various topic numbers, namely 8, 10, 13, 15, 20, and 25 (Fig. 2). We decided to take 10 after this processing. Refer to [37], which stated, that “stability analysis using topic coherence and the Xie-Beni statistic also favored large models (K = 100 topics)”.

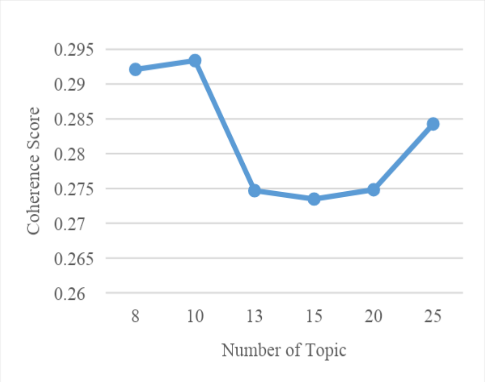


Fig 2. Result of Coherence Score Calculation

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.22058

D. Knowledge Representation

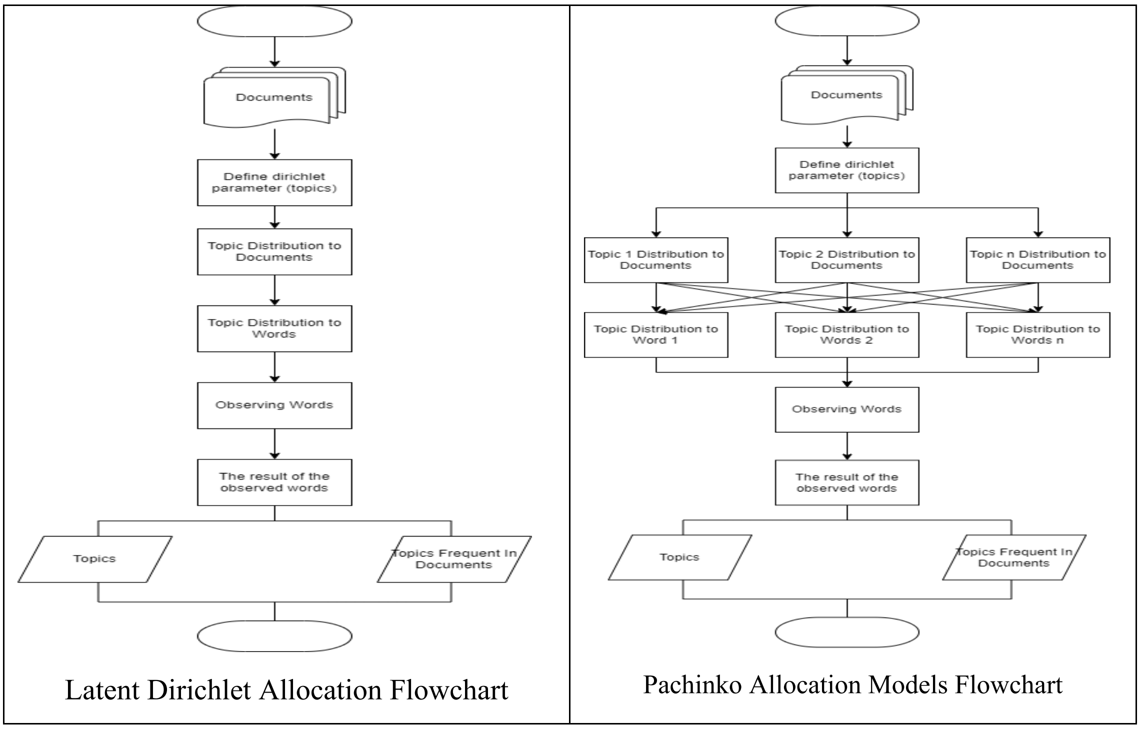
According to the data illustrated in Figure 2, the optimal choice for the number of topics is 10, achieving a coherence score of 0.293. The other knowledge representation as visual shows on Fig 4. 5, and 6.

E. Modeling Techniques LDA and PAM

According to [38], modeling technique has four aspects, i.e., neural, fuzzy, algebraic, and probabilistic. LDA is mostly preferred and applied to paper research [7][22][31][39][40]. The flowchart demonstrates the step and process of LDA and PAM for topic modeling (Fig. 3). In this research, we employ two distinct Python libraries for implementing LDA [21]: Gensim and Tomotopy. The discrepancy between these two libraries lies in the functions they utilize.

F. Evaluation Metrics

The extension process in Gensim operates in a streamed manner, implying that the training of documents occurs sequentially rather than randomly. Additionally, Gensim operates within constant memory constraints, enabling the processing of documents larger than the available RAM size. Meanwhile, LDA with Tomotopy Extension, is an extension for topic modeling that relies on Gibbs Sampling methodology [4]. It optimizes speed by leveraging the vectorization capabilities of contemporary CPUs. This extension is versatile and applicable not only for LDA but also for various other topic modeling methods including PAM [9], Advanced LDA, and Hierarchical Dirichlet Processes (HDP) [4][21]. The Tomotopy extension was employed to execute the PAM, similar to its utilization in an LDA model. The output from the PAM model will showcase details such as word distribution within sub-topics and the distribution of sub-topics across super-topics. According to [37] it is ideal if a few model evaluation criteria identify mid-sized topic models ($25 \leq K \leq 75$). Even human judgment indicated that mid-sized topic models provided evocative, low-dimensional summaries of the corpus. We agree with [41] [26][42] and [41] that several studies on various topics have shown that there is no single metric to validate clustering results, and metric performance suffers significantly when there is noise or overlap across clusters.


Documents
Define dirichlet parameter (topics)
Topic Distribution to Documents
Topic Distribution to Words
Observing Words
The result of the observed words
Topics
Topics Frequent In Documents
Documents
Define dirichlet parameter (topics)
Topic 1 Distribution to Documents
Topic 2 Distribution to Documents
Topic n Distribution to Documents
Topic Distribution to Word 1
Topic Distribution to Words 2
Topic Distribution to Words n
Observing Words
The result of the observed words
Topics
Topics Frequent In Documents

Latent Dirichlet Allocation Flowchart

Pachinko Allocation Models Flowchart

Fig 3. Modeling Techniques Flowchart

The procedure (Fig. 3) for how LDA and PAM work is as follows: (1) document collected from various resources research database; (2) define and initialize several parameters,; in LDA, the most important parameter is the number of topics, including the number of documents, topics, and iterations; (3) the topics will be separated into words for 1 to n and it process iterations, singular for LDA and parallel for PAM (Fig. 3); (4) assign words to certain topics randomly according to LDA and PAM distribution; (5) Observe words and repeat each process flow for all words in the corpus. Parameters used during the process of LDA and PAM calculations are as follows: (a) Random state: 100; (b) Update Every: 1; (c) Chunk Size: 10; (d) Passes: 10; (e) Alpha: Symmetric; (f) Iterations: 100; (g) Per Word Topics: True. Finally, in determining the number of measuring instrument topics, we used coherence value.

G. Interpretation of Results

The word cloud presents the frequency distribution of words within each topic, utilizing the data output showcased which visualization (Fig. 4).

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.22058

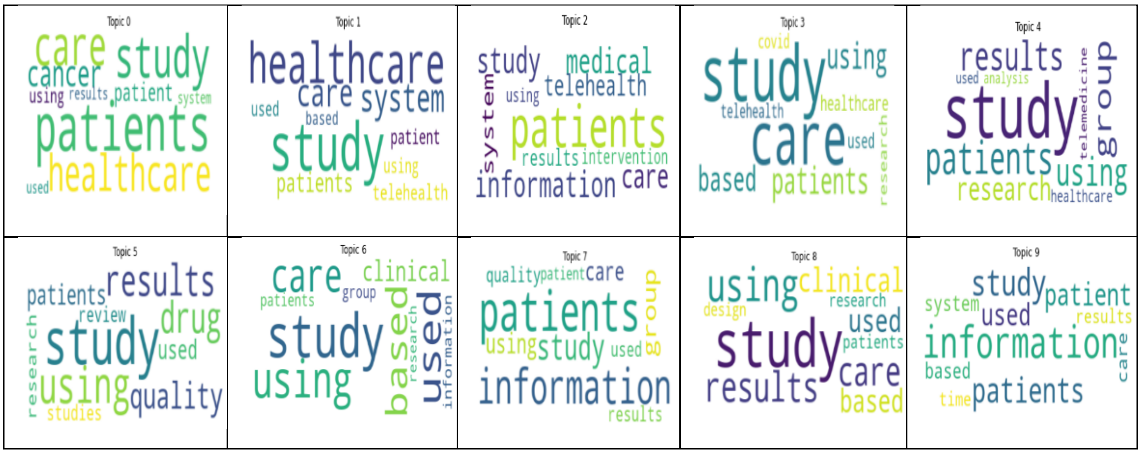


Fig 4. LDA Gensim Topic Word cloud

Figure 4 illustrates the top 10 words that appear most frequently within each topic. For instance, in Topic 0, the word "patients" is the most frequently occurring. The program execution time is measured during runtime, where, for LDA Gensim, it requires 0.368770837783 seconds to process 404 documents and generate the topic models. The assignment of topics to the respective documents is also conducted. According to the data displayed, it is evident that Topic Number 2 occupies or defines Document 0 with a portion of 0.993, which represents 99.3% of the document.

The output of the word cloud generated by LDA Tomotopy is presented in Figure 5. The processing time required for running LDA with the Tomotopy extension is 0.0001575946807 seconds.

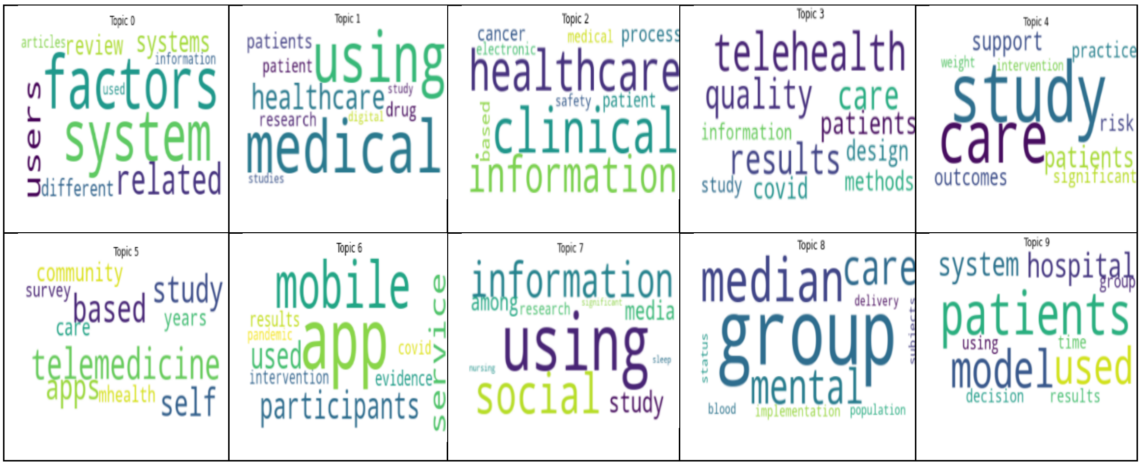


Fig 5. LDA Tomotopy Topic Word cloud

The output from the PAM model will showcase details such as word distribution within sub-topics and the distribution of sub-topics across super-topics (Fig. 6). Figure 6 displays the word cloud alongside the visual representation of the percentage for the 10 most frequently occurring words within each topic derived from the PAM output. The execution time is 0.000118970870 seconds. Refer to [37], validation that is used in topic model selection and evaluation, in which subject matter experts and data scientists iteratively review learned topic models and subjectively determine an appropriate fitting model for the corpus at hand, is frequently criticized for lacking empirical rigor, and advocates frequently suggest using one of the potentially many available topic model quality indices to guide model selection.

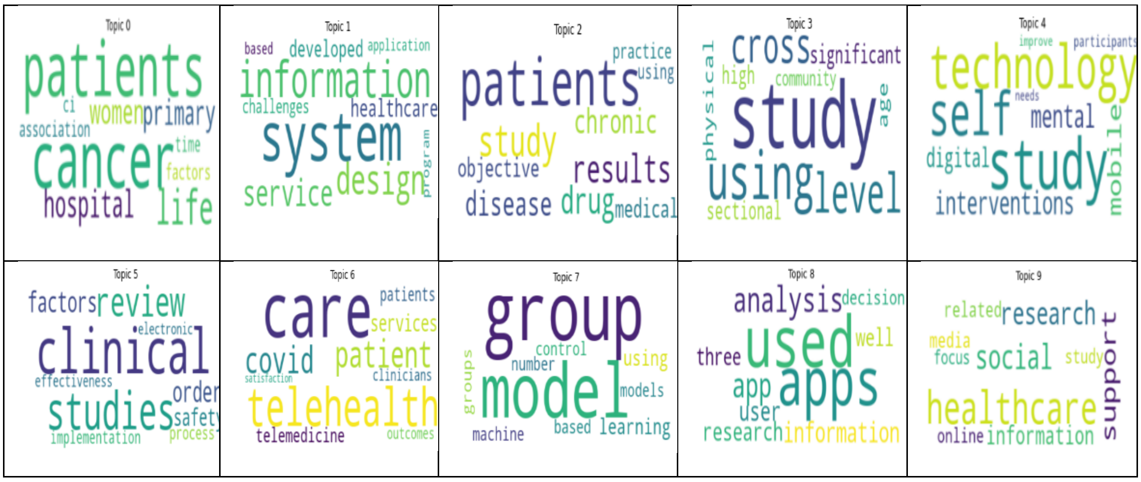


Fig 6. Pam Topic Word Cloud

H. Knowledge Integration, Documentation, and Communication

The results depicting the topic distribution for documents across years (Fig. 7). The data portrayed in highlights Prevention as the consistently dominant topic, followed by treatment as the second most prevalent. In contrast, diagnosis and Lifestyle seem to be less frequently addressed topics. However, it's important to acknowledge a potential bias in document distribution, notably with a majority of publications focused on the years 2020 and 2021. Despite this discrepancy, the chart remains informative, providing insights into prevailing topics. Future studies might benefit from a more equitable distribution of documents across publication years, enabling a fairer comparison and enhancing result accuracy.

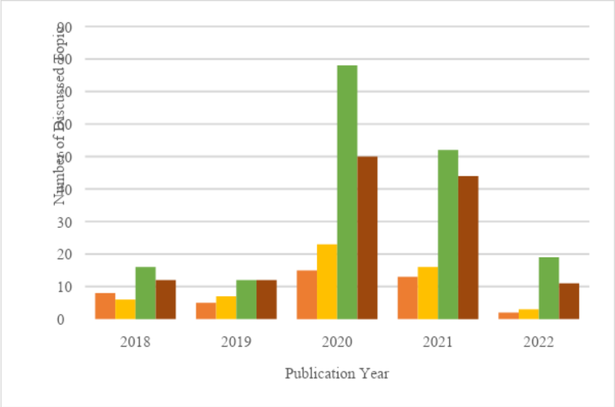


Fig 7. Mapping of Topics to Document Across Year

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v8i1.22058

From the data presented in Tabel 2, it's evident that PAM exhibits the highest number of unique tokens among the compared methods. Conversely, LDA Gensim demonstrates the fewest unique tokens, displaying repeated words across different topics. The presence of repeated tokens within topics can impact topic diversity; higher uniqueness in tokens enhances topic reliability. It also highlights the notable performance of the PAM in terms of execution time, efficiently processing 404 documents in 0.000118970870 seconds. In contrast, the LDA Gensim model records the longest execution time of 0.368770837783, showcasing the least favorable performance in terms of processing speed among the models compared. This is reinforced by [43], the best timing to apply the LDA approach depends on the particular situation in which it is being employed. The LDA performance is affected by the best number of topics to choose and other model parameters, including the Dirichlet prior parameters. In practice, the number of topics should be selected with a high degree of topic separation and good prediction capacity. As a result, the best time to use the LDA approach varies depending on the particular activity and the environment in which it is being used. Next by [44], the best time to use the PAM also depends on the particular application and the available computational resources. It is renowned for its capacity to represent topic connections and for offering greater expressive freedom and power in comparison to other methods like LDA. Also refer to [45], "the total training time for the NIPS dataset (as described in Section 3.2) is approximately 20 hours on a 2.4 GHz Opteron machine with 2GB memory".

Table 2. Model Comparison

Parameter	**Measurement**		
	LDA Gensim	***LDA Tomotophy***	***PAM***
Unique Token	27	71	87
Execution Time(s)	0. 368770837783	0.000157594680	0.000118970870

IV. CONCLUSION

Through experiments using the Knowledge Discovery Framework, topic modeling—a useful method for grouping topics within a corpus—was carefully assessed and proven. The results clearly show that the Pachinko Allocation Model (PAM) outperforms other approaches in topic modeling, as seen by the speed of execution and the uniqueness of words in each subject. The result shows that the best time for LDA is 0.368770837783 seconds for 404 documents to process and generate the topic models. Meanwhile, as for PAM, the execution time is 0.000118970870 seconds. For each topic modeling, we have displayed the word cloud as a visual representation of the knowledge for the 10 most frequently occurring words within each topic derived from the LDA and PAM output. Moreover, our word cloud results were examined across multiple models.

Document the entire process, including methodologies, results, and insights. Furthermore, based on the result of knowledge integration, it will provide and rise into new knowledge and enrich the sharing of knowledge with others. Adapting this framework will involve tailoring each step to the nuances of comparing topic modeling techniques in digital health research.

Future Suggestions for Research: Future study in this area should concentrate on improving pre-processing methods to lessen the frequency of repeated words in different formats and increase token uniqueness. Furthermore, the completeness and correctness of the findings can be greatly enhanced by streamlining the document distribution process during data scraping, particularly when it comes to the annual release of documents. Simplifying the analysis and increasing the usefulness of topic modeling.

Author Contributions: *Siti Rohajawati*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Puji Rahayu*: Software, Investigation, Data Curation, Writing - Original Draft. *Afny Tazkiyatul Misky*: Investigation, Data Curation. *Khansha Nafi Rasyidatus Sholehah*: Investigation. *Normala Rahim*: Investigation, Data Curation. *R.R. Hutanti Setyodewi*: Review.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data cannot be openly shared for the protection of study participant privacy.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Siti Rohajawati: http://orcid.org/0000-0002-6775-8997
Puji Rahayu: http://orcid.org/0000-0002-6684-9774
Afny Tazkiyatul Misky: http://orcid.org/0009-0005-4555-5901
Khansha Nafi Rasyidatus Sholehah: http://orcid.org/0009-0008-1259-5837
Normala Rahim: http://orcid.org/0000-0002-2094-7694
R.R. Hutanti Setyodewi: http://orcid.org/0009-0008-3937-652X

REFERENCES

[1] A. Adhikari and J. Adhikari, *Advances in Knowledge Discovery in Databases*, Intelligen. New York Dordrecht London: Springer International Publishing Switzerland, 2015. doi: 10.1007/978-3-319-13212-9.

[2] M. Furner, M. Z. Islam, and C.-T. Li, “Knowledge Discovery and Visualisation Framework using Machine Learning for Music Information Retrieval from Broadcast Radio Data,” *Expert Syst. Appl.*, vol. 182, p. 115236, 2021, doi:

https://doi.org/10.1016/j.eswa.2021.115236.

[3] V. Vasilaki, V. Conca, N. Frison, A. L. Eusebi, F. Fatone, and E. Katsou, “A Knowledge Discovery Framework to Predict the N2O Emissions in the Wastewater Sector,” *Water Res.*, vol. 178, p. 115799, 2020, doi: https://doi.org/10.1016/j.watres.2020.115799.

[4] H. Jelodar *et al.*, “Latent Dirichlet Allocation (LDA) and Topic modeling: Models, Applications, a Survey,” *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, Mar. 2003, doi: https://doi.org/10.1007/s11042-018-6894-4.

[5] A. Ahmed, R. Charate, N. V. K. Pothineni, S. K. Aedma, R. Gopinathannair, and D. R. Lakkireddy, “Role of Digital Health During Coronavirus Disease 2019 Pandemic and Future Perspectives,” *Card. Electrophysiol. Clin.*, vol. 14, pp. 115–123, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:240230974

[6] K. R. Jongsma, M. N. Bekker, S. Haitjema, and A. L. Bredenoord, “How Digital Health Affects the Patient-Physician Relationship: An Empirical-Ethics Study into the Perspectives and Experiences in Obstetric Care,” *Pregnancy Hypertens.*, vol. 25, pp. 81–86, 2021, doi: https://doi.org/10.1016/j.preghy.2021.05.017.

[7] A. Nurlayli and M. A. Nasichuddin, “Topic Modeling Penelitian Dosen JPTEI UNY pada Google Scholar Menggunakan Latent Dirichlet Allocation,” *Elinvo (Electronics, Informatics, Vocat. Educ.*, vol. 4, no. 2, pp. 154–161, 2019, doi: 10.21831/elinvo.v4i2.28254.

[8] X. Cheng, Q. Cao, and S. S. Liao, “An Overview of Literature on COVID-19, MERS and SARS: Using Text Mining and Latent Dirichlet Allocation,” *J. Inf. Sci.*, vol. 48, no. 3, pp. 304–320, Aug. 2020, doi: 10.1177/0165551520954674.

[9] J. Tuke *et al.*, “Pachinko Prediction: A Bayesian method for event prediction from social media data,” *Inf. Process. Manag.*, vol. 57, no. 2, p. 102147, 2020, doi: https://doi.org/10.1016/j.ipm.2019.102147.

[10] Y. A. Alsahafi and V. Gay, “An Overview of Electronic Personal Health Records,” *Heal. Policy Technol.*, vol. 7, no. 4, pp. 427–432, 2018, doi: https://doi.org/10.1016/j.hlpt.2018.10.004.

[11] L. M. Ganiem, “Efek Telemedicine pada Masyarakat (Kajian Hukum Media McLuhan: Tetrad),” *Interak. J. Ilmu Komun.*, vol. 9, no. 2, pp. 87–97, 2021, doi: 10.14710/interaksi.9.2.87-97.

[12] C. Schaefer and A. Makatsaria, “Framework of Data Analytics and Integrating Knowledge Management,” *Int. J. Intell. Networks*, vol. 2, pp. 156–165, 2021, doi: https://doi.org/10.1016/j.ijin.2021.09.004.

[13] X. Shu and Y. Ye, “Knowledge Discovery: Methods from Data Mining and Machine Learning,” *Soc. Sci. Res.*, vol. 110, p. 102817, 2023, doi: https://doi.org/10.1016/j.ssresearch.2022.102817.

[14] A. Ciapetti, G. Ruggiero, and D. Toti, “A Semantic Knowledge Discovery Framework for Detecting Online Terrorist Networks,” in *MultiMedia Modeling*, 2019, pp. 120–131.

[15] A. Jahani, P. Akhavan, M. Jafari, and M. Fathian, “Conceptual model for knowledge discovery process in databases based on multi-agent system,” *VINE J. Inf. Knowl. Manag. Syst.*, vol. 46, no. 2, pp. 207–231, Jan. 2016, doi: 10.1108/VJIKMS-01-2015-0003.

[16] A. Halder and M. Kannadhasan, “Knowledge Structure, Progression and Emergent Areas of Corporate Bankrupty: A Blibliiometric and Topic Modelling Analyses,” *SSRN Electr.*, pp. 1–25, 2022, doi: https://dx.doi.org/10.2139/ssrn.4193714.

[17] H. Kim, I. Cho, and M. Park, “Analyzing genderless fashion trends of consumers’ perceptions on social media: using unstructured big data analysis through Latent Dirichlet Allocation-based topic modeling,” *Fash. Text.*, vol. 9, no. 1, p. 6, 2022, doi: 10.1186/s40691-021-00281-6.

[18] 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, 2016, doi: 10.1186/s40064-016-3252-8.

[19] M. Thompson, "The Geographies of Digital Health – Digital Therapeutic Landscapes and Mobilities," *Health Place*, vol. 70, p. 102610, 2021, doi: https://doi.org/10.1016/j.healthplace.2021.102610.

[20] A. P. Sunjaya, "Potensi, Aplikasi dan Perkembangan Digital Health di Indonesia," *J. Indones. Med. Assoc.*, vol. 69, no. 4, pp. 167–169, 2019, doi: 10.47830/jinma-vol.69.4-2019-63.

[21] I. Vayansky and S. A. P. Kumar, "A Review of Topic Modeling Methods," *Inf. Syst.*, vol. 94, p. 101582, 2020, doi: https://doi.org/10.1016/j.is.2020.101582.

[22] K. R. Nastiti, A. F. Hidayatullah, and A. R. Pratama, "Discovering Computer Science Research Topic Trends using Latent Dirichlet Allocation," *J. Online Inform.*, vol. 6, no. 1, p. 17, 2021, doi: 10.15575/join.v6i1.636.

[23] S. Yamasaki, K. Yaji, and K. Fujita, "Knowledge Discovery in Databases for Determining Formulation in Topology Optimization," *Struct. Multidiscip. Optim.*, vol. 59, no. 2, pp. 595–611, 2019, doi: 10.1007/s00158-018-2086-0.

[24] T. Y. Choi and V. Cho, "Towards a knowledge discovery framework for yield management in the Hong Kong hotel industry," *Int. J. Hosp. Manag.*, vol. 19, no. 1, pp. 17–31, 2000, doi: 10.1016/S0278-4319(99)00053-5.

[25] R. J. Roiger, "The Knowledge Discovery Process," *Data Min.*, pp. 199–220, 2018, doi: 10.1201/9781315382586-6.

[26] A. T. Jebb, S. Parrigon, and S. E. Woo, "Exploratory Data Analysis as a Foundation of Inductive Research," *Hum. Resour. Manag. Rev.*, vol. 27, no. 2, pp. 265–276, 2017, doi: 10.1016/j.hrmr.2016.08.003.

[27] P. Chakri, S. Pratap, Lakshay, and S. K. Gouda, "An Exploratory Data Analysis Approach for Analyzing Financial Accounting Data using Machine Learning," *Decis. Anal. J.*, vol. 7, no. January, p. 100212, 2023, doi: 10.1016/j.dajour.2023.100212.

[28] M. O. Adeniyi *et al.*, "Dynamic Model of COVID-19 Disease with Exploratory Data Analysis," *Sci. African*, vol. 9, p. e00477, 2020, doi: 10.1016/j.sciaf.2020.e00477.

[29] A. Patel and S. Jain, "Formalisms of Representing Knowledge," *Procedia Comput. Sci.*, vol. 125, pp. 542–549, 2018, doi: 10.1016/j.procs.2017.12.070.

[30] M. M. Abdul Jalil, C. P. Ling, N. M. Mohamad Noor, and F. Mohd, "Knowledge Representation Model for Crime Analysis," *Procedia Comput. Sci.*, vol. 116, pp. 484–491, 2017, doi: 10.1016/j.procs.2017.10.067.

[31] C. Palma, V. Morgado, and R. J. N. B. da Silva, "Top-down evaluation of matrix effects uncertainty," *Talanta*, vol. 192, pp. 278–287, 2019, doi: 10.1016/j.talanta.2018.09.039.

[32] J. Rossmann, R. Gurke, L. D. Renner, R. Oertel, and W. Kirch, "Evaluation of the matrix effect of different sample matrices for 33 pharmaceuticals by post-column infusion," *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, vol. 1000, pp. 84–94, 2015, doi: 10.1016/j.jchromb.2015.06.019.

[33] X. Zhang, "Knowledge integration in interdisciplinary research teams: Role of social networks," *J. Eng. Technol. Manag.*, vol. 67, p. 101733, 2023, doi: https://doi.org/10.1016/j.jengtecman.2023.101733.

[34] K. Gugerell, V. Radinger-Peer, and M. Penker, "Systemic knowledge integration in transdisciplinary and sustainability transformation research," *Futures*, vol. 150, no. May, p. 103177, 2023, doi: 10.1016/j.futures.2023.103177.

[35] M. Furner, M. Z. Islam, and C. T. Li, "Knowledge discovery and visualisation framework using machine learning for music information retrieval from broadcast radio data," *Expert Syst. Appl.*, vol. 182, no. May, p. 115236, 2021, doi: 10.1016/j.eswa.2021.115236.

[36] K. Ogunsina, I. Bilionis, and D. DeLaurentis, "Exploratory data analysis for airline disruption management," *Mach. Learn. with Appl.*, vol. 6, no. July, p. 100102, 2021, doi: 10.1016/j.mlwa.2021.100102.

[37] C. Meaney, T. A. Stukel, P. C. Austin, R. Moineddin, M. Greiver, and M. Escobar, "Quality indices for topic model selection and evaluation: a literature review and case study," *BMC*

Med. Inform. Decis. Mak., vol. 23, no. 1, pp. 1–18, 2023, doi: 10.1186/s12911-023-02216-1.

[38] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, p. 102131, 2023, doi: https://doi.org/10.1016/j.is.2022.102131.

[39] C. C. Silva, M. Galster, and F. Gilson, "Topic modeling in software engineering research," *Empir. Softw. Eng.*, vol. 26, no. 6, 2021, doi: 10.1007/s10664-021-10026-0.

[40] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, "Prediction of research trends using LDA based topic modeling," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 298–304, 2022, doi: 10.1016/j.gltp.2022.03.015.

[41] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrista-Salas, T. Hernandez-Boussard, and J. Bian, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artif. Intell. Med.*, vol. 117, no. May, p. 102096, 2021, doi: 10.1016/j.artmed.2021.102096.

[42] V. Alekseev, E. Egorov, K. Vorontsov, A. Goncharov, K. Nurumov, and T. Buldybayev, "TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation," *Data Knowl. Eng.*, vol. 135, p. 101921, 2021, doi: 10.1016/j.datak.2021.101921.

[43] J. Gan and Y. Qi, "Selection of the optimal number of topics for LDA topic model—Taking patent policy analysis as an example," *Entropy*, vol. 23, no. 10, 2021, doi: 10.3390/e23101301.

[44] T. Huynh-The, O. Banos, B. V. Le, D. M. Bui, Y. Yoon, and S. Lee, "Traffic behavior recognition using the pachinko allocation model," *Sensors (Switzerland)*, vol. 15, no. 7, pp. 16040–16059, 2015, doi: 10.3390/s150716040.

[45] W. Li; and A. McCallum, "Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations," 2006.