Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 7, No. 4, pp. 641-653, November 2025

RESEARCH PAPER **OPEN ACCES**

Topic Mining-Based Knowledge Discovery of User Health Information Needs

Dayana Khoiriyah Harahap, Ken Ditha Tania, and Putri Eka Sevtiyuni
Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia

ABSTRACT

Understanding the user's need for health information has become increasingly important as the use of digital health services continues to grow. However, the unstructured data of user-generated questions presents challenges in accurately capturing and analyzing these needs. This study contributes to addressing SDG 3 (Good Health and Well-being) by utilizing topic mining-based knowledge discovery to identify the primary topics emerging from user questions submitted through the "Tanya Dokter" feature on the Alodokter platform. A total of 8,550 questions were obtained through web scraping between July 2024 and June 2025. The collected data were preprocessed and subsequently analyzed using seven topic modeling approaches: Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), BERTopic, Top2Vec, and ProdLDA. To assess model performance, the coherence metric (c_v) was employed to identify the most effective method. Among these techniques, NMF achieved the best results, producing the highest coherence score of 0.67 with six well-defined topics. The findings show six primary areas of concern: pregnancy; menstruation and contraceptive management; general health and minor ailments; infant care; dermatological conditions; and musculoskeletal and other physical complaints. General health-related issues occurred most frequently, particularly during seasonal transitions, while menstruation and contraceptive management received the least attention, despite menstruation contributing to women's health risks and the use of contraceptives helping to reduce maternal mortality in Indonesia. These findings offer valuable insights for digital health platforms like Alodokter to enhance information delivery and health literacy, ultimately improving online health services and supporting the achievement of SDG 3.

Paper History

Received August 10, 2025
Revised Oct. 10, 2025
Accepted Oct. 18, 2025
Published Nov. 30, 2025

Keywords

Topic Mining;
Topic Modeling;
Knowledge Discovery;
Online Health Communities;
Alodokter

Author Email

harahapdayana01@gmail.com
kenya.tania@gmail.com
putrieka@unsri.ac.id

I. Introduction

Online searches for health information are increasingly popular, along with the rapid growth of information technology [1][2]. People increasingly use the Internet as their primary source of information on nutrition, diseases, symptoms, and treatments. [3][4]. Digital platforms have changed the way people find and share health-related information. A survey by the Association of Indonesian Internet Service Providers (APJII) in 2024 showed that health is one of the top three most visited content categories, reaching 27.79% [5]. The significant attention underscores the great potential of health data generated from online interactions on these platforms. Sources for acquiring health information encompass official hospital websites, health news portals, social media, and online health communities.

Online Health Communities (OHCs) are digital platforms that provide accessible and efficient healthcare services [6]. These platforms let people interact, share information about health issues, exchange experiences, and offer mutual support [7][8][9][10]. The Q&A services in OHCs give users with various health concerns access to consultations, advice, and help from healthcare professionals [11]. Alodokter is an example of a health platform in Indonesia. Alodokter offers several services, including an online consultation service known as "Tanya Dokter" (Ask a Doctor) [12]. This tool lets users ask doctors about their concerns or symptoms. These Q&A sessions produce unstructured textual data encompassing a diverse range of health issues, symptoms, and users' experiences related to health [13][14]. These questions are considered to reflect users' genuine health concerns and information needs [15], as previous work has shown that analyzing online question texts can identify patients' real concerns, including drug side effects and unmet medical needs. In the Indonesian context, public health knowledge can be extracted from existing data, as demonstrated in previous work that identified genuine needs using mining approaches [16]. This data offers significant insights for understanding public health needs and contributes to progress toward SDG 3 (Good Health and Well-being).

Knowledge discovery plays a strategic role in supporting SDG 3. It enables the identification of patterns, trends, and public health needs that are hidden within data [17]. Through a systematic data analysis process, researchers and health practitioners can gain new insights into disease risk factors, the effectiveness of interventions, and users' health information needs. Thus, knowledge

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

discovery not only enhances the accuracy of health decision-making but also facilitates the design of more targeted, sustainable, and inclusive health policies and services, thereby advancing the objectives of SDG 3 [18].

Topic modeling approaches help the knowledge discovery process [19][20][21]. In [19], topic modeling was employed to gain insight into trends and developments in the field of cybersecurity. In [20], the method mapped knowledge structure and research trends in data warehousing. In [21], it provided practical solutions in assembly workshops by using past problem-solving records. This flexibility demonstrates that topic modeling is effective for knowledge discovery across various fields, including online health.

In this context, topic modeling often supports topic mining to identify and map health information needs from user queries. Previous research has demonstrated that topic mining can effectively identify health information needs from user-generated content in online health communities. For example, [22] studied data from xywy.com to investigate the health information needs of Chinese adolescents. The research combined statistical analysis with Sentence-BERT topic modeling and clustering. It revealed a predominance of topics in obstetrics, internal medicine, and dermatology, focusing on symptoms and treatments, while also noting the impact of demographic factors. A study by [23] examined COVID-19 information needs from six Chinese online health communities. Researchers used the CL-LDA topic model to sort material into four topics: symptoms, preventive, examination, and therapy. The study found that information needs varied according to the progression of the epidemic and user demographics. Both studies demonstrate the strong potential of topic mining in understanding health information requirements based on user interactions in online health communities.

Topic modeling has also been applied within the health context in Indonesia. In [24], Latent Dirichlet Allocation (LDA) topic modeling was utilized in combination with sentiment analysis to analyze Twitter data related to stunting, finding that negative sentiment was predominant and identifying frequently discussed terms such as “children,” “prevention,” “nutrition,” “decrease,” and “number”. In [25], LDA and sentiment analysis were employed on Twitter discussions related to COVID-19, revealing a predominance of negative sentiment and topics reflecting public concerns. In [26], LDA was applied to investigate research trends in Indonesian health studies indexed in SINTA, with 94.1% of respondents rating the topic modeling results as very good, thereby providing valuable references for future health research in Indonesia. Previous studies have demonstrated the potential of topic modeling to identify key health-related themes in the Indonesian context. However, these studies are limited by their reliance on a single topic modeling method and a focus on specific diseases or research publications. There remains a lack of studies that explore Indonesian-language health queries submitted by the public through online consultation platforms, particularly in identifying major health issues.

To address this gap, this study aims to examine user-submitted health inquiries from Alodokter's "Tanya Dokter" service using topic mining. Text mining supports the SDGs [27][28][29]. The Topic Mining-Based Knowledge Discovery approach can help SDG 3 by identifying the most in-demand health topics. This approach identifies key topics, examines temporal trends, and explores their correlations with health events in Indonesia. The topic model was evaluated using the coherence score (c_v) as the primary metric, which has been shown to exhibit the highest correlation with human interpretation among available measures [30]. To ensure that the identified topics are both meaningful and contextually relevant, this quantitative evaluation was complemented with a manual assessment of interpretability [31]. This study seeks to achieve the following goals: 1) Conduct a detailed analysis of Indonesian-language health queries from Alodokter, 2) Apply and compare various topic modeling methods to evaluate their effectiveness, 3) Identify the dominant health topics, analyze their temporal patterns, and examine their associations with health events in Indonesia, and 4) Provide insights to help online health platforms improve information relevance, health literacy, disease prevention, and service quality.

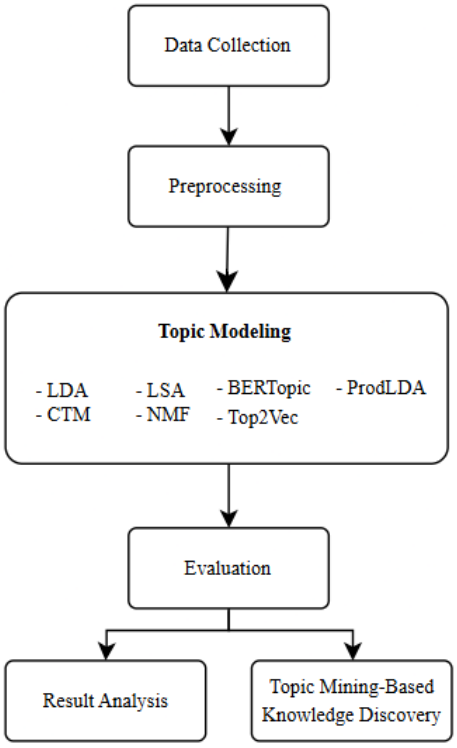

Fig. 1. Research Method of Topic Mining-Based Knowledge Discovery of User Health Information Needs.

This paper is organized into several key sections to present the methodology and research findings in a systematic manner. Following this introductory section, Section II describes the dataset, text preprocessing steps, and the topic modeling methods (LDA, CTM, LSA, NMF, BERTopic, Top2Vec, and ProdLDA), as well as the evaluation approach. Section III covers results, including coherence scores, topic interpretation, topic distribution,

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

temporal trends, and knowledge discovery. Section IV discusses the findings, compares them to prior studies, and outlines the limitations. Section V provides key insights and proposes directions for future research.

II. Materials And Methods

As shown in Fig. 1, the study workflow begins with data collection and preprocessing. Topic modeling methods, namely LDA, CTM, LSA, NMF, BERTopic, Top2Vec, and ProdLDA, are applied to analyze the processed data. Model performance is evaluated, followed by results analysis and topic mining-based knowledge discovery. In the following subsection, each stage is explained in detail.

A. Data Collection

Data were collected via web scraping from the Alodokter website's "Tanya Dokter" page. Ethical procedures were followed: only publicly accessible content was used, usernames were anonymized by the platform (e.g., "Bo****a"), and no attempts were made to identify individual users. Information was gathered from 570 pages, including username, question titles, question content, and timestamps. This produced a dataset of 8,550 questions submitted from July 1, 2024, to June 30, 2025. The year-long collection period provided comprehensive coverage of an annual cycle, facilitating analysis of topic variation and potential seasonal trends. The collected data was then recorded in CSV format for further processing.

B. Processing

Preprocessing refers to the stage of cleaning the data by removing unnecessary information [32]. In this study, the preprocessing workflow consisted of cleaning, case folding, normalization, tokenization, and removal of stop words. These steps are essential for minimizing noise and reducing imbalance to ensure coherent and interpretable topic modeling outcomes [33][34].

1) Cleaning
   Duplicate queries and missing entries were eliminated to preserve the dataset's integrity and reliability.

Table 1. Steps of Text Preprocessing Applied to User-Generated Health Questions Dataset.

Prepocessing	Text
Original Question	Does Snakehead Fish Oil Really Help Speed Up Wound Healing? Alodokter, Doc I want to ask about snakehead fish oil. Lately I have been trying to take snakehead fish oil because they say it is good for wound recovery and boosting the immune system. I just recovered from dengue fever, and now I still often feel weak and my appetite hasn't fully returned. So, I'm curious, does snakehead fish oil really help speed up recovery? Also, if taken long term, are there any side effects? I take 2 capsules a day, is that still safe? Thanks in advance doc.
Cleaning	Does Snakehead Fish Oil Really Help Speed Up Wound Healing Alodokter Doc I want to ask about snakehead fish oil Lately I have been trying to take snakehead fish oil because they say it is good for wound recovery and boosting the immune system I just recovered from dengue fever and now I still often feel weak and my appetite hasn't fully returned So I'm curious does snakehead fish oil really help speed up recovery Also if taken long term are there any side effects I take capsules a day is that still safe Thanks in advance doc
Case Folding	does snakehead fish oil really help speed up wound healing alodokter doc i want to ask about snakehead fish oil lately i have been trying to take snakehead fish oil because they say it is good for wound recovery and boosting the immune system i just recovered from dengue fever and now i still often feel weak and my appetite hasn't fully returned so i'm curious does snakehead fish oil really help speed up recovery also if taken long term are there any side effects i take capsules a day is that still safe thanks in advance doc
Normalization	does snakehead fish oil really help speed up wound healing alodokter doc i want to ask about snakehead fish oil lately i have been trying to take snakehead fish oil because they say it is good for wound recovery and boosting the immune system i just recovered from dengue fever and now i still often feel weak and my appetite has not fully returned so i am curious does snakehead fish oil really help speed up recovery also if taken long term are there any side effects i take capsules a day is that still safe thanks in advance doc
Tokenization	['does', 'snakehead', 'fish', 'oil', 'really', 'help', 'speed', 'up', 'wound', 'healing', 'alodokter', 'doc', 'i', 'want', 'to', 'ask', 'about', 'snakehead', 'fish', 'oil', 'lately', 'i', 'have', 'been', 'trying', 'to', 'take', 'snakehead', 'fish', 'oil', 'because', 'they', 'say', 'it', 'is', 'good', 'for', 'wound', 'recovery', 'and', 'boosting', 'the', 'immune', 'system', 'i', 'just', 'recovered', 'from', 'dengue', 'fever', 'and', 'now', 'i', 'still', 'often', 'feel', 'weak', 'and', 'my', 'appetite', 'has', 'not', 'fully', 'returned', 'so', 'i', 'am', 'curious', 'does', 'snakehead', 'fish', 'oil', 'really', 'help', 'speed', 'up', 'recovery', 'also', 'if', 'taken', 'long', 'term', 'are', 'there', 'any', 'side', 'effects', 'i', 'take', 'capsules', 'a', 'day', 'is', 'that', 'still', 'safe', 'thanks', 'in', 'advance', 'doc']
Stopword Removal	snakehead fish oil speed wound healing snakehead fish oil try take snakehead fish oil good wound recovery boost immune system recovered dengue fever feel weak appetite returned curious snakehead fish oil help speed recovery taken long term side effects take capsules day safe

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

Cleaning addressed data noise by eliminating irrelevant characters and patterns. URLs, hashtags, and escape characters were removed. Punctuation and numeric characters were also eliminated. Excess whitespace was standardized to ensure consistency.

2) Case Folding
Casefolding converts text to lowercase so that words with different capitalizations, such as 'Demam' and 'demam,' are recognized as the same token.

3) Normalization
Normalization focused on correcting vocabulary inconsistencies by mapping non-standard or informal Indonesian words to their formal. This was achieved by utilizing a custom dictionary containing pairs of non-standard and standard words (e.g., replacing 'gimana' with 'bagaimana' and 'udh' with 'sudah').

4) Tokenization
Tokenization divides text into words by splitting on whitespace. For instance, the sentence “anak demam” was transformed into the tokens [“anak”, “demam”].

5) Stopword Removal
Stopword Removal aims to reduce the data dimensionality by filtering out words that are common but hold little semantic value in determining the topic or meaning of a document. Common Indonesian stopwords from the NLTK Indonesian stopword list were removed. Additionally, domain-specific stopwords that frequently appear in medical Q&A contexts were excluded (e.g., dokter, alodokter, tanya, halo). These custom removals ensured that the remaining tokens carried the most relevant semantic information for topic extraction.

Several challenges were encountered during preprocessing, including handling informal language, spelling variations, and unnecessary terms. These issues required careful handling to ensure that the dataset remained clean, consistent, and suitable for topic modeling. After preprocessing, 8,425 entries remained. The detailed outcomes of the preprocessing procedures are summarized in Table 1.

C. Topic Modeling

Topic modeling is a method used to find topics within a collection of texts based on patterns of co-occurring words [35]. This technique is used to reveal hidden issues in documents [36]. In this study, seven topic modeling methods were evaluated, covering four categories of approaches: probabilistic generative models (LDA and CTM), classical matrix factorization methods (LSA and NMF), embedding-based methods (BERTopic and Top2Vec), and modern neural-based models (ProdLDA). This categorization ensures that both traditional and modern approaches are represented, providing a comprehensive evaluation of topic modeling performance. The number of topics tested ranged from 5 to 15.

1. LDA

Latent Dirichlet Allocation (LDA) discovers topics in documents using an unsupervised probabilistic approach, estimating both the distribution of words across issues and the distribution of topics within each document [37][38]. The joint probability distribution of the observed words w and the latent variables (θ, ϕ, z) given hyperparameters α and β are defined in Eq. (1):

$$p(\theta, \phi, z \mid w, \alpha, \beta) = \frac{p(w,z,\theta,\phi \mid \alpha,\beta)}{p(w \mid \alpha,\beta)} \quad (1)$$

In this study, LDA was applied to data represented in a TF–IDF matrix. The dataset was prepared by loading preprocessed text. Each document was tokenized, and a Gensim dictionary and corpus were constructed. The documents were then converted into a TF–IDF matrix, using max_df=0.8 and min_df=3 to filter out terms that were either too frequent or too rare. LDA models with different numbers of topics were trained using scikit-learn’s LatentDirichletAllocation. Each model was set to run for ten iterations, employ a batch learning algorithm, and use a random state of 42.

2. CTM

The Correlated Topic Model (CTM) extends LDA by applying a logistic-normal distribution to represent the correlations between topics within documents [39][40]. The joint probability distribution of words w, topic assignments z, document-level latent variables η, and topic–word distributions ϕ given hyperparameters μ, Σ, β is defined in Eq. (2):

$$p(w, z, \eta, \phi \mid \mu, \Sigma, \beta) = \prod_{k=1}^{K} p(\phi_k \mid \beta) \prod_{d=1}^{D} p(\eta_d \mid \mu, \Sigma)$$
$$\prod_{n=1}^{N_d} p(z_{dn} \mid \theta_d) p(w_{dn} \mid \phi_{z_{dn}}) \quad (2)$$

In this study, CTM was implemented using the tomotopy package. The preprocessed text was stored in the Tomotopy corpus format, with each document as a list of tokens. The model was trained by fitting CTM models with different numbers of topics and iteratively updating each document to estimate topic–word and document–topic distributions.

3. LSA

Latent Semantic Analysis (LSA) identifies topics in documents by using Singular Value Decomposition (SVD) to capture semantic similarities in the term–document matrix [41][42]. The top-k singular values are used to approximate the document–term matrix X using matrices U_k, $\sum_k$, and V_k^T, as shown in Eq. (3):

$$X \approx U_k \sum_k V_k^T \quad (3)$$

In this study, LSA was conducted by first converting the preprocessed documents into a TF–IDF matrix, applying max_df=0.8 and min_df=3. The documents were tokenized to construct a Gensim dictionary and corpus. Multiple LSA models with varying numbers of topics were then trained using TruncatedSVD. The models used a randomized algorithm for the decomposition, were iterated 10 times, and had the random state fixed at 42 for reproducibility.

4. NMF

Non-Negative Matrix Factorization (NMF) is a statistical approch that breaks down document–term matrices into two smaller positive matrices to find and evaluate topics in documents through an iterative optimization process [43]. The approximation between the document–term matrix V

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

and the factorized matrices W and H is expressed in Eq. (4):

$$V_{m\times n} \approx W_{m\times k} \cdot H_{k\times n}, \quad W,H \geq 0 \tag{4}$$

In this study, the preprocessed texts were first converted into a TF–IDF representation, with a maximum document frequency (max_df) of 0.8 and a minimum document frequency (min_df) of 3. Documents were tokenized to generate a Gensim dictionary and corpus. Several NMF models, each with a different number of topics, were then trained using scikit-learn's NMF implementation, with the initialization set to 'nndsvd', a maximum of 500 iterations, and random_state fixed at 42.

5. BERTopic

BERTopic is a technique for generating meaningful topics from text by employing a class-based TF-IDF approach to produce coherent topic representations [44]. The process begins with document embedding, where each document d_i is converted into a dense vector e_i using a BERT-based model Eq. (5). These embeddings are mapped into a lower-dimensional space using UMAP, producing the reduced vectors Z Eq. (6). HDBSCAN is then applied to these vectors to cluster semantically similar documents, assigning each document to a cluster c_i Eq. (7). Finally, representative words for each topic are extracted using a class-based TF–IDF weighting scheme, where the importance of term t in cluster k is calculated as shown in Eq. (8). In this study, the preprocessed text was embedded using the SentenceTransformer model (all-MiniLM-L6-v2), reduced with UMAP, clustered with HDBSCAN, and finalized with class-based TF–IDF to obtain coherent topic representations.

Document embedding:

$$e_i = f_{BERT}(d_i) \tag{5}$$

Dimensionality reduction:

$$Z = UMAP(E) \tag{6}$$

Clustering:

$$c_i = HDBSCAN(Z_i) \tag{7}$$

Topic representation:

$$c-TF-IDF_{t,k} = \frac{f_{t,k}}{n_k} \cdot \log\frac{N}{f_t} \tag{8}$$

6. Top2Vec

Top2Vec is a topic modeling algorithm that automatically discovers and represents topics in text by generating vectors for topics and documents while reducing the influence of noisy data [45]. Each document d_i is mapped into an embedding vector e_i Eq. (9), and each word w_j is similarly represented as an embedding vector v_j Eq. (10). The document embeddings are then clustered into groups c_i that represent potential topics, Eq. (11). For each cluster C_k, a topic vector t_k is defined as the centroid of the document embeddings contained within it Eq. (12). Finally, the most representative words of each topic are identified by selecting those whose word embeddings maximize the cosine similarity with the topic vector, as shown in Eq. (13). Top2Vec was executed using the all-MiniLM-L6-v2 transformer model, with the number of topics refined through hierarchical topic reduction.

Document embedding:

$$e_i = f_{embed}(d_i) \tag{9}$$

Words embedding:

$$v_j = f_{embed}(w_j) \tag{10}$$

Clustering:

$$c_i = Cluster(e_i) \tag{11}$$

Topic centroid:

$$t_k = \frac{1}{|C_k|}\sum_{e_i\in C_k} e_i \tag{12}$$

Topic words:

$$Topic_k = \underset{w_j}{\operatorname{argmax}} \cos(t_k, v_j) \tag{13}$$

7. ProdLDA

ProdLDA is a neural-based topic model derived from LDA, which leverages Autoencoded Variational Inference for Topic Models (AVITM) and replaces the word-level mixture model with a product of experts [46]. In this framework, each document d is associated with a latent variable z_d drawn from a standard normal distribution, which is then mapped to topic proportions θ_d through a softmax transformation Eq. (14). The decoder models the word distribution for a document as a multinomial conditioned on the topic proportions and topic–word matrix β Eq. (15). Training involves maximizing the Evidence Lower Bound (ELBO), which balances reconstruction accuracy while enforcing regularization by penalizing divergence between the variational posterior and the prior Eq. (16). In this study, ProdLDA was implemented by converting preprocessed text into vector representations with max_df=0.8 and min_df=3. The encoder–decoder neural network was optimized using Stochastic Variational Inference (SVI) in Pyro using the Adam optimizer across a range of topic numbers.

Logistic normal:

$$z_d \sim \mathcal{N}(0,I), \ \theta_d = \text{softmax}(z_d) \tag{14}$$

Decoder:

$$p(w_d|\,\theta_d,\beta) = \text{Multinomial}(N_d, \text{softmax}(\beta\theta_d)) \tag{15}$$

Training objective (ELBO):

$$\mathcal{L} = \mathbb{E}_{q(z_d|w_d)}[\log p(w_d|z_d)] - \text{KL}(q(z_d|w_d)||P(z_d)) \tag{16}$$

D. Evaluation

Topic coherence evaluates the quality of topics in topic modeling by emphasizing their interpretability and comprehensibility for humans [47][48]. Among the available measures, the c_v score is widely applied. This metric evaluates topics by constructing word vectors from co-occurrence patterns and measuring their similarity through cosine similarity and Normalized Pointwise Mutual Information (NPMI) [49].

Topic coherence was employed as the quantitative metric to evaluate and compare topic modeling approaches. For each model, topics were extracted, and the most representative words were used to calculate the coherence score (c_v). Consistent with previous studies [50][51] that evaluated topic modeling performance across varying topic numbers, this study explored a range from 5 to 15 to determine the optimal number for the dataset. The model configuration achieving the top coherence score was identified as the optimal one. The coherence score, which ranges from 0 to 1, quantifies the frequency with

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

Table 2. Topic Interpretation Based on Ten Representative Words from User Questions.

Topic	Representative Words	Topic Interpretation	User Question
0	pregnant, belly, pregnancy, fetus, age, sign, normal, mark, sex, weeks	pregnancy	What Are the Signs That a Fetus Has Died in the Womb? Hi Doc, I'm currently 7 months pregnant, and lately I've been thinking a lot because I feel my baby's movements are becoming less frequent. Usually very active, but for the past two days, I haven't felt any movement at all. I also feel like my belly isn't as firm as it usually is. There's no spotting or bleeding, but I'm scared that maybe something bad happened, like my baby isn't developing or has even died in the womb. Are there any definite signs that the fetus has died in the womb, doc? And when should I see a doctor immediately?
1	period, blood, kb, late, finish, spotting, injection, cycle, brown, menstruation	menstruation and contraceptive management	Can Kb Implants Cause Irregular Periods? Good morning doc. I had a KB implant inserted about 4 months ago. Since then, my menstrual cycle has become irregular. Sometimes I don't have a period for 2 months, then suddenly there's spotting for a few days, and then it stops again. Before using KB, my period was quite regular. I'm a bit worried. Is it normal for KB implants to have this kind of effect? Will it return to normal once the body gets used to it, or will this effect persist throughout my use of the implant? Also, if my periods become rare, are there any other health impacts?
2	medicine, eat, body, take, child, food, cough, weight, fever, sleep	general health and minor ailments	How to Treat a Hot Body But No Fever. Doc, the unpredictable weather right now makes my body feel strange too. I feel like my body is hot, but when I check the temperature, it's normal. What should I do, doc, to treat it? If I take paracetamol, I don't have a fever anyway. please give me your advice, doc :))
3	baby, child, age, breastfeed, milk, sign, mpasi, defecate, normal, formula	infant care	What Kind of Fish Is Safe for Baby's Mpasi? My baby is now 8 months old, and has started MPASI. I would like to know what types of fish are safe and suitable for babies. I'm afraid of choosing the wrong one, because I heard some fish are high in mercury and not ideal for babies.
4	itchy, skin, spots, face, red, use, acne, remove, appear, dry	dermatological conditions	How to Treat Sunburn After Being Exposed to Sunlight? My skin just got burned after being exposed to sunlight during my beach vacation yesterday. At that time, I forgot to use sunscreen and spent about 3 hours outside around noon. Now my shoulders and back are very red, sore to the touch, and slightly peeling. I've tried cold compresses and using moisturizer, but it still feels hot and uncomfortable. Are there any products that can help relieve the pain and heal the burned skin? I want it to heal quickly because I have an essential event soon and I need to wear a sleeveless outfit.
5	eye, lump, left, pain, right, tooth, ear, head, swollen	musculoskeletal and other physical complaints	Light Exercise to Relieve Back Pain doc for the past 2 years I've often had back pain, and I also don't move much doc. what kind of light exercise can I do that doesn't require too much movement to help relieve my back pain?

which the most representative topic words co-occur within the corpus. Higher coherence scores demonstrate that topic words frequently co-occur across documents, indicating greater semantic coherence and interpretability. Similar to [31], in addition to the c_v evaluation metric, the top 10 keywords of each topic were manually reviewed to assess interpretability, ensuring that the resulting topics were coherent, meaningful, and distinct in representing health-related themes.

III. Results

A. Evaluation

Based on the coherence score analysis, the NMF model demonstrates the best performance among all models tested, as illustrated in Fig. 2. NMF achieved its peak coherence score of 0.67 using six topics as the optimal number. This result indicates that NMF is capable of generating the most coherent and interpretable topics, with strong semantic relationships among the words within each topic. In comparison, CTM also performs well, yielding stable and relatively high coherence values, with its highest score of 0.60 achieved at 15 topics. ProdLDA likewise shows competitive performance, although its

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

coherence values are less stable than those of NMF and CTM.

The LDA and LSA methods demonstrate weaker performance as the number of topics increases. LDA achieves its highest coherence score of 0.55 at five topics, which subsequently declines to 0.42 as the number of topics grows. LSA consistently exhibits the lowest performance, with coherence scores ranging from 0.38 to 0.44. In contrast, embedding-based models, such as BERTopic and Top2Vec, display relatively stable results across different topic numbers, although their coherence scores remain lower than those of NMF. Overall, NMF emerges as the most suitable method for this study, as it consistently produces the most meaningful and coherent topics.

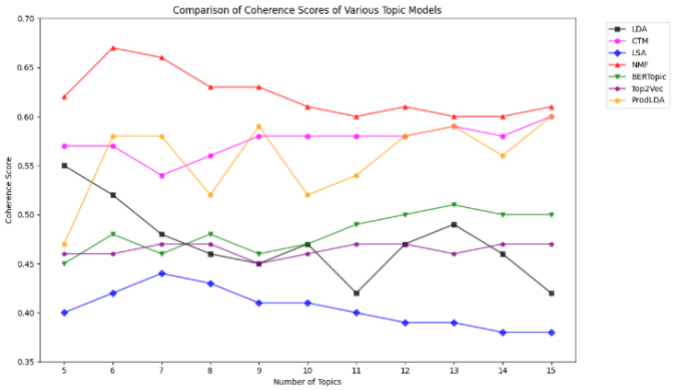

Fig. 2. Comparative Performance of Topic Models Based on Coherence Score Across a Range of Topics.

Given the NMF model's superior performance, the study conducted topic analysis using its outcomes, extracting the top ten words from each topic to characterize the thematic content. Based on these keywords, six topics were identified and subsequently categorized into broader health themes. As shown in Table 2, the topics include: pregnancy (e.g., hamil [pregnant], kehamilan [pregnancy], janin [fetus]), menstruation and contraceptive management (e.g., haid [period], KB [contraception], siklus [cycle]), general health and minor ailments (e.g., obat [medicine], batuk [cough], demam [fever]), infant care (e.g., bayi [baby], susu [milk], MPASI [complementary foods]), dermatological conditions (e.g., gatal [itchy], kulit [skin], jerawat [acne]), and musculoskeletal and physical complaints (e.g., benjolan [lump], nyeri [pain], bengkak [swollen]).

B. Result Analysis

The distribution of users' health information needs, as shown in Fig. 3, indicates that discussions concerning general health and minor ailments dominate online conversations, accounting for the most significant proportion at 31.7% of the total documents. This means strong public interest in health complaints, daily concerns, and medication-related inquiries. Topics on dermatological conditions (21.6%) and musculoskeletal and other physical complaints (19.2%) also represent a substantial share of the data. Collectively, these three categories account for more than 70% of the dataset, confirming that general health issues are the primary focus of user discussions. By contrast, topics such as pregnancy (10.6%), infant care (9.5%), and menstruation and contraceptive management (7.4%) appear less frequently, reflecting comparatively lower emphasis in online conversations.

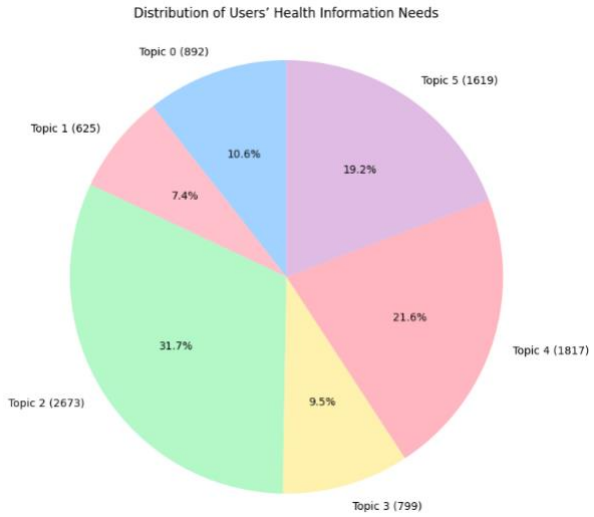

Fig. 3. Distribution of Users' Health Information Needs Across Six Discovered Health Topics.

The monthly trend illustrated in Fig. 4 shows that discussions of general health and minor ailments consistently dominate across the observation period. Notably, a marked increase is observed in late September 2024 and early April 2025, coinciding with seasonal transitions in Indonesia. This suggests that changes in weather conditions directly influence the rise in health complaints, particularly those related to the respiratory and immune systems, such as coughing and fever. Discussions on dermatological conditions decline in early 2025 but surge again between April and June 2025, likely due to temperature and humidity fluctuations associated with seasonal changes. Fluctuations in musculoskeletal and other physical complaints are observed. However, these variations are difficult to associate with seasonal patterns, in contrast to general health, minor ailments, and dermatological conditions. Meanwhile, topics such as pregnancy, menstruation, and contraceptive management, and infant care remain relatively stable, indicating that these topics represent routine concerns that are less affected by seasonal variations.

Beyond overall distributions and seasonal variations, this study identifies detailed patterns within specific sub-topics, correlations, and temporal trends. Pregnancy-related discussions primarily focus on fetal development and pregnancy symptoms. Menstruation and contraceptive management cover irregular cycles and hormonal side effects. General health and minor ailments are characterized by concerns such as fever and cough. Musculoskeletal and other physical complaints manifest as localized pain or swelling, toothaches, earaches, and lumps in a specific body part. Infant care subtopics include breastfeeding and complementary foods, while dermatological conditions are primarily characterized by acne and rashes. There is a notable correlation between

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

pregnancy and infant care, as prenatal concerns frequently transition into planning for infant care. Temporal analysis reveals that minor ailments are most prevalent during the rainy season, whereas dermatological issues tend to increase during months with higher temperatures.

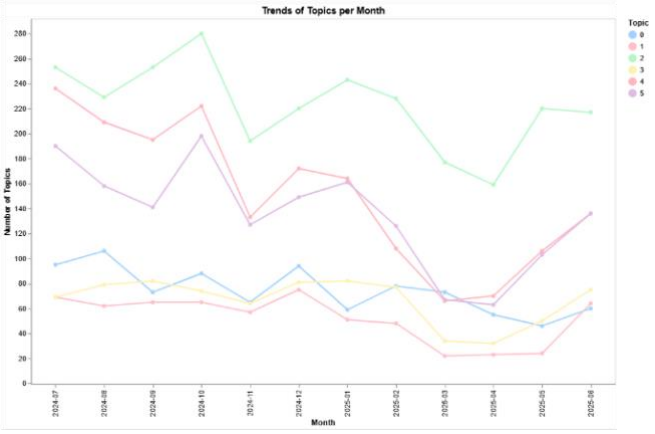

Fig. 4. Trends of Topics per Month Regarding User Health Information Needs.

C. Topic Mining-Based Knowledge Discovery

At this stage, this study identifies the knowledge discovery dimension based on the topic mining results. Rather than limiting the analysis to a list of topics, the study also explores the potential knowledge that can inform healthcare development. The identified topics reveal several actionable insights as follows.

1) Knowledge Discovery of Health Information Needs Patterns

Analysis of the generated topics (pregnancy, menstruation and contraceptive management, general health and minor ailments, infant care, dermatological conditions, musculoskeletal and other physical complaints) demonstrates that users predominantly seek practical, specific information directly related to their daily health conditions. This finding suggests that online health services should prioritize content focused on self-care, managing minor ailments, and facilitating preventive consultations.

2) Knowledge Discovery of Health Service Gaps

The topic mining results also revealed information gaps that have not been optimally addressed in conventional health services. For example, the topics of dermatological conditions, menstruation, and contraceptive management indicate that many users still seek alternative solutions in online forums before going to the doctor. The topic of musculoskeletal and other physical complaints frequently emerges in connection with sedentary lifestyles and work postures, yet these issues are often underemphasized in general medical consultations. This insight suggests that online health platforms have the potential to address service gaps by offering educational modules, interactive FAQs, or dedicated chatbots.

3) Knowledge Discovery of Inter-Topic Relationships

Topic mining can also generate relational insights. For example, pregnancy topics are often related to infant care, while menstruation and contraceptive management sometimes overlap with dermatological conditions (e.g., the side effects of hormonal contraceptives on the skin). This knowledge is helpful in designing integrated information flows; for example, when a user searches for pregnancy information, the system also recommends articles about infant immunizations or maternal nutrition.

4) Knowledge Discovery of Topic Trends and Evolution

If data are analyzed temporally, it can be identified when a topic is more dominant. Temporal analysis reveals that general health and minor ailments tend to peak during seasonal transitions in Indonesia, whereas dermatological conditions increase in response to changes in temperature and humidity. These patterns suggest opportunities for adaptive content strategies, including seasonal health campaigns, preventive education, and timely recommendations tailored to users' changing needs.

IV. Discussion

This study analyzed question–answer interactions from the "Tanya Dokter" feature of the Alodokter platform, spanning the period from July 1, 2024, to June 30, 2025. Multiple topic modeling methods were compared, including Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM), Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), BERTopic, Top2Vec, and ProdLDA. Based on coherence scores, NMF outperformed all other methods (c_v = 0.67), indicating its superior capability to produce coherent and interpretable topics. A coherence score within this range indicates robust topic quality, as prior studies demonstrate that values around 0.60 are associated with efficient, interpretable, and semantically coherent topics [52][53][54]. The NMF-based analysis revealed six primary topics that reflect the health information needs of Alodokter users: pregnancy, menstruation and contraceptive management, general health and minor ailments, infant care, dermatological conditions, and musculoskeletal and other physical complaints.

General health and minor ailments dominated the discussions, with notable peaks observed in September 2024 and April 2025. These peaks coincide with Indonesia's seasonal transitions, suggesting that environmental fluctuations may influence public demand for health information, particularly regarding respiratory and immune-related complaints. This aligns with findings by [55], which highlight the prevalence of respiratory and immune system illnesses such as influenza, commonly presenting with symptoms like cough, fever, and runny nose. The study also reported that peaks in influenza incidence coincided with periods of higher precipitation, further emphasizing the link between climatic conditions and health concerns during transitional seasons in Indonesia. This insight underscores the potential for digital health platforms to develop seasonally adaptive preventive education and for public health authorities to implement more responsive communication strategies that address health risks associated with general health and minor ailments. These efforts align with Indonesia's National Medium-Term Development Plan (RPJMN) (2025–2029), Priority 4, which emphasizes enhancing the

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

quality of primary care, the equitable distribution of healthcare workers, and preventive services [56].

Dermatological conditions represent a significant public health concern that affects both quality of life and productivity [57]. The study also noted that the burden of skin and subcutaneous diseases in Indonesia varies geographically and is influenced by socioeconomic disparities, access to care, and environmental factors such as weather and climate, which may explain the increased occurrence of skin-related inquiries during seasonal transitions. These findings align with Minister of Health Regulation No. 71 of 2015, which classifies skin and subcutaneous diseases as part of non-communicable diseases (NCDs) that require comprehensive prevention and control measures through health promotion, early risk detection, and protective interventions.

The emergence of the pregnancy topic indicates growing public awareness and concern regarding maternal health issues. However, Indonesia's maternal mortality ratio (MMR) remains among the highest in Southeast Asia [58]. In response, the Indonesian government has implemented policies and programs to reduce maternal mortality and ensure that all deliveries take place in standardized healthcare facilities, in line with the Policy Intervention for Achieving National Priority Target 4 [56].

Queries related to infant care demonstrate increasing public attention to child health, with examples found in discussions about breastfeeding and complementary feeding (MPASI). In Indonesia, persistent cases of chronic malnutrition that lead to child stunting remain a serious public health challenge [59]. The findings underscore the importance of providing age-appropriate nutrition, strengthening healthcare support, and ensuring regular growth monitoring. In line with this, Indonesia's RPJMN includes ongoing national programs that aim to reduce stunting by 2029 through integrated interventions during the first 1,000 days of life [56].

Musculoskeletal and other physical complaints reflect growing public awareness of lifestyle-related and occupational health issues. These conditions are prevalent across various occupational groups in Indonesia, with common symptoms including neck, back, waist, and shoulder pain, as well as muscle fatigue [60][61][62]. The government's policy response to these issues is reflected in the Healthy Living Community Movement (GERMAS) initiative, which promotes physical activity and preventive healthcare. This initiative highlights the growing public awareness of occupational and lifestyle-related health issues, emphasizing the importance of health promotion, regular exercise, and ergonomic education [56].

Menstruation and contraceptive management emerged as the least discussed topics, representing only 7.4% of the dataset (625 out of 8,425 questions). Despite their underrepresentation compared to general health (31.7%) and dermatological conditions (21.6%), these topics remain significant. This is notable given national data, including the Health Statistics Profile 2023, which highlights that menstruation contributes to women's health risks and that safe and effective contraceptive use is essential for reducing maternal mortality and preventing pregnancy-related complications [63]. This finding reveals a discrepancy between the urgency of national health issues and the public's interest in seeking related information. For Alodokter and similar platforms, this highlights the need to strengthen literacy on menstruation and contraceptive management through targeted educational campaigns. For policymakers, it underscores the importance of developing comprehensive public communication strategies that promote awareness and preventive care related to menstruation and contraceptive management.

These findings demonstrate that digital health platforms serve not only as channels for addressing immediate user needs but also as diagnostic tools for identifying gaps in health literacy and public awareness. In alignment with Sustainable Development Goal (SDG) 3, user-generated health queries represent a valuable resource that can be leveraged to enhance the relevance of health services, inform adaptive communication strategies, and contribute to improving health outcomes. Integrating such user-driven insights into both platform design and policy development holds significant potential for advancing national health literacy and promoting sustainable health behaviors.

In comparison to the previous study, this study confirms that topic modeling is well-suited for analyzing health-related data. For example, [64] applied LDA to news articles at the onset of the COVID-19 outbreak and successfully extracted key themes. More recently, [65] compared NMF and LDA on maternal health data and found that NMF achieved substantially higher coherence scores. Consistent with these findings, our results show that NMF obtained the highest coherence score (c_v = 0.67) across the seven tested models. Collectively, these studies underscore both the applicability of topic modeling in healthcare and the superior performance of NMF in producing coherent and interpretable topics.

Certain limitations in this study warrant consideration. The analysis is based on user-generated questions from Alodokter, a single online health platform. While the survey is intentionally focused on this platform, extending the approach to other platforms may reveal additional topics and trends. Second, the platform maintains user anonymity, resulting in the unavailability of demographic information, including age, gender, education, and geographic location. This absence of demographic data limits the depth of interpretation, as the distribution of health topics may be shaped by the user composition on the platform. The user composition on the platform may, in turn, shape the distribution of health topics. For example, a higher proportion of female users may result in more questions about menstruation and pregnancy, while younger users may focus on general health and daily ailments. Socioeconomic differences could also affect interest in topics such as dermatological conditions. These patterns remain insufficiently examined.

While NMF produced the most coherent and interpretable topics in this study, several limitations should be acknowledged. The normalization process for non-formal Indonesian words may not capture all linguistic

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

variations, which could potentially affect token consistency. The removal of stopwords, although effective for reducing noise, may have excluded words with subtle contextual significance, potentially impacting topic interpretation. The choice of TF–IDF thresholds and NMF hyperparameters, although yielding satisfactory results, may affect the balance between topic specificity and generality. Although the six-topic configuration achieved optimal coherence in this research, it is important to acknowledge that NMF's performance is sensitive to the number of components, which may alter topic granularity in other contexts.

V. Conclusion

This study identified patterns of public health information needs using user-generated questions from the Alodokter "Tanya Dokter" feature. Seven topic modeling methods were evaluated, with NMF achieving the highest coherence value (0.67), making it the most effective approach for this dataset. The analysis revealed six dominant topics; general health and minor ailments were the most frequently discussed, while menstruation and contraceptive management were the least represented. These findings provide actionable insights for Alodokter to enhance user services by prioritizing high-demand areas while proactively addressing underrepresented but important topics, such as menstruation and contraceptive management, through targeted educational initiatives.

This study advances the understanding of online health information needs in the Indonesian context and highlights how topic modeling enables knowledge discovery from user interactions in online health communities, and contributes to SDG 3. Future research could incorporate demographic factors to capture variations in information needs across user groups and employ sentiment or emotion analysis to understand the psychological dimensions of health-related questions better.

References

[1] X. Jia, Y. Pang, and L. S. Liu, "Online Health Information Seeking Behavior: A Systematic Review," Healthcare, vol. 9, no. 12, p. 1740, Dec. 2021, doi: 10.3390/healthcare9121740.

[2] Y. Hua, W. Shujuan, and W. Fucheng, "Online health community—An empirical analysis based on grounded theory and entropy weight TOPSIS method to evaluate the service quality," Digit. Heal., vol. 9, Jan. 2023, doi: 10.1177/20552076231207201.

[3] E. Vega, M. Zepeda, E. Gutierrez, M. Martinez, S. Gomez, and S. Caldera, "Internet Health Information on Patient's Decision-Making: Implications, Opportunities and Challenges," Med. Res. Arch., vol. 11, no. 7.2, 2023, doi: 10.18103/mra.v11i7.2.4066.

[4] K. Stifjell, T. M. Sandanger, and C. Wien, "Exploring Online Health Information–Seeking Behavior Among Young Adults: Scoping Review," J. Med. Internet Res., vol. 27, p. e70379, Sep. 2025, doi: 10.2196/70379.

[5] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), "Survei Penetrasi Internet Indonesia 2024." [Online]. Available: https://survei.apjii.or.id/

[6] Z. Sun, K. Wang, Y. Jin, Z. Wang, and R. Yang, "Why are you? Exploring patients' behavior in selecting physicians in online health communities," Inf. Manag., vol. 62, no. 6, p. 104176, Sep. 2025, doi: 10.1016/j.im.2025.104176.

[7] S. Sanger, S. Duffin, R. E. Gough, and P. A. Bath, "Use of Online Health Forums by People Living With Breast Cancer During the COVID-19 Pandemic: Thematic Analysis," JMIR Cancer, vol. 9, p. e42783, Feb. 2023, doi: 10.2196/42783.

[8] H. E. Wood et al., "Moderators' Experiences of the Safety and Effectiveness of Patient Engagement in an Asthma Online Health Community: Exploratory Qualitative Interview Study," J. Med. Internet Res., vol. 27, p. e58167, Apr. 2025, doi: 10.2196/58167.

[9] J. Gao, Y. Zhao, D. Yang, Y. Liu, and L. Zhao, "Dynamic recommender system for chronic disease-focused online health community," Expert Syst. Appl., vol. 258, p. 125086, Dec. 2024, doi: 10.1016/j.eswa.2024.125086.

[10] Y. Zhao and L. Zhang, "Getting better? Examining the effects of social support in OHCs on users' emotional improvement," Inf. Process. Manag., vol. 61, no. 4, p. 103754, Jul. 2024, doi: 10.1016/j.ipm.2024.103754.

[11] R. Bongelli, A. Bertolazzi, M. Paolanti, and I. Riccioni, "Exploring online patient-doctor interactions. An epistemic and pragmatic analysis of Q&A patterns in an Italian 'Ask to the doctor' medical forum," Patient Educ. Couns., vol. 134, p. 108662, May 2025, doi: 10.1016/j.pec.2025.108662.

[12] Alodokter, "Tanya Dokter", [Online]. Available: https://www.alodokter.com/komunitas/diskusi/penyakit

[13] L. Nie, J. Xu, and R. Wang, "Health information needs and feedback of users in the online TCM community," PLoS One, vol. 19, no. 3, p. e0301536, Mar. 2024, doi: 10.1371/journal.pone.0301536.

[14] K. Mermin-Bunnell et al., "Use of Natural Language Processing of Patient-Initiated Electronic Health Record Messages to Identify Patients With COVID-19 Infection," JAMA Netw. Open, vol. 6, no. 7, p. e2322299, Jul. 2023, doi: 10.1001/jamanetworkopen.2023.22299.

[15] M. Kamba et al., "Medical Needs Extraction for Breast Cancer Patients from Question and Answer Services: Natural Language Processing-Based Approach," JMIR Cancer, vol. 7, no. 4, p. e32005, Oct. 2021, doi: 10.2196/32005.

[16] H. Sufriyana, Y.-W. Wu, and E. C.-Y. Su, "Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia," EBioMedicine, vol. 54, p. 102710, Apr. 2020, doi:

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

10.1016/j.ebiom.2020.102710.

[17] Q. Zhang, T. Guan, and Y. Liao, "Knowledge of and policy support for the SDGs: An inverted U-shaped relationship," J. Environ. Manage., vol. 368, p. 122117, Sep. 2024, doi: 10.1016/j.jenvman.2024.122117.

[18] A. Ficko, S. Sarkki, Y. S. Gultekin, A. Egli, and J. Hiedanpää, "Reflective thinking meets artificial intelligence: Synthesizing sustainability transition knowledge in left-behind mountain regions," Geogr. Sustain., vol. 6, no. 1, p. 100257, Feb. 2025, doi: 10.1016/j.geosus.2024.100257.

[19] F. Alqurashi and I. Ahmad, "A data-driven multi-perspective approach to cybersecurity knowledge discovery through topic modelling," Alexandria Eng. J., vol. 107, pp. 374–389, Nov. 2024, doi: 10.1016/j.aej.2024.07.044.

[20] T. Timakum, S. Lee, and M. Song, "Exploring the research landscape of data warehousing and mining based on DaWaK Conference full-text articles," Data Knowl. Eng., vol. 135, p. 101926, Sep. 2021, doi: 10.1016/j.datak.2021.101926.

[21] W. Ning, J. Liu, and H. Xiong, "Knowledge discovery using an enhanced latent Dirichlet allocation-based clustering method for solving on-site assembly problems," Robot. Comput. Integr. Manuf., vol. 73, p. 102246, Feb. 2022, doi: 10.1016/j.rcim.2021.102246.

[22] J. Wang, X. Wang, L. Wang, and Y. Peng, "Health Information Needs of Young Chinese People Based on an Online Health Community: Topic and Statistical Analysis," JMIR Med. Informatics, vol. 9, no. 11, p. e30356, Nov. 2021, doi: 10.2196/30356.

[23] J. Wang, L. Wang, J. Xu, and Y. Peng, "Information Needs Mining of COVID-19 in Chinese Online Health Communities," Big Data Res., vol. 24, p. 100193, May 2021, doi: 10.1016/j.bdr.2021.100193.

[24] A. Muhaimin et al., "Social Media Analysis and Topic Modeling: Case Study of Stunting in Indonesia," Telematika, vol. 20, no. 3, p. 406, Nov. 2023, doi: 10.31315/telematika.v20i3.10797.

[25] M. Habibi, A. Priadana, and M. Rifqi Ma'arif, "Sentiment Analysis and Topic Modeling of Indonesian Public Conversation about COVID-19 Epidemics on Twitter," IJID (International J. Informatics Dev., vol. 10, no. 1, pp. 23–30, Jun. 2021, doi: 10.14421/ijid.2021.2400.

[26] Y. Sahria and Dhomas Hatta Fudholi, "Analysis of Health Research Topics in Indonesia Using the LDA (Latent Dirichlet Allocation) Topic Modeling Method," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 4, no. 2, pp. 336–344, Apr. 2020, doi: 10.29207/resti.v4i2.1821.

[27] B. Krishna and P. Puram, "The impact of trust-based challenges on user satisfaction in food sharing platforms: A text mining approach," Technol. Forecast. Soc. Change, vol. 216, p. 124159, Jul. 2025, doi: 10.1016/j.techfore.2025.124159.

[28] N. Falah, N. Falah, J. Solis-Guzman, and M. Marrero, "An indicator-based framework of circular cities focused on sustainability dimensions and sustainable development goal 11 obtained using machine learning and text analytics," Sustain. Cities Soc., vol. 121, p. 106219, Mar. 2025, doi: 10.1016/j.scs.2025.106219.

[29] N. Strelkovskii and N. Komendantova, "Integration of UN sustainable development goals in national hydrogen strategies: A text analysis approach," Int. J. Hydrogen Energy, vol. 102, pp. 1282–1294, Feb. 2025, doi: 10.1016/j.ijhydene.2025.01.134.

[30] E. Rijcken, K. Zervanou, M. Spruit, P. Mosteiro, F. Scheepers, and U. Kaymak, "Exploring Embedding Spaces for more Coherent Topic Modeling in Electronic Health Records," in 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, Oct. 2022, pp. 2669–2674. doi: 10.1109/SMC53654.2022.9945594.

[31] E. Navarro and H. Homayouni, "Topic Modeling in Cardiovascular Research Publications," vol. 1, no. 1, 2023.

[32] M. Mujahid et al., "Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19," Appl. Sci., vol. 11, no. 18, p. 8438, Sep. 2021, doi: 10.3390/app11188438.

[33] M. Razavi et al., "Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review," JMIR Ment. Heal., vol. 11, p. e53714, Aug. 2024, doi: 10.2196/53714.

[34] E. W. D'Souza, A. J. MacGregor, R. R. Markwald, T. A. Elkins, and J. M. Zouris, "Investigating insomnia in United States deployed military forces: A topic modeling approach," Sleep Heal., vol. 10, no. 1, pp. 75–82, Feb. 2024, doi: 10.1016/j.sleh.2023.09.014.

[35] C. Lalk et al., "Measuring Alliance and Symptom Severity in Psychotherapy Transcripts Using Bert Topic Modeling," Adm. Policy Ment. Heal. Ment. Heal. Serv. Res., vol. 51, no. 4, pp. 509–524, Jul. 2024, doi: 10.1007/s10488-024-01356-4.

[36] J. W. Lee, Y. Kim, and D. H. Han, "LDA-based topic modeling for COVID-19-related sports research trends," Front. Psychol., vol. 13, Nov. 2022, doi: 10.3389/fpsyg.2022.1033872.

[37] B. Karas, S. Qu, Y. Xu, and Q. Zhu, "Experiments with LDA and Top2Vec for embedded topic discovery on social media data—A case study of cystic fibrosis," Front. Artif. Intell., vol. 5, Aug. 2022, doi: 10.3389/frai.2022.948313.

[38] S. Ying, "Guests' Aesthetic experience with lifestyle hotels: An application of LDA topic modelling analysis," Heliyon, vol. 10, no. 16, p. e35894, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35894.

[39] Z. Chen and B. Zaman, "In case players were wondering: A topic modelling and sentiment

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

analysis study of community discussions on weapon cases in the CS:GO game," Entertain. Comput., vol. 54, p. 100936, Jun. 2025, doi: 10.1016/j.entcom.2025.100936.

[40] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," Inf. Syst., vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[41] A. Meddeb and L. Ben Romdhane, "Using Topic Modeling and Word Embedding for Topic Extraction in Twitter," Procedia Comput. Sci., vol. 207, pp. 790–799, 2022, doi: 10.1016/j.procs.2022.09.134.

[42] W. Jo, Y. Kim, M. Seo, N. Lee, and J. Park, "Online information analysis on pancreatic cancer in Korea using structural topic model," Sci. Rep., vol. 12, no. 1, p. 10622, Jun. 2022, doi: 10.1038/s41598-022-14506-1.

[43] T. Kekere, V. Marivate, and M. Hattingh, "Exploring COVID-19 public perceptions in South Africa through sentiment analysis and topic modelling of Twitter posts," African J. Inf. Commun., no. 31, Jun. 2023, doi: 10.23962/ajic.i31.14834.

[44] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[45] X. Gao and C. Sazara, "Discovering Mental Health Research Topics with Topic Modeling," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.13569

[46] A. Srivastava and C. Sutton, "Autoencoding Variational Inference For Topic Models," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.01488

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

[48] S. Tunca, "Algorithms of emotion: A hybrid NLP analysis of neurodivergent Reddit communities," Acta Psychol. (Amst)., vol. 260, p. 105519, Oct. 2025, doi: 10.1016/j.actpsy.2025.105519.

[49] A. Cheddak, T. Ait Baha, Y. Es-Saady, M. El Hajji, and M. Baslam, "BERTopic for Enhanced Idea Management and Topic Generation in Brainstorming Sessions," Information, vol. 15, no. 6, p. 365, Jun. 2024, doi: 10.3390/info15060365.

[50] E. Cheese et al., "Using Natural Language Processing to Explore Patient Perspectives on AI Avatars in Support Materials for Patients With Breast Cancer: Survey Study," J. Med. Internet Res., vol. 27, p. e70971, Jun. 2025, doi: 10.2196/70971.

[51] X. Li, X. Liu, C. Yin, S. Collins, and E. Alanazi, "Impact of a Virtual Reality Video ('A Walk-Through Dementia') on YouTube Users: Topic Modeling Analysis," JMIR Form. Res., vol. 9, pp. e67755–e67755, Apr. 2025, doi: 10.2196/67755.

[52] S. O. Korkut, O. O. Kaymak, A. Onan, E. Ulker, and F. Yalcin, "A Roadmap of Emerging Trends Discovery in Hydrology: A Topic Modeling Approach," pp. 1–16, 2023, [Online]. Available: https://arxiv.org/abs/2310.15943v1

[53] K. Georgiou, N. Mittas, A. Chatzigeorgiou, and L. Angelis, "An empirical study of COVID-19 related posts on Stack Overflow: Topics and technologies," J. Syst. Softw., vol. 182, p. 111089, Dec. 2021, doi: 10.1016/j.jss.2021.111089.

[54] A. Krishnan and I. M. Ghebrehiwet, "GCD-TM: Graph-Driven Community Detection for Topic Modelling in Psychiatry Texts," in Proceedings of the 1st Workshop on NLP for Science (NLP4Science), Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 47–57. doi: 10.18653/v1/2024.nlp4science-1.6.

[55] D. Agustian, K. Mutyara, C. Murad, T. M. Uyeki, C. B. Kartasasmita, and E. A. Simoes, "Epidemiology and population-based incidence of influenza in two communities, Bandung district, West Java, Indonesia, 2008–2011," Narra J, vol. 4, no. 3, p. e981, Oct. 2024, doi: 10.52225/narra.v4i3.981.

[56] Kementerian PPN/Bappenas, "Rencana Pembangunan Jangka Menengah Nasional Tahun 2025-2029." [Online]. Available: https://www.bappenas.go.id/datapublikasishow?q=Rencana Pembangunan dan Rencana Kerja Pemerintah

[57] F. U. Prameswari, F. R. Muharram, T. Setyaningrum, and C. R. S. Prakoeswa, "Burden of Skin and Subcutaneous Diseases in Indonesia 1990 to 2019," Acta Derm. Venereol., vol. 103, p. adv18291, Dec. 2023, doi: 10.2340/actadv.v103.18291.

[58] M. Syairaji, D. S. Nurdiati, B. S. Wiratama, Z. D. Prüst, K. W. M. Bloemenkamp, and K. J. C. Verschueren, "Trends and causes of maternal mortality in Indonesia: a systematic review," BMC Pregnancy Childbirth, vol. 24, no. 1, p. 515, Jul. 2024, doi: 10.1186/s12884-024-06687-6.

[59] I. Siramaneerat, E. Astutik, F. Agushybana, P. Bhumkittipich, and W. Lamprom, "Examining determinants of stunting in Urban and Rural Indonesian: a multilevel analysis using the population-based Indonesian family life survey (IFLS)," BMC Public Health, vol. 24, no. 1, p. 1371, May 2024, doi: 10.1186/s12889-024-18824-z.

[60] T. A. E. Prasetya, N. I. A. Samad, A. Rahmania, D. A. Arifah, R. A. A. Rahma, and A. Al Mamun, "Workstation Risk Factors for Work-related Musculoskeletal Disorders Among IT Professionals in Indonesia," J. Prev. Med. Public Heal., vol. 57, no. 5, pp. 451–460, Sep. 2024, doi: 10.3961/jpmph.24.214.

[61] K. A. Akbar, P. Try, P. Viwattanakulvanid, and K. Kallawicha, "Work-Related Musculoskeletal Disorders Among Farmers in the Southeast Asia Region: A Systematic Review," Saf. Health Work, vol. 14, no. 3, pp. 243–249, Sep. 2023, doi: 10.1016/j.shaw.2023.05.001.

[62] E. Kholinne, X. Azalia, E. P. Rahayu, I. J. Anestessia, N. Agil, and Muchtar, "The prevalence and risk factors of musculoskeletal disorders

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270

among Indonesian dental professionals," Front. Rehabil. Sci., vol. 6, Feb. 2025, doi: 10.3389/fresc.2025.1513442.

[63] Badan Pusat Statistik (BPS), "Profil Statistik Kesehatan 2023." [Online]. Available: https://www.bps.go.id/id/publication/2023/12/20/feffe5519c812d560bb131ca/profil-statistik-kesehatan-2023.html

[64] Q. Liu et al., "Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach," J. Med. Internet Res., vol. 22, no. 4, p. e19118, Apr. 2020, doi: 10.2196/19118.

[65] J. Muragijemariya, V. Ihogoza, and E. T. Luhanga, "Scope of Online Maternal Health Information in Kinyarwanda and Opportunities for Digital Health Developers," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.03805

Author Biography

Dayana Khoiriyah Harahap is a final-year Information Systems student at Sriwijaya University. She is deeply interested in utilizing technology, particularly machine learning, to uncover patterns and insights in complex datasets. She is committed to bridging the gap between theory and practice, frequently applying academic knowledge to real-world problems. This dedication to experiential learning is evident in her contribution to a published research project, where she investigated the practical applications of data analysis. This experience solidified her commitment to the field and honed her skills in transforming raw data into actionable intelligence. As she approaches the completion of her undergraduate studies, she intends to pursue a career in technology, where she can apply her analytical abilities and continue to contribute to the evolving field of data science.

Ken Ditha Tania is a lecturer in the Department of Information Systems at Sriwijaya University. She holds a Bachelor of Computer Science from Indo Global Mandiri University (2007), a Master of Computer Science from the University of Indonesia (2010), and a Doctor of Philosophy from the University of Technology Malaysia (2024). Her academic endeavors are directed toward advancing the fields of information systems and intelligent systems. Her body of work, comprising numerous journal articles and conference proceedings, consistently explores knowledge management, data science, and the utility of sophisticated machine learning algorithms and computational modeling in transforming raw data into strategic assets, thereby enhancing complex decision-making processes.

Putri Eka Sevtiyuni is a lecturer in the Department of Information Systems at Sriwijaya University. She completed her Bachelor's degree in Information Systems at Sriwijaya University in 2012 and earned her Master of Engineering from the Bandung Institute of Technology in 2016. In her academic capacity, she applies machine learning and data mining techniques to address complex challenges within the field of information systems. She contributes to the development of intelligent, data-driven systems that can process and analyze large-scale information. Her work involves exploring various computational methods to extract valuable knowledge from complex datasets, demonstrating her expertise in developing data-centric solutions.

Corresponding author: Ken Ditha Tania, kenya.tania@gmail.com, Department of Information Systems, Faculty of Computer Science, Sriwijaya University, Jl. Raya Palembang - Prabumulih No.KM. 32, Indralaya Indah, Indralaya, Ogan Ilir Regency, South Sumatra 30862, Indonesia
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.270