eISSN 2783-6851

VILNIUS TECH Vilnius Gediminas Technical University

NEW TRENDS in COMPUTER SCIENCES

2026
Volume 4
Issue 1
Pages 111–130
https://doi.org/10.3846/ntcs.2026.25980

DISCOVERING DISEASE RISK TRENDS BY APPLYING LATENT DIRICHLET ALLOCATION SENTIMENT ANALYSIS TO SOCIAL MEDIA HEALTH DISCUSSIONS

Sandeep BHATTACHARJEE

Amity school of Business, Amity University, Kolkata, India

Article History:
- received 3 February 2026
- accepted 7 July 2026

Abstract. Social media platforms generate large volumes of health-related content that can provide valuable insights into emerging disease concerns and public perceptions. This study explores disease-risk trends through topic modeling and sentiment analysis of social media discussions. We used Latent Dirichlet Allocation (LDA) to extract topics and VADER sentiments from 50,000 statements related to diseases. The model performance was evaluated in terms of coherence, topic diversity, and stability metrics. The comparison was made with NMF and BERTopic. Three main topics were identified: COVID-19 symptoms and diagnosis; chronic disease management; and disease risk factors. The LDA model obtained coherence and topic diversity scores of 0.402 and 0.77, respectively. The sentiment analysis shows that the discussions are overall neutral, with some sentiments related to negative concerns. The results indicate that the use of topic modeling along with sentiment analysis is a useful approach to extract major disease-related topics from social media data. Future studies using transformer-based models, multilingual data, and longitudinal studies may improve digital disease surveillance and enable proactive public health monitoring.

Keywords: disease risk surveillance, social media analytics, Latent Dirichlet Allocation, sentiment analysis, topic modeling, public health informatics, health intelligence.

Corresponding author. E-mail: *sandeepbitmba@gmail.com*

1. Introduction

Topic modeling is the new revolution in text mining. It is a statistical technique for revealing the underlying semantic structure in a large collection of documents (Kherwa & Bansal, 2019). Topic modeling is a text-mining technique, whereby one determines the existence of latent themes in massive datasets of textual information. For disease research, topic modeling can analyze huge volumes of clinical notes, scientific literature, and patient-generated data (e.g., social media) to find patterns associated with diseases, symptoms, and the outcome of treatment. This is done through the grouping of words into topics, which can help researchers discuss recurring themes within voluminous texts without manual sifting. Some of the healthcare applications from topic models include identifying emerging issues related to health, understanding what patients are concerned about, and trending diseases. For example, it was possible to identify topics where people talk online about mental health using topic models and understand patient sentiment and experience (De Choudhury et al., 2013). This application has also been used in handling large volumes of EHRs in disease progression patterns (Chen et al., 2017).

2. Conceptual background

The rapid growth of digital content on social media platforms provides unique opportunities for the extraction of useful health-related information from unstructured text data. Two of the most popular text mining techniques are topic modeling and sentiment analysis, which are used to understand public opinion, identify underlying themes and track emerging trends in large-scale datasets. Topic modeling is a form of unsupervised machine learning that detects the hidden thematic structure within a collection of documents. Latent Dirichlet Allocation (LDA) is one of the most popular probabilistic models for topic modeling among various techniques. LDA treats each document as a mixture of topics and each topic as a distribution over words. This method allows the researchers to reveal latent topics by identifying repeating word patterns without the need for the manual labeling of the data (Blei et al., 2003). It has been widely applied in health care research to analyze clinical records, scientific articles and social media conversations about diseases and emerging health trends.

Sentiment analysis is a natural language processing technique that can be used to determine the emotional tone of textual data and classify the texts as positive, negative, or neutral. This classification allows for the measurement of public sentiment and opinions on different subjects (Liu, 2012). In the healthcare space, sentiment analysis may help understand patients' opinions, opinions about treatment, public sentiment about health policies, and opinions about disease outbreaks. An approach to health-related text analytics is to integrate topic modeling and sentiment analysis. Topic modeling can help reveal what the main topics are that users are discussing, and sentiment analysis can help reveal what emotions are attached to those topics. This hybrid approach has great potential for disease surveillance, public health monitoring and risk assessment because it allows researchers to identify emerging disease concerns and the sentiments that accompany them. Therefore, this study uses LDA based topic modeling and sentiment analysis to explore disease-related discussions on social media and identify important disease-risk trends.

3. Related work

3.1. COVID-19 surveillance and public health monitoring

The COVID-19 pandemic highlights the importance of monitoring systems in the detection, monitoring and response to infectious disease outbreaks. Epidemiological monitoring enables public health officials to track the spread of diseases, identify new hot spots, and implement appropriate intervention strategies in time. Despite the difficulties of poor data quality, delayed reporting, and scarce resources, Ibrahim (2020) emphasizes the importance of efficient surveillance systems for managing pandemics. Similarly French and Monahan (2020) argue that current surveillance is a tool of disease monitoring but also a digital tracking and societal response method. Genomic surveillance has also emerged as an essential tool in the monitoring of viral mutations and support of evidence-based public health decision-making in rapidly evolving health emergencies (Robishaw et al., 2021). Chen et al. (2025) analyzed social media discussions to understand public attitudes toward COVID-19 vaccination and misinformation. Their study demonstrated the usefulness of social media analytics for monitoring health perceptions and disease-related behaviors.

3.2. Health informatics and data-driven healthcare

Health informatics has transformed health care using information technologies, data analytics, and clinical decision support systems to improve the health of patients. According to Imhoff et al. (2001), health informatics is a multidisciplinary field that helps to improve the delivery of health care by effectively managing information. The rise of big data and artificial intelligence has also given rise to health analytics. Herland et al. (2014) focused on the use of data mining methods to extract meaningful patterns from large healthcare data repositories. Deep learning has been successful in medical diagnosis and disease prediction (Ravì et al., 2017). Holzinger (2016) indicated the importance of machine learning methods for the development of intelligent healthcare systems and personalized medicine is growing. Recent developments in health informatics have demonstrated the growing integration of intelligent communication systems, artificial intelligence, and deep learning in healthcare. Priya and Malhotra (2023) proposed a QoE-aware framework for 5G-enabled healthcare networks, while Yu et al. (2024) introduced an AI-agent-based platform for intelligent data exchange and analytics. Akila and Bhushan (2026) highlighted the effectiveness of machine learning and deep learning techniques in brain tumor diagnosis. Similarly, Kumari et al. (2024) emphasized the potential of ChatGPT and generative AI in clinical support and medical knowledge dissemination, reflecting the expanding role of AI-driven technologies in healthcare analytics and decision-making.

3.3. Disease surveillance using social media platforms

Social media sites have become a primary source of real-time health data, enabling researchers to rapidly monitor disease outbreaks and public health issues. Lee et al. (2013) demonstrated the usefulness of Twitter data to follow the flu and cancer discussions and thus improve the disease surveillance. Likewise, Broniatowski et al. (2013) found large correlations between Twitter activity and the influenza prevalence during the 2012–2013 epidemic. A systematic review conducted by Charles-Smith et al. (2015) also confirms that social media analytics can support actionable disease surveillance and outbreak management by providing rapid insights into public health events. These studies highlight the growing need for user-generated content to support traditional surveillance systems.

3.4. Natural language processing in healthcare

Healthcare analytics are insufficient without Natural Language Processing (NLP), which helps in deriving useful information from unstructured text data. According to Iroju and Olaleke (2015), NLP techniques improve clinical decision-making, information retrieval and knowledge discovery in healthcare records. This is the premise on which Velupillai et al. (2018) demonstrated the potential of clinical NLP in health outcomes research by converting the text-based information into useful knowledge. Natural language processing (NLP) in smart healthcare systems helps in illness diagnosis, patient monitoring and healthcare automation (Zhou et al., 2022). The recent advances in this field show the importance of Natural Language Processing (NLP) in modern health informatics.

3.5. BERTopic applications in healthcare research

BERTopic is a new topic modeling technique using transformers insertions and clustering for interpretable topic generation. Grootendorst (2022) introduced BERTopic, a neural topic modeling framework that combines transformer embeddings with class-based TF-IDF to generate semantically coherent topics. By leveraging contextual representations, BERTopic offers improved topic interpretability compared with traditional probabilistic approaches such as LDA. The need for transformer-based topic modeling is increasing in healthcare analytics as revealed by different studies. Digital healthcare trends in metabolic syndrome research were identified by means of topic modeling (Lee et al., 2024). Academic and public perception of medical AI was studied using BERTopic and sentiment analysis (Li & Hu, 2025). The LDA approach in topic modeling identified perception in cardiovascular risk (Ma et al., 2025).

3.6. Transformer-based topic modeling

Recent advances in transformer-based topic modeling have substantially improved semantic understanding and topic discovery from textual data. Devlin et al. (2019) introduced BERT, which established the foundation for contextual language representations, while Liu et al. (2021) further enhanced semantic learning through self-supervised transformer architectures. Hassija et al. (2023) highlighted the transformative role of conversational AI and transformer models in knowledge extraction and natural language understanding. Building on these developments, Mersha and Kalita (2024) proposed semantic-driven topic modeling using transformer embeddings and clustering techniques to improve topic coherence. More recently, Reuter et al. (2025) integrated transformer representations within probabilistic topic models, demonstrating enhanced interpretability and semantic consistency. Collectively, these studies illustrate the evolution from conventional topic modeling to sophisticated transformer-based approaches for extracting meaningful patterns from large text corpora.

Transformer architecture has significantly advanced semantic representation and contextual understanding in topic modeling. The transformer embeddings are more efficient in topic classification and theme interpretation than the traditional methods. Mersha and Kalita (2024) propose semantically driven topic modeling with transformer embeddings and clustering algorithms to get meaningful topics. Reuter et al. (2025) combined transformer representations with probabilistic topic models. The improvements were in topic quality, interpretability and scalability. Therefore, transformer-based approaches are gaining popularity for large healthcare and social media datasets.

Previous research has investigated disease surveillance, health informatics, natural language processing (NLP), and advanced topic modeling techniques independently, but very little research has been done regarding the utility of LDA-based topic modeling and sentiment analysis for identifying disease-risk trends from large-scale social media health conversations. Moreover, there is a lack of comparative analysis of traditional probabilistic models and new transformer-based methods in public health surveillance. Thus, the objective of this study is to fill this gap by applying LDA and sentiment analysis on social media health conversations to reveal disease-risk trends, public perceptions and emerging health issues that are of interest to researchers, health care providers and policymakers.

Considering the research gap identified and the increasing importance of social media analytics in disease surveillance, this study aims to answer the following research questions:

RQ1: What disease-related topics are extracted from social media health discussions by Latent Dirichlet Allocation (LDA)?

RQ2: What are the sentiments in the identified disease-related topics, and how do they reflect the public perceptions and concerns towards health conditions?

RQ3: How does LDA perform in comparison to other approaches for topic modeling, like NMF and BERTopic, in terms of topic coherence, diversity and interpretability?

In such a context, the proposed development of an analytical framework will provide a systematic basis for assessing the effectiveness of topic modeling and sentiment analysis in identifying disease-risk trends from social media data in light of these research questions.

4. Research methods

In the first phase, 50,000 statements on disease research in social media were identified using python 3.12.13 and stored into a text file (Sandeepbitmba, 2026). Table 1 summarizes the characteristics of the analyzed corpus comprising 50,000 disease-related social media statements. After preprocessing, the corpus retained a vocabulary of 1,654 unique terms, indicating adequate lexical diversity for topic modeling. The average document length was 4.81 words, reflecting the concise nature of social media expressions. Topic prevalence analysis revealed that Topic 0 (20,012 documents) and Topic 1 (19,987 documents) dominated the corpus, whereas Topic 2 accounted for 10,001 documents. This distribution suggests that disease-related discussions were concentrated around a few major themes while still preserving semantic variability across the dataset.

Table 1. Dataset characteristics (source: author)

Characteristic	Value
Documents	50000
Vocabulary Size	1654
Average words/document	4.81
Topic 0 prevalence	20012
Topic 1: prevalence	19987
Topic 2 prevalence	10001

In the later stage, Latent Dirichlet Allocation (LDA) and VADER-based sentiment analysis (Hutto & Gilbert, 2014) were employed to extract disease-related topics and investigate the associated sentiment distributions. The analysis generated topic representations, keyword visualizations, topic prevalence statistics, and sentiment mappings that facilitated the interpretation of disease-risk trends in social media discussions. Latent Dirichlet Allocation was used to identify latent disease-related themes in the collected social media health discussions. This unsupervised probabilistic topic modeling technique assumes each document is a mixture of topics and each topic is a probability distribution over words (Blei et al., 2003). LDA attempts to find the latent topic structures via analyzing the co-occurrence of terms in a corpus of text. Before the model development, the textual data was preprocessed to generate tokens and convert to lowercase, remove punctuation, remove stop words, and lemmatize. The clean text was then used to construct a document-term matrix. Then, the LDA model was used to estimate the hidden topic distribution for each document and the word distribution for each topic.

In a corpus of (D) documents and (K) topics, LDA assumes that each document is generated by a probabilistic process with Dirichlet priors. The probability of a document can be computed as

$$P(d) = \prod_{n=1}^{N} P(w_n|d) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(w_n|z_k) P(z_k|d), \tag{1}$$

where w_n represents the observed word, z_k denotes the latent topic, $P(w_n \mid z_k)$ is the probability of a word belonging to a topic, $P(z_k \mid d)$ is the probability of a topic occurring within a document.

The best number of topics was chosen according to the coherence score that indicates how semantically consistent are the words within each topic. Higher coherence values mean better interpretability and topic quality. The model fit and generalization performance were also checked using perplexity analysis. Following topic extraction, the dominant topic for each document was selected based on the highest posterior topic probability. The highest-ranking keywords from the model were analyzed to interpret the topics, allowing the identification of disease-related topics such as diagnosis, treatment, chronic disease management, and emerging health issues. LDA was chosen due to its interpretability, scalability and widespread use in healthcare text mining and disease surveillance studies. Its probabilistic framework allows it to discover meaningful thematic patterns from large-scale unstructured social media discussions. Thus, it is suitable for identifying disease-risk trends and public health concerns.

4.1. Experimental setup

To ensure reproducibility, the implementation details, preprocessing libraries, vectorization approaches, model-specific hyperparameters, random seed settings, and hardware environment are summarized in Appendix Tables A1 and A2. These settings were used consistently across the LDA, NMF, and BERTopic experiments.

4.2. Sentiment analysis

Sentiment analysis was performed to examine the emotional orientation of disease-related discussions obtained from social media platforms. Topic modeling identifies the main topics in textual data, while sentiment analysis provides further insight into users' opinions, attitudes, and emotional responses toward specific health conditions, treatments, and disease-related issues. Sentiment analysis integrated with topic modeling provides a more holistic insight into the content and sentiment found within public health discussions.

In this work, we used the Valence Aware Dictionary and sEntiment Reasoner (VADER) algorithm for sentiment classification. VADER is a lexicon and rule-based sentiment analysis tool specifically designed for analyzing social media text and short informal messages. It handles slang, abbreviations, punctuation, capitalisation and emoticons in online discussions particularly well. VADER has become popular in social media analytics and healthcare-related text mining research because it is computationally efficient and well suited to user-generated content.

The VADER algorithm for each text document produces four sentiment metrics i.e. positive, negative, neutral and compound scores. The compound score is a normalized and weighted composite sentiment score ranging from –1 to +1 and derived as

$$\text{Compound Score} = \frac{x}{\sqrt{x^2 + \alpha}}, \tag{2}$$

where x represents the cumulative valence score of the text and α is a normalization constant (typically 15).

Based on the compound score, each document was classified into one of three sentiment categories:

- **Positive Sentiment:** Compound Score ≥ 0.05.
- **Neutral Sentiment:** –0.05 < Compound Score < 0.05.
- **Negative Sentiment:** Compound Score ≤ –0.05.

After classifying the sentiments, we analyzed the sentiment distributions for the discovered disease-related topics. This analysis made it possible to assess the public's perceptions about different health conditions and treatments. Thus, a positive sentiment may indicate confidence in the efficacy of treatments or disease management strategies, whereas a negative sentiment may reflect anxiety, dissatisfaction, or concerns about the severity of the disease and the quality of healthcare services.

The combination of LDA-based topic modeling and sentiment analysis offers a complementary analysis framework. LDA finds hidden topics related to diseases from a large amount of text data, and sentiment analysis provides contextual information about the emotions that are related to these topics. The integrated approach assists in identifying new disease concerns, public awareness trends and health risk perception and thereby enhances disease surveillance, influences healthcare decision making and provides input for public health policy development. Sentiment analysis was performed on the collected social media discussions to better understand disease-related topics. This approach enables us to assess public opinion and emotional responses to identified health themes.

Disease-related social media posts were obtained from publicly accessible online platforms and the corresponding GitHub repository (Sandeepbitmba, 2026). Data collection was performed using keyword-based queries related to symptoms, diseases, diagnosis, and treatment. Only English-language posts meeting predefined inclusion criteria were retained, while duplicate and irrelevant entries were removed to improve corpus quality and consistency.

4.3. Integration of topic modeling and sentiment analysis

The combination of topic modeling and sentiment analysis builds a robust framework for discussions on social media. Latent Dirichlet Allocation (LDA) helps in determining the latent topics of the health-related discussions, while sentiment analysis determines the sentiment polarity corresponding to each topic. The combination of the two approaches enables us to uncover not only emerging concerns about diseases but also insights into public perceptions, attitudes and awareness of risks associated with specific health conditions (see Figure 1).

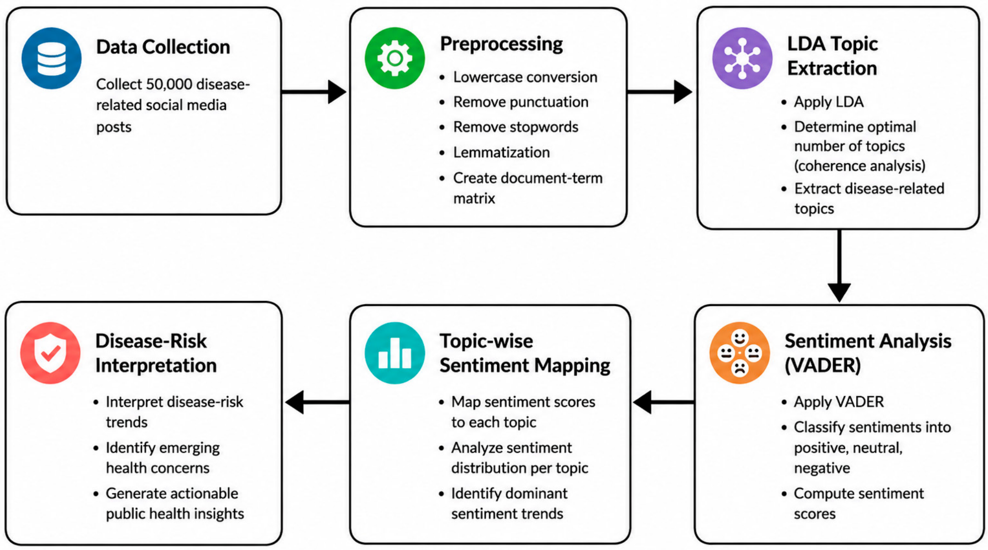

Figure 1. Workflow diagram (source: author analysis)

Figure 1 illustrates the proposed analytical framework for identifying disease-risk trends from large-scale social media discussions. The novelty of the proposed approach lies in its integration of probabilistic topic modelling and lexicon-based sentiment analysis within a single, reproducible analytical workflow for digital disease surveillance. Unlike previous studies that primarily focus on topic extraction or sentiment analysis independently, the proposed framework systematically links latent disease-related topics with their corresponding sentiment patterns, thereby providing a more comprehensive understanding of public health concerns and disease-risk perceptions. Furthermore, the framework incorporates model validation through coherence, perplexity, topic diversity, and stability analyses, together with comparative evaluation against NMF and BERTopic, ensuring the robustness and interpretability of the extracted topics. Consequently, the proposed methodology offers a transparent, scalable, and evidence-based approach that can support researchers and public health authorities in monitoring emerging disease trends and facilitating timely health-related decision-making.

4.4. BERTopic hyperparameter tuning and validation

To ensure robust topic generation, BERTopic hyperparameters were optimized using five bootstrap runs. Different combinations of minimum topic size (50, 75, and 100) and target numbers of topics (10, 15, and 20) were evaluated. Model performance was assessed using topic coherence (Cv), topic diversity, and the average number of topics identified across bootstrap samples. The evaluation aimed to balance semantic coherence with topic distinctiveness and to minimize variability across runs. The configuration with a minimum topic size of 50 and 15 target topics provided a suitable trade-off and was selected for subsequent analysis (see Table 2).

Table 2. BERTopic hyperparameter tuning strategy and validation settings

Component	Parameter	Candidate Values / Setting	Purpose
Embedding Model	Sentence Transformer	all-MiniLM-L6-v2	Generate contextual embeddings
Dimensionality Reduction	UMAP	Default settings	Preserve semantic structure.
Clustering Algorithm	HDBSCAN	Default settings	Discover topic clusters
Minimum Topic Size	min_topic_size	50, 75, 100	Control cluster granularity
Target Number of Topics	nr_topics	10, 15, 20	Topic reduction and interpretability
Top Words per Topic	top_n_words	10	Topic representation
Random Seed	random_state	42	Ensure reproducibility.
Validation Method	Bootstrap Runs	5 runs	Assess robustness and stability
Evaluation Metric 1	Topic Coherence (Cv)	Average over bootstrap runs	Measure semantic consistency
Evaluation Metric 2	Topic Diversity	Average over bootstrap runs	Measure topic distinctiveness
Evaluation Metric 3	Number of Topics Found	Average over bootstrap runs	Evaluate topic stability.
Selection Criterion	Optimal Configuration	Best trade-off between coherence and diversity	Final model selection

4.5. Ethical considerations

The study utilized publicly available social media posts and did not collect any personally identifiable information. All data were anonymized and analyzed at an aggregate level to ensure compliance with ethical research practices and GDPR privacy requirements.

5. Results

5.1. Optimal topic selection

Figure 2 shows coherence scores for different numbers of topics (K = 2–10) obtained during the tuning process of the LDA model. The best coherence score was obtained at 0.402 with three topics, which indicated a semantically meaningful and interpretable topic structure. Coherence refers to how semantically similar words are within a topic. The results therefore suggest that a three-topic solution offers the best compromise between topic distinctiveness and interpretability. Therefore, the optimal number of topics (K) for the subsequent topic modeling analysis was chosen to be three.

Figure 2 confirms that the highest coherence score was achieved when the number of topics was set to three, indicating the most semantically coherent and interpretable topic structure. Therefore, the three-topic configuration was selected for all subsequent analyses to ensure robust and reliable identification of disease-related themes.

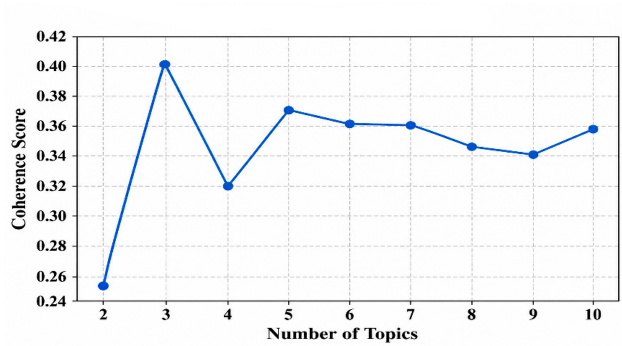

Figure 2. Determination of the optimal number of topics using coherence scores

5.2. Topic extraction and interpretation

Topic modeling identified three main themes in the discussions about diseases on social media. The first topic was about COVID-19 symptoms, diagnoses, and clinical experiences. This may be due to the public's concern about infectious disease surveillance and healthcare consultations. Topic 2 included discussions of experiences with chronic disease and treatment practices, with a particular focus on diabetes and arthritis management. Topic 3 covered disease risk factors, including diet, cardiovascular health and new medical research. Taken together, these topics indicate that social media platforms can be a useful resource for exploring public health concerns, treatment experiences, and perceptions of disease risk (See Table 3).

Table 3. Identified disease-related topics, interpretations, and public health risk implications derived from LDA topic modeling

Topic	Top Keywords	Interpretation	Disease-Risk Implication
Topic 1: COVID-19 Symptoms and Clinical Diagnosis	symptom, new, diagnosed, diabetes, case, mild, covid, might, said, doctor	The use of terms like "COVID," "symptom," "doctor," and "diagnosed" indicates that the public's focus is on the diagnosis of diseases and their clinical experiences concerning reported COVID-19 cases.	Frequent references to infectious diseases, their symptoms, and health care indicate a significant public interest, highlighting the essential role of digital disease monitoring systems.
Topic 2: Chronic Disease Experiences and Treatment Discussions	diabetes, new, im, experiencing, flu, symptom, arthritis, anyone, treatment, tried	The topic of personal experiences in health and treatment discussions on chronic conditions includes diabetes, arthritis, and flu symptoms. Users post reviews on the effectiveness of various treatments, share their experiences, and seek advice.	The topic shows a growing public concern about the management of chronic diseases and self-care practices, thus showing the need for patient-centered healthcare interventions and awareness programs.
Topic 3: Disease Risk Factors and Medical Research	experiencing, diabetes, diagnosed, link, heart, diet, disease, research, suggests, new	The topic relates to health status discussions, lifestyle factors, cardiovascular health, diet, and recent medical research findings. Keywords like "heart," "diet," "research," and "link" show a high interest in understanding the causes of diseases and their risk factors.	The topic is related to growing social awareness of ways of preventing diseases and reducing risk, especially for cardiovascular diseases and metabolic disorders, which are a major public health issue.

5.3. Topic prevalence analysis

Topic 0, corresponding to COVID-19 symptoms and diagnosis, represents the largest proportion of the corpus (20,012 documents), indicating that discussions surrounding symptoms, clinical diagnosis, and medical consultations dominate disease-related conversations. Words such as "symptom," "new," "diagnosed," "covid," and "doctor" reflect concerns about acute case detection and clinical reporting. Topic 2 is the least discussed, and keywords like "link," "heart," "diet," "disease," "research," and "suggests" show that chronic, mechanistic links between diabetes, cardiovascular health and lifestyle factors are not discussed. This imbalance greatly affects disease surveillance. The 'symptom-case' framework of Topic 0, if overused, may lead to a reactive system, detecting illness at presentation. But if we ignore the preventive insights from Topic 2 (e.g., diet-heart-disease links), long-term metabolic epidemics have no early warning signals. Topic 2: Routine alerts combining nutritional and comorbidity data to train algorithms for enhancing rare but high-value etiological signals. Without this shift, the public health systems will be geared to treating cases instead of anticipating disease drivers (see Figure 3).

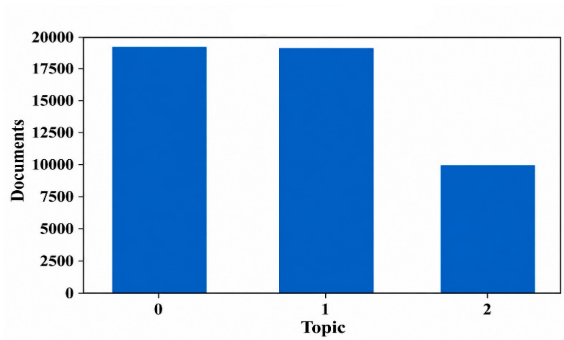

Figure 3. Topic prevalence reflects the relative frequency of each disease theme within the corpus (source: author analysis)

Figure 3 demonstrates that disease-related discussions are concentrated around a few dominant themes, with COVID-19 symptoms and chronic disease management accounting for the largest share of the corpus. This topic distribution highlights the effectiveness of the LDA model in identifying prominent disease-related trends and provides a foundation for the subsequent model validation analysis.

5.4. Topic stability and validation

The results for validating the selected LDA model are presented in Table 4. The coherence score of 0.402 indicates a moderate level of semantic consistency between generated topics, which suggests that the disease-related topics extracted are relatively interpretable. The topic diversity score of 0.77 indicates that 77% of the top words across topics are unique. This reflects good topic distinctiveness with little overlap in themes. The model fit for the data set is an acceptable one for the perplexity value (12.20). The average stability coherence score of 0.3247, along with a small standard deviation of 0.0484, also indicates that the model produces stable topic structures regardless of random initialization. These measures show the robustness and reliability of the LDA model in identifying disease-related trends in social media discussions.

Table 4. Topic validation metrics (source: author's analysis)

Metric	Value
Optimal Number of Topics	3
Coherence Score	0.402
Perplexity	12.2
Topic Diversity	0.77
Mean Stability Coherence	0.3247
Stability Std. Dev.	0.0484

To ensure the robustness of the LDA model, we performed topic stability analysis under five different random seeds. The coherence scores range from 0.2563 to 0.4040 with a mean of 0.3247 and a standard deviation of 0.0484. The standard deviation is relatively low, which indicates satisfactory stability, i.e., the model produces similar topic structures for different initializations. Probabilistic topic modeling is expected to exhibit some variability in coherence, but the overall results indicate that the disease-related topics identified are robust and not highly sensitive to the random seed selected. This stability verifies the accuracy and reproducibility of the results of trend analysis on disease risk from topic modeling (see Table 5).

Table 5. Topic stability across seeds (source: author analysis)

Seed	Coherence
1	0.31425382
2	0.342478272
3	0.30659218
4	0.25631693
5	0.403957885
Mean Coherence: 0.3247	
Standard Deviation of Coherence: 0.0484	

5.5. Comparative analysis of topic models

Table 6 and Figure 4 show the performance of LDA, NMF and BERTopic based on topic coherence and diversity. BERTopic obtained the best value of 0.5841, which indicates high semantic consistency. LDA was close, with a coherence score of 0.4027 and the highest topic diversity score of 0.7667, which implies broader coverage and better topic distinctiveness. Furthermore, LDA provides probabilistic topic distributions and interpretable word-topic associations that are useful for interpreting disease-related discussions and public health. However, BERTopic's embedding-based architecture adds more computational complexity, while LDA provides a simpler, more computationally efficient, and highly interpretable framework. Thus,

Table 6. Comparative performance of topic modeling approaches

Model	Coherence	Topic Diversity	Perplexity
LDA	0.4027	0.7667	12.2072
NMF	0.3811	0.7333	NA
BERTopic	0.5841	0.7180	NA

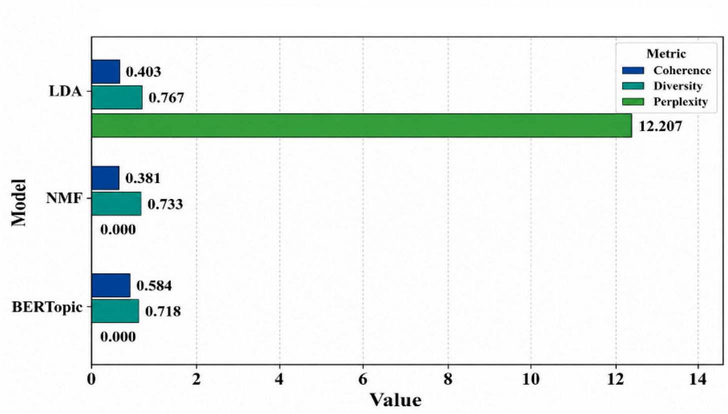

Figure 4. Comparison of topic models on coherence and topic diversity metrics (source: author analysis)

although BERTopic shows better semantic capabilities, LDA was chosen as the best method for this study because it achieves the best trade-off between interpretability, diversity, reproducibility, and practical applicability in the context of disease surveillance. Consequently, although the semantic coherence of BERTopic was better and validated with fivefold cross-validation, the interpretability, higher diversity of topics, lower computational requirements, and suitability for transparent disease-risk analysis led to the adoption of LDA as the preferred model.

Figure 4 illustrates the comparative performance of LDA, NMF, and BERTopic across topic coherence and diversity metrics. Although BERTopic achieved the highest coherence score, LDA provided a better balance between interpretability, topic diversity, and computational efficiency, supporting its selection as the primary model for this study.

5.6. BERTopic performance evaluation

Table 7 shows the BERTopic parameter tuning results averaged over five bootstrap runs. Different combinations of minimum topic size and target number of topics were evaluated using topic coherence and topic diversity metrics. The configuration with a minimum topic size of 50 and 15 target topics achieved an average coherence score of 0.5841 and a topic diversity score of 0.7180, while consistently identifying approximately 15 topics. This combination provided the best balance between semantic interpretability and topic distinctiveness and was therefore selected for subsequent analysis. The consistency of results across bootstrap samples demonstrate the robustness of the BERTopic framework (see Table 7).

Table 7. BERTopic parameter tuning results averaged over five bootstrap runs

min_topic_size	nr_topics	Avg_Coherence	Avg_Topic_Diversity	Avg_Num_Topics_Found
50	15	0.5841	0.718	14.6
75	15	0.5732	0.7511	14.4
50	10	0.612	0.669	9.6
50	20	0.5694	0.7093	19.4
75	10	0.5976	0.6908	9.4
100	15	0.551	0.7933	14.6

5.7. Sentiment analysis of topics

Table 8 presents the sentiment distribution in the three identified disease-related topics. The results show that the neutral sentiment is the most prevalent across all the topics, accounting for almost 90% of the discussions. Topic 1 had the highest number of neutral posts (n = 18,017), followed by Topic 0 (n = 18,014), and Topic 2 included 8,969 neutral statements (see Figure 5). Negative sentiment was relatively low, with 1,998, 1,970 and 1,032 posts reported for Topics 0, 1 and 2 respectively. Importantly, no positive sentiment instances were identified in the three topics. This pattern suggests that social media discussions about disease tend to be more about sharing information and experiences than positive emotions.

Table 8. Topic-wise sentiment distribution (source: author's analysis)

Topic	Positive	Neutral	Negative
0	0	18014	1998
1	0	18017	1970
2	0	8969	1032

The issues of symptoms, diagnosis, treatment outcomes, and disease management received negative sentiment. Overall, the results suggest that the public discourse on diseases is, on the whole, largely neutral, while a smaller but significant portion of the discussion reveals health-related anxieties and risk perceptions (see Figure 5).

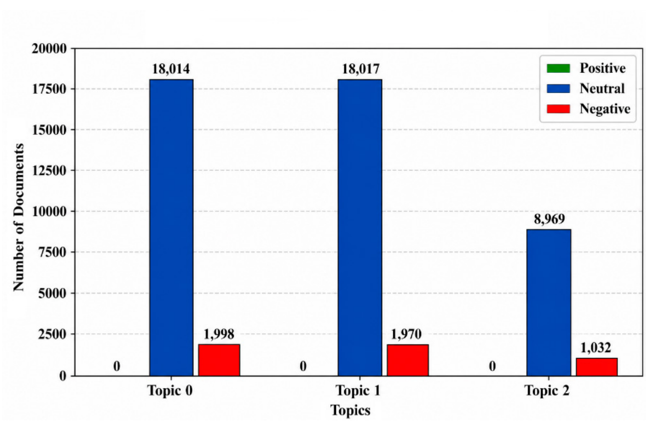

Figure 5. Topic-wise sentiment distribution (source: author's analysis)

Figure 5 shows that neutral sentiment predominates across all identified disease-related topics, while negative sentiment reflects public concerns regarding symptoms, diagnosis, and treatment experiences. These findings demonstrate that integrating sentiment analysis with topic modelling provides deeper insights into public health perceptions and supports more informed disease surveillance and risk assessment.

6. Discussion and findings

Novel insights

- Chronic disease discussions occupied a substantial portion of the corpus, indicating sustained public concern beyond pandemic-related issues.
- Negative sentiments associated with disease-risk topics suggest heightened anxiety and uncertainty among users.
- Topic prevalence analysis provides a quantitative view of public attention and disease awareness.

Public health implications

- Social media analytics can complement traditional epidemiological surveillance.
- Early detection of emerging concerns may support timely intervention.
- Topic-wise sentiment mapping can assist health agencies in identifying misinformation and public anxiety.

Consistent with previous disease surveillance studies, the present findings demonstrate that social media discussions can capture evolving health concerns. However, unlike earlier studies that focused primarily on outbreak monitoring, the proposed framework integrates topic modeling with sentiment analysis to provide additional insights into disease-related perceptions and risk awareness.

The study analyzed 50,000 social media statements on diseases using Latent Dirichlet Allocation (LDA) and sentiment analysis to explore health concerns and public perceptions. The most common topics were clinical diagnosis and symptoms of COVID-19, personal experiences with chronic diseases and treatment discussions, and risk factors and medical research on diseases. Analysis showed that some 80% of the conversations were about COVID-19 symptoms and the management of chronic diseases, reflecting a significant public interest in these health issues. The LDA model was robust (coherence score 0.402, topic diversity score 0.77) and consistent with multiple random seed initializations.

In comparison, while BERTopic exhibited the highest semantic coherence, LDA achieved a more favorable balance of coherence, diversity, and interpretability. Furthermore, sentiment analysis showed that most of the disease-related discussions were neutral, while a small proportion of the discussions were negative in terms of diagnosis, treatment and disease management. These findings emphasize the usefulness of social media analytics in real-time detection of public health concerns and disease risk discussions.

The sentiment distribution revealed that neutral and negative sentiments dominated across all identified topics, whereas positive sentiments were absent. This pattern is consistent with the nature of disease-related social media discussions, which primarily focus on symptoms, diagnosis, treatment experiences, and perceived health risks. Consequently, most posts tend to convey informational content or concern rather than optimism. Moreover, the use of the VADER lexicon-based sentiment analyzer may limit the detection of subtle positive expressions in healthcare-related texts. Therefore, the findings should be interpreted in the context of both the underlying corpus characteristics and the methodological limitations of sentiment classification.

7. Conclusions

We analyze disease-related discussions on social media using a combined approach of LDA-based topic modeling and sentiment analysis. We identified three dominant themes: symptom reporting, chronic disease management and disease-risk awareness. Validation measures proved the reliability and stability of the LDA model, while a comparative analysis revealed that BERTopic outperformed LDA in semantic coherence and that LDA provided a good trade-off between interpretability and topic diversity. The sentiment analysis results indicated that most of the discussions about the diseases were neutral, which were mainly information sharing and personal health experiences. Conversely, fewer of the conversations were about concerns regarding diagnosis and treatment. Overall, the study demonstrates the utility of using topic modeling and sentiment analysis to analyze large-scale health-related social media data and to discover important trends related to disease risk. Future work can explore larger multilingual datasets, transformer-based language models, and longitudinal analyses to improve the accuracy and scalability of digital disease surveillance systems.

The proposed framework can be incorporated into public health surveillance systems to continuously monitor disease-related discussions and sentiment trends. Healthcare organizations and policymakers may utilize these insights to prioritize awareness campaigns, identify emerging concerns, and strengthen data-driven decision-making for disease prevention and risk communication.

8. Theoretical and practical implications

From a theoretical perspective, this study contributes to the increasing literature on health informatics and disease surveillance by demonstrating the combined use of topic modeling and sentiment analysis to extract meaningful information from unstructured social media data. Moreover, the study empirically compares LDA, NMF, and BERTopic and highlights their respective strengths in semantic interpretation and topic diversity.

It might be used as supplementary sources for public health intelligence. Healthcare practitioners, researchers and public health agencies can use topic modeling techniques to identify emerging discussions about diseases, follow public concerns and analyze sentiment patterns related to health conditions. Such insights may be useful for early awareness initiatives, risk communication strategies and evidence-based decision-making. However, social media analytics should be considered a complementary tool for surveillance but not a substitute for traditional epidemiological surveillance systems.

Furthermore, the reliance on the VADER sentiment analyzer may not fully capture nuanced emotional expressions in health-related discussions, potentially underrepresenting positive sentiments. Future studies may employ transformer-based sentiment models such as BERT or BioBERT to improve contextual understanding and sentiment classification accuracy.

Disclosure statement

The author reports there are no competing interests to declare.

References

Akila, V., & Bhushanm, K. (2026, May). A systematic review of machine learning and deep learning techniques for brain tumor detection and classification in MRI images (2020–2026). In *2026 International Conference on Signal, Systems, and Computing for Next-Gen Automation* (*ICSSCNA*) (pp. 1–9). IEEE. https://doi.org/10.1109/ICSSCNA68616.2026.11546721

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. *PLoS ONE*, *8*(12), Article e83672. https://doi.org/10.1371/journal.pone.0083672

Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., Robinson, S. J., & Corley, C. D. (2015). Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PLoS ONE*, *10*(10), Article e0139701. https://doi.org/10.1371/journal.pone.0139701

Chen, J., Wei, W., Guo, C., Tang, L., & Sun, L. (2017). Textual analysis and visualization of research trends in data mining for electronic health records. *Health Policy and Technology*, *6*(4), 389–400. https://doi.org/10.1016/j.hlpt.2017.10.003

Chen, N., Chen, X., Zhong, Z., & Pang, J. (2025). "Double vaccinated, 5G boosted!": Learning attitudes towards COVID-19 vaccination from social media. *ACM Transactions on the Web*, *19*(1), 1–24. https://doi.org/10.1145/3731205

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

French, M., & Monahan, T. (2020). Disease surveillance: How might surveillance studies address COVID-19? *Surveillance* & *Society*, *18*(1), 1–11. https://doi.org/10.24908/ss.v18i1.13985

Grootendorst, M. (2022). *BERTopic*: *Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://doi.org/10.48550/arXiv.2203.05794

Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., & Sikdar, B. (2023). Unleashing the potential of conversational AI: Amplifying ChatGPT's capabilities and tackling technical hurdles. *IEEE Access*, *11*, 143657–143682. https://doi.org/10.1109/ACCESS.2023.3339553

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, *1*(1), Article 2. https://doi.org/10.1186/2196-1115-1-2

Holzinger, A. (2016). Machine learning for health informatics. In A. Holzinger (Ed.), *Machine learning for health informatics*: *State-of-the-art and future challenges* (pp. 1–24). Springer. https://doi.org/10.1007/978-3-319-50478-0_1

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 216–225). https://doi.org/10.1609/icwsm.v8i1.14550

Ibrahim, N. K. (2020). Epidemiologic surveillance for controlling the COVID-19 pandemic: Types, challenges and implications. *Journal of Infection and Public Health*, *13*(11), 1630–1638. https://doi.org/10.1016/j.jiph.2020.07.019

Imhoff, M., Webb, A., & Goldschmidt, A. (2001). Health informatics. *Intensive Care Medicine*, *27*(1), 179–186. https://doi.org/10.1007/PL00020869

Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science*, *8*(8), 44–50. https://doi.org/10.5815/ijitcs.2015.08.07

Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), Article e2. https://doi.org/10.4108/eai.13-7-2018.159623

Kumari, K., Pahuja, S. K., & Kumar, S. (2024). A comprehensive examination of ChatGPT's contribution to the healthcare sector and hepatology. *Digestive Diseases and Sciences*, *69*(11), 4027–4043. https://doi.org/10.1007/s10620-024-08659-4

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time disease surveillance using Twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1474–1477). Association for Computing Machinery. https://doi.org/10.1145/2487575.2487709

Lee, K., Chung, Y., & Kim, J. S. (2024). Research trends on metabolic syndrome in digital healthcare using topic modeling: Systematic search of abstracts. *Journal of Medical Internet Research*, *26*, Article e53873. https://doi.org/10.2196/53873

Li, C., & Hu, X. (2025). Medical artificial intelligence in scholarly and public perspective: A BERTopic-based analysis of topic-sentiment collaborative mining. *Data Science and Informetrics*, *5*(1), 33–42. https://doi.org/10.1016/j.dsim.2025.05.001

Liu, A. T., Li, S. W., & Lee, H. Y. (2021). TERA: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, *29*, 2351–2366. https://doi.org/10.1109/TASLP.2021.3095662

Liu, B. (2012). *Sentiment analysis and opinion mining*. Springer. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Ma, L., Chen, R., Ge, W., Rogers, P., Lyn-Cook, B., Hong, H., & Zou, W. (2025). AI-powered topic modeling: Comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Experimental Biology and Medicine*, *250*, Article 10389. https://doi.org/10.3389/ebm.2025.10389

Mersha, M. A., & Kalita, J. (2024). Semantic-driven topic modeling using transformer-based embeddings and clustering algorithms. *Procedia Computer Science*, *244*, 121–132. https://doi.org/10.1016/j.procs.2024.10.185

Priya, B., & Malhotra, J. (2023). 5GhNet: An intelligent QoE-aware RAT selection framework for 5G-enabled healthcare networks. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 8387–8408. https://doi.org/10.1007/s12652-021-03606-x

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Reuter, A., Thielmann, A., Weisser, C., Säfken, B., & Kneib, T. (2025). Probabilistic topic modeling with transformer representations. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(8), 14551–14565. https://doi.org/10.1109/TNNLS.2025.3538262

Robishaw, J. D., Alter, S. M., Solano, J. J., Shih, R. D., DeMets, D. L., Maki, D. G., & Hennekens, C. H. (2021). Genomic surveillance to combat COVID-19: Challenges and opportunities. *The Lancet Microbe*, *2*(9), e481–e484. https://doi.org/10.1016/S2666-5247(21)00121-X

Sandeepbitmba. (2026). *Disease statements dataset* [Data set]. GitHub. https://github.com/sandeepbitmba/socialmediastatements/blob/main/Disease%20statements.docx

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, R. D., Stewart, R., & Dutta, R. (2018). Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, *88*, 11–19. https://doi.org/10.1016/j.jbi.2018.10.005

Yu, Y., Yao, S., Zhou, T., Fu, Y., Yu, J., Wang, D., & Lin, Y. (2024, June). Data on the move: Traffic-oriented data trading platform powered by AI agents with common sense. In *2024 IEEE Intelligent Vehicles Symposium* (*IV*) (pp. 521–526). IEEE. https://doi.org/10.1109/IV55156.2024.10588800

Zhou, B., Yang, G., Shi, Z., & Ma, S. (2022). Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*, *17*, 4–18. https://doi.org/10.1109/RBME.2022.3210270

APPENDIX

Sample posts were randomly selected from the corpus and are shown only for illustration. Note: The sample statements shown in Appendix A1 are representative excerpts obtained after preprocessing and normalization of larger social media discussions. They are provided only for illustrative purposes to enhance interpretability and do not correspond to the complete original posts (see Figure A1).

My doctor said I might have a mild case of COVID-19.

I'm experiencing symptoms of the flu.

New research suggests a link between diet and heart disease.

I'm experiencing symptoms of the flu.

My doctor said I might have a mild case of COVID-19.

New research suggests a link between diet and heart disease.

Has anyone tried this new treatment for arthritis?

Just diagnosed with diabetes.

My doctor said I might have a mild case of COVID-19.

Has anyone tried this new treatment for arthritis?

My doctor said I might have a mild case of COVID-19.

My doctor said I might have a mild case of COVID-19.

I'm experiencing symptoms of the flu.

Just diagnosed with diabetes.

New research suggests a link between diet and heart disease.

My doctor said I might have a mild case of COVID-19.

Has anyone tried this new treatment for arthritis?

Has anyone tried this new treatment for arthritis?

Just diagnosed with diabetes.

Has anyone tried this new treatment for arthritis?

Figure A1. Representative statements

Table A1. Model configuration and experimental settings

Category	LDA	NMF	BERTopic
Vectorization	Bag-of-Words	TF-IDF	Sentence Embeddings
Number of Topics	3	3	15
Random Seed	42	42	42
Iterations	100	200	Default
Alpha	Auto	–	–
Beta (Eta)	Auto	–	–
Initialization	–	nndsvd	all-MiniLM-L6-v2
min_topic_size	–	–	50
Bootstrap Runs	–	–	5
Evaluation Metrics	Coherence, Diversity	Coherence, Diversity	Coherence, Diversity

Table A2. Software and hardware environment

Component	Specification
Programming Language	Python 3.12
Libraries	Gensim, Scikit-learn, BERTopic, NLTK, and spaCy
Operating System	Windows 11
Processor	Intel Core i7
RAM	16 GB
Random Seed	42